HyVis: A Hybrid Visualization Assessment for Balancing Human Readability and Machine Comprehension

Anonymous ACL submission

Abstract

Data visualization has evolved from a purely human-centric craft to a dual-purpose tool consumed by both humans and machine-driven models. However, most existing evaluations focus primarily on aesthetics and clarity for human users, overlooking machine interpretability. To bridge this gap, this study introduces HyVis (A Hybrid Visualization assessment for balancing human readability and machine comprehension), a framework for evaluating visualization quality by combining human preference criteria and model interpretability. Unlike prior studies focused on human perception, HyVis integrates model readability, ensuring visualizations are interpretable for machine-driven analysis. Experimental results demonstrate that HyVis improves human preference-based evaluations by up to 16% and achieves a 3.14% higher accuracy in machine-readable assessments compared to large-scale models.

1 Introduction

Data visualization has been essential in science and other fields, helping people understand complex data and share insights(Yu and Silva, 2019; Ouyang, 2024). Recently, multimodal Large Language Models (LLMs) have made it possible to automatically create various types of charts and graphs (Hu et al., 2024; Han et al., 2023). This advancement has changed how we create and understand visualizations.

Traditional frameworks for evaluating visualizations are based on how humans see and understand visual information(Munzner, 2014), focusing on making visuals easy to read and visually appealing(Barcellos et al., 2022; Andreou et al., 2023).

These guidelines have become standard in the field(Tufte and Graves-Morris, 1983; Schwabish, 2021). However, with the rise of advanced LLMs, we can no longer assume that visualizations are consumed *solely* by humans. Recent studies (Wu et al., 2021; Yuan et al., 2021), show that many visualizations are now processed and interpreted by machines, meaning that the "end user" of a visualization isn't always human.

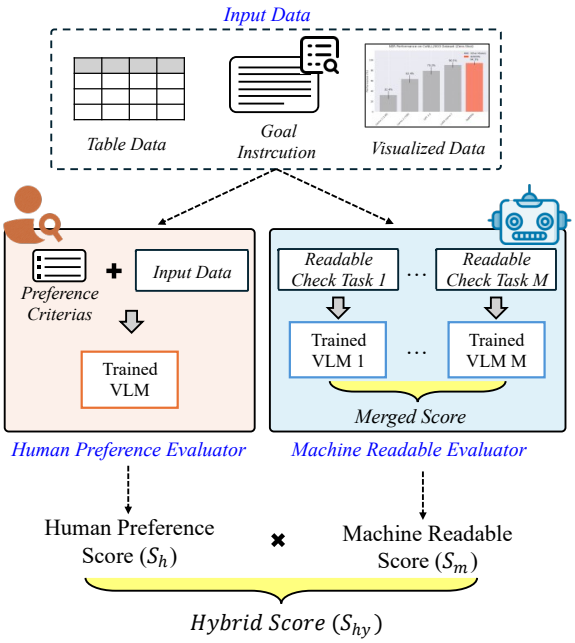
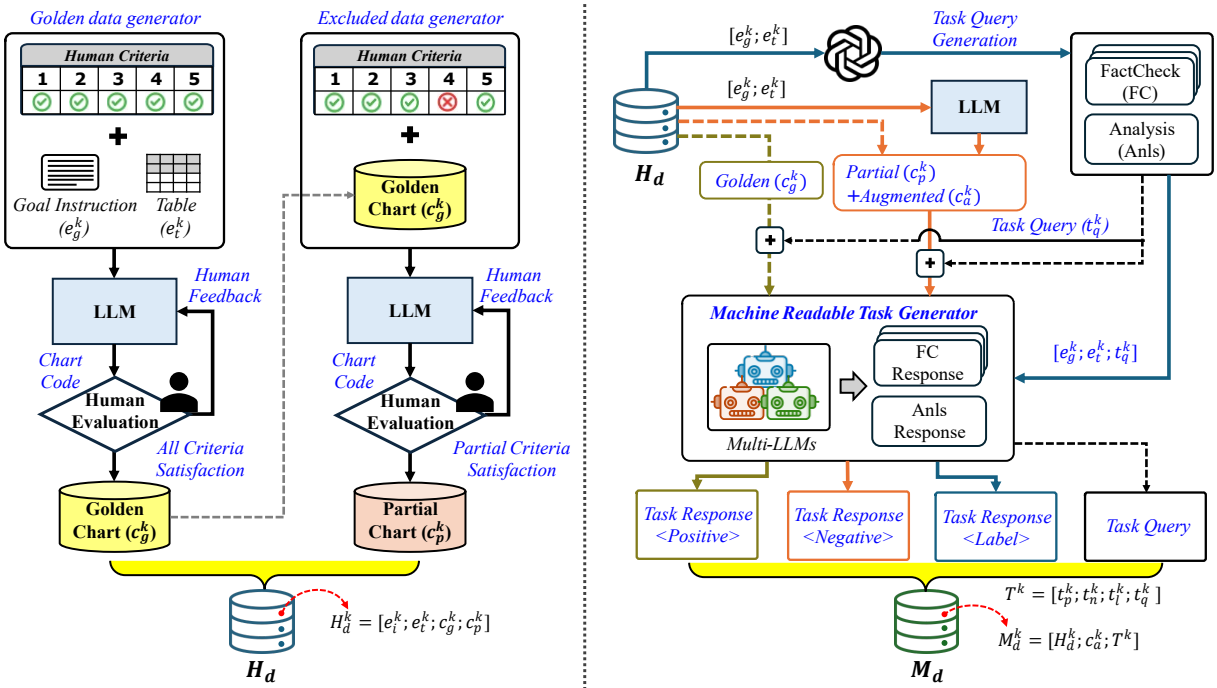


Figure 1: Illustration of our method's contributions: We propose an evaluation approach that assesses data visualization quality based on human preferences and machine performance.

Given this change, we need to update our visualization standards to consider both human and machine interpretation. To address this, we present the **HyVis** (A Hybrid Visualization assessment for balancing human readability and machine comprehension) framework, which combines criteria for both human and machine understanding. Our goal is to help create visualizations that communicate effectively to both human and machine audiences, leading to better visualization practices overall.

As illustrated in Figure 1, HyVis takes chart images and analytical objective descriptions as inputs, diagnosing whether the charts meet predefined cri-



(a) Human Preference Dataset (b) Machine Readable Dataset

Figure 2: Human preference and machine readable dataset construction. (a) The human preference dataset (H_d) consists of two categories: golden charts that satisfy all preference criteria and partial charts derived from golden charts that fail to satisfy certain human preferences. (b) The machine readable dataset (M_d) is designed to evaluate chart understanding based on goal instructions. It comprises query/response pairs(T^k) for FactCheck (FC) and Analysis (Anls) tasks. The generated chart inference labels are used for evaluator training, while table inference labels are utilized for scoring machine readable task results.

teria and scoring their alignment with the model's goals. This approach defines chart evaluation criteria, including model readability, and enables HyVis to automatically assess chart quality.

To address this, our study introduces a framework comprising:

- **Human Preference Evaluation**: Assessing how clear and useful a chart is from a human viewpoint.
- **Machine Readable Evaluation**: Introduces new criteria to evaluate model understanding of charts and their ability to achieve goals, shifting from human-centered approaches.
- **Hybrid Score Calculation**: Combining both scores to provide a comprehensive evaluation of visualization quality for both humans and machines.

This approach aims to enhance the effectiveness of data visualizations in environments where humans and machine systems co-work, ensuring that visual information is accessible and interpretable by both.

2 Related Works

This research is grounded in three key areas: vision-language models (VLMs) for chart understanding, a multi-LLM collaborative framework, and machine readability assessment. The relevant literature is reviewed, highlighting the problems this study aims to address.

2.1 Vision-Language Model For Chart Understanding

Recent advances in chart QA systems demonstrate the potential of VLMs in multimodal reasoning. ChartInstruct(Masry et al., 2024) proposed a hybrid evaluation combining human preference and machine readability scores, establishing new chart analysis standards. Based on UniChart (Masry et al., 2023), it leverages visual element extraction and data table reconstruction for state-of-the-art performance.

DePlot+FlanPaLM (Liu et al., 2023a) pioneered chart-to-table conversion for numerical reasoning, while MatCha (Liu et al., 2023b) enhanced visual encoding through derendering pretraining. However, recent evaluations (Moritz et al., 2019a; Li

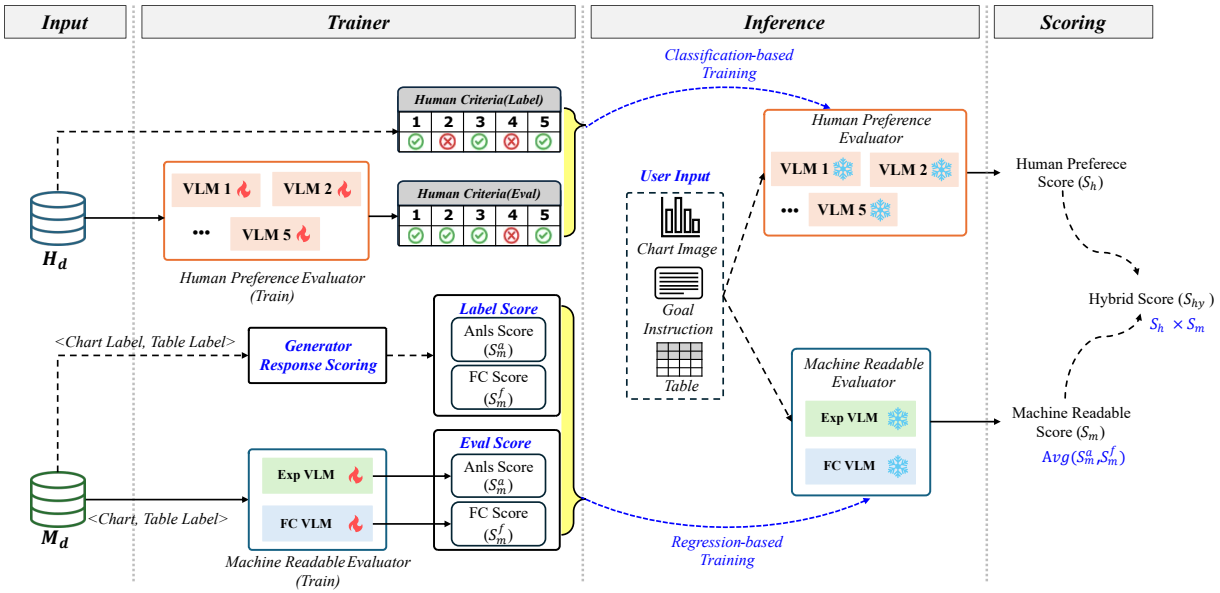


Figure 3: Framework of a hybrid visualization assessment for balancing human readability and machine comprehension (HyVis). The preprocessing stage constructs datasets H_d and M_d, consisting of chart images and seed data for training the human preference and machine readable evaluator. The trained evaluator produces Human preference Score (S_h) and Machine readable Score (S_m) ranging from 0 to 1, which are then combined to compute the Hybrid Score(S_{hy}).

et al., 2023) indicate that existing VLMs struggle with complex numerical operations and maintaining consistency across diverse chart types.

2.2 Multi-LLM Collaboration Framework

Recent studies on multi-LLM systems suggest that collaboration among specialized agents can enhance performance on complex tasks. (Cao, 2024) introduced cooperative interactions between analysis and execution agents, while the Multi-Agent-Debate framework (Liang) improved reasoning capabilities through iterative debate processes.

2.3 Machine Readable Evaluation

Interpretability research offers essential methods for evaluating model reasoning. The CLEAR corpus (Munzner, 2014) established readability metrics based on human judgment, while contrastive explanations (Moritz et al., 2019a) highlighted the value of comparative analysis. For chart evaluation, (Masry et al., 2023) introduced a hybrid score combining factual consistency (S_m^f) and explanation quality (S_m^e).

3 HyVis: A Hybrid Visualization Assessment for Balancing Human Readability and Machine Comprehension

In this study, we propose a Chart data quality assessment framework that considers both human preferences and model readable evaluator. To achieve this, the framework is structured into three components: (1) *the Human Preference Evaluator*, which assesses human preferences for chart visualization, (2) *the Machine Readable Evaluator*, which evaluates the analytical suitability of charts based on task-specific requirements, and (3) *the Hybrid Score Calculation*, which integrates both evaluations.

3.1 Human Preference Evaluator

The Human preference evaluator quantifies whether a chart's components align with intuitive human evaluation criteria, ensuring ease of interpretation.

Human Preference Criteria The core evaluation criteria for chart visualization are derived from existing research on visualization quality assessment (Munzner, 2014; Borkin et al., 2016). To reflect effective user preferences in data analysis, we selected the following five criteria:

- **Completeness:** Ensuring that all necessary

visual elements (axes, legends, titles, etc.) represent the table categories effectively.

- **Optimization:** Structuring the chart effectively to align with the analytical purpose.
- **Emphasis:** Highlighting key information relevant to the analysis goal.
- **Representation:** Accurately representing chart data without distortion or confusing visual elements.
- **Numericalization:** Using intuitive chart labels and item value representation.

Human Preference Data Generation As illustrated in Figure 2-(a), the Human Preference Dataset (H_d) is created using table data and goal instructions for criteria classification training. H_d consists of two primary datasets: the golden dataset (c_g^k) and the partial dataset (c_p^k). The golden dataset is generated by inputting prompts containing human preference criteria into an LLM, which then produces chart generation code. A feedback loop incorporating human evaluation based on the criteria ensures that the golden dataset meets all the specified criteria, as shown in Figure 4.

The partial dataset is generated using a similar process but with specific criteria deliberately restricted at the prompt input stage. Human feedback then verifies whether the exclusion conditions have been met. Ultimately, the golden dataset and partial dataset are combined to form H_d.

Human Preference Evaluator Training The human preference evaluator consists of a vision encoder and a text decoder model, which takes a chart image and instruction as input to predict compliance with each criterion. The model is trained using the golden dataset and partial dataset, with criteria treated as multi-class classification labels. During inference, the trained model assigns a Human Preference Score (S_h), where each criterion contributes 0.2 points, yielding a total score between 0 and 1. The model's learning and evaluation structure is illustrated in Figure 3.

3.2 Machine Readable Evaluator

The machine readable evaluator assesses a chart's analytical suitability from the perspective of models that analyze and interpret charts.

Machine Readable Criteria The machine readable evaluator aims to assess a model's ability to understand and analyze data. To evaluate visualization quality based on goal instructions, we selected two tasks as machine-readable criteria: Chart FactCheck(Liu et al., 2023b) and Chart Analysis(Hu et al., 2024). These tasks verify the accuracy of information required by the goal instruction and assess whether the chart structure is appropriate for analysis. The factcheck task verifies whether the chart accurately reflects table data and instructions, while the analysis task involves summarizing the chart's key contents. These tasks are commonly used to evaluate chart generation models, enabling a quantitative assessment of visualization quality (Masry et al., 2023; Liu et al., 2023b).

Golden Data Generation Prompt

Based on the **TUBULAR** data mentioned below, create a chart in a format that effectively represents the **KEY_CONTEXT** information.
When visualizing, ensure that the chart satisfies all the **CONDITIONS** listed below.
The following content outlines the conditions that must be met to create a structure conducive to effective analysis when performing visualization using charts.
Additionally, provide the Python code for generating the chart.
===
TUBULAR###
{tubular}

KEY_CONTEXT###
{instruction}

CONDITIONS###
{Human Preferance Criteria}

Figure 4: Golden data generation prompt configuration.

Machine Readable Data Generation The machine readable evaluator should be dependent on input data quality rather than model performance. To ensure this quality, we construct a Positive/Negative task dataset as shown in Figure 2-(b). In addition, for robust task scoring, we generate table-based task responses for each input chart. Formally, let be the golden, partial, and augmented charts(c_g^k, c_p^k,c_a^k) for the k-th instance, respectively. We define a machine readable task generator function:

$$G_{\text{task}}(\cdot),$$

which transforms the input charts into machine-readable tasks:

$$t_p^k = G_{\text{task}}(c_g^k), \quad t_n^k = G_{\text{task}}(c_p^k,\ c_a^k).$$

Here, t_p^k represents the *positive* response derived

Type	Complexity	Tables	Golden chart	Partial chart
Continuous	Low	18	36	3,531
	high	11	22	2,246
Time-series	Low	21	42	4,124
	high	8	16	1,503
Total	-	**58**	**116**	**11,404**

Table 1: Example Table Data and Human Preference Dataset Composition.

from the golden chart c_g^k, whereas t_n^k is a *negative* response constructed from the partial and augmented charts $\{c_p^k,\ c_a^k\}$. Similarly, label response (e.g., t_ℓ^k) can be generated as needed.

As a result of this process, an *evaluation task dataset* T^k containing the *task query* is generated, and the entire set M_d is constructed as follows:

$$T^k = \{\, t_p^k,\ t_n^k,\ t_\ell^k,\ t_q^k \}$$

$$M_d = \{\, H_d^k,\ c_a^k,\ T^k \}$$

Machine Readable Evaluator Training The machine readable evaluator is a VLM-based structure designed to analyze and evaluate charts. During training, the model is fed with chart images, task queries, and golden answers. The training data is derived from M_d, including labels for FactCheck and Analysis tasks.

During evaluation, the model generates task queries based on table and instruction inputs and uses them to compute the FactCheck Score (S_m^f) and Analysis Score (S_m^a). The final Machine Readable Score (S_m) is then calculated as the average of these two scores. S_m serves as a quantitative measure of how well the model aligns chart data with table information and interprets it according to the analytical objective.

3.3 Hybrid Score Calculation

Hybrid score calculation is the stage that integrates the evaluation results from the Human preference evaluator and the machine readable evaluator to compute the overall chart assessment score (S_{hy}). The overall score is calculated as the product of S_h and S_m, ensuring that both human intuitive evaluation and the model's interpretative capabilities are considered simultaneously. This metric serves as a quantitative measure of how well a chart aligns with the intended analytical objectives. Through this approach, Hybrid Critics establishes a comprehensive evaluation framework that considers both human and model-based criteria for assessing overall chart quality.

Parameters	HP Evaluator	MR Evaluator (FC / Anls)
Epochs	10	12
Batch Size	32	32
Learning Rate	3×10^{-5}	5×10^{-5}
Optimizer	AdamW	AdamW
Mixed Precision	FP16	FP16

Table 2: Implementation Details of Evaluators

4 Experimental Setup

In this section, we outline the experimental setup designed to evaluate the effectiveness of the HyVis framework.

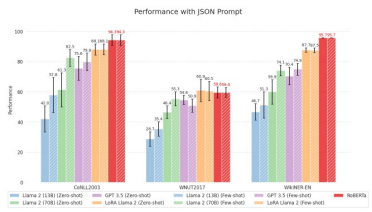
4.1 Dataset

The chart generation data used for training and evaluating the proposed evaluator was obtained from the TATQA dataset (Zhu et al., 2021) and table data extracted from ArXiv. Subsequently, we performed data augmentation using LLMs to enhance the dataset, ensuring greater diversity and robustness in the evaluation process. The table data collection was based on two primary criteria as outlined in Table 1: (1) Data type (continuous, time-series), and (2) Data complexity (Single/Multi-head, Error rate, Data scale). The appropriate visualization method differs based on the type of data, which has a direct relationship with human preference (Wongsuphasawat et al., 2015), (Moritz et al., 2019b). Additionally, when visualizing complex data within the same type, it is essential to apply methods that mitigate visual confusion (Rougier et al., 2014). Using these two criteria, we conducted experiments to evaluate chart quality across diverse chart types.

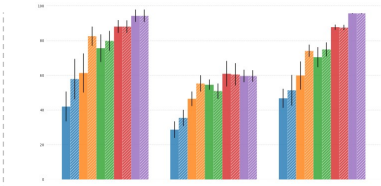
To account for variations in visualization based on analytical objectives, we created golden charts aligned with two randomly chosen analytical purposes. Additionally, to construct the partial dataset, we generated five charts that fail to meet one of the five criteria per analytical purpose and 20 charts that fail to meet two criteria. This resulted in a total dataset of 11,404 chart data samples.

4.2 Model Configuration

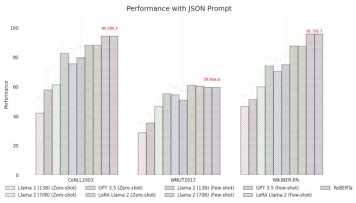
HyVis is built upon the Qwen/Qwen2-VL-7B-Instruct model, utilizing LoRA adapters to train both the human preference and machine-readable evaluation models. The model training and evaluation were conducted using eight NVIDIA A6000 GPUs, with hyperparameter settings detailed in Ta-



Golden Data



Case 1.
w/o *Completeness, Numericalization*



Case 2.
w/o *Represenetation*

Model	Criteria	Completeness	Optimization	Emphasis	Represenetation	Numericalization	Score
GPT-4o	*Golden*	✓	✓	✓	✓	✓	1.0
	Case 1	✗	✓	✗	✗	✗	0.6
	Case 2	✓	✓	✗	✓	✓	0.6
Claude-3.5 -sonnet	*Golden*	✓	✓	✓	✓	✓	1.0
	Case 1	✗	✓	✗	✓	✗	0.8
	Case 2	✓	✓	✓	✓	✗	0.6
Ours	*Golden*	✓	✓	✓	✓	✓	**1.0**
	Case 1	✗	✓	✓	✓	✗	**1.0**
	Case 2	✓	✓	✓	✗	✓	**1.0**

Figure 5: Example results of the Human preference evaluator. The highlighted criteria indicate evaluations that do not match the label.

Models	**Criterias (*Acc* ↑)**				
	Completeness	Optimization	Emphasis	Representation	Numericalization
Prompt-based eval					
GPT-4o	0.72	0.68	0.62	0.65	0.65
Claude-3.5-sonnet	0.75	0.64	0.66	0.70	0.71
Trained eval					
Ours	**0.91**	**0.88**	**0.74**	**0.70**	**0.86**

Table 3: Comparison of accuracy for each criterion of human preference. GPT-4o and Claude-3.5-sonnet use prompt-based evaluation, while our model is specifically trained for evaluation.

ble 2. To compare the performance of the trained model, we evaluated chart quality using open-source models capable of chart analysis (UniChart, ChartInstruct, TinyChart (Zhang et al., 2024) as well as closed-source models (GPT-4o, Claude-3.5-sonnet).

5 Experimental Results

In this section, we introduce the experiments designed to validate the effectiveness of the HyVis framework in assessing the quality of data visualization. The objectives of our experiments are as follows:

- Demonstrate the reflection of human preference by evaluating the rank exact match between the Human preference evaluator and human rankings.
- Compare baseline VLM performance with S_m to verify the superiority of the machine readable evaluator.
- Analyze the correlation between the Hybrid Score, S_h, and S_m.
- Investigate the relationship between Chart type, complexity, and S_m to validate the reliability of the Machine Readable Score.

5.1 Performance of the Human Preference Evaluator

To evaluate the HyVis framework's capacity for quality determination, it is essential to compare human preference assessments alongside model-based metrics. As shown in Table 3, the human preference evaluator effectively classifies charts based on user-oriented criteria. Figure 5 compares

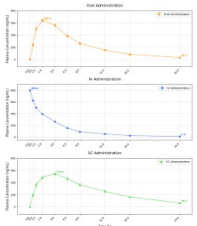
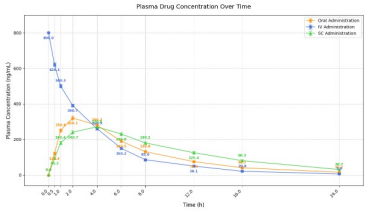
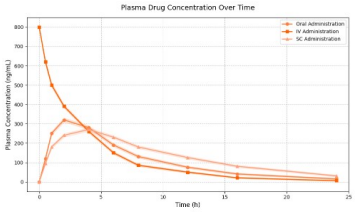
Model	Chart Image (H_d)		
	 (a) Golden Data	 (b) Partial Data <*w/o Optimization*>	 (c) Partial Data < *w/o Represenetation, Numericalization*>
Machine Readable Score (Residual Error)			
Multi-LLMs (Label Score)	*0.913 (-)*	*0.864 (-)*	*0.463 (-)*
GPT-4o	**0.869 (0.044)**	0.847 (0.017)	0.798 (0.335)
Claude-3.5-sonnet	0.861 (0.052)	**0.849 (0.015)**	0.726 (0.263)
ChartInstruct	0.839 (0.074)	0.842 (0.022)	0.612 (0.149)
TinyChart	0.854 (0.059)	0.810 (0.054)	0.637 (0.174)
ChartGemma	0.841 (0.072)	0.831 (0.033)	0.663 (0.200)
Ours	0.842 (0.071)	0.848 (0.016)	**0.579 (0.116)**

Figure 6: Comparison of machine readable scores across models and error differences with Multi-LLMs' label scores. The **bold** scores indicate the lowest error values.

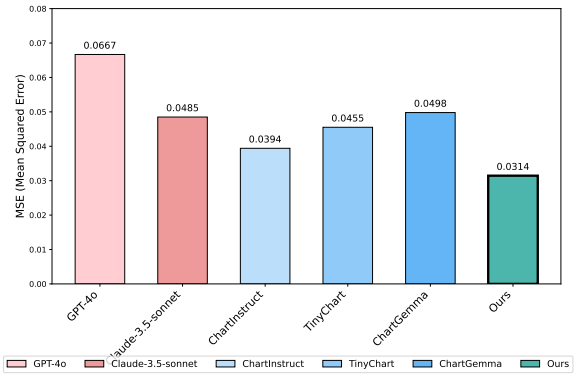


Figure 7: Comparison of the performance between the machine readable evaluator and baseline models.

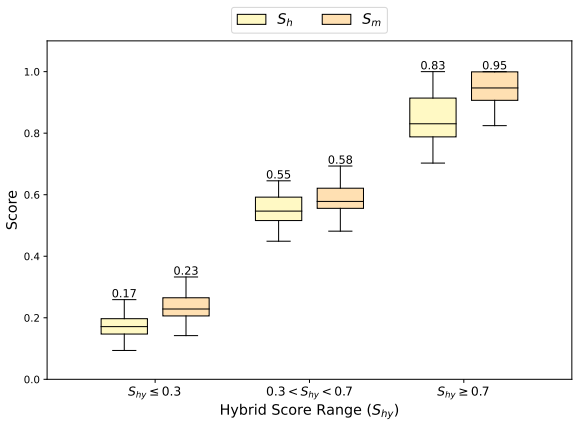


Figure 8: Verification of the proportional relationship between the hybrid score(S_{hy}) and the human preference(S_h)/machine readable(S_h) scores.

the evaluation results across different criteria for the baseline models. While all models accurately classified the golden dataset, partial data assessments revealed shortcomings in evaluating the *Emphasis* and *Numericalization* criteria. Across both the human preference dataset and augmented data, our proposed model achieved the highest alignment with human evaluations, outperforming the closed-model baseline. Details of the evaluation prompts are provided in Appendix A.

5.2 Performance of Machine Readable Evaluator

The machine readable evaluator was assessed by measuring the agreement between responses to FC/Anls task queries generated by GPT-4o and the original table data. As shown in Figure 7, our model achieved the lowest *mean squared error* (MSE) at 3.14%, outperforming other models by up to 3.53%. Figure 6 illustrates the S_m scores across different data structures. While the closed-LLM models exhibited the lowest error rates for high-label-score datasets (a) and (b), our proposed model outperformed them in dataset (c), where both closed-LLM models produced higher errors. This result confirms that our model's ability to learn negative factors enables more precise evaluations

Models	Parameters	Δ=0.6		Δ=0.2	
		ChartQA	Chart-to-Text	ChartQA	Chart-to-Text
Close-source models					
GPT-4o	-	0.7729	0.9138	0.7762	0.8629
Claude-3.5-sonnet	-	0.7972	0.8664	0.7695	0.8021
Open-source models					
Unichart	1B	0.6123	0.7112	0.5983	0.7051
ChartInstruct	7B	0.6234	0.9324	0.6357	0.8073
Qwen2-VL	2B	0.7105	0.9487	0.6902	0.8260
ChartGemma	3B	0.7356	0.8932	0.7123	0.8305
TinyChart	3B	0.7321	0.9621	0.7158	0.8378

Table 4: Exact matching of aligned machine-readable scores with chart understanding task performance. The symbol Δ denotes the difference in the machine-readable score.

in machine-readable chart quality assessment.

5.3 Hybrid Evaluation Analysis

Finally, the hybrid score must demonstrate that it produces chart outputs which are both easily interpreted by models and comprehensible to humans. In Figure 8, we quantitatively show that S_{hy} is proportional to both S_h and S_m, confirming that data points with high human preference also possess high machine-readable quality. This underscores the synergy between human interpretability and machine interpretability in chart design.

5.4 Validity of the Machine Readable Score

Our experimental results reveal a strong correlation between the machine-readable score and overall performance on the chart understanding task, thus validating the 'machine readable' metric proposed under the HyVis framework. The dataset constructed through this experiment, along with the evaluation tasks used to assess machine readability, confirms that this scoring approach effectively captures meaningful differences.

In the experiment summarized in Table 4, when the difference in S_m reached 0.8, all models achieved an Exact Match score of at least 70%. Notably, ChartQA recorded a lower match rate compared to Chart-to-Text, likely because it attained high accuracy for simpler questions regardless of variations in S_m, resulting in an overall lower match rate.

6 Conclusion

This study proposed HyVis, a novel framework for evaluating data visualization quality. Unlike traditional human-centered chart evaluation approaches, HyVis incorporates both human interpretability and machine readability, focusing on how AI models analyze charts. By leveraging a multi-LLM collaborative structure, HyVis provides a comprehensive assessment of chart quality and calculates a hybrid score to verify whether the chart aligns with the data analysis objectives.

Experimental results show that HyVis outperforms existing chart evaluation methods using LLMs, offering higher quality assessments from both human preference and model interpretation perspectives. The framework ensures that charts provide optimal visual information for both humans and AI models. Furthermore, the hybrid score was validated as a meaningful metric, integrating human evaluation standards and machine analytical performance.

Future research could explore expanding HyVis to generate and evaluate charts optimized for human preferences using generative AI. This study offers a new direction for improving the quality of visual information that AI models can use, contributing to the expansion of the data visualization paradigm from human-centered to AI-human collaborative models.

Limitation

Although the HyVis framework evaluates visualization quality by integrating both human preference and machine-readable criteria, its reliance on existing chart understanding models restricts the range of chart types it can effectively assess. Specifically, the machine-readable component is derived from the performance of models on predetermined tasks, limiting adaptability to novel or less common visualization formats. One promising avenue

for expanding chart coverage involves adopting a reinforcement learning approach to validate task queries and labels. Future work should therefore explore the development of reinforcement learning-based machine-readable evaluators, aiming to both broaden the array of chart types and enhance overall evaluation performance.

References

Panayiotis Andreou, Christos Amyrotos, Panagiotis Germanakos, and Irene Polycarpou. 2023. Human-centered information visualization adaptation engine. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 25–33.

Raissa Barcellos, José Viterbo, and Flavia Bernardini. 2022. A process for improving the quality and interpretability of data visualizations. *Univers. Access Inf. Soc.*, 23(2):779–794.

Michelle A. Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2016. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528.

Yue; Chen Xuanjing; Huang Xuanjing Cao, Yixin; Zhang. 2024. Llm-collab: Cooperative ai agents for complex task solving. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page TBD.

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. *arXiv preprint arXiv:2311.16483*.

Linmei Hu, Duokang Wang, Yiming Pan, Jifan Yu, Yingxia Shao, Chong Feng, and Liqiang Nie. 2024. Novachart: A large-scale dataset towards chart understanding and generation of multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3917–3925.

Zijie Li, Yifei Zhang, Yifan Wei, Yixuan Wu, and Qianwen Yang. 2023. Towards automatic data visualization: A survey of generative models. *arXiv preprint arXiv:2305.02618*.

Wu;MingChen Liang, YaoBo;Tong. Multi-agent debate for complex reasoning tasks. In *NAACL Conference Proceedings*, page TBD.

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.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023b. Matcha: Enhancing visual language pretraining with math reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language model for chart comprehension. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10387–10409, Bangkok, Thailand.

Dominik Moritz, Chenglong Wang, Greg L. Nelson, Halden Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer. 2019a. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448.

Dominik Moritz, Chenglong Wang, Gregory Nelson, Halden Lin, Adam Smith, Bill Howe, and Jeffrey Heer. 2019b. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press.

Wenyi Ouyang. 2024. Data visualization in big data analysis: Applications and future trends. *Journal of Computer and Communications*, 12(11):76–85.

Nicolas P Rougier, Michael Droettboom, and Philip E Bourne. 2014. Ten simple rules for better figures.

Jonathan Schwabish. 2021. *Better data visualizations: A guide for scholars, researchers, and wonks*. Columbia University Press.

Edward R Tufte and Peter R Graves-Morris. 1983. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT.

Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658.

Aoyu Wu, Yun Wang, Xinhuan Shu, Dominik Moritz, Weiwei Cui, Haidong Zhang, Dongmei Zhang, and Huamin Qu. 2021. Ai4vis: Survey on artificial intelligence approaches for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5049–5070.

Bowen Yu and Cláudio T Silva. 2019. Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE transactions on visualization and computer graphics*, 26(1):1–11.

Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2021. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7:3–36.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1898, Miami, Florida, USA. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Closed-LLM Evaluation Prompt

This appendix provides the detailed prompt structure used in the evaluation tasks described in the main text. Each prompt is designed to capture both human preference criteria and machine interpretability aspects, ensuring consistency across different dataset splits.

By incorporating these elements, the evaluator consistently measures the extent to which a chart meets the human-centered requirements outlined in Section 5.1. In combination with the model interpretability prompts, these instructions form the basis of the Hybrid Score discussed throughout the paper.

Human Preference Evaluation Prompt

For data analysis using the **CHART**, the following five requirements must be satisfied:
1.Completeness: Reflect all categories from the **TABLE** data in the necessary chart elements (axes, legend, title, etc.).
2.Optimization: Design a chart structure that effectively addresses the objective of the data analysis.
3.Emphasis: Highlight the key information related to the **GOAL_INSTRUCTION**.
4.Representation: Accurately represent the chart data without distortion or confusing visual elements.
5.Numericalization: Present chart labels and item values in an intuitive form.
Please evaluate whether the attached **CHART** image meets each of these five criteria.
===
###TABLE###
{table}

###CHART###
{chart_image}

###GOAL_INSTRUCTION###
{goal_instruction}

Figure 9: Human preference evaluation prompt configuration.