Unraveling the Truth: Do VLMs really Understand Charts? A Deep Dive into Consistency and Robustness

Srija Mukhopadhyay[*], **Adnan Qidwai**[*], **Aparna Garimella**[†], **Pritika Ramu**[†]
Vivek Gupta[‡], **Dan Roth**[§]
[*]IIIT Hyderabad, [†]Adobe Research, [‡]Arizona State University, [§]University of Pennsylvania
{srija.mukhopadhyay@research, adnan.qidwai@students}.iiit.ac.in,
{garimell,pramu}@adobe.com; vgupt140@asu.edu; danroth@seas.upenn.edu

Abstract

Chart question answering (CQA) is a crucial area of Visual Language Understanding. However, the robustness and consistency of current Visual Language Models (VLMs) in this field remain under-explored. This paper evaluates state-of-the-art VLMs on comprehensive datasets, developed specifically for this study, encompassing diverse question categories and chart formats. We investigate two key aspects: 1) the models' ability to handle varying levels of chart and question complexity, and 2) their robustness across different visual representations of the same underlying data. Our analysis reveals significant performance variations based on question and chart types, highlighting both strengths and weaknesses of current models. Additionally, we identify areas for improvement and propose future research directions to build more robust and reliable CQA systems. This study sheds light on the limitations of current models and paves the way for future advancements in the field.

1 Introduction

Chart question answering (CQA) (Masry et al., 2022; Chaudhry et al., 2020) has emerged as a critical area within the field of Visual Language Understanding (VLU) (Lee et al., 2023; Ghosh et al., 2024), aiming to equip machines with the ability to comprehend and answer questions based on data visualizations. While recent advancements in Vision Language Models (VLMs) and Multimodal Large Language Models (MLLMs) have yielded impressive performance improvements in CQA (Liu et al., 2023b; Masry et al., 2023; Xia et al., 2024; Xu et al., 2024; Team et al., 2023; Achiam et al., 2023; Meng et al., 2024), their true capabilities remain obscure in uncertainty. This paper delves into an insightful analysis of the robustness and consistency of state-of-the-art CQA models, exposing their limitations and guiding future research directions.

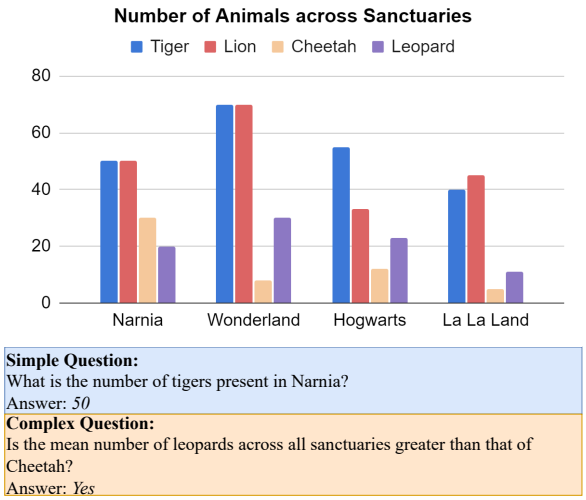


Figure 1: Simple and Complex Questions on a Complex chart

We address several key questions regarding the current state of CQA: **Are existing models truly effective, or do their impressive average scores mask significant weaknesses?** For instance, in Figure 1, one can ask that if the model's performance remains consistent across two distinct question types? The first type, *Simple Questions* like *"What is the number of tigers present in Narnia?"*, involves straightforward value extraction. In contrast, *Complex Questions* such as *"Is the mean number of leopards across all sanctuaries greater than that of cheetah?"* require extracting multiple values, aggregating them, and making boolean comparisons. It's evident that complex questions pose challenges even for humans; understanding how models handle these complexities provides valuable insights into their capabilities.

How do models perform on specific aspects of chart understanding, such as question complexity and chart type? Figure 2 shows the different types of charts across which the performance of a model can be evaluated—specifically, *Simple Charts* and *Complex Charts*—along with the differ-

* contributed equally, ‡ corresponding author (work done while at UPenn)

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16696–16717

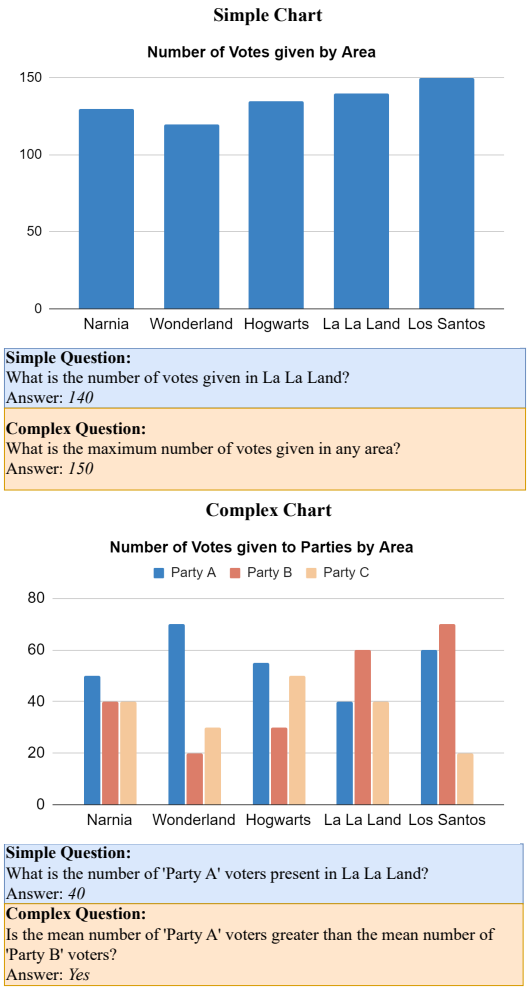


Figure 2: Example of simple chart and complex chart, along with simple and complex questions.

ent possible question types, including *Simple Questions* and *Complex Questions*. Complex Charts, such as grouped-bar charts that compare multiple attributes side by side present information in a more intricate manner compared to Simple Charts, which depict data about a single attribute using a single bar. Similarly, questions can range from complex tasks like identifying maximum values, performing aggregations, and making comparisons, to simpler queries focused on straightforward value extraction. Investigating how models handle these varied chart types and question complexities provides crucial insights into their performance and adaptability.

Furthermore, is the robustness of these models, their ability to generalize across diverse variations, adequately explored? The same data can be depicted in multiple visual formats. For instance, Figure 3 demonstrates how an original chart can be transformed into stair plots, bar charts, stacked representations, and many more. These variations can differ in aspects such as color schemes, patterns, legend positioning, and even details specific to each chart type like legend orientation and grid sizes on the x-axis and y-axis. Exploring the effect of these variations could provide deeper insights into the data and enhance the comprehensibility of the visualizations for models.

To answer these questions, we present a rigorous evaluation of leading CQA models on a meticulously curated dataset. This dataset includes diverse chart types and question categories, allowing for a thorough assessment of model performance across varying levels of complexity. We examine how well the models generalize across diverse visual representations of identical data, assessing their robustness against perturbations.

Our findings reveal significant performance discrepancies, particularly when transitioning from simple to complex chart-question combinations. Moreover, we demonstrate that even the highest-performing models exhibit a substantial drop in accuracy when subjected to diverse perturbations, highlighting the critical need for improved robustness in CQA. This paper makes the following contributions:

- Providing a thorough analysis of the strengths and weaknesses of current VLMs and MLLMs for chart understanding.
- Introducing a new evaluation set with fine-grained splits across chart types and question complexities, facilitating a deeper understanding of model performance.
- Performing a detailed robustness analysis to uncover the shortcomings of current models, emphasizing the necessity for additional research in this domain.

Our research sheds light on the current state of CQA, offering crucial insights. Our datasets, along with all the associated scripts, are available at https://robustcqa.github.io/

2 Initial Dataset

This section highlights the dataset preparation process employed to analyze the performance of CQA models across a spectrum of chart types and question complexities.

2.1 Dataset Selection

To ensure a comprehensive evaluation of CQA models, we selected the ChartQA dataset (Masry et al., 2022) as our primary benchmark. This

dataset is widely used in CQA benchmarking, covering diverse domains from sources like Our World in Data, Statista, OECD, and Pew Research.

ChartQA includes two distinct question categories: *"Human"* and *"Augmented"*. *"Human"* questions were generated by human annotators, while *"Augmented"* questions were machine-generated, ensuring a diverse spectrum of question styles. Another important aspect which motivated our choice of ChartQA dataset was the presence of underlying tables. This feature enabled us to generate controlled visual perturbations for the later section of our study. Our experiments were conducted exclusively on the test set of ChartQA, comprising questions, charts and the corresponding tables.

2.2 Chart and Question Labelling

To facilitate a more granular analysis of model performance, we categorized both charts and questions according to their complexity levels. This categorization was applied to the entire ChartQA test set, resulting in a modified evaluation dataset tailored for our experiments.

Chart Categorization. The tables provided by ChartQA were loaded as a pandas dataframe. We classify charts as either simple or complex, based on column count in the dataframe: two columns indicate a simple chart, while more than two columns signify a more complex chart. We leverage this fact to classify them using a python script.

- **Simple Charts:** The tables of these charts contain two columns to represent the dependent and independent variable. The charts thus formed represent a single entity and exhibit no overlaps or complex visual elements. Figure 2 shows an example of such chart titled *"Number of Votes given by Area"*.

- **Complex Charts:** The tables of these charts feature more than two columns, often encompassing multiple dependent variables. Thus, the charts formed have increased visual complexity. These charts usually depict multiple entities over a common series. Figure 2 shows an example of such chart titled *"Number of Votes given to Parties by Area"*.

Question Categorization. Human annotators cleaned and categorized the questions from the ChartQA dataset into two categories based on their complexity:

- **Simple Questions:** These questions primarily focus on data extraction, and typically involve a **single step of reasoning**. A human annotator can ideally answer such a question in a single step. Figure 2 shows an example of such questions *"What is the number of votes given in La La La Land?"*. One can simply answer the question by fetching the value from the chart.

- **Complex Questions:** These questions require multi-step reasoning along with data extraction, and often involve comparisons and logical inferences. If it takes multiple steps for the human annotator to answer a question, it would be classified as a complex question. Figure 2 shows an example of such questions *"Is the mean number of 'Party A' voters greater than the mean number of 'Party B' voters?"*. For this question, one would require multiple calculations to reach the final answer.

We introduced these categorizations while preserving the existing division of question generation types (human-generated and augmented questions) which was present in the original dataset, resulting in eight categories. The number of unique question-chart pairs in each category is presented in Table 1. We call this modified dataset ChartQA-Split. The detailed catagorization in our dataset allows us to isolate the impact of chart and question complexity on model performance, providing a deeper understanding of their capabilities and limitations.

	Human		Augmented	
	Simple	Complex	Simple	Complex
Simple	149	450	876	165
Complex	143	419	133	38

Table 1: Dataset statistics. Rows represent the type of Chart, Columns represent the type of Question and its Generation method.

3 Experiments

Models. To rigorously assess the performance of CQA models, we selected a diverse range of state-of-the-art models, varying in architecture, size, and training setup. All models were evaluated using a zero-shot Chain-of-Thought (Wei et al., 2022) prompting approach. An example of our prompt can be found in Figure 4. It is important to note that no additional reasoning aids were provided to any of the models. For the sake of clarity and analysis, we grouped the models into three broad categories:

Chart-based VLMs. This category contains open-source VLMs specifically adapted for chart reasoning. *MatCha (282M)* (Liu et al., 2023b) is a transformer based model which enhances the capabilities of Pix2Struct (Lee et al., 2023) models

through pre-training on mathematical reasoning and chart derendering tasks. *UniChart (201M)* (Masry et al., 2023) is another similar model which achieves chart understanding by leveraging pre-training on tasks such as data table generation, numerical and visual reasoning, and open-ended question answering. *DePlot (282M)* (Liu et al., 2023a) is a model which specializes on extracting tabular data from a given chart. The extracted table is subsequently passed to a Language Model (LM), e.g. *Flan UL2 (20B)* (Tay et al., 2022), for reasoning via Chain-of-Thought prompting (Wei et al., 2022).

Generalist VLMs. This category comprises open-source VLMs trained on general visual comprehension tasks. Notably, these models were not specifically trained or adapted for chart reasoning. *QwenVL* (Bai et al., 2023b) is a generalist 7-billion-parameter VLM built on top of *Qwen-LM* (Bai et al., 2023a) through the integration of visual encoders and the use of general and multi-task pre-training. *CogAgent VQA* (Hong et al., 2024) is an 18-billion-parameter VLM specializing in Graphical User Interface (GUI) understanding and navigation. *InternLM-XComposer2 (8B)* (Dong et al., 2024) is an adaptation of *InternLM2-7B* (Cai et al., 2024), excelling in producing high-quality long-text multi-modal content and reasoning within visual-language understanding contexts.

Large MLLMs. This category features state-of-the-art closed-source Multimodal Large Language Models (MLLMs) pre-trained on extensive visual and language data. For this category, we utilized *Gemini 1.5 Flash* (Team et al., 2023), and *GPT-4o* (Achiam et al., 2023), renowned for their capabilities in reasoning and visual understanding.

Evaluation To evaluate our models, we decided to utilize the Relaxed Accuracy metric owing to the objective nature of the expected answers. To improve on the Relaxed Accuracy metric, we introduce extra checks for precise and accurate answer matching. This metric, similar to Relaxed Accuracy, provides a 5% leverage for numerical answer matching. However, it includes the following checks:

Alphanumeric String Matching: Removing comma and spaces from the given answer and gold label to ensure an exact alphanumeric string comparison.

Strict Year Matching: For questions specifically asking for a "Year" as an answer, the 5% relaxation is disabled, forcing a strict string match. This ensures that the model accurately identifies the correct year.

Unordered Exact List Matching: For questions requiring multiple answers, an unordered exact list matching is applied, to ensure that the model correctly identifies all the expected elements in answer list, regardless of their order.

Furthermore, to validate the accuracy of our proposed evaluation metric, we manually verified the answers obtained using this metric. Our metric is usable and applicable for general large-scale model evaluation in question-answering based tasks.

Smaller VLMs. We noticed that smaller models (QwenVL, CogAgent, InternLM) struggled to produce answers in the correct format. This might be due lack of complex instruction following abilities. We addressed this by using Gemini 1.5 Flash to extract answers from their outputs in a favourable format, hence using the *LLM as an extractor*. **Manual verification** of 150 samples confirmed that Gemini 1.5 Flash primarily acted as a formatting tool, preserving the original model's answer in 149 cases and performing rounding in the one remaining instance. This demonstrates Gemini's effectiveness in enhancing the usability of smaller models without significantly altering their intent. The prompt used, can be found in Figure 5.

4 Can models reasons consistently?

This section presents our findings and analysis on the performance of various chart question answering (CQA) models across different chart types and question complexities.

4.1 Results and Discussion

Table 2 gives an overview of all results obtained for this section.

$(Q1)$ **Does any model excel across all categories?** While no single model dominates all categories, GPT-4o and Gemini 1.5 Flash consistently demonstrate impressive performance, with GPT-4o leading in most cases. Among open-source models, InternLM stands out as the top performer.

Models specifically trained on chart reasoning tasks (MatCha, UniChart) show exceptional performance while answering augmented questions, as highlighted in previous work as well ((Liu et al., 2023b; Masry et al., 2023). This likely stems from their exposure to similar question formats during

Type	Chart-based VLMs			Generalist VLMs			MLLMs	
	Mat-Cha	Uni-Chart	DePlot + Flan UL2	Qwen VL	CogAgent VQA	Intern LM	Gemini 1.5 Flash	GPT 4o
Human								
SS	57.00	49.60	51.60	66.40	81.20	79.90	87.92	**88.59**
SC	30.22	32.00	32.80	44.20	55.50	58.60	81.11	**88.22**
CS	45.40	47.50	30.60	60.10	58.00	74.10	80.42	**81.82**
CC	25.29	25.00	25.20	35.00	42.40	51.30	74.46	**83.29**
Augmented								
SS	91.40	87.20	76.10	86.50	80.90	82.50	91.32	**94.18**
SC	65.40	66.00	72.70	72.10	76.90	68.40	80.61	**88.48**
CS	78.10	69.20	48.10	61.60	47.30	68.40	**81.20**	80.45
CC	34.20	44.70	52.60	36.80	55.20	47.30	65.79	**71.05**

Table 2: Model accuracy across different categories. *S* denotes '*S*imple' and *C* denotes '*C*omplex'. The first and second letter represents chart and question type respectively.

training, which is particularly evident in simple questions from the augmented set, where MatCha achieves a high accuracy of 91.40%, followed by UniChart at 87.20%. However, they struggle significantly with reasoning-based questions, achieving as low as 25% accuracy for complex chart and complex question pairs, highlighting the need for enhancement in the reasoning abilities of such models.

$(Q2)$ **How do models perform across various chart types?** Across all models, a consistent trend emerged: performance was consistently better on simple charts compared to complex charts, while comparing with the same question type. This behavior is likely attributable to the inherent difficulty in understanding and extracting values from complex charts. Factors like overlapping data points and the requirement of precise color resolution contributes to challenges in data extraction, increasing the difficulty of reasoning on such charts.

$(Q3)$ **How do models perform across various question types?** For the same chart type, models consistently perform better on simple questions compared to complex questions. This significant difference in scores highlights the limitations of certain models in fine-grained data extraction and reasoning. GPT-4o and Gemini 1.5 Flash exhibit the smallest decrease in scores, indicating strong reasoning capabilities along with commendable data extraction skills. Smaller models, particularly those specifically trained on charts, struggle with questions requiring mathematical reasoning, despite their competence in basic data extraction.

$(Q4)$ **Do models struggle more with complex charts or complex questions?** To further assess model capabilities, we compared performances of models on two categories: "Simple Charts, Complex Questions" and "Complex Charts, Simple Questions." This analysis reveals whether a model excels at visual data extraction (complex charts) or reasoning (complex questions).

Our results show that MLLMs like GPT-4o demonstrate strong reasoning skills, excelling in answering complex questions. Gemini 1.5 Flash on the other hand performs consistently across both categories. Generalist and chart-based VLMs tend to favor the complex chart, simple question pair over the simple chart, complex question pair, suggesting limitations in reasoning abilities. This insight allows for targeted model fine-tuning to enhance specific domains where they lack dexterity.

$(Q5)$ **Are there charts and questions where all models consistently fail to answer accurately?** We focused on identifying patterns of model failure across all categories. In total, we found that 181 questions could not be answered by any model that we tested on. Given below are a few recurring difficulties that models faced:

36/181 - Questions involving counting: Models consistently struggled to accurately count objects when the number exceeded ten.

30/181 - Charts containing similar colours: Models struggled with charts which required discrimination between highly similar colors or shades of the same color.

17/181 - Identifying colours from name: Models struggle to accurately identify the color of chart elements when prompted to do so.

17/181 - Charts involving summary statistics: Models struggle to interpret charts with summary statistics, often confusing presented values with the need for recalculation. For example, given a chart of "Average of company revenues," they struggle to answer questions about "company A's average revenue," unsure whether to extract the value or recalculate it. This highlights a key limitation in their understanding of statistical representations.

7/181- Tight pie charts: In some instances, models incorrectly assigned labels to categories in pie charts with narrow slices, hence failing to identify the correct association.

A more detailed analysis on this topic has been presented in the Appendix.

$(Q6)$ **How well do the models attend to the provided image for reasoning?** To investigate the extent to which models rely on visual information

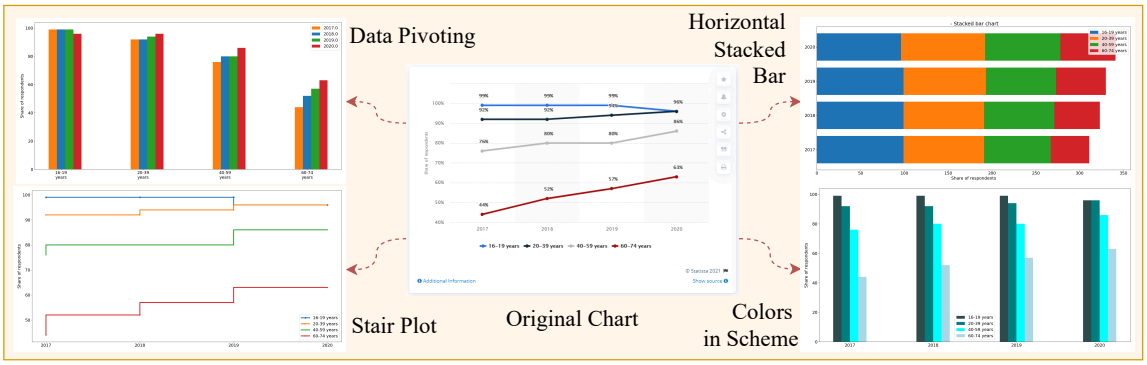


Figure 3: Examples of different types of perturbations on the same original chart and data.

versus their internal knowledge base, we conducted an experiment using *blank images* and *irrelevant charts*. We sampled 100 questions from each category and tested the top-performing models on their reasoning skills.

Surprisingly, even when models were presented with irrelevant or blank images, some models successfully answered the questions, indicating a reliance on their pre-existing knowledge. This observation suggests potential leaks in testing data. Models were able to provide the correct answer even if the answer was factually incorrect from a real-world point-of-view, highlighting the need for masked evaluation sets for visual reasoning tasks.

Model	Blank Charts				Irrelevant Charts			
	SS	SC	CS	CC	SS	SC	CS	CC
Gemini 1.5 Flash	0	0	0	0	0	2	2	4
GPT-4o	0	3	0	3	0	2	1	6
InternLM-XComposer2	2	3	8	6	1	5	3	2
CogAgent-VQA	**11**	5	13	9	5	7	**20**	8
Qwen-VL	7	**9**	**21**	**17**	**9**	**8**	13	**14**

Table 3: Performance of models when probed with blank and irrelevant charts. *S* denotes '*S*imple' and *C* denotes '*C*omplex'. The first letter represents chart type and the second letter represents question type.

Our analysis, detailed in Table 3, reveals that even large models like Gemini 1.5 Flash and GPT-4o were capable of answering few questions based on irrelevant charts, highlighting the needs of developing models that integrate visual information for robust visual reasoning capabilities.

While our analysis reveals that models face challenges with certain categories of questions and charts, it also underscores the significant progress achieved in chart question answering (CQA) performance across various models.

5 Are models robust on CQA?

Another crucial aspect of our analysis involves investigating the robustness and consistency of these models across different visual representations of the same underlying data. Through the help of this probing, we aim to understand if model performance remains stable when presented with variations in chart types, styles, or aesthetics while conveying the same information.

Figure 3 illustrates how an original chart can be converted into stair plots, bar charts, stacked representations, and more. These variations may differ in color schemes, patterns, legend positioning, and other chart-specific details like legend orientation and grid sizes on the x-axis and y-axis. Examining these variations can offer deeper insights into the data and improve the clarity of the visualizations.

5.1 Our RobustCQA Dataset

Following the initial dataset preparation, a perturbation dataset was created to rigorously assess the robustness of the top-performing models across diverse chart variations. We refer to this dataset as the RobustCQA dataset, which systematically manipulates various chart elements while preserving the underlying data.

Creation We identified 75 unique perturbation types for both simple and complex charts. These perturbations cover a broad spectrum of visual variations, including:

- **Color Scheme Changes:** Modifying color palettes, gradients and hues.
- **Chart Type Variations:** Experimenting with line plots, bar plots, stair plots, stem plots and other less commonly used chart types.
- **Legend and Axis Modification:** Altering label position, formatting, and positioning of legend and axis elements.

The perturbed charts were generated using the Matplotlib library. We ensured that **only one element is altered per perturbation** keeping the

	Simple Questions					Complex Questions				
	MLLMs		Generalist VLMs			MLLMs		Generalist VLMs		
Category	Gemini 1.5 Flash	GPT-4o	Qwen VL	CogAgent VQA	InternLM XComposer2	Gemini 1.5 Flash	GPT-4o	Qwen VL	CogAgent VQA	InternLM XComposer2
original_chart	***94***	*89*	*62*	*60*	*77*	*71*	***82***	*36*	*45*	*49*
annotations	86	**90**	34	37	59	61	**77**	31	40	43
annotated_bars	83	**89**	35	31	64	68	**74**	26	28	38
basic	**67**	43	17	17	51	**55**	46	28	31	37
color_random	**66**	45	15	14	51	**53**	49	21	29	31
color_scheme	**56**	45	16	13	47	**55**	52	26	31	40
data_pivot	**56**	43	11	9	46	**44**	23	28	27	38
font	**67**	49	16	18	34	**51**	43	26	28	33
grid	**67**	48	18	16	51	**52**	51	21	24	34
hatching	**57**	37	11	9	42	**49**	42	28	29	37
horizontal_grouped	**60**	32	19	14	49	**51**	42	29	29	40
horizontal_stacked	**30**	20	16	11	22	**59**	46	22	32	43
legend_position	**52**	44	15	19	49	**47**	46	28	30	28
line_representation	**52**	44	13	18	35	**42**	40	29	27	33
log_scale	**38**	41	11	9	5	**55**	45	27	30	38
only_data_color_scheme	**62**	44	17	18	53	**51**	50	25	27	39
replacing_legend_with_labels	**59**	48	19	14	41	45	**56**	30	28	31
scaling_size	**63**	43	11	13	31	**47**	41	30	25	28
scatter_representation	**43**	38	12	14	37	**45**	44	23	27	29
stacked	**36**	28	14	13	**36**	**47**	38	26	33	32
stacked_area	**34**	24	19	13	31	**45**	41	26	34	34
stair_plot_normal	**49**	41	14	16	41	**52**	43	17	29	20
stair_plot_with_marker	**55**	48	13	20	43	**57**	51	24	27	45
stem_plot	47	36	12	12	**52**	**55**	41	22	28	35
tick_orientation	**66**	51	19	14	43	**42**	33	29	27	27
tick_position	**56**	48	21	16	47	**49**	42	30	31	30

Table 4: Model Performance on various perturbations on Complex Charts

rest of the elements the same. The tables from the ChartQA dataset served as the source for the underlying data.

Human Verification To ensure the quality and relevance of our dataset, a rigorous manual annotation process was employed. Expert evaluators meticulously verified each perturbed chart, assessing how easily comprehensible and answerable each perturbed chart was. They also evaluated the relevance of each perturbation to the specific chart type, refining the perturbation set to include only meaningful variations. The underlying tables were also thoroughly verified to confirm that the generated questions remained answerable based on the chart data. This comprehensive evaluation was facilitated by a custom-built annotation platform, specifically designed to streamline the manual annotation process and ensure high-quality data.

Final Dataset The original 75 perturbations were then grouped into categories of related perturbations to create the final dataset. This set consists of 22 unique perturbation categories for simple charts and 25 such categories for complex charts, covering a wide range of visual variations.

To ensure a fair analysis of model robustness across perturbations, 100 questions were sampled for each chart type and question type pair. This resulted in a total of 400 unique table and QA pairs for our final evaluation. This standardized question set allows for direct comparison of model performance across different visual representations. We finally compare the results of all perturbations against the basic or default Matplotlib chart.

A detailed breakdown of the perturbation categories along with examples has been included in the Appendix.

5.2 Methodology

To delve deeper into the performance and limitations of leading chart question answering models, we evaluated Qwen-VL, CogAgent-VQA, InternLM-XComposer2 (open-source VLMs) and Gemini 1.5 Flash, GPT-4o (closed-source MLLMs) using our RobustCQA dataset. We employed a similar evaluation metric as described previously, leveraging an extractor LLM for smaller models to ensure consistent output format, and analyzed all models through Zero-Shot Chain-of-Thought prompting.

5.3 Results and Discussion

The results obtained for perturbations on complex charts have been highlighted in table 4. The re-

sults for perturbations on simple charts has been presented in table 9 in the Appendix.

($Q1$) **Does model performance stay consistent with perturbed charts?** The results reveal a significant performance degradation for most models when confronted with perturbations. While performance generally decreases across all models, some exhibit more drastic drops. Among open sourced models, InternLM-XComposer2, and among closed source models, Gemini 1.5 Flash proved to be the most consistent across various perturbations. Open sourced models like CogAgent-VQA and Qwen-VL and even closed source models like GPT-4o displayed relatively low accuracy with most perturbations, highlighting a potential lack of robust data extraction skills. Manual analysis of the responses from the models highlighted the importance of improving data extraction for non-annotated charts to enhance model robustness in chart-based tasks.

($Q2$) **Are there specific perturbations that help enhance model performance?** Our experiments highlighted several perturbations that improved model performance. Across all models, annotated data points consistently boosted accuracy. While the most beneficial plot type varied across models and question/chart categories, annotated bar graphs emerged as a consistently positive influence.

In addition to that, grids to act as reference points for data extraction and better tick orientation also contributed positively. Furthermore, labelling the lines to reduce the complexity of color resolution and placing the legend optimally to ensure that it doesn't obscure crucial data points also helped the models. We also noticed that increasing the font size played an important role in aiding all models, especially smaller models. The results for the same have been presented in Table 5.

($Q3$) **Are there perturbations which are always detrimental to the model performance?** Our analysis reveals that while models demonstrate promising performance on standard chart datasets, they struggle with robustness when faced with visual perturbations. While annotations generally help improve model performance, most other perturbations negatively impacted model accuracy.

Notable ones among them include logarithmic scales which can be challenging for humans as well. Additionally, models also struggle significantly with horizontal chart variations, particularly horizontal stacked charts. In general it was noticed that models struggled to reason on stacked charts, possibly due to the requirement of additional mathematical reasoning for data extraction. Stair plots also caused significant trouble as models could not identify the precise data point to refer to. Our findings emphasize the need of more diverse datasets along with more robust models that can effectively interpret visual information beyond just simple visual cues.

($Q4$) **Are there certain perturbations which are more effective for certain question types?** Our analysis suggests chart type effectiveness varies by question type. For instance, line charts help in visualizing trends and correlations. Stacked bar charts are generally unsuitable except for questions that require data aggregation. Bar charts, while useful for comparing individual values within a certain group, prove to not be good for showcasing correlations across different groups or entities. Our analysis helps with understanding and creating suitable charts for domain specific tasks.

($Q5$) **Does the effect of each perturbation type vary across models?** The impact of each perturbation on model performance exhibits significant variation. While question and chart type play a role, for a given model, certain perturbations consistently prove more helpful or harmful. This nuanced effect of perturbation type on the model performance is detailed in Table 7, 8. We believe that our analysis allows us to identify specific areas for helping improve each model through targeted model fine-tuning.

Additional insights and details obtained from our experiments have been presented in the Appendix.

6 Related Work

Chart comprehension and question answering (CQA) are critical domains with a growing body of research. While existing CQA datasets assess models' advanced reasoning capabilities over charts, many face significant limitations. These include small dataset sizes (Kim et al., 2020), reliance on template-based questions and synthetically generated charts (Methani et al., 2020; Chaudhry et al., 2020; Kafle et al., 2018; Han et al., 2023), restriction to specific domains (Methani et al., 2020; Ahmed et al., 2023; Li and Tajbakhsh, 2023), or focusing solely on open-domain question answering (Kantharaj et al., 2022). Even the current state-of-

the-art dataset, Chart QA (Masry et al., 2022), has limitations due to the lack of classification labels for more meaningful analysis, and limited variation in chart types.

More recent datasets, such as ChartX (Xia et al., 2024), have expanded the range of chart types analyzed. ChartBench (Xu et al., 2024) and MMC (Liu et al., 2024) focus on large-scale datasets with more diverse chart types.

A very recent work, CharXiv (Wang et al., 2024), provides extensive evaluations across a range of charts and questions, including both reasoning-based and descriptive queries. They also perform ablation studies by modifying charts and questions.

However, to the best of our knowledge, RobustCQA is the first dataset to systematically perturb all elements within a chart, enabling fine-grained analysis of factors affecting model performance. Additionally, we conduct a detailed analysis of the Chart QA dataset based on question and chart complexity, which has not been done at this level of detail before.

Modeling approaches for charts Various approaches have been developed for chart modeling. This includes models specifically designed for chart comprehension and reasoning, built with the end-to-end goal of reasoning over charts (Liu et al., 2023b; Masry et al., 2023; Singh and Shekhar, 2020), as well as models that convert charts into intermediate table formats (Liu et al., 2023a), enabling reasoning by generalized large language models (LLMs) through Chain of Thought (Wei et al., 2022) or Program of Thought (Chen et al., 2023) prompting. Additionally, there are generalized models used for multi-modal reasoning tasks, including chart comprehension (Team et al., 2023; Achiam et al., 2023; Bai et al., 2023b; Dong et al., 2024; Hong et al., 2024). Recent efforts have also focused on developing smaller, yet accurate models for this task (Wang et al., 2024).

While these approaches have shown significant progress, their specific failure points remain unclear. A recent study (Islam et al., 2024) analyzed the performance of GPT-4v and Gemini, providing a broad evaluation of these models across various chart comprehension tasks, including question answering, summarization, and fact-checking. In contrast, our work focuses specifically on CQA across a broader range of models, offering an in-depth analysis of the question and chart types contributing to model failures. Our contribution identifies the exact question types and chart elements that lead to model errors, offering key insights to improve model performance.

Vision-Language Model Robustness Recent studies have highlighted the vulnerability of models to attacks and perturbations (Ma et al., 2024; Zhao et al., 2023), raising concerns about their robustness in real-world deployment. Motivated by this, we developed a robustness benchmark specifically for chart question answering (CQA). While previous work (Gupta et al., 2024) analyzed models like DePlot and MatCha on perturbed charts, focusing on questions related to structural and visual context, our study examines general reasoning questions. This approach helps us assess how variations in the visual representation of the same data affect model performance.

7 Conclusion

This research introduces ChartQA-Split and RobustCQA, the first datasets dedicated to understanding model consistency across complexities and robustness to visual perturbations in chart question answering. Our evaluation of SOTA models, including baselines and VLMs/MLLMs, using a zero-shot chain-of-thought setting, reveals significant challenges in both areas. We perform an in-depth analysis of model weaknesses and identify key areas for improvement, such as enhancing data extraction for non-annotated charts and developing models that can effectively interpret complex visual information, taking every possible visual cue into consideration. Our work provides a foundation for future research in developing more robust and reliable chart question answering systems.

Future Directions. Our perturbation analysis provides a nuanced understanding of model performance by revealing both universal and model-specific vulnerabilities and strengths. This insight drives targeted improvements: **Model Pretraining:** Focusing on perturbations that affect models allows for effective fine-tuning to address weaknesses. **Perturbation-Aware Training:** Integrating specific perturbations during training enhances overall robustness, helping models develop resilience against challenges. **Interpretable Models:** Understanding the impact of perturbations aids in debugging and building explainable models, fostering the development of reliable and transparent chart understanding and reasoning systems.

Limitations

The presented work exhibits a few limitations. First, our data was obtained from a singular dataset, and we used only one plotting software for testing the perturbations. Expanding the dataset to include diverse sources and exploring various plotting libraries would strengthen the findings and improve generalizability. Second, the dataset is limited to English, while models are developed and evaluated on a wide variety of languages. Future research is required to expand the domain beyond English. Third, we were not able to cover a few chart types in the course of our analysis in order to make a more generalized perturbation set. This included pie and doughnut charts, pyramid and funnel charts as well as radar charts. Due to metadata limitations and the complexity of adapting data for chart representation, these charts were excluded. Fourth, inconsistent metadata of the original dataset sometimes lacked visual captions present in the original charts, which could not be replicated in the perturbed charts. Because of this, we were not able to identify attributes pertaining to chart elements, for example, colour.

Ethics Statement

This research adheres to the ACL code of ethics, acknowledging and addressing potential ethical implications. While LLMs assisted in writing and presentation, all ideas and conclusions are solely attributed to the authors. The research promotes responsible and fair use of methodologies, ensuring transparency and reproducibility. We plan to release all scripts, resources, comprehensive documentation, evaluation metrics, datasets, model specifications, and prompting methods to enable others to build upon our work. We strive to present our findings clearly and accurately, avoiding exaggerated claims or misinterpretations.

Acknowledgement

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0080. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This work was partially funded by ONR Contract N00014-23-1-2365. Lastly, we acknowledge the generous gift from Adobe.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Saleem Ahmed, Bhavin Jawade, Shubham Pandey, Srirangaraj Setlur, and Venu Govindaraju. 2023. Realcqa: Scientific chart question answering as a test-bed for first-order logic. In *International Conference on Document Analysis and Recognition*, pages 66–83. Springer.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *CoRR*.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode & attend for figure question answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui

He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions.

Ashim Gupta, Vivek Gupta, Shuo Zhang, Yujie He, Ning Zhang, and Shalin Shah. 2024. Enhancing question answering on charts through effective pretraining tasks.

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.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning? an extensive investigation into the capabilities and limitations of lvlms.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. OpenCQA: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs.

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. Association for Computational Linguistics.

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 and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310, Mexico City, Mexico. Association for Computational Linguistics.

J. Ma, P. Wang, D. Kong, Z. Wang, J. Liu, H. Pei, and J. Zhao. 2024. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Transactions on Pattern Analysis I&; Machine Intelligence*, 46(08):5575–5594.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

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

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Hrituraj Singh and Sumit Shekhar. 2020. 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)*, pages 3275–3284, Online. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier García, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2022. Ul2: Unifying language learning paradigms. In *International Conference on Learning Representations*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2024. Chartbench: A benchmark for complex visual reasoning in charts.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Appendix

Effect of Font-size on Models Table 5 illustrates the significant impact of font size on model performance. Increasing the font size improves the OCR capabilities of visual language models (VLMs), suggesting that larger font sizes can enhance model accuracy in chart comprehension tasks. This finding indicates that adjusting font size could serve as an effective preprocessing step for boosting performance in such tasks.

	Gemini 1.5 Flash		Qwen VL	
Perturbation types	**Small Font**	**Big Font**	**Small Font**	**Big Font**
Normal line plot	62	63	6	27
Colors in a given scheme (line)	56	63	7	26
Colors random (scatter)	46	57	5	21
Line Represntation	47	50	10	18
Stem Plot	45	52	6	15
Stair Plot	42	48	10	26
Ablation - removing Y axis	63	64	7	22
Rotated X axis Tick	56	62	8	22
Annotated Bar Graph	77	82	12	42
Horizontal Bar Graph	54	65	11	19

Table 5: Effect of increasing font size

Model Scores Alongside the scores of models across various perturbations for complex charts provided in Table 4, we have also presented the model scores for different perturbations in simple charts in Table 9.

Where can models not answer? Continuing with the analysis of *Q5* in section 4.1, we provide a further breakdown of cases where models failed to answer correctly, as shown in Table 6. In addition to previously discussed issues, we found that in some instances (4/181), models did not fully comprehend the entire chart before answering.

Type of chart	Wrong	Total	Error Rate
Bar	121	1842	6.56%
Line	44	380	**11.57**%
Pie	16	151	10.59%
Total	181	2373	7.62%

Table 6: Error distribution among different chart types

For example, in a chart with four columns labeled "No (low confidence), No (high confidence), Yes (low confidence), and Yes (high confidence)," models were asked to calculate the percentage of 'No'. However, they failed to recognize that the question required summing the percentages from both 'No' columns and instead provided the percentage from only one column. This error highlights a potential gap in the models' ability to fully integrate image encoding with language decoding, suggesting improvements could be made to better interpret such visual data.

Models and Perturbation Types				
Gemini 1.5 Flash	**GPT-4o**	**Qwen-VL**	**CogAgent-VQA**	**InternLM-XComposer2**
Annotations on individual points	Annotations on individual points	Annotations on Bar Graphs	Annotations on individual points	Annotations on bar charts
Annotations on Bar Graphs	Annotations on Bar Graphs	Annotations on individual points	Annotations on Bar Graphs	Annotations on individual points
Random Color Scheme in Chart	Placing Legend Elements with Line	Basic Matplotlib Charts	Random Color Scheme in Chart	Area Plot
Placing Legend Elements with Line	Random Markers and Line Styles	Placing Legend Elements with Line	Axes Transposition	Horizontal Bar Charts
Basic Matplotlib Charts	Basic Matplotlib Charts	Changing Font Size	Basic Matplotlib Charts	Random Color Scheme

Table 7: Top 5 best performing perturbations for each model

Models and Perturbation Types				
Gemini 1.5 Flash	**GPT-4o**	**Qwen-VL**	**CogAgent-VQA**	**InternLM-Xcomposer2**
Stacked Area Chart	Horizontally Stacked Bars	Stacked Bar Graphs	Horizontal Bar Charts	Horizontally Stacked Bars
Horizontally Stacked Bars	Stacked Area Chart	Changing Horizontal and Vertical Dimension	Stacked Area Chart	Changing Horizontal and Vertical Dimension
Stacked Bar Graphs	Stacked Bar Graphs	Log Scale	Horizontally Stacked Bars	Stacked Area Chart
Log Scale	Horizontal Grouped Bar Charts	Random Representation of Scatter Plots	Horizontal Grouped Bar Charts	Stacked Bar Graphs
Normal Stair Plot	Hatched Pattern in Bar Charts	Stair Plots with Marker	Log Scale	Changing Font Size

Table 8: Top 5 worst performing perturbations for each model

	Simple Questions					Complex Questions				
	MLLMs		Generalist VLMs			MLLMs		Generalist VLMs		
Category	Gemini 1.5 Flash	GPT-4o	Qwen VL	CogAgent VQA	InternLM XComposer2	Gemini 1.5 Flash	GPT-4o	Qwen VL	CogAgent VQA	InternLM XComposer2
original_chart	***96***	*94*	*76*	*79*	*83*	*85*	***89***	*56*	*64*	*69*
annotations	90	**91**	62	66	65	**74**	64	42	42	47
area_plot	**73**	42	21	16	61	**71**	64	39	42	48
annotated_bars	**93**	91	71	63	73	78	**90**	45	58	48
basic	**73**	43	24	18	54	**73**	61	38	38	46
color_random	**79**	43	20	22	63	**72**	64	32	40	42
color_scheme	**78**	50	18	23	58	**68**	64	30	36	36
data_pivot	**74**	55	13	13	56	**71**	62	42	40	41
font	**79**	53	19	28	28	**65**	52	33	26	51
grid	**79**	58	23	24	57	**66**	64	31	42	44
hatching	**75**	44	20	18	67	**72**	69	39	42	44
horizontal	**73**	33	19	14	58	**67**	64	35	45	43
legend_position	**78**	57	14	23	54	59	**62**	32	41	43
line_representation	**84**	59	19	25	56	**67**	63	31	34	40
log_scale	**42**	36	12	12	14	**78**	21	32	37	45
replacing_legend_with_labels	**78**	59	18	22	53	**72**	61	27	41	40
scaling_size	**73**	53	17	25	31	**62**	55	34	38	39
scatter_representation	**75**	48	15	17	47	**64**	57	34	43	44
stair_plot_normal	**59**	53	20	23	52	**65**	61	39	38	48
stair_plot_with_marker	**64**	51	15	24	60	**68**	64	25	41	31
stem_plot	**72**	41	12	17	70	75	**86**	36	54	48
tick_orientation	**76**	57	18	22	50	**72**	60	41	46	49
tick_position	**69**	54	27	27	51	53	**61**	35	42	41

Table 9: Model Performance on various perturbations on Simple Charts.

Sample prompt for GPT-4o and Gemini 1.5 Flash

Task: You will be given a chart and a question. Answer the given question from the chart given to you.
Instructions:
0) Look carefully at the chart, think about the type of chart, before answering the question directly.
1) If a question asks about a column name, give the full and exact name for the column as it is written in the chart.
2) If a question required multiple outputs, give it in the form: [<output1>, <output2> ..] where outputs are in sorted order. For example, if the output is 'Australia and India' give the answer as [Australia, India]. Please dont use this with column names invloving 'and' keyword.
3) If a question requires doing arithmetic operations, calculate till the final number.
4) If a question asks for what column a certain value is in, give the full and exact name of the column and not the value.
5) If a question asks how many times a certain value appears, give the count and not the name of the columns where it appears.
6) Answer without taking account of the units or scale given in chart. For example, if the chart has values in millions, you should ignore the scale and account absolute numbers. Remove the unit from your final answer and reason based on the absolute values obtained directly from the chart. Example: If your answer is 10 million USD, you should write 10 as your answer.
7) It is known that the answer is obtainable from the chart given to you.
8) Write your intermediate steps.

The chart might not have exact values written on it, therefore you might need to find the exact value in that case with the help of the axes.
Think step by step and append the answer at the last of your response in the form: "... . The answer is: <answer>"
Question:

Figure 4: Prompt for testing chart question answering

Sample prompt for Extraction through Gemini 1.5 Flash

You are an expert in getting the answers from a given long answer with steps. These questions were asked about a chart.
Task: Extract the final answer based on the given long sequence of reasoning with answer, given the question.

Instructions:
Append to your response and reasoning: 'The answer is: <final_answer>'.

If a question asks about a column name, give the full and exact name for the column as it is written in answer.
If a question required multiple outputs and the output contains multiple outputs as well, give it in the form: [<output1>, <output2> ..] where outputs are in sorted order. For example, if the output is 'Australia and India' give the answer as [Australia, India].
Ignore percentage signs.
Remove the units from the answer. For example, if the answer is '10 million', give the answer as '10'.

A few examples:

Question: What is the value of the blue column?
Given Answer: The blue column has the name 'XXX' and the value is 10.
Your Answer: <reasoning>. The answer is: 10

Question: What is the share of people above 65+ years in the small business category?
Given Answer: To find the share of SME owners in small business over 65 years, we need to add the percentages for the '65-69 years' and '70-74 years' age groups. The calculation is as follows: 26.1% (65-69 years) + 11.8% (70-74 years) = 37.9%. So, the share of SME owners in small business over 65 years is 37.9%.
Your Answer: <reasoning>. The answer is: 37.9

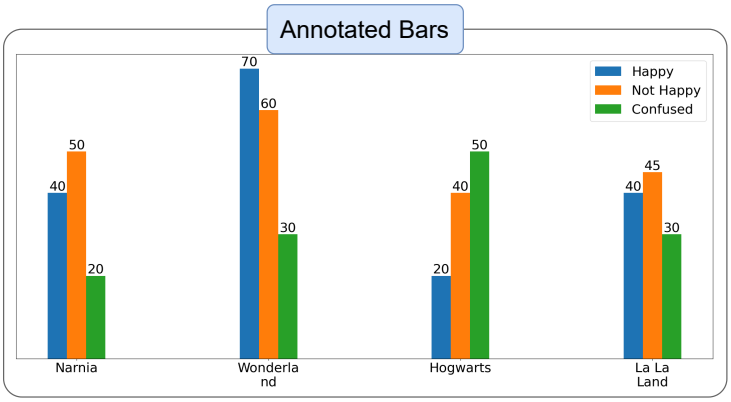
Where <reasoning>. is your reasoning and your chain of thought to get to the answer.

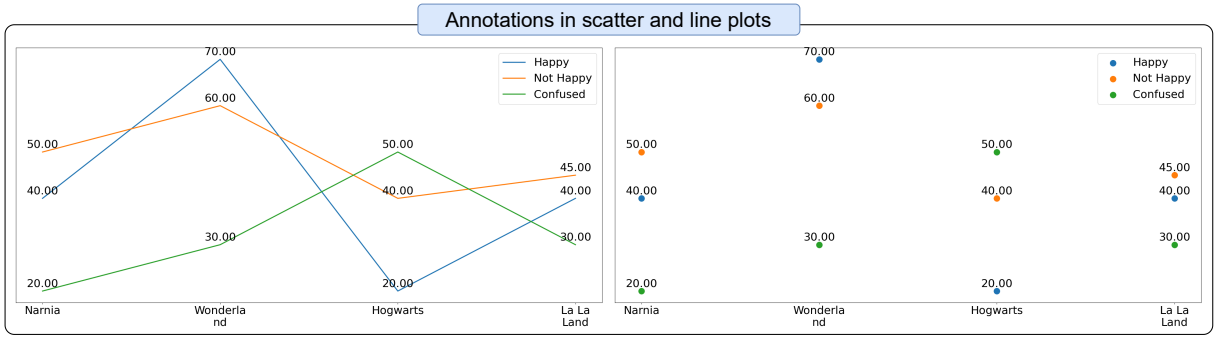
You need to carefully look at the question and the given answer. Think step by step.

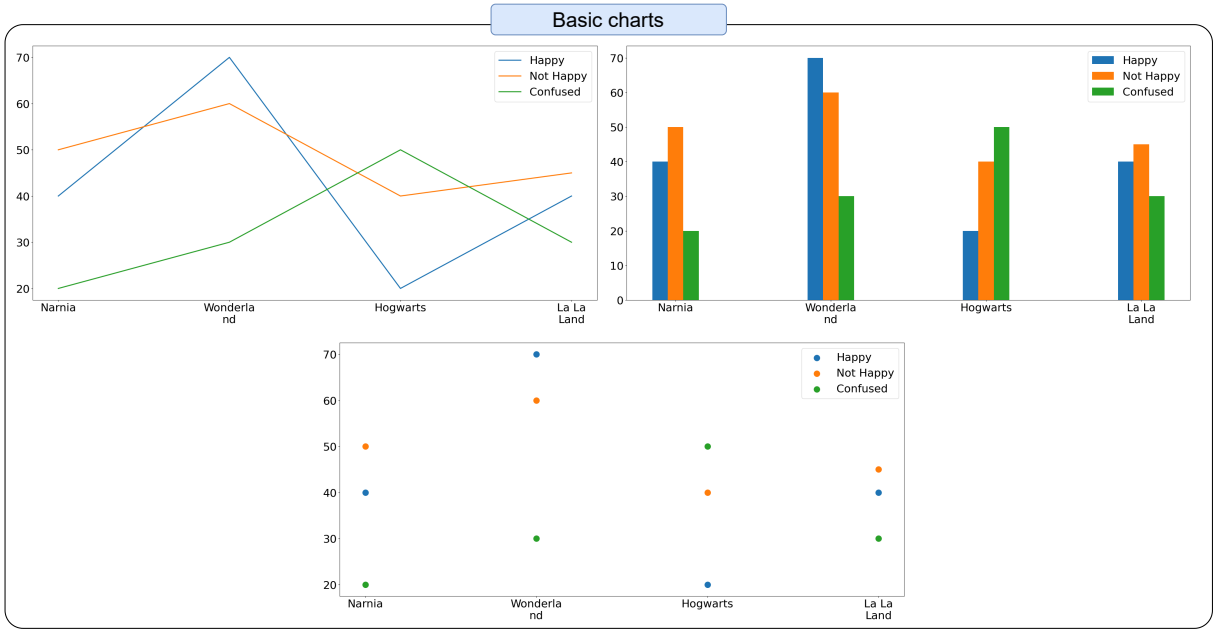
Question: {question}
Given Answer: {answer}

Figure 5: Prompt for extracting answers through an LLM from a different LLM

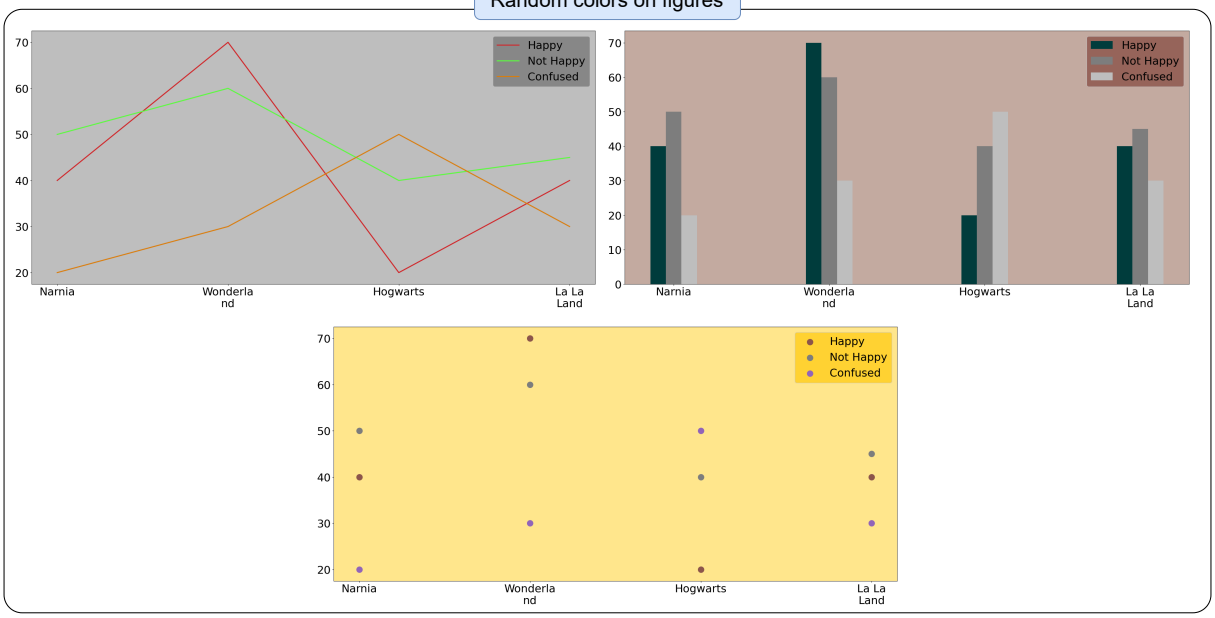
Visual examples of perturbation types



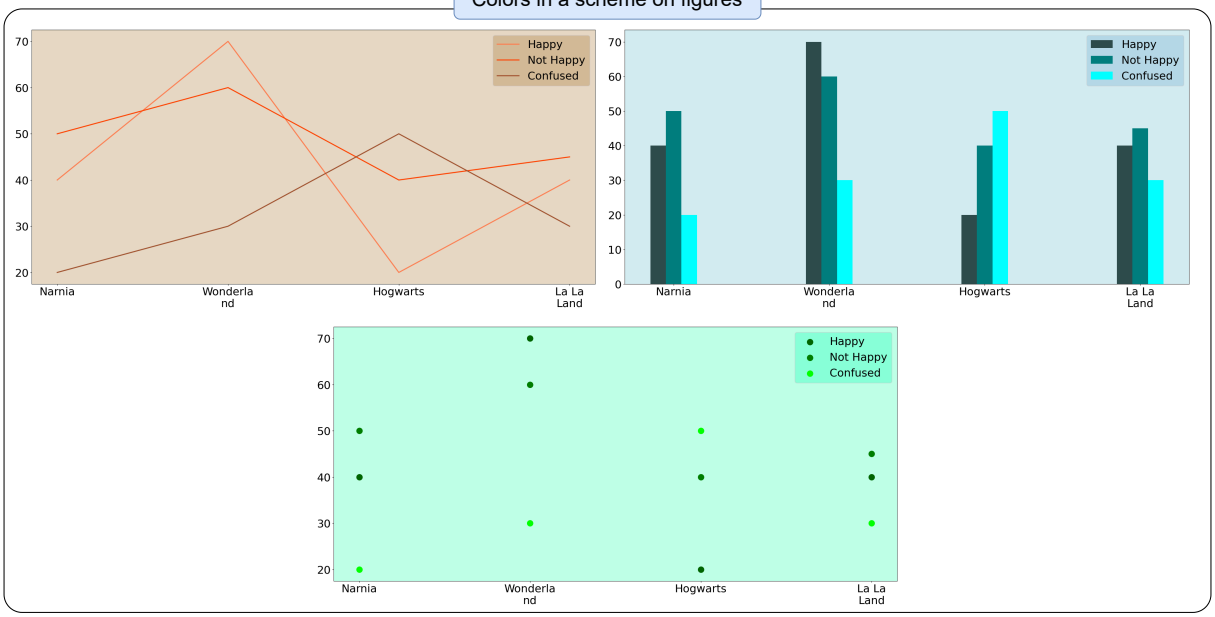




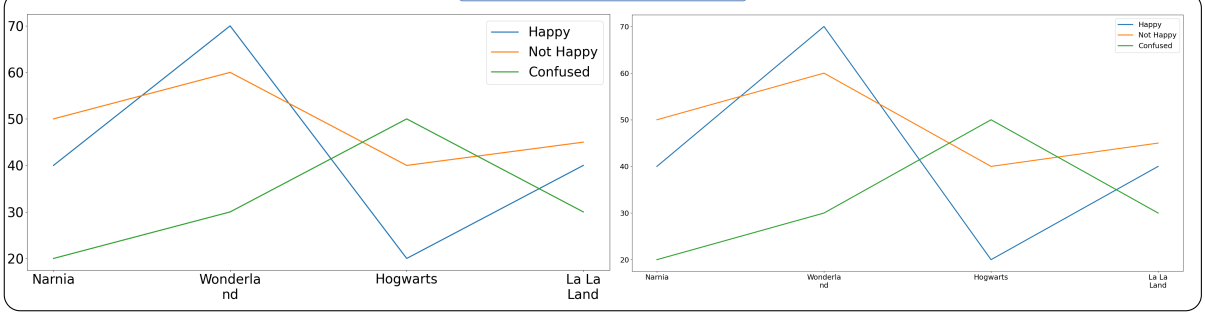
Random colors on figures

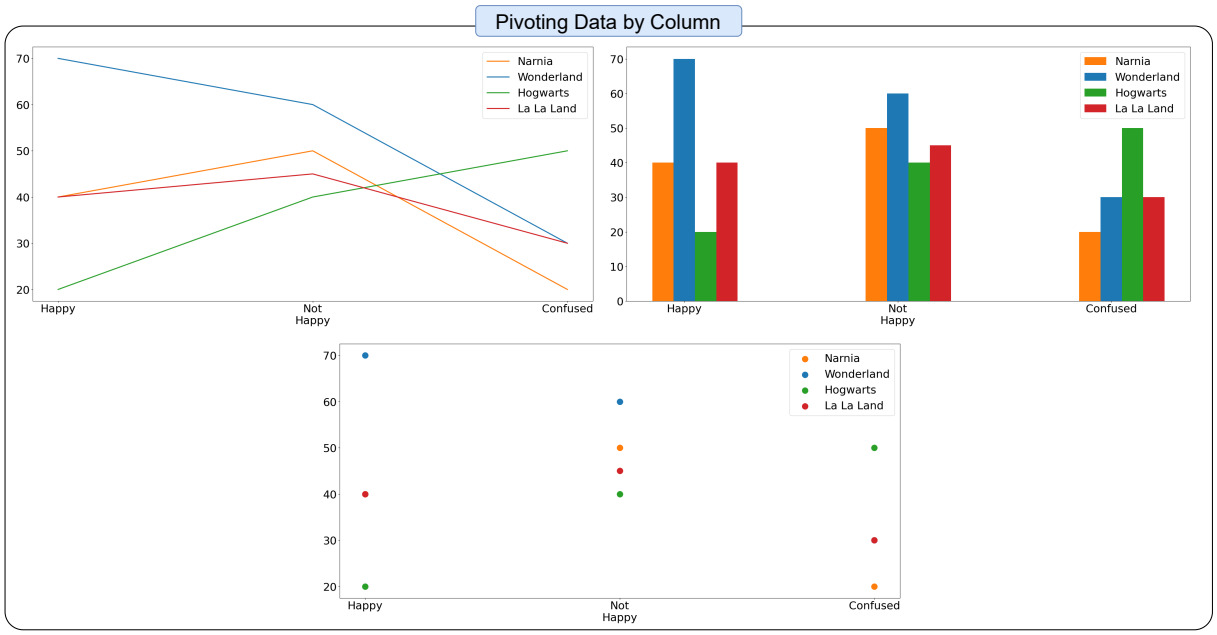


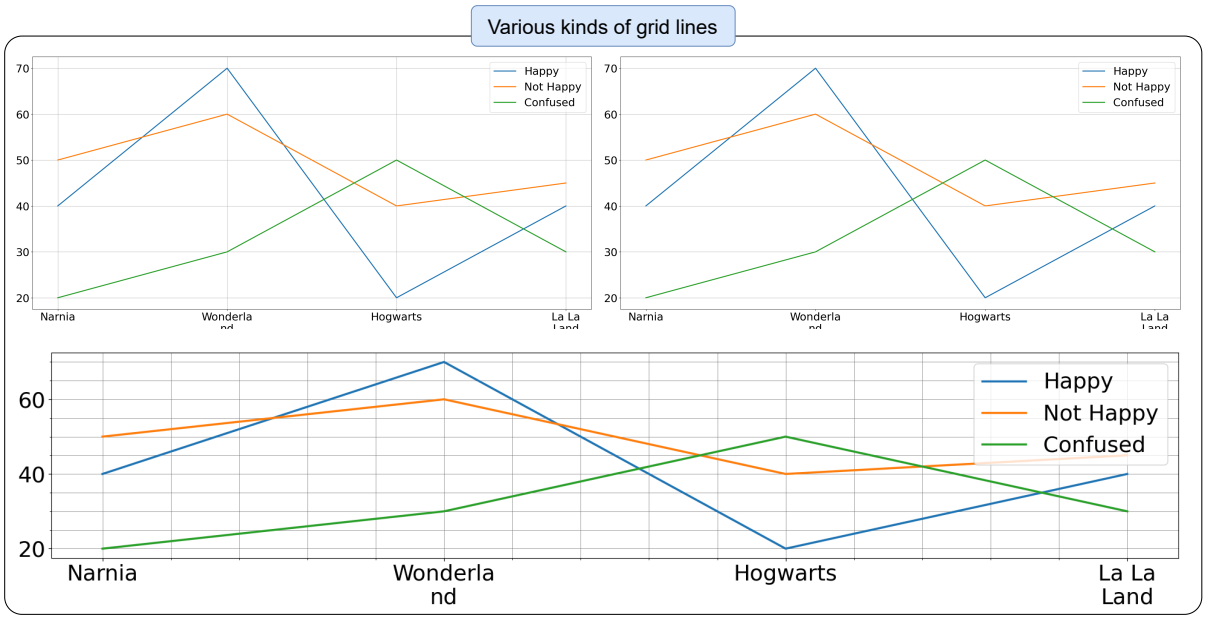
Colors in a scheme on figures

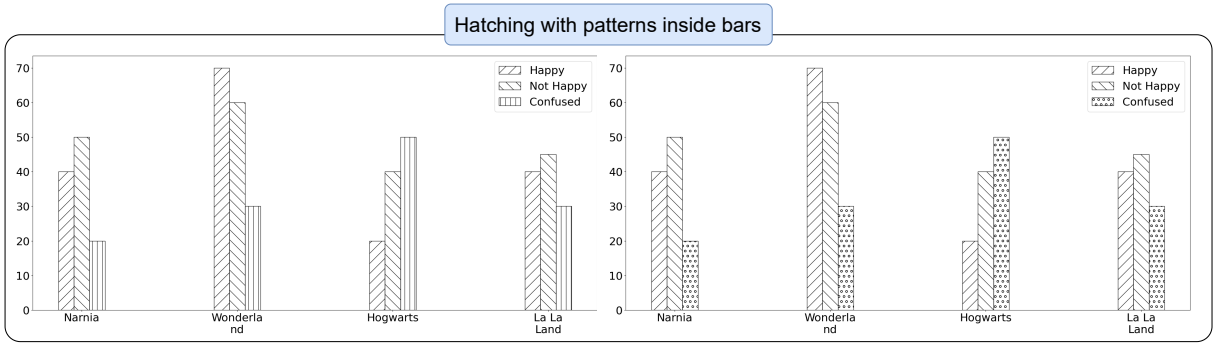


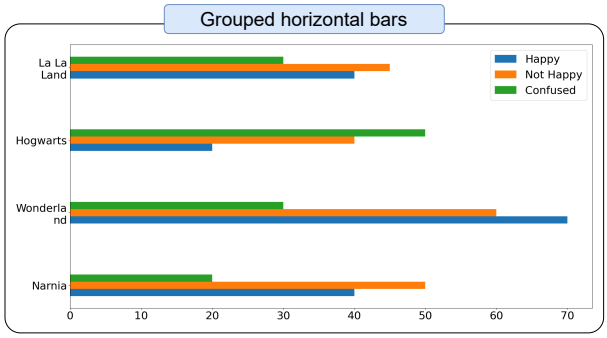
Changing Font Size

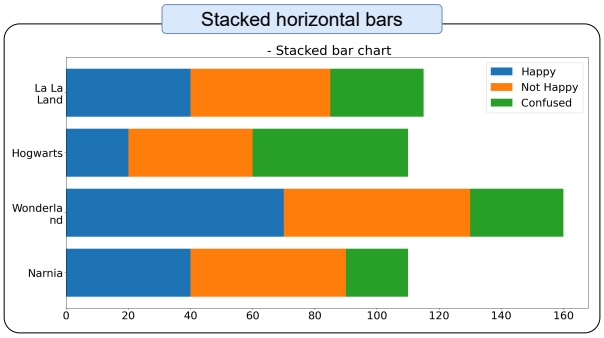


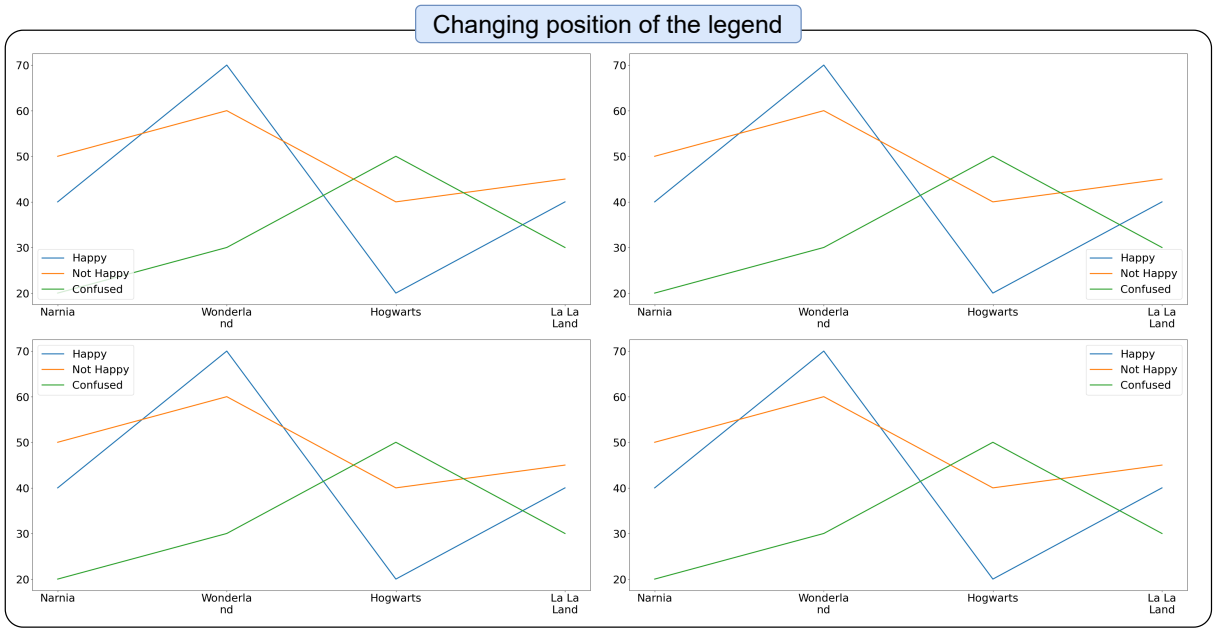

Pivoting Data by Column
Narnia
Wonderland
Hogwarts
La La Land
Happy
Not Happy
Confused

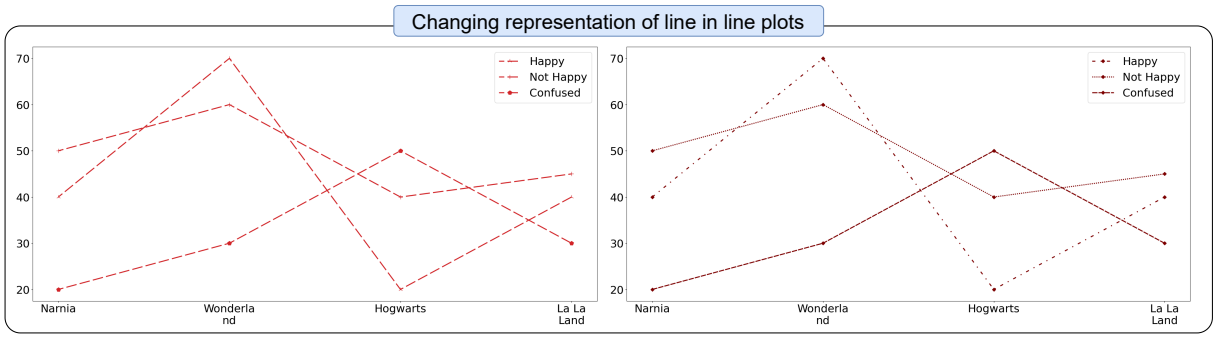

Various kinds of grid lines
Happy
Not Happy
Confused
Narnia
Wonderland
Hogwarts
La La Land


Hatching with patterns inside bars
Happy
Not Happy
Confused
Narnia
Wonderland
Hogwarts
La La Land

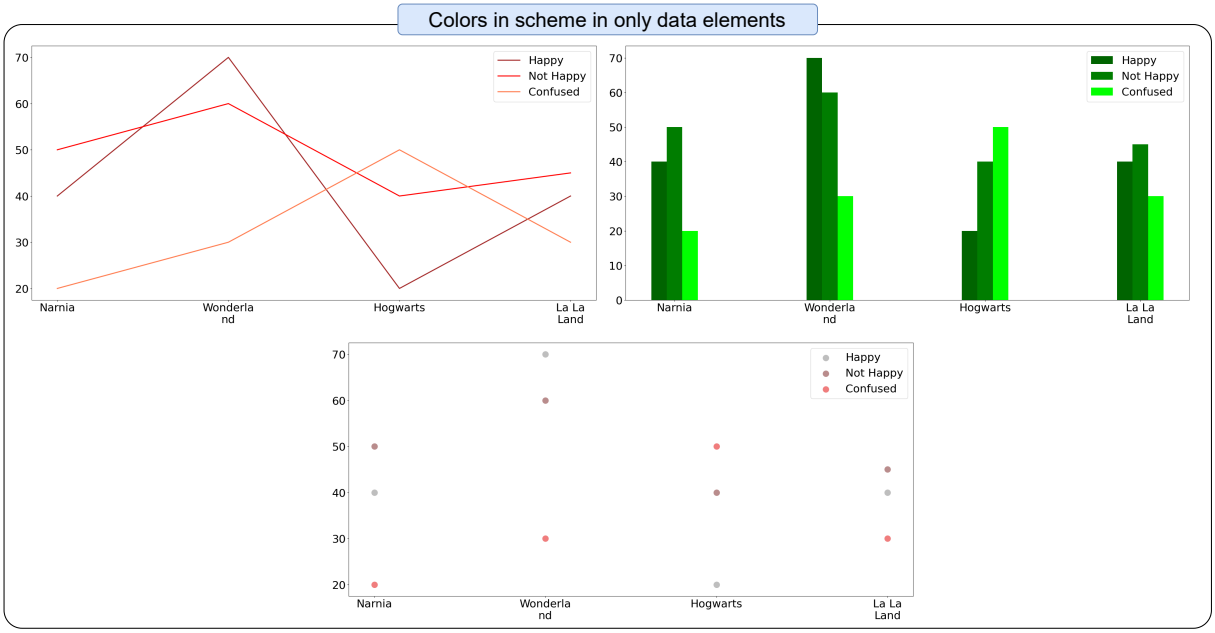

Grouped horizontal bars
La La Land
Hogwarts
Wonderla nd
Narnia
Happy
Not Happy
Confused
0
10
20
30
40
50
60
70


Stacked horizontal bars
- Stacked bar chart
La La Land
Hogwarts
Wonderla nd
Narnia
Happy
Not Happy
Confused
0
20
40
60
80
100
120
140
160

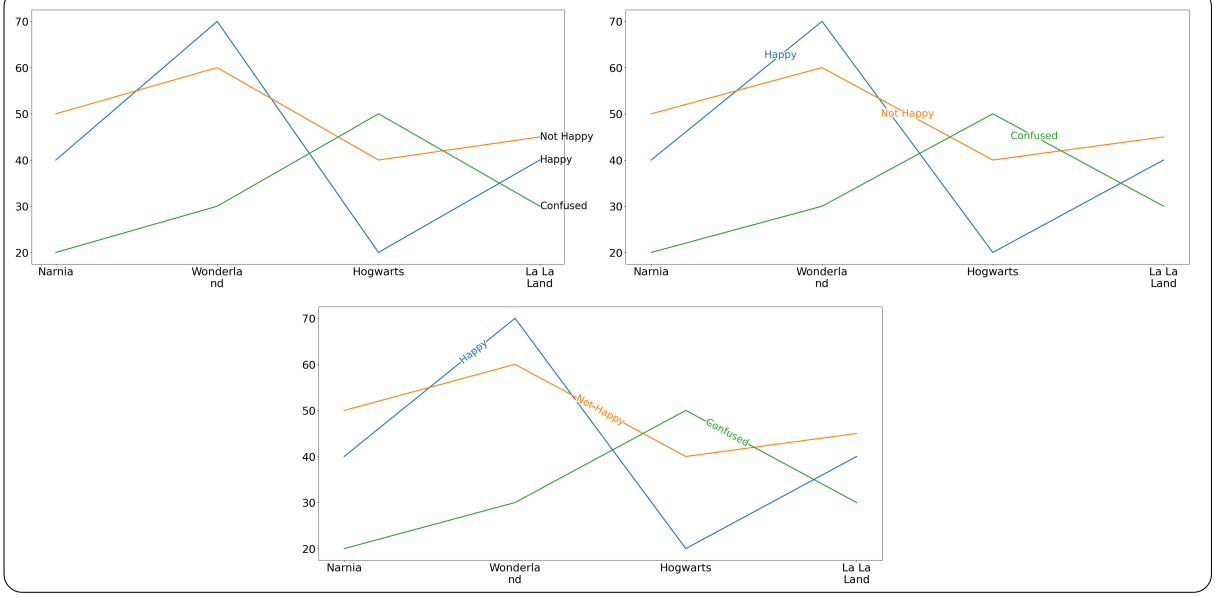

Changing position of the legend
Happy
Not Happy
Confused
Narnia
Wonderla nd
Hogwarts
La La Land
20
30
40
50
60
70


Changing representation of line in line plots
Happy
Not Happy
Confused
Narnia
Wonderla nd
Hogwarts
La La Land
20
30
40
50
60
70

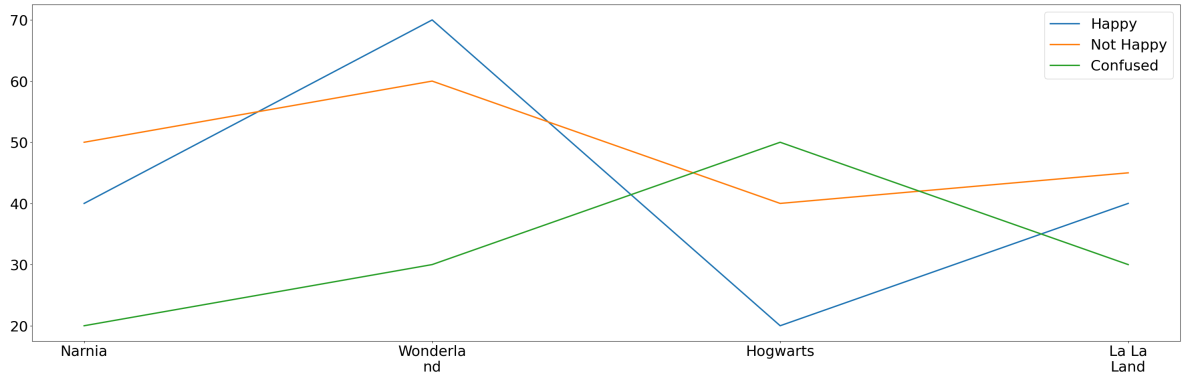
Colors in scheme in only data elements

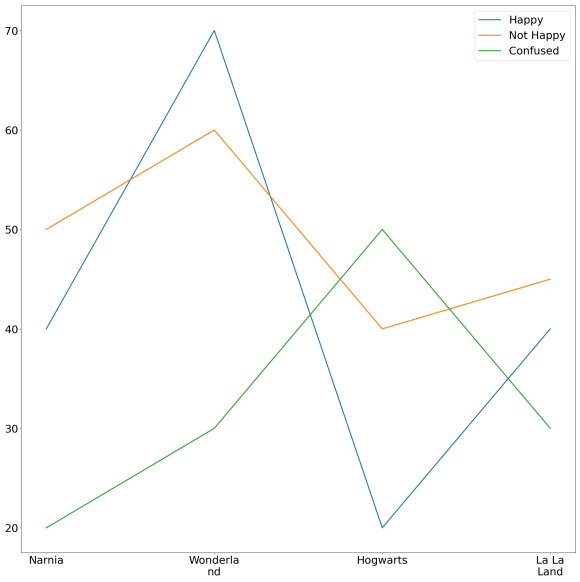


Replacing legend with labels in lines

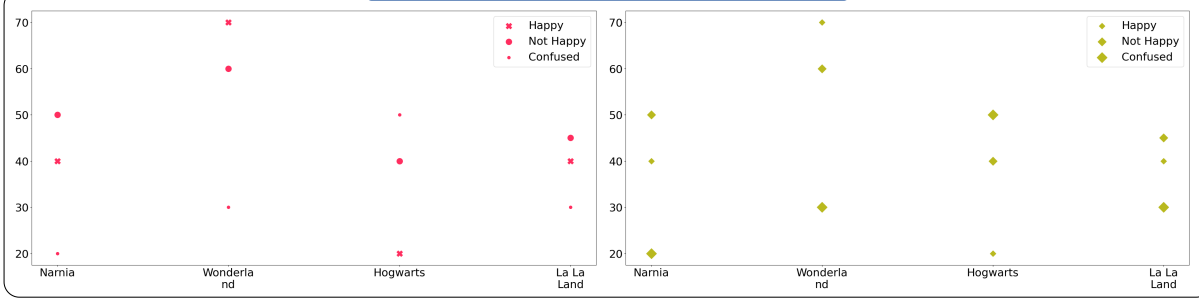


Changing the scale of the figure





Different representation for scatter plots



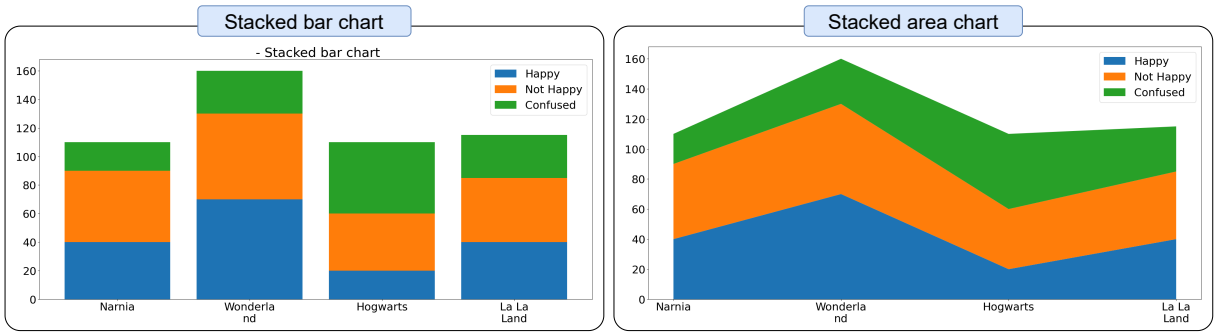
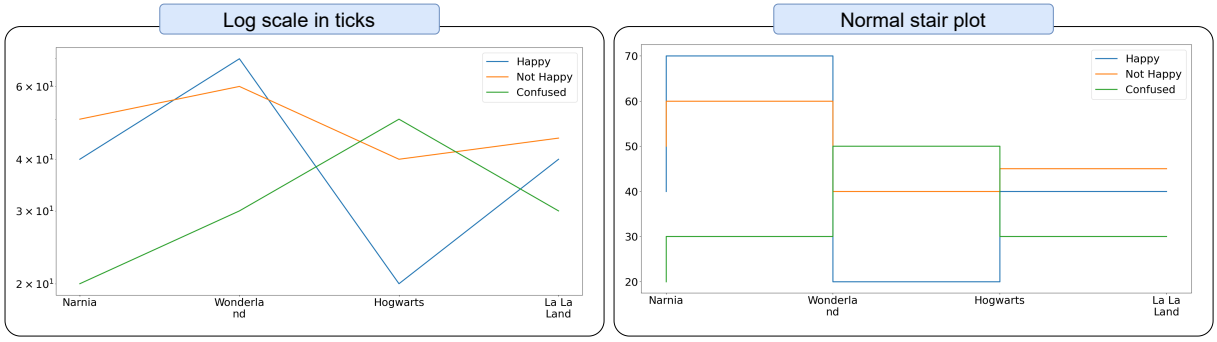
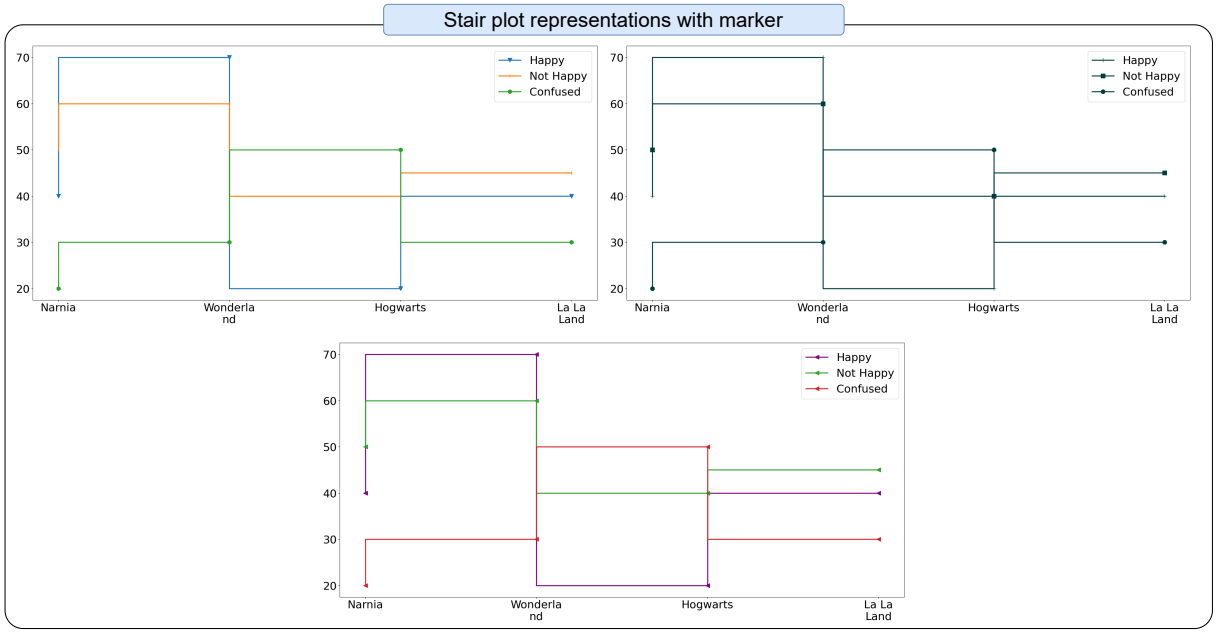
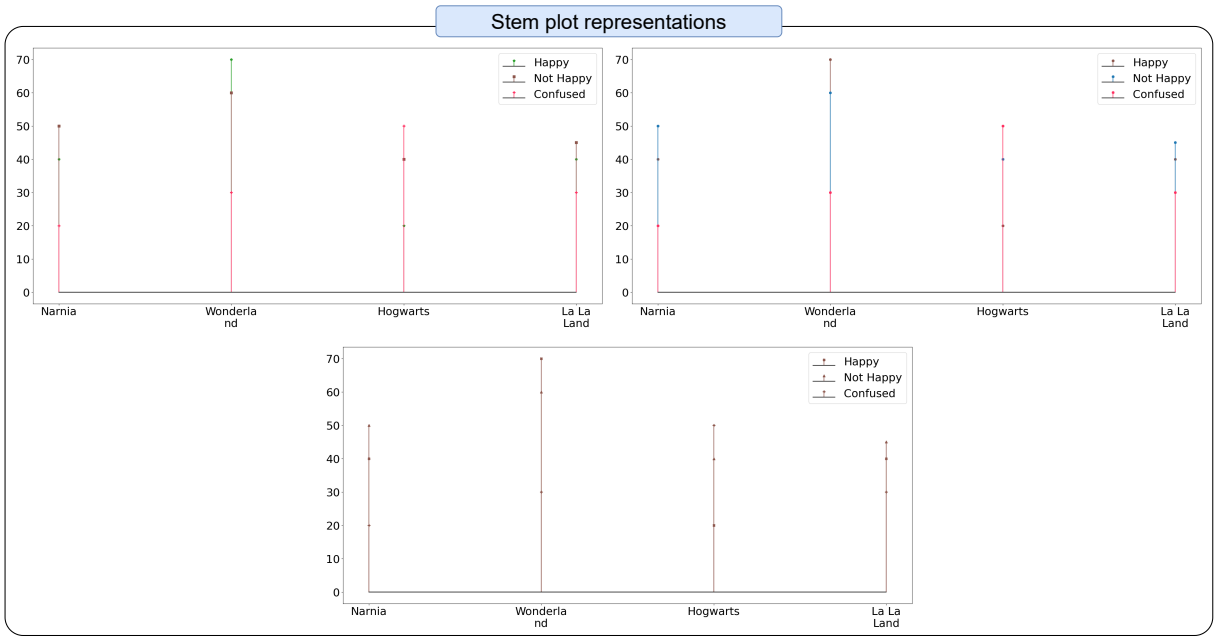
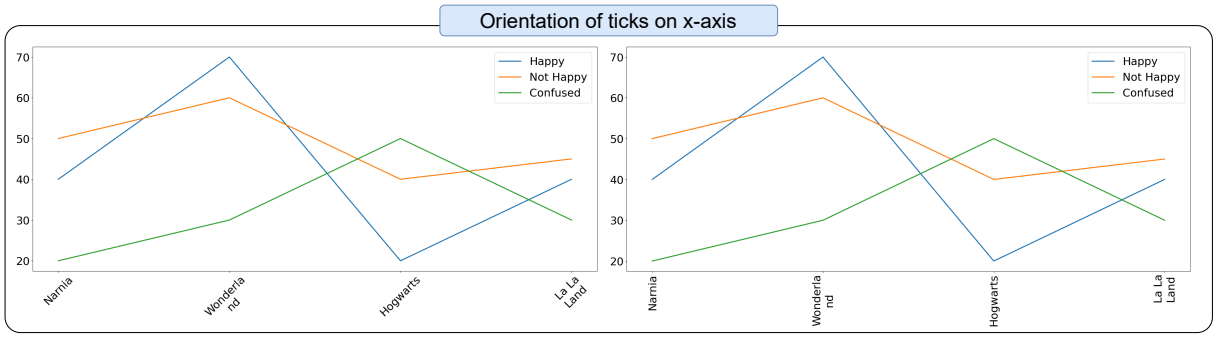


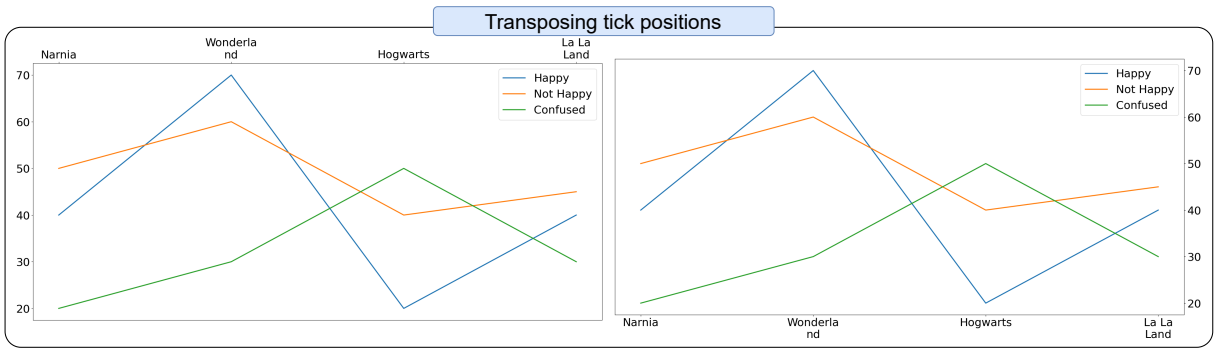
Figure 6: Stacked: Stacked bar graphs






Stem plot representations
Happy
Not Happy
Confused
0
10
20
30
40
50
60
70
Narnia
Wonderland
Hogwarts
La La Land


Orientation of ticks on x-axis
Happy
Not Happy
Confused
20
30
40
50
60
70
Narnia
Wonderland
Hogwarts
La La Land


Transposing tick positions
Happy
Not Happy
Confused
20
30
40
50
60
70
Narnia
Wonderland
Hogwarts
La La Land