Show and Tell: Exploring Large Language Model's Potential in Formative Educational Assessment of Data Stories

Naren Sivakumar* University of Maryland Baltimore County
Lujie Karen Chen† University of Maryland Baltimore County
Pravalika Papasani‡ University of Maryland Baltimore County
Vigna Majmundar § University of Maryland Baltimore County
Jinjuan Heidi Feng ¶ Towson University
Louise Yarnall ‖ SRI International
Jiaqi Gong ** University of Alabama

Abstract

Crafting accurate and insightful narratives from data visualization is essential in data storytelling. Like creative writing, where one reads to write a story, data professionals must effectively "read" visualizations to create compelling data stories. In education, helping students develop these skills can be achieved through exercises that ask them to create narratives from data plots, demonstrating both "show" (describing the plot) and "tell" (interpreting the plot). Providing formative feedback on these exercises is crucial but challenging in large-scale educational settings with limited resources. This study explores using GPT-4o, a multimodal LLM, to generate and evaluate narratives from data plots. The LLM was tested in zero-shot, one-shot, and two-shot scenarios, generating narratives and self-evaluating their depth. Human experts also assessed the LLM's outputs. Additionally, the study developed machine learning and LLM-based models to assess student-generated narratives using LLM-generated data. Human experts validated a subset of these machine assessments. The findings highlight the potential of LLMs to support scalable formative assessment in teaching data storytelling skills, which has important implications for AI-supported educational interventions.

1 Introduction

Data storytelling [22], or communicating with data [18], is a critical skill for data scientists or data analysts. In addition to the technical skills of programming and analysis, they must know how to effectively communicate the complex meanings of data through data visualization and craft compelling narratives. Data storytelling extends beyond traditional data visualization [12], aiming to effectively communicate data insights to stakeholders with the intent to prompt actions [3].

The data storytelling workflow involves several interconnected subprocesses that are closely aligned with the data science workflow [5, 13, 18]. Typically, this process begins with a data question, which guides the creation of visualizations designed to answer these inquiries. Analysts must then interpret these visualizations, making sense of the data to determine whether further investigation is warranted. The product of the sense-making process can be narratives or data stories that accurately communicate the meaning of the data visualization. Data storytelling involves several steps [18], including filtering, grouping, and ordering, all of which rely heavily on the analyst's ability to understand data visualizations at various stages [4]. This understanding ranges from a shallow level of "reading the data" (i.e., simply describing what is shown) to a more involved deeper cognitive process of "reading between the data" and "reading beyond the data" (i.e., interpreting the meaning that the data tells). Sense-making, particularly through exploratory data analysis, is a driving force in the data science and data storytelling workflow, which enables the creation of compelling narratives.

Similar to the strategy of "read to write" in writing training - where students achieve a certain level of reading comprehension before developing writing skills - a similar approach could be adopted in data storytelling training. Sense-making, or chart-reading skills, can be viewed as a "reading" comprehension skill essential for supporting the "production skills" of creating complete sets of data stories. In recent years, there has been growing interest in training data storytelling skills, and courses on data storytelling or data visualization have become popular. However, there is still no principled framework for effectively supporting students in developing these skills. There is little discussion on how component skills such as sense-making can be developed and supported in data science or data analysis training curricula. One challenge educators face is providing timely, targeted feedback or formative assessment on the sense-making or chart-reading work products in the form of narratives that students produce.

In this paper, we present a computational framework for automatically assessing narratives or data stories generated from data plots as the result of the sense-making process. This assessment provides feedback on both the depth and quality of students' sense-making outputs. To achieve this, we first instructed GPT-4o [1](using zero-shot, one-shot, or two-shot configurations) to generate narratives or data stories sentence-by-sentence for a series of 12 data plots with varied chart types and difficulty levels. Expert evaluations show that these narratives consistently maintain high quality. Furthermore, we observed that GPT-4o could perform self-evaluation regarding the depth of the narrative and quality assessment with a certain degree of reliability.

Based on these results, we constructed a training dataset for machine learning models using the LLM-generated narratives for the 12 data plots. We used labels generated from self-evaluation results (for narrative depth, due to their high reliability) or expert annotations (for quality assessment, due to the lower reliability of self-evaluation). This model was then used as an educational assessment tool to evaluate students' responses to five of the data plots collected from a sense-making or story-reading exercise designed for an undergraduate introductory data science course. Independent human evaluation on a sample of the responses demonstrated promising results with an accuracy of around 80% and an Area Under the Curve (AUC) of 77% for providing feedback on narrative depth and about 90% accuracy for quality assessment.

*e-mail: narens1@umbc.edu
†e-mail: lujiec@umbc.edu
‡e-mail: io11937@umbc.edu
§e-mail:vignam1@umbc.edu
¶e-mail: jfeng@towson.edu
‖louise.yarnall@sri.com
**e-mail: jiaqi.gong@ua.edu

[1]https://openai.com/index/hello-gpt-4o/

Our study fills a gap in research by exploring a scalable solution to support the educational assessment of narratives or data stories generated from data visualization. This work focuses on the critical cognitive process of sense-making, which underlies many data storytelling sub-processes. Once further validated, the machine learning framework explored in this study has the potential to offer automatic formative assessments of students' responses to data story-reading exercises in real-world educational contexts and significantly contribute to the robust development of students' data storytelling skills on a large scale.

2 Related Work

2.1 Sense-making Stages of Data Visualization

The ability to understand or make sense of data visualization is essential in the digital age, both in the workplace and in everyday life. Crucial decisions regarding personal health, financial management, and career choice heavily involve the understanding and interpretation of data visualization [6]. However, roughly 29% of Americans have poor data visualization literacy [21]. It is also a critical component of the data storytelling competency.

The sense-making process of data visualization includes three stages that were broadly described as reading the data, reading between the data, and reading beyond the data [4].

- In the first stage, viewers identify the various elements presented in the visualization and make sense of each element (e.g., title, X-axis, Y-axis, labels, values, symbols). They also identify and interpret the value encoded elements such as shape, size and position. In addition, viewers need to understand the type of the visualization that determines the overall approach to interact with and interpret the visualization. Saliency is particularly important in this stage because it directs readers' attention and cognitive effort to different elements of the visualization [19].
- In the second stage, reading between the data, viewers compare and connect multiple elements in the visualization to identify patterns, differences, and outliers. This may include activities such as dentifying extremes, finding exact values, anomalies, or clusters, making estimates, or noting ranges of values. This stage involves an iterative process of searching through the data, visually encoding it, and then mapping the information presented to the viewer's internal mental model of the visualization [15].
- The third stage -reading beyond the data – is the most demanding stage of the sense-making process because it is heavily driven by previous knowledge. In this stage, viewers need to draw more heavily upon information in their long-term memory (e.g., domain knowledge related to the topic of the visualization, more advanced numeracy skills) to further engage with the data visualization and perform inference-making tasks. It is during this stage that the viewers leverage the information in the visualization to reflect upon their previous view or behavior and gain potential insights for decision making and future actions.

It is important to note that the three stages of reading a data visualization are not necessarily a linear process. Rather, viewers may move back and forth between stages as they as they take in more information and make new connections both within and beyond the visualization. This three stage sense-making model serves as the foundation for the evaluation of both the LLM- and human generated descriptions of data visualization.

In this paper, we refer to the first stage as **"show,"** while the second and third stages are referred to as **"tell"** the stories. In the data story-reading exercises, students are encouraged to move beyond just "showing" or describing the elements of the plot and strive to "tell" or explain the stories or messages embedded in the given data visualization or data plot.

2.2 LLM-related work in sense-making of data visualization

In recent years, there has been a surge of interest in leveraging LLMs to automate certain components of the data storytelling pipeline. A recent survey by He et al. [8] summarizes the nine tasks across four stages, including data, narration, visualization, and presentation. Among these nine tasks, those relevant to sense-making are those belonging to interpretation and comprehension. Specific related tasks addressed in the research include:

- **Chart Summarization** [20, 10, 14], which takes a chart as input and outputs a paragraph summary that could include a mixture of sense-making outputs, including both "show" and "tell.";
- **Chart Question and Answer** [16, 23], where the LLM is asked to answer specific questions. Whether or not the output is related to "show" vs. "tell" is highly dependent on the questions;
- **Chart Captioning** [24] which requires the LLM to generate concise captions to highlight insights that can be of flavor either "show" or "tell";
- **Fact-checking** [2, 1], a task focused on checking factual correctness, which can be at either the "show" or "tell" level.

There are several recent datasets or models focused on chart understanding and reasoning. For example, the OpenCQA dataset [9] crowdsources questions of various types; a random sample of 100 questions shows that only 20% involve comparisons, which is a type of "tell," while the rest are at the "show" level. ChartBench [25] is a benchmarking dataset that includes 66.6k charts and 600k question-answer pairs, categorizing Q&A tasks into five categories that fall into two groups at the perception level and conceptual level, similar to the notions of "show" and "tell," respectively. While three of the categories—chart recognition, value extraction, and number Q&A—belong to the "show" task, value comparison, and global conception are similar to "tell." ChartInstruct[17] uses an instruction tuning approach to generate a dataset of additional tasks beyond those existing tasks. Some tasks have a clear notion of "telling," e.g., data correlation, future forecasting, anomaly detection, while others remain at the "show" level, such as chart title extraction. ChartLlama [7] is a recent LLM evaluated on several related tasks. Among the tasks demonstrated by ChartLlama relevant to our use case are chart description and Q&A. However, these tasks do not seem to explicitly differentiate between surface level of understanding (i.e., show) and deep understanding (i.e., tell). From the examples, the sense-making appears to operate at stage 1 (or "show" level).

In summary, there are emerging lines of research using LLMs for sense-making tasks or chart understanding, and several patterns emerge: (1) Most of the chart understanding tasks do not explicitly differentiate between "show" and "tell," ; (2)Comprehensive, relevant evaluations of LLMs regarding the quality of sense-making output in terms of "show" and "tell" have yet to be conducted; (3)The open-ended task of chart summarization is most similar to our use case; however, the answer is not returned at the sentence level, which makes it challenging to provide targeted feedback to students.

3 METHOD

3.1 Data plots used in this study

Table 1 gives an overview of the 12 plots that were used to prompt the LLM to generate narratives or data stories. This set of plots includes a mixture of different plot types, such as bar charts, line plots, and scatter plots. There are a few more complex data plots, like heat maps and grouped bar charts. Those chart types are selected as they are commonly encountered in introductory data science classes. The topics of these data plots cover a variety of domains, such as public health, climate science, education, and scientific studies, catering to students' diverse interests. Most of the data plots are extracted from news articles. These plots cover a range of complexity or difficulty levels based on factors such as the number of data points or data series encoded in the plot, the chart type complexity, and the subtleties of the message. Please refer to the supplemental for details of the 12 plots. The instructor selected five plots for data story-reading exercises in the data science courses, considering the course load and students' growing competency in story-reading exercises. Most of the chosen plots fall within the medium or moderate difficulty range. This selection reflects a level of difficulty that is adequately challenging for students, thus providing ample opportunities for constructive feedback and improvement.

Table 1: Data Visualization/Plot used in this study

Plot Name	Student Use	Level	Level Desc.	Chart Type
Vaccine	No	1	easy	bar
Youtube	Yes	1	easy	heatmap
Degree	No	1	easy	bar
Meaningful Life	No	1	easy	dot
Solar	No	2	easy/medium	line
Walk Dog	Yes	3	medium	line
hurricane	No	3	medium	bar
Time Use	Yes	3	medium	area
STEM	Yes	4	medium/difficult	bar
Wealth	Yes	4	medium/difficult	line
MAP	Yes	4	medium/difficult	dot
Vulnerability	No	5	difficult	dot

3.2 Student Data Collection

The student story-reading exercise data was collected from three cohorts of undergraduate introductory data science courses from 2023 to 2024 at a public university on the East Coast of the U.S. The data collection was approved by the university's Institutional Research Board. The objective of this exercise is to help students enhance their skills in sense-making data visualization toward the overall learning objective of mastering data storytelling. These exercises are given throughout the semester, where students are asked to write a paragraph of narratives to both "show" and "tell" the "data stories" given a data plot, following the definition described in section 2.1. We collected 435 responses in the form of narratives or data stories to five of the twelve data plots in Table 1. We annotated about 20% of the students' data stories, sentence-by-sentence, on depth ("show" vs. "tell") and quality of narratives.

3.3 LLM used in this study

In our experiments, we chose GPT-4o, OpenAI's most recent multimodal LLM model, which can return a JSON response with a maximum length of 300 tokens through API calls. We selected this LLM because it is the latest and fastest multimodal model that can process image files. The GPT API was used to ensure reliable communication and fast responses and to facilitate the large-scale experimentation required in this study.

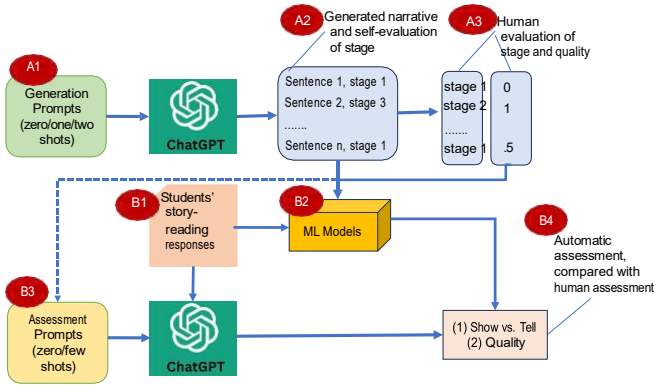

Figure 1: Overview of Experiments

3.4 Overview of Experiments

Fig. 1 provides an overview of the experiments conducted in this study. We choose to generate and evaluate the narrative on a sentence-by-sentence level rather than full text. Evaluating "show" versus "tell" on a sentence-by-sentence basis offers the advantage of providing specific, targeted feedback. In contrast, when examining a full text, the intermingling of "show" and "tell" can make it difficult to distinguish between narrative elements that reflect shallow versus deep cognitive processes.

There are two categories of experiments:

The first category (A1-A3) evaluates the LLM's capacity to generate high-quality narratives to both "show" and "tell" the stories. The workflow begins with prompting the LLM to generate stories (A1), given a data plot. We experimented with three versions of generation prompts with various configurations of instructions and examples, as illustrated in Table 2. The LLM is instructed to generate data stories sentence-by-sentence to support feedback at a fine-grained level. After each sentence is generated, the LLM is also prompted to self-evaluate at one of the three stages, as defined in section 2.1 (A2).

To validate the LLM-generated output, two human experts independently evaluated it with respect to the stage (blind to the LLM's self-evaluation of the stage) and the quality of the narrative. The evaluation was conducted on a scale of 0, 0.5, and 1, which assesses the degree of consistency between the narrative and the data plot. A score of 0 indicates a completely wrong narrative, 1 indicates a completely correct narrative, and 0.5 indicates a partially correct narrative.

The second category (B1-B4) includes steps to train models to eventually provide automatic assessments of students' responses to story-reading exercises. Based on this evaluation, we assessed the performance of two models: a machine learning-based model (B2) and an LLM-based model (B3). Both models will estimate the depth of the narrative regarding whether each sentence is "show" vs. "tell" and the quality of the narrative. The machine-generated assessment was then validated against human evaluation.

To train a machine learning-based assessment model that can differentiate between "show vs tell", we utilized the LLM-generated narratives and self-evaluation of stages as ground-truth labels (binarized into "show" vs. "tell" to be consistent with what students are instructed to do in the classroom settings). For each sentence, we first converted it into lowercase and removed punctuation to eliminate unnecessary complications. The sentences were then split into tokens, and stop words were removed. Following tokenization and

Table 2: Three types of generation prompts used to prompt LLM to generate data stories.

Generation Prompts	**Instruction** (Definition of three-Stage)	**Shots** (Show & tell example)
Zero Shot	Yes	None
One Shot	Yes	One
Two Shots	Yes	Two

lemmatization, feature extraction was performed using TF-IDF vectorization. We extracted additional NLP features, followed by classical ML classification models such as logistic regression, Support Vector Machine (SVM), or Random Forest(RF).

Additionally, the ML model was trained to estimate the quality of the narrative using the labels generated through human experts' annotation, as described in A3. In the preprocessing step, we handled the missing values, encoding the categorical data and then the text vectorization to convert the text into numerical features. Machine learning models used for training are Logistic Regression, SVM, and Naive Bayes, and the models were evaluated using 5-fold cross-validation and Leave One Plot Out setups. The models are then tested using the students' data.

To train an LLM-based assessment model, we designed a series of assessment prompts (see Appendix 7.2) to instruct the LLM to provide feedback on whether a given sentence is "show" vs. "tell" and the quality of the narrative, as previously described. We experimented with a few-shot configuration, providing examples of assessments drawn from the GPT-generated stories and self-evaluation (for show and tell, A2) and human evaluation (for quality assessment, A3). We specifically focused on "far transfer" use cases where the example plot did not overlap with the test plot. This approach more closely resembles real-world educational settings, where models are preferred to be used off-the-shelf rather than needing to be retrained with new plots.

4 Results

4.1 Overview of LLM-generated narrative sentences

The LLM generated a total of 351 narrative sentences from three prompting strategies (Table 2) using 12 data plots as described in Table 1. Among these, 118 sentences were generated from zero-shot, 130 from one-shot, and 113 from two-shot configurations. Figure 2 summarizes the proportion of LLM-generated sentences belonging to each stage, comparing shot configurations (left plot) and data plot and chart complexity levels (right plot).

We noted that, in general, about half of the sentences generated are at stage 1 (or "show" level), while a smaller proportion is generated at stage 3 (advanced "tell" level). The zero-shot configuration, where no specific examples were given, tends to generate relatively more "tell" sentences than other shot configurations. From the right plot, we observed that the proportion of show vs. tell sentences varies from plot to plot. For example, some plots like "Solar" generated sentences mainly at the "show" level, while for the plot named "STEM," the majority of the sentences are at Stage 2 or 3 (or "tell" level). We did not observe particular patterns related to chart complexity levels.

4.2 Human Evaluations of LLM-Generated Stories

4.2.1 Evaluating LLM's self-evaluation of three stages

Figure 3 summarizes the inter-rater reliability (Cohen's Kappa) calculated from ratings on narrative sentences generated with various prompting strategies (zero-shot, one-shot, or two-shot) as well as those pooled from all shot configurations. As shown, the overall agreement between LLM and human evaluators is high; similarly, the agreement between evaluators is high. In addition, there is a

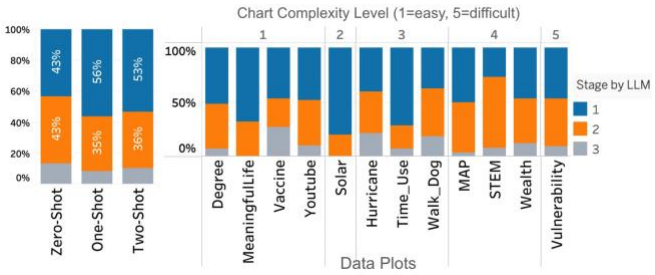

Figure 2: Percentage of LLM-generated narrative sentences belonging to each of the three stages (per LLM self-evaluation), categorized by shots configuration in generation (left plot) and by chart complexity and data plots (right plot).

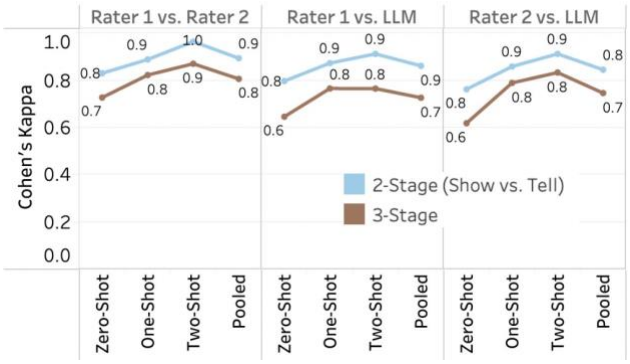

Figure 3: Inter-rater reliability (Cohen's Kappa) calculated from ratings on narrative sentences generated with various prompting strategies, categorized by pairwise comparison among two raters and with LLM, comparing ratings with respect to three stages (Stage 1, 2, or 3) or Show vs. Tell (Stage 1 vs. Stages 2 & 3).

notable trend that an increasing number of shots corresponds to the improved consistency of ratings between raters and LLM and between raters. We also noted that all pairs of inter-rater reliabilities improve when ratings are evaluated at two categories by collapsing stages 2 and 3 into the "tell" category.

4.2.2 Evaluating the quality of LLM-generated stories

Figure 4 summarizes the results of quality ratings of LLM-generated narrative sentences. The overall quality rating is high, within the range of .85 and 1. As shown in the left plot, the quality rating seems to decrease with the increase in difficulty, particularly for plots with difficulty levels beyond the easy level. This pattern is consistent across the different shots. The inter-rater reliability (as shown in the right plot) is relatively low, with the two-shot configuration showing the highest reliability. The squared-weighted method gave slightly higher reliability than the linear-weighted method.

4.3 Performance of Machine Learning-based Assessment Model

4.3.1 ML model performance in discriminating between "show" and "tell"

Figure 5 and Figure 6 present the machine learning model performance with respect to accuracy and Area Under Curve, respectively, from various setups of the experiment. The "LLM-generated stories" version refers to internal validation results using LLM-generated stories and the LLM's self-evaluation as described in Step

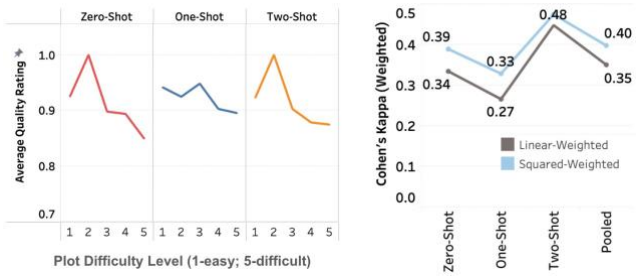

Figure 4: Quality Rating (averaged between two raters) by shots and plot difficulty level (left plot) and the inter-rater reliability (Cohen's weighted Kappa) by shots, comparing weighting methods (linear vs. squared).

Version	Experiment	Model	Training Data: Combined	Zero Shot	One Shot	Two Shot
LLM Generated Stories	Cross Validation	Logistic Regression	0.85	0.76	0.65	0.73
		Naive Bayes	0.84	0.69	0.70	0.74
		SVM	0.87	0.73	0.65	0.75
	Leave One Plot Out	Logistic Regression	0.87	0.77	0.65	0.87
		Naive Bayes	0.86	0.86	0.76	0.85
		SVM	0.85	0.76	0.63	0.86
Student Generated Stories	Train/Validate	Logistic Regression	0.81	0.73	0.57	0.75
		Naive Bayes	0.75	0.77	0.58	0.70
		SVM	0.77	0.74	0.56	0.73

Figure 5: ML model performance (accuracy) in discriminating whether a given narrative sentence is "show" versus "tell".

A2 in section 3.4. This training set includes 351 sentences generated from 12 data plots. We experimented with two versions of cross-validation: the 5-fold cross-validation ignores the fact that samples are grouped according to data plots, thus its performance is likely overestimated due to potential data leakage [11] from the split. In the "Leave One Plot Out" experiment, the training and validation plots do not overlap. This emulates application scenarios where the model will be used to estimate a new plot that the model has never been exposed to before, which is a more relevant use case in real-world educational settings. Models of this kind can be used off-the-shelf without needing to be retrained with new plots given to students.

The lower part of the table presents the results where the models are trained on LLM-generated stories and self-evaluation of stages and validated on 20% of students' generated stories, sentence-by-sentence. Additionally, we explored how performance varies with the amount of training data provided to the model. We experimented with three representative machine-learning models based on the extracted NLP features.

Several interesting patterns were noted: (1) Overall, there is little difference in performance across different machine learning models, with no single model distinctly better than others. (2) When the model is trained on a combined dataset generated by zero-shot, one-shot, and two-shot settings, it performs the best. (3) Performance does not significantly deteriorate (and sometimes even increases) when the leave-one-plot-out experiment protocol is applied. (4) Model performance slightly deteriorates when applied to real-world student data. Despite this, the model performance is reasonably good; for example, the best-performing model, Logistic Regression, achieves an accuracy of 0.81 and an AUC score of 0.77, both with the combined training set. This level of performance is acceptable considering the low-stakes formative assessment use cases.

Version	Experiment	Model	Training Data: Combined	Zero Shot	One Shot	Two Shot
LLM Generated Stories	Cross Validation	Logistic Regression	0.93	0.85	0.79	0.86
		Naive Bayes	0.92	0.79	0.78	0.82
		SVM	0.94	0.82	0.79	0.83
	Leave One Plot Out	Logistic Regression	0.95	0.94	0.92	0.95
		Naive Bayes	0.95	0.94	0.93	0.95
		SVM	0.95	0.94	0.94	0.95
Student Generated Stories	Train/Validate	Logistic Regression	0.77	0.68	0.69	0.78
		Naive Bayes	0.77	0.67	0.70	0.77
		SVM	0.76	0.67	0.68	0.77

Figure 6: ML model performance (Area Under Curve) in discriminating whether a given narrative sentence is "show" versus "tell"

Version	Experiment	Model	Training Data: Combined	Zero Shot	One Shot	Two Shot
LLM Generated Stories	Cross Validation	Logistic Regression	0.86	0.81	0.92	0.90
		Naive Bayes	0.83	0.74	0.86	0.79
		SVM	0.81	0.80	0.92	0.88
	Leave One Plot Out	Logistic Regression	0.87	0.83	0.92	0.88
		Naive Bayes	0.60	0.60	0.79	0.64
		SVM	0.83	0.83	0.92	0.89
Student Generated Stories	Train/Validate	Logistic Regression	0.89	0.92	0.91	0.91
		Naive Bayes	0.89	0.92	0.92	0.89
		SVM	0.80	0.82	0.89	0.84

Figure 7: ML model performance (accuracy) in estimating the quality of the stories

4.3.2 ML model performance in estimating the quality of the stories

Figure 7 summarizes the quality evaluation performance of three different models (Logistic Regression, SVM, Naive Bayes) over two categories of experiment, those based on LLM-generated stories and student-generated stories. There are several notable observations: 1) For internal validation experiments with LLM-generated stories, Logistic Regression and SVM slightly outperform other models in cross-validation and Leave-One-Plot-Output experiment setup, achieving the highest accuracy with one-shot training data. (2) There is no notable performance deterioration for Logistic Regression and SVM. (3) In evaluating student-generated stories, Logistic Regression and Naive Bayes models performed similarly in all scenarios, achieving the highest accuracy of 0.92 with both one-shot and zero-shot training data. The overall results suggest that the one-shot trained model performed consistently well, with the highest accuracy scores across all the models and experiments. Those results indicate that those are feasible models for quality estimation.

4.4 Performance of LLM-based Assessment model

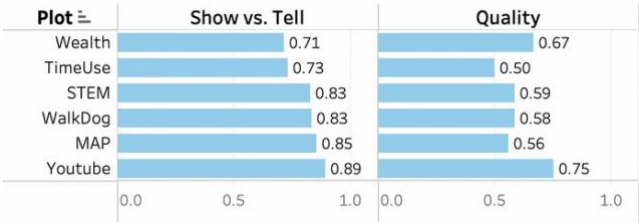

Figure 8: Performance of LLM-based assessment model on a subset of students' data stories, with respect to two assessment tasks: (1) discrimination between "show" vs. "tell" (left plot) and (2) quality assessment (right plot) for a given sentence-by-sentence narrative, aggregated by data plot.

The overall accuracy is 82% when the LLM is used to assess

whether a given sentence is "show" vs. "tell." As summarized in Figure 8, accuracy varies by data plot, with the highest accuracy close to 90% and the lowest about 70%. On the other hand, the overall reliability of quality assessment, as measured by accuracy, is 62%, with the lowest at 50% and the highest at 75%. This set of results suggests that the LLM-based assessment model is not a better option than the ML-based assessment model as described above.

5 Discussion

In this study, we conduct an empirical evaluation of GPT-4o's capability to produce high-quality data stories in the form of narrative sentences given a data plot. The LLM is specifically instructed to demonstrate sense-making abilities to both "show" (describe data faithfully) and "tell" (explain or interpret correctly) the information embedded in the data plot, following a framework of three-stage chart-reading outlined in section 2.1. We also explored approaches to utilize these LLM-generated narratives as training data to build machine learning or LLM-based models to give feedback on students' data story-reading exercises. In this section, we will briefly summarize the results and discuss the implications of this study for scalable assessment in sense-making exercises in real-world educational settings to support the training for data storytelling skills at a large scale.

Firstly, we noted that GPT-4o could generate high-quality narratives consistent with the data plot, although the quality slightly decreases when the chart complexity or difficulty level increases. We also noted that GPT-4o produced fairly reliable self-evaluations of the depth of the narrative concerning the three stages, particularly accurate in differentiating between "show" and "tell." This suggests that GPT-4o can generate useful data story-reading contrasted examples of "show" and "tell" for students while they are learning to make sense of data visualizations and craft narratives, a critical component of data storytelling skills.

Secondly, we observed that the relatively reliable generation of narratives in both "show" and "tell" forms makes it feasible to generate training data for machine learning at a low cost. The machine learning model we built with these data showed reasonable performance, with an 80% accuracy and a 77% AUC score when evaluated on a subset of student-generated stories, and the quality assessment reached a level of 90%. The LLM-based assessment using few-shot "far-transfer" prompting showed a similar level of performance in "show" vs. "tell" discrimination while slightly worse quality-assessment performance compared to the ML-based model. This set of experiments suggests the promise of using LLM to generate training datasets and then using machine learning models to provide formative assessments to students without having to depend on LLM for assessment, which could be expensive when used at a large scale.

When further validated, we envision this generation-assessment pipeline integrating machine learning and LLM as a viable solution to provide timely formative feedback on students' sense-making exercises at a large scale. For example, one use case is to provide support to students' on-demand sense-making requests for given data visualizations; on the other hand, we could use the tool to give students timely feedback on story-reading exercises sentence-by-sentence, for example, on whether students are demonstrating deeper understanding by "telling" the stories rather than merely "showing" the stories. Future work will be useful in understanding the reasoning of the LLM and machine-learning models in sense-making and extracting deeper insights from the data visualizations, which becomes vital to providing useful and actionable feedback to students in developing sense-making skills critical for data storytelling.

6 Conclusion

Sense-making, the cognitive process that involves the understanding or interpreting of data visualizations, also called chart interpretation or comprehension [8], is an essential foundational skill for data storytelling. It enables the crafting of accurate and compelling narratives from data. Sense-making is closely related to other sub-processes, such as the generation of data visualizations, which form an integral part of data exploration and analysis, and the story construction process, all of which support the data storytelling process. The LLM and machine-learning-based automatic assessment pipeline explored in this study may open an avenue for designing AI-inspired educational interventions and formative assessments that support the development of data storytelling skills at a large scale.

7 Appendix

7.1 Generation Prompts

7.1.1 Zero-shot Prompt

This is the zero-shot prompt used to prompt GPT-4o to generate narrative sentences following the definition of the three stages of chart-reading (and Show vs Tell) in section 2.1.

```
In this task, we will give you a plot and your job is to generate a narrative to demonstrate your understanding of the plot. A good narrative will include both show and tell which describe as below:

Stage 1: Show or Describe: this stage is equal to the stage of ''reading the data", and involves various sense-making processes (i.e., taking in and interpreting visual information) to identify the various parts that make up the visualization as a whole

Stage 2 & 3: Tell or explain: this stage is equal to the stages of 'reading between the data' and 'reading beyond the data'. It can be further broken down into Stage 2 (basic tell): Reading between the data: involves comparing and connecting information to identify patterns, differences, and outliers Stage 3 (advanced tell): Reading beyond the data: viewers draw upon and employ their information held in their long-term memory to inform their learning, inference-making, and internal mental model as they engage with a new data visualization.

Please generate your output sentence by sentence, try to generate stage 2 or 3 statement as much as possible. The first sentence will identify chart type as much as you can, if unsure, type ''I am unsure of the chart type" For each sentence, label the stage as integer number 1, 2, 3 as defined above, if unsure, type -1 For each sentence, if applicable, identify the elements of the chart that support your statement, this could include chart title, x-axis, y-axis, data series, color, size, annotation, etc.
```

7.1.2 Few-shot Prompt

In the few-shot prompt configuration, we give the LLM instance more data to work with. We begin by giving it a plot, followed by either a human-generated show and tell, or one generated by a separate instance of GPT. In either case, we aim to give GPT an example of what a show and a tell is, apart from the detailed prompt above, to see if it improves performance in any significant way.

7.2 Assessment prompts

7.2.1 Zero-shot Prompt

This is the zero-shot prompt used to prompt GPT 4o to generate the assessment of a given narrative sentence with regard to (1) the three stages of chart-reading (and Show vs Tell) as defined in section 2.1, and (2) the quality rating with similar scale as provided by human evaluation as described in Section 3.4 (A3).

```
Given the following levels of a show and tell:
Stage 1: Show or Describe: this stage is equal
to the stage of ``reading the data", involves
various sense-making processes (i.e., taking in
and interpreting visual information) to identify
the various component parts that make up the
visualization as a whole
Stage 2 & 3: Tell or explain: this stage
is equal to the stages of 'reading between the
data' and 'reading beyond the data'. It can be
further broken down into Stage 2 (basic tell)
: Reading between the data: involves comparing
and connecting information to identify patterns,
differences, and outliers; Stage 3 (advanced
tell): Reading beyond the data: viewers draw
upon and employ their information held in their
long-term memory to inform their learning,
inference-making, and internal mental model
as they engage with a new data visualization.
Generate a stage number.
Also given that a zero is a factually wrong
statement, a 0.5 is partially correct with respect
to the context, and a 1 is fully correct with
the context given, rate each sentence on this
scale too. Classify the following data as per
the instructions above. Split each entry into
singular sentences based on full stops and give
me an analysis for each sentence. Don't add extra
escape characters either.
```

7.2.2 Few-Shot Prompt

In few-shot learning scenarios, in addition to the prompts above, we also provide shots/examples randomly selected from the training set (i.e., those generated by LLM labeled either self-evaluated by LLM (for stage) or by human evaluators (for quality).

ACKNOWLEDGMENTS

This project is partially supported by National Science Foundation grant # 2302794 and #2302795

REFERENCES

[1] M. Akhtar, O. Cocarascu, and E. Simperl. Reading and reasoning over chart images for evidence-based automated fact-checking. *arXiv preprint arXiv:2301.11843*, 2023. 2

[2] M. Akhtar, N. Subedi, V. Gupta, S. Tahmasebi, O. Cocarascu, and E. Simperl. Chartcheck: An evidence-based fact-checking dataset over real-world chart images. *arXiv preprint arXiv:2311.07453*, 2023. 2

[3] R. Bhargava. Data storytelling studio climate change, 2017. 1

[4] F. R. Curcio. Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education JRME*, 18(5):382 – 393, 1987. doi: 10.5951/jresematheduc.18.5.0382 1, 2

[5] B. Dykes. Data storytelling: The essential data science skill everyone needs. *Forbes Magazine*, 2016. 1

[6] S. El-Toukhy, A. Me´ndez, S. Collins, and E. J. Pe´rez-Stable. Barriers to patient portal access and use: Evidence from the health information national trends survey. *The Journal of the American Board of Family Medicine*, 33(6):953–968, 2020. doi: 10.3122/jabfm.2020.06.190402 2

[7] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 2

[8] Y. He, S. Cao, Y. Shi, Q. Chen, K. Xu, and N. Cao. Leveraging large models for crafting narrative visualization: A survey, 2024. 2, 6

[9] S. Kantharaj, X. L. Do, R. T. K. Leong, J. Q. Tan, E. Hoque, and S. Joty. Opencqa: Open-ended question answering with charts. *arXiv preprint arXiv:2210.06628*, 2022. 2

[10] S. Kantharaj, R. T. K. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022. 2

[11] S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023. 5

[12] R. Kosara and J. Mackinlay. Storytelling: The next step for visualization. *Computer*, 46(5):44–50, 2013. 1

[13] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale. More than telling a story: Transforming data into visually shared stories. *IEEE computer graphics and applications*, 35(5):84–90, 2015. 1

[14] M. Liu, D. Chen, Y. Li, G. Fang, and Y. Shen. Chartthinker: A contextual chain-of-thought approach to optimized chart summarization. *arXiv preprint arXiv:2403.11236*, 2024. 2

[15] U. Ludewig. *Understanding graphs: modeling processes, prerequisites and influencing factors of graphicacy*. PhD thesis, Universita¨t Tu¨bingen, 2018. 2

[16] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2

[17] A. Masry, M. Shahmohammadi, M. R. Parvez, E. Hoque, and S. Joty. Chartinstruct: Instruction tuning for chart comprehension and reasoning, 2024. 2

[18] D. Nolan and S. Stoudt. *Communicating with data: The art of writing for data science*. Oxford University Press, 2021. 1

[19] K. L. Norman. *Cyberpsychology: An introduction to human-computer interaction*. Cambridge university press, 2017. 2

[20] J. Obeid and E. Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*, 2020. 2

[21] C. Oguguo, F. A. Nannim, A. O. Okeke, R. I. Ezechukwu, G. A. Christopher, and C. O. Ugorji. Highlights of the program for the international assessment of adult competencies, u.s. results. https://nces.ed.gov/surveys/piaac/national_results.asp, 2017. Accessed: July 9, 2023. 2

[22] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010. 1

[23] H. Singh and S. Shekhar. Stl-cqa: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3275–3284, 2020. 2

[24] B. J. Tang, A. Boggust, and A. Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 2

[25] Z. Xu, S. Du, Y. Qi, C. Xu, C. Yuan, and J. Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023. 2