# Analysis of LLMs for Visualization Understanding

*Thesis submitted by*

**Vinit Chandak**
**2022EET2109**

*under the guidance of*

**Prof. Sougata Mukherjea, Indian Institute of Technology Delhi**

*in partial fulfilment of the requirements*
*for the award of the degree of*

**Master of Technology**

**Department Of Electrical Engineering**
**INDIAN INSTITUTE OF TECHNOLOGY DELHI**

**June 2024**

# THESIS CERTIFICATE

This is to certify that the thesis titled **Analysis of LLMs for Visualization Understanding**, submitted by **Vinit Chandak (2022EET2109)**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

**Sougata Mukherjea**
Professor
Dept. of Electrical Engineering
IIT-Delhi, 110016

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to IIT Delhi and the Department of Electrical Engineering for granting me this invaluable opportunity to work on this project. Their forward-thinking approach in promoting project-based learning has provided an exceptional platform for my personal and professional growth. I am truly grateful to my project advisor, Prof. Sougata Mukherjea, for their invaluable guidance and support throughout the project. Their expertise has enabled me to develop a profound understanding of the key concepts and fundamentals involved. I would also like to thank our course coordinator Prof. Tapan K. Gandhi for his timely guidance throughout the course duration.

I would also like to extend my gratitude to Mr. Saadiq Rauf Khan for sharing valuable ideas that have greatly contributed to the successful completion of my work. Lastly, but certainly not least, I would like to express my heartfelt thanks to my parents and friends for their constant inspiration and unwavering support during times of need.

# ABSTRACT

A large language model (LLM) is a computational model notable for its ability to achieve general-purpose language generation and other natural language processing tasks such as classification. Based on language models, LLMs acquire these abilities by learning statistical relationships from vast amounts of text during a computationally intensive self-supervised and semi-supervised training process. LLMs can be used for text generation, a form of generative AI, by taking an input text and repeatedly predicting the next token or word.

Scientific figures are essential tools for communicating complex information concisely and intuitively. They visually represent data, revealing trends, relationships, and patterns between variables that might be difficult to grasp from text alone. Figures allow researchers to efficiently convey findings, support arguments, and facilitate understanding across disciplines. Additionally, they can enhance the aesthetic appeal of scientific publications and presentations, making the content more engaging and memorable.

Machine understanding of this structured visual information could assist human analysts in extracting knowledge from the vast documentation produced by modern science. Besides immediate applications, machine understanding of plots is interesting from an artificial intelligence perspective, as most existing approaches simply revert to reconstructing the source data, thereby inverting the visualization pipeline.

LLMs can be powerful tools for understanding scientific figures. They can analyze the underlying data, identify trends and relationships between variables, and answer questions about the information presented. By processing the visual elements alongside textual descriptions or captions, LLMs can provide a comprehensive understanding of the figure's meaning and implications. They can also generate textual summaries or explanations, making the figures more accessible to a wider audience. This capability is particularly beneficial for individuals with visual impairments or those who are not experts in the specific field of study.

This study focuses on analysis of the state of the art general purpose LLMs with respect to scientific visualizations and evaluates the level to which they understand them. To achieve our goal, we have used various approaches such as benchmarking with datasets, direct questioning and summary generation.

# Contents

# List of Tables

# List of Figures

# ABBREVIATIONS

| | |
|---|---|
| **LLM** | Large Language Model |
| **VQA** | Visual Question Answering |
| **CQA** | Chart Question Answering |

# Chapter 1

# INTRODUCTION

## 1.1 Motivation

### 1.1.1 Why scientific charts?

Scientific visualizations are graphical representations of data or concepts from scientific research, designed to make complex information more comprehensible. These visualizations include charts, graphs, maps, simulations, and 3D models, which transform raw data into a visual context, making patterns, trends, and correlations easier to identify and understand.

They are helpful because they offer a quick and easy approach to understand big datasets, make it easier to share research results, and improve understanding of complex scientific processes. They facilitate rapid comprehension of key findings and decision-making for scientists, researchers, and the general public by transforming numerical or abstract data into a visual format.

Some of the advantages that make them indispensable in research and communication are as follows:

1. Simplifying Complex Information: They transform intricate data sets into intuitive visual representations, making it easier to grasp patterns, trends, and relationships that might be difficult to discern from raw numbers alone.
2. Enhancing Understanding: Visualizations provide a quick and comprehensive overview of the data, allowing researchers to identify key insights and draw conclusions more efficiently.
3. Facilitating Communication: They serve as a universal language, effectively conveying findings to both experts and non-experts, bridging knowledge gaps and promoting collaboration.
4. Supporting Decision-Making: By presenting data clearly and concisely, visualizations empower researchers and policymakers to make informed decisions based on evidence.
5. Revealing Hidden Patterns: Through visual exploration, visualizations can uncover hidden patterns, correlations, or anomalies in data that might be overlooked in traditional analysis.
6. Enhancing Engagement: Visualizations are more engaging and memorable than plain text or tables, captivating the audience's attention and fostering a deeper understanding of the subject matter.

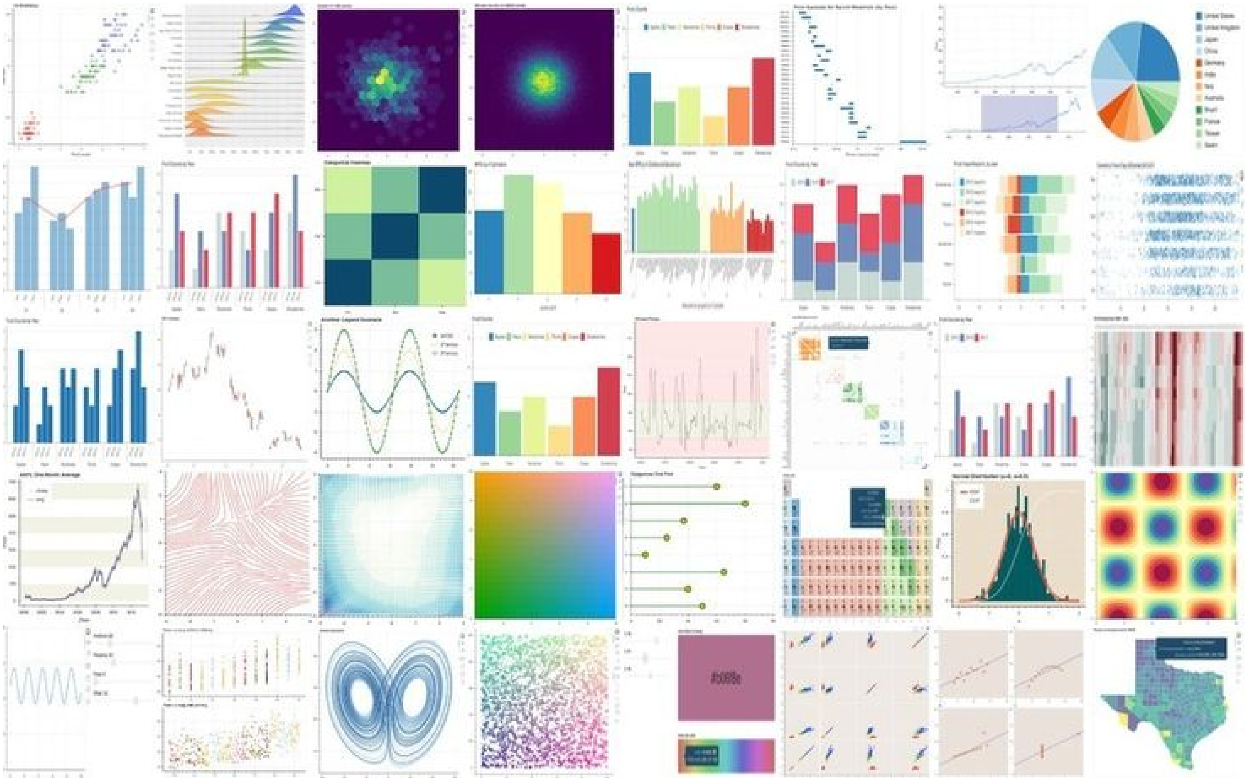


Figure 1.1: Some of the scientific figures/visulizations.

7. Promoting Transparency: By visualizing data, researchers can provide transparency and accountability in their work, allowing others to scrutinize and validate their findings.

### 1.1.2 Why do we want machines to understand them?

Machine understanding of these scientific visualizations is essential for several reasons. First, it accelerates data analysis by allowing machines to process and interpret large volumes of visual data quickly, which reduces the time researchers need to spend on manual interpretation. This increased efficiency is crucial in fields where timely insights can drive progress. Second, it enhances accuracy by providing consistent and objective extraction of information from visualizations, thereby minimizing human error and subjective biases. Third, machines can identify subtle patterns, correlations, and anomalies in the data that might be overlooked by human observers, leading to potential new scientific discoveries. Fourth, understanding visual data enables machines to integrate information from various sources, offering a more comprehensive and cohesive view of the scientific landscape. Finally, machine interpretation helps translate complex visual data into more accessible formats, making scientific findings more understandable to a broader audience, including those without specialized expertise. This democratization of knowledge can foster greater engagement and collaboration across different fields and communities.

### 1.1.3 Current methods and their drawbacks

Firstly, the current methods often rely on predefined algorithms and rule-based systems, which can be rigid and unable to adapt to the diversity and complexity of scientific data. This lack of flexibility makes it challenging to interpret visualizations that deviate from standard formats or incorporate novel representations.

Secondly, they lack the ability to generalize across different domains and types of visualizations. They are typically designed for specific tasks or datasets, requiring extensive reconfiguration or retraining to handle new types of visual data, which is both time-consuming and resource-intensive.

Additionally, these methods might not effectively handle noisy or ambiguous data. Scientific visualizations can sometimes contain imperfections, unclear labels, or overlapping elements, which traditional methods might misinterpret or fail to process accurately. Also, most existing approaches simply revert to reconstructing the source data, thereby inverting the visualization pipeline.

In contrast, LLMs offer more adaptability, contextual understanding, and generalization capabilities, making them better suited for the complex task of interpreting scientific figures and automating this.

### 1.1.4 How can LLMs help?

LLMs have the potential to greatly improve machine comprehension of scientific visualizations through several key capabilities. They excel at understanding natural language, enabling them to process textual elements surrounding a figure, such as captions and legends, and integrate this information with the visual data. This holistic approach allows for a more complete understanding of the visualization. Additionally, LLMs can dynamically learn and adapt to new patterns and structures within visual data, enhancing their flexibility to handle diverse and complex scientific figures. They can also leverage pre-existing knowledge from vast scientific literature, providing contextually relevant interpretations. Furthermore, LLMs can generate explanations, summaries, and insights in natural language, making the extracted information more accessible and usable for researchers. By bridging the gap between visual and textual data, LLMs offer a more nuanced and accurate interpretation, enhancing the overall utility and accessibility of scientific visualizations.

## 1.2 A brief overview of Large Language Models(LLMs)

Large Language Models (LLMs) represent a groundbreaking advancement in artificial intelligence and natural language processing. These sophisticated AI systems are trained on massive datasets of text and code, enabling them to understand, interpret, and generate human-like text.

### 1.2.1 How do LLMs work?

Large language models or LLMs typically have three architectural elements:

1. Encoder: After a tokenizer converts large amounts of text into tokens, which are numerical values, the encoder creates meaningful embeddings of tokens that put words with similar meanings close together in vector space.

2. Attention mechanisms: These algorithms are used in LLMs that enable the model to focus on specific parts of the input text, for related words of text. This is not separate from the encoder and decoder.

3. Decoder: The tokenizer converts the tokens back into words so we can understand. In this process, the LLM predicts the next word, and the next word, for millions of words. Once the models complete their training process, they can now accomplish new tasks such as answering questions, doing language translations, semantic search and more.[Dat23]

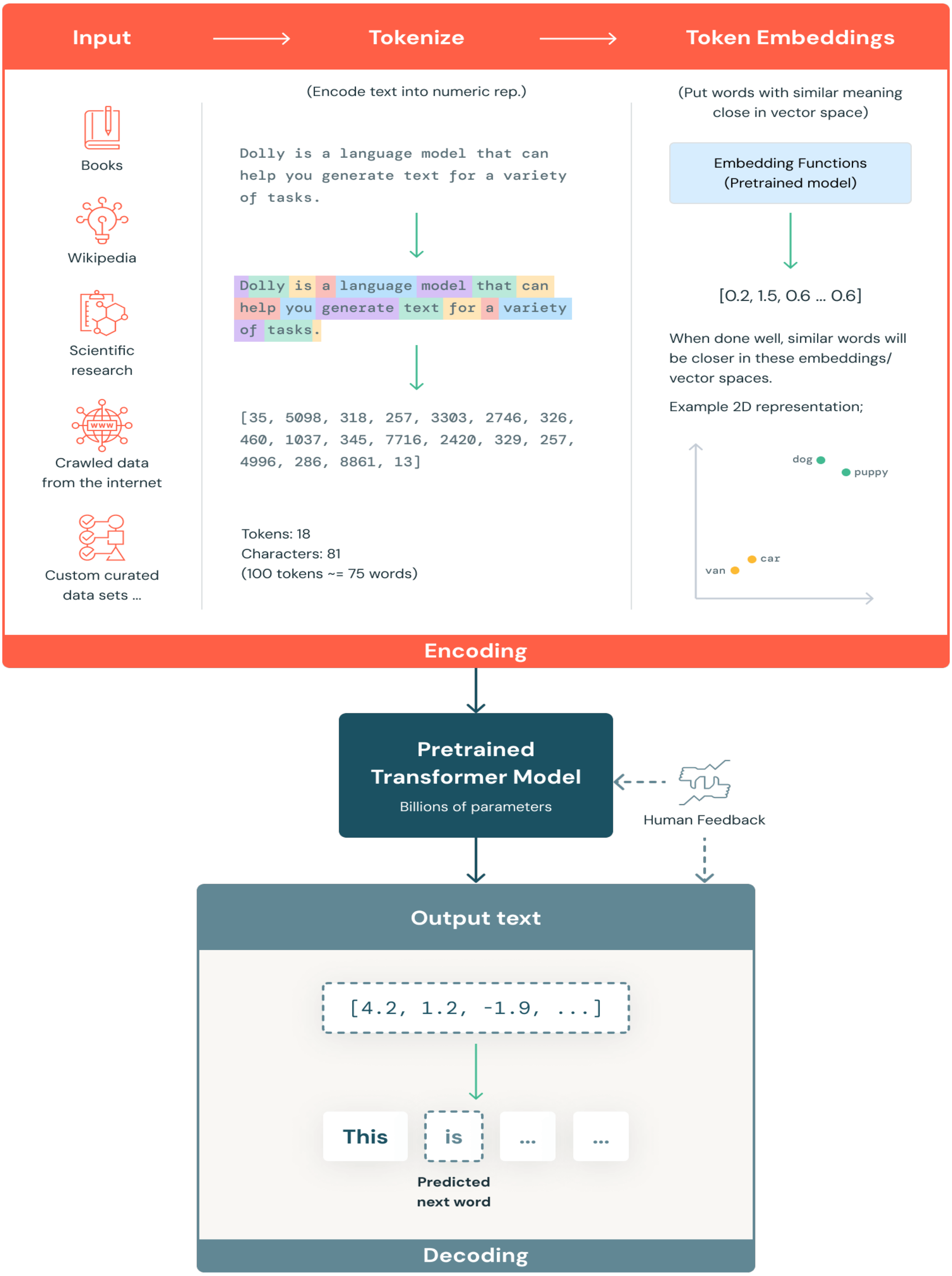



Figure 1.2: Simplified version of how LLMs work.[Dat23]

### 1.2.2 Capabilities

LLMs showcase an impressive range of capabilities:

- Text Generation: They can generate creative and informative text formats like essays, poems, code, scripts, and even musical pieces.
- Translation: LLMs can translate text between multiple languages with high accuracy.
- Summarization: They can condense lengthy documents into concise summaries, extracting key points and essential information.
- Question Answering: LLMs can provide relevant answers to a wide range of questions based on their extensive knowledge base.
- Chatbots and Conversational AI: They can engage in natural and dynamic conversations with users, offering information, assistance, or creative interactions.
- Content Creation: LLMs can assist in drafting emails, social media posts, marketing copy, and other forms of content.
- Code Generation: They can generate code snippets and even complete programs based on natural language descriptions.

### 1.2.3 Applications

The potential applications of LLMs span across various industries:

- Customer Service: Chatbots powered by LLMs can handle customer queries and provide support efficiently.
- Education: LLMs can serve as personalized tutors, answering questions and providing explanations.
- Healthcare: LLMs can analyze medical records and assist in diagnosis and treatment planning.
- Research: LLMs can accelerate research by summarizing literature, generating hypotheses, and analyzing data.
- Creative Writing: LLMs can assist authors and artists in generating ideas and even co-creating works.

### 1.2.4 Challenges and Limitations

While LLMs hold immense promise, they also present challenges. Occasionally, they could produce information that sounds plausible but is inaccurate or incomprehensible. THey are also prone to baises in the training data, leading to inaccurate or harmful outputs.

Furthermore, their large size and computational requirements pose practical challenges for deployment and use.

Also, the development and deployment of LLMs raise important ethical considerations. Ensuring fairness, transparency, and accountability is crucial to prevent misuse and mitigate potential harm.

It should also be true that this is a rapidly evolving field with massive investments being made, so the information which might be true today might become redundant or false tomorrow.

We've compared three of the leading LLMs for our analysis. All of these were the best offerings from these companies that were available publicly. They are as follows:

1. OpenAI's **GPT-4o** with the context window size of 128K tokens.
2. Google's **Gemini 1.5 Pro** with the context window size of 1 Million tokens.
3. Anthropic's **Claude 3 Opus** with the context window size of 200K tokens.

## 1.3 Organization of the thesis

The thesis has been thoughtfully structured into multiple chapters, allowing for a systematic and comprehensive investigation and analysis of the LLMs ability to comprehend and understand scientific figures and visualizations.

**Chapter 1: Introduction**

- Motivation
- A brief overview of LLMs
- Organization of thesis

**Chapter 2: Literature Review**

- Automatically Parsing Bar Charts
- Reasoning, Synthetic Scenes, and Diagrams
- Brief review of some of the latest papers on CQA and VQA in general.

**Chapter 3: Dataset and Preprocessing**

- Dataset Details
- Dataset Limitations

- Preprocessing

**Chapter 4: Methodology, results and discussion**

- Automated Analysis using FigureQA
- Need for manual analysis.
- Manual Analysis

**Chapter 5: Conclusion**

**Chapter 5: Future Work**

# Chapter 2

# LITERATURE REVIEW

This section reviews recent advancements in machine understanding scientific figures, exploring their potential for enhancing research, education, and data analysis. We focus on studies that address visual question answering (VQA) tasks on figures, multi-modal understanding of visualizations, and the challenges and opportunities presented by this rapidly evolving field. In this section, we will do an in-depth examination of the literature and relevant studies, analyzing their methodologies, contributions, and limitations. The goal is to provide a comprehensive overview of the current state-of-the-art and identify promising avenues for future research in this emerging field.

## 2.1 Automatically Parsing Bar Charts

Extracting data from bar charts using computer vision has been extensively studied [[AZG15], [SCZ11], [KSB13], [PH17], [SKC+11]]. Some focus on extracting the visual elements from the bar charts [PH17], while others focus on extracting the data from each bar directly [[SKC+11], [KSB13]]. Most of these approaches use fixed heuristics and make strong simplifying assumptions, e.g., [SKC+11] made several simplifying assumptions about bar chart appearance (bars are solidly shaded without textures or gradients, no stacked bars, etc.). Moreover, they only tested their data extraction procedure on a total of 41 bar charts.

## 2.2 Reasoning, Synthetic Scenes, and Diagrams

While VQA is primarily studied using natural images, several datasets have been proposed that use synthetic scenes or diagrams to test reasoning and understanding [[JHvdM+16], [KSK+16], [KSS+17]]. The CLEVR [JHvdM+16] dataset has complex reasoning questions about synthetically created scenes, and systems that perform well on popular VQA datasets perform poorly on CLEVR. The TQA [KSS+17] and AI2D [KSK+16] datasets both involve answering science questions about text and images. Both datasets are relatively small, e.g., AI2D only contains $15,000$ questions. These datasets require more than simple pattern matching and memorization. Their creators showed that state-of-the-art VQA systems for natural image datasets performed poorly on their datasets. However, there are key differences between these datasets and DVQA[KCPK18]. First, none of these datasets contain questions specific to bar charts. Second, their datasets use multiplechoice schemes that reduce the problem to a ranking problem, rather than the challenges posed by having to generate open-ended answers. Studies have shown that multiplechoice schemes have biases that models will learn to exploit [JJvdM16].

## 2.3 DVQA: Understanding Data Visualizations via Question Answering

Kafle et al. (2018) introduce DVQA, a groundbreaking dataset and associated models aimed at advancing machine understanding of bar charts. This work is directly relevant to the broader objective of my thesis, which seeks to assess and enhance Large Language Models' (LLMs) ability to interpret scientific figures.

**Key Contributions:**

- DVQA Dataset: The DVQA dataset is a significant contribution to the field, comprising over 3 million question-answer pairs based on 300,000 synthetic bar charts. The diversity in chart styles and question types (structure understanding, data retrieval, and reasoning) provides a comprehensive benchmark for evaluating models' abilities to interpret visual data and extract meaningful information.
- Challenges to VQA: The authors highlight the shortcomings of existing Visual Question Answering (VQA) models when applied to bar charts. They emphasize the need for algorithms capable of handling out-of-vocabulary (OOV) words, dealing with arbitrary label semantics, and adapting to the sparse and precise visual elements in bar charts. This analysis underscores the unique challenges posed by scientific figures and reinforces the importance of specialized approaches.
- Novel Models (MOM and SANDY): The paper proposes two novel models, MOM (Multi-Output Model) and SANDY (SAN with DYnamic Encoding), which outperform baseline VQA methods on DVQA. MOM leverages OCR to handle chart-specific answers, while SANDY employs a dynamic encoding model to address OOV words and generate chart-specific answers. These models demonstrate the potential for combining visual and textual information for more effective figure comprehension.

While DVQA is a significant step forward, it has some limitations such as the dataset being restricted to bar charts and the evaluation metrics used focus on exact string matching, which might not fully capture the nuances of understanding complex visualizations.[KCPK18]

## 2.4 Chart-to-Text: A Large-Scale Benchmark for Chart Summarization

Kantharaj et al. (2022) presented Chart-to-Text, a significant benchmark for chart summarization, comprising two large-scale datasets (Statista and Pew) with a total of 44,096 charts covering various topics and types. They proposed two problem variations: one with access to the underlying data table and another where the model extracts data from chart images. Several state-of-the-art neural models were adapted as baselines, utilizing image captioning and data-to-text generation techniques.

Their evaluation revealed that the best models, while generally fluent, struggled with hallucinations, factual errors, and conveying complex patterns. Specifically, the model's performance was better on the Statista dataset, which had accompanying data tables, compared to the Pew dataset, which relied on OCR-extracted data. Models with large-scale pretraining performed better, and the inclusion of bounding box information for OCR-extracted text improved performance slightly.

Limitations of their work include the focus primarily on summarization, a limited range of chart types compared to the broader spectrum in scientific literature, the reliance on synthetic data for the Statista dataset, and evaluation metrics that might not fully capture the nuances of understanding complex visualizations. Despite these limitations, Chart-to-Text serves as a valuable resource for advancing research in chart summarization and understanding scientific visualizations.[KLL+22]

## 2.5 RealCQA: Scientific Chart Question Answering as a Test-bed for First-Order Logic

Ahmed et al. (2023) focused on addressing the lack of annotated real-world data for chart visual question answering (QA) by introducing RealCQA, a new benchmark dataset derived from scientific literature. This dataset stands out for its inclusion of complex, real-world charts with diverse visual styles and structures, moving beyond the limitations of synthetic datasets used in previous works. RealCQA introduces a novel "list" answer type, offering both ranked and unranked variations, increasing the complexity and practicality of the task. Furthermore, the dataset includes new chart types like scatter and box plots, along with a variety of question categories based on a structured taxonomy.

The authors meticulously annotated a dataset of real-world charts from scientific papers, ensuring the diversity and complexity of the included visualizations. This involved extracting not only chart images but also their underlying data tables where available, allowing for both image-based and table-based QA tasks. The dataset's questions cover a wide range of topics and difficulties, ranging from simple factual questions to those requiring complex reasoning and first-order logic (FOL). The annotation process involved careful curation and validation of question templates to ensure a thorough evaluation of models' FOL capabilities.

However, while RealCQA is a significant step forward, it does have limitations. The dataset size, although larger than previous real-world chart QA datasets, might still be limited for training highly complex models. Additionally, the focus on template-based QA may not fully capture the diverse and nuanced nature of real-world questions asked about scientific figures. This limitation could be addressed in future work by incorporating more open-ended questions or using natural language processing techniques to generate more varied and complex queries.[AJP+23]

## 2.6 Answering Questions about Data Visualizations using Efficient Bimodal Fusion

Kafle et al. (2019) proposed the Parallel Recurrent Fusion of Image and Language (PReFIL) algorithm for Chart Question Answering (CQA). Their work focused on:

- PReFIL Algorithm: Developing a novel algorithm that combines low- and high-level image features with question embeddings to learn bimodal representations for answering complex questions about data visualizations.
- Benchmarking: Evaluating PReFIL on two synthetic datasets, FigureQA and DVQA, demonstrating superior performance compared to state-of-the-art methods and human baselines in most cases.
- Table Reconstruction: Introducing the novel task of table reconstruction by iteratively asking questions about a chart, showcasing PReFIL's potential in practical applications.

PReFIL's ability to fuse image and question features effectively allows it to outperform existing methods and human baselines on synthetic CQA datasets. The use of both low- and high-level image features enables PReFIL to capture both fine-grained details and global context in data visualizations. The dynamic OCR integration is crucial for handling out-of-vocabulary words and chart-specific labels, improving PReFIL's performance on DVQA. Limitations:

Reliance on Synthetic Datasets: The study focuses on synthetic datasets, FigureQA and DVQA, which may not fully represent the complexities and variations found in real-world charts. Dependence on OCR Accuracy: PReFIL's performance on DVQA is significantly influenced by the accuracy of the OCR system, highlighting the need for improved OCR methods for real-world applications. Limited Scope of Evaluation: The evaluation focuses primarily on accuracy, neglecting other aspects like explainability and quality of answers, which are important for real-world usability. Limited Chart Types: The study primarily focuses on bar charts, leaving the generalizability of PReFIL to other chart types, such as line graphs and pie charts, unexplored.[KSP+20]

## 2.7 SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs

Li and Tajbakhsh (2023) introduced SciGraphQA, a large-scale synthetic multi-turn question-answering dataset for scientific graphs. Their work focused on:

- Dataset Creation: Leveraging 290,000 scientific papers from arXiv and the Palm-2 language model to generate 295K question-answer dialogues around scientific graphs. These dialogues are enriched with contextual information like titles, abstracts, captions, and relevant paragraphs.
- Evaluation: Assessing the zero-shot capabilities of various Multimodal Large Language Models (MLLMs) on SciGraphQA, finding LLaVA-13B to be the most performant. Additionally, they fine-tuned LLaVA-13B on the dataset, further improving its performance.
- Quality Assessment: Utilizing GPT-4 to rate the quality of generated responses, indicating a high average rating of 8.7/10, confirming the dataset's validity.

Limitations:

- Synthetic Question Generation: Despite the large scale and use of real-world graphs, the questions and answers are still synthetically generated, potentially limiting their diversity and real-world applicability compared to human-generated questions.
- Evaluation Metrics: The evaluation relies on standard NLP metrics (BLEU-4, ROUGE, CIDEr), which may not fully capture the nuances of understanding complex scientific visualizations and reasoning.
- Limited Model Exploration: The study focuses primarily on evaluating existing MLLMs rather than developing novel architectures specifically tailored for scientific graph understanding.
- Restricted to Graphs: The dataset is limited to graphs, excluding other types of scientific visualizations like charts, diagrams, and tables.
- Potential for Bias: As the dataset is generated using a language model (Palm-2), there's a potential for inheriting biases present in the model's training data.[LT23]

## 2.8 Chart-based Reasoning: Transferring Capabilities from LLMs to VLMs

Carbune et al. (2024) propose a method to enhance the reasoning capabilities of Vision-Language Models (VLMs) by transferring knowledge from Large Language Models (LLMs). Their work focuses on:

- Improving Chart Representation: They refine the way charts are represented within the VLM by continuing the pre-training stage using an improved chart-to-table translation task. This allows the model to better understand the underlying data structure.
- Synthesizing Reasoning Traces: They create a 20x larger synthetic dataset with reasoning traces using LLMs. This helps the VLM learn complex reasoning patterns and numerical operations on the tabular representation of charts.
- Multitask Fine-tuning: The VLM is fine-tuned using a multitask loss that combines answer prediction and rationale generation, improving the model's ability to explain its reasoning process.

Conclusions:

- State-of-the-Art Performance: Their approach, applied to the PaLI-3 VLM, achieves state-of-the-art performance on the ChartQA benchmark, outperforming even larger models.
- Improved Generalization: Their method demonstrates improved performance on other chart QA datasets like PlotQA and FigureQA, indicating its effectiveness across different types of visualizations.
- Reasoning Enhancement: The use of synthetic data with reasoning traces and multitask fine-tuning significantly enhances the VLM's reasoning capabilities.
- Efficiency: Their approach maintains the inference time of the baseline PaLI-3 model, demonstrating its efficiency despite improved performance.
- Rationale Refinement: Refining rationales with program-of-thought prompts further boosts performance, surpassing even Gemini Ultra and GPT-4V on ChartQA.

Limitations:

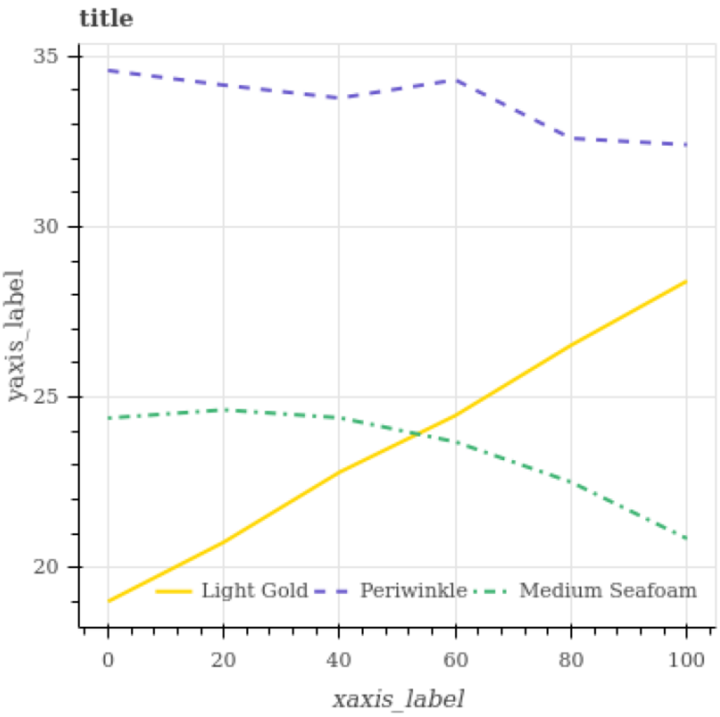
- Reliance on Table Representation: The synthetic data generation process requires access to a tabular representation of the charts, which may not always be available in real-world scenarios.
- Limited Color Reasoning: Their synthetic data lacks color metadata, resulting in the model struggling with reasoning tasks that involve color information.

- Challenges with Complex Reasoning: While the approach improves reasoning abilities, it still faces difficulties with highly complex reasoning tasks requiring intricate numerical computations or semantic understanding of chart elements.
- Limited Evaluation Metrics: The evaluation focuses primarily on accuracy, neglecting other aspects like the quality and explainability of the answers, which are crucial for real-world applications.
- Proprietary Model: Their work relies on the PaLI-3 model, a proprietary VLM not accessible to the broader research community, potentially limiting the reproducibility and extensibility of their findings.[CML+24]

# Chapter 3

# DATASET AND PREPROCESSING

For a part of our analysis, we've used the FigureQA dataset[KMA+18]. FigureQA consists of common scientific-style plots accompanied by questions and answers concerning them. The corpus is synthetically generated at large scale: its training set contains 100,000 images with 1.3 million questions; the validation and test sets each contain 20,000 images with over 250,000 questions.




**Q:** Does Medium Seafoam intersect Light Gold?
**A: Yes**

**Q:** Is Medium Seafoam the roughest?
**A: No**

**Q:** Is Light Gold less than Periwinkle?
**A: Yes**

**Q:** Does Periwinkle have the maximum area under the curve?
**A: Yes**

**Q:** Does Medium Seafoam have the lowest value?
**A: No**

Figure 3.1: Sample line plot figure with question-answer pairs.

The corpus represents numerical data according to five figure types commonly found in analytical documents, namely, horizontal and vertical bar graphs, continuous and discontinuous line charts, and pie charts. These figures are produced with white background and the colors of plot elements (lines, bars and pie slices) are chosen from a set of 100 colors. Figures also contain common plot elements such as axes, gridlines, labels, and legends. The authors have generated question-answer pairs for each figure from its numerical source data according to predefined templates. There are 15 questions types, given in Table 3.1, that compare quantitative attributes of two plot elements or one plot element versus all others.

In particular, questions examine properties like the maximum, minimum, median, roughness, and greater than/less than relationships. All are posed as a binary choice between yes and no. In addition to the images and question-answer pairs, we are provided with both the source data and bounding boxes for all figure elements, and supplement questions with the names, RGB codes, and unique identifiers of the featured colors. These are for optional use in analysis or to define auxiliary training objectives.

Table 3.1: Question Templates and Applicable Figure Types

| **Template** | **Question** | **Figure Types** |
|---|---|---|
| 1 | Is X the minimum? | bar, pie |
| 2 | Is X the maximum? | bar, pie |
| 3 | Is X the low median? | bar, pie |
| 4 | Is X the high median? | bar, pie |
| 5 | Is X less than Y? | bar, pie |
| 6 | Is X greater than Y? | bar, pie |
| 7 | Does X have the minimum area under the curve? | line |
| 8 | Does X have the maximum area under the curve? | line |
| 9 | Is X the smoothest? | line |
| 10 | Is X the roughest? | line |
| 11 | Does X have the lowest value? | line |
| 12 | Does X have the highest value? | line |
| 13 | Is X less than Y? | line |
| 14 | Is X greater than Y? | line |
| 15 | Does X intersect Y? | line |

### 3.0.1 Source data and Figures

The many parameters used to generate our source data and figures are summarized in Table 3.2. These constrain the data-sampling process to ensure consistent, realistic plots with a high degree of variation. Generally, data values are drawn from uniform random distributions within parameter-limited ranges. They further constrain the "shape" of the data using a small set of commonly observed functions (linear, quadratic, bell curve) with additive perturbations. A figure's data points are identified visually by color; textually (on axes and legends and in questions), we identify data points by the corresponding color names. For this purpose they've chosen 100 unique colors from the X11 named color set4, selecting those with a large color distance from white, the background color of the figures.

FigureQA's training, validation, and test sets are constructed such that all 100 colors are observed during training, while validation and testing are performed on unseen color-plot combinations. This is accomplished using a methodology consistent with that of the CLEVR dataset ([JHvdM$^+$16]), as follows. They've divided the 100 colors into two disjoint, equally-sized subsets (denoted A and B). In the training set, they color a particular figure type by drawing from one, and only one, of these subsets (see Table 3.2). When generating the validation and test sets, they draw from the opposite subset used for coloring the figure in the training set, i.e., if subset A was used for training, then subset B is used for validation and testing. This coloring for the validation and test sets is defined as the “alternated color scheme.” The placement of the legend within or outside the plot area is determined by a coin flip, and they select its precise location and orientation to cause minimal obstruction by counting the occupancy of cells in a $3 \times 3$ grid. Figure width is constrained to within one to two times its height, there are four font sizes available, and grid lines may be rendered or not – all with uniform probability.

### 3.0.2 Questions and Answers

The authors have generated questions and their answers by referring to a figure’s source data and applying the templates given in Table 3.1. One yes and one no question is generated for each template that applies. Once all question-answer pairs have been generated, they are filtered to ensure an equal number of yes and no answers by discarding question-answer pairs until the answers per question type are balanced. This removes bias from the dataset to prevent models from learning summary statistics of the question-answer pairs. Note that since they’ve provide source data for all the figures, arbitrary additional questions may be synthesized. This makes the dataset extensible for future research.

### 3.0.3 Plotting

The figures are generated from the synthesized source data using the open-source plotting library *Bokeh*. Bokeh was selected for its ease of use and modification and its expressiveness. The library’s web-based rendering component was modified to extract and associate bounding boxes for all figure elements. Figures are encoded in three channels (RGB) and saved in Portable Network Graphics (PNG) format.

Table 3.2: FigureQA: Synthetic Data Parameters

| **Figure Type** | **Elements** | **Points** | **Shapes** | **Color Scheme** (Training/Alternated) |
|---|---|---|---|---|
| Vertical Bar | 1 | 2-10 | Uniform random, linear, bell-shape | A / B |
| Horizontal Bar | 1 | 2-10 | Uniform random, linear, bell-shape | B / A |
| Line | 2-7 | 5-20 | Linear, linear with noise, quadratic | A / B |
| Dot Line | 2-7 | 5-20 | Linear, linear with noise, quadratic | B / A |
| Pie | 2-7 | 1 | N/A | A / B |

### 3.0.4 Limitations

A major limitation of the FigureQA dataset stemming from the binary nature of its questions is the susceptibility to random guessing. Since each question has only two possible answers (yes or no), models can achieve a 50% accuracy by simply making random choices without any actual understanding of the figure's content. This can lead to misleading performance evaluations, as models might appear to perform well despite lacking true comprehension of the visual information.

This issue is particularly problematic when assessing the reasoning capabilities of machine learning models. the depth and complexity required to fully evaluate a model's ability to interpret intricate relationships from visual data is often lacked by the question that are binary in nature. The oversimplification of answers to "yes" or "no" responses may mask underlying flaws in the model's reasoning process.

In the methodology chapter, we'll see how tried to overcome this limitation by using custom questions for our analysis which are not binary in nature. This helps for a more complete analysis and helps us to determine if the LLMs actually understand the underlying charts.

### 3.0.5 Preprocessing

Some basic preprocessing had to be done on the original FigureQA dataset. The *.json* file had some irregularities which had to be rectified. For some of the entries, the *image_index* field had incorrect data type which we fixed using a python script.

# Chapter 4

# METHODOLOGY, RESULTS AND DISCUSSION

## 4.1 Automated Analysis on FigureQA

Firstly, the models that we've selected for the analysis are all the best publicly available models from the leading AI companies. Hence, the APIs for all the models are paid and expensive. Working with the open source models or outdated ones would not have been meaningful as that would not allow us to accurately judge the current visualization capacities of the LLMs. Initially, we wanted to test the models for the entire FigureQA dataset i.e all of $100,000$ images and 1.3 million questions. But this would have been very very expensive, so we decided to use the sample train data which had 1000 images and $13,350$ questions.

We could only do this for the models provided by google since this was also very expensive and we only had credits for Gemini API. Below, we've shown the pseudo code for the script and some explanation for the same. Some major problems faced in this were w.r.t the API use, passing images to the model, extracting the answer from the model response and compliance with the API limits.

**Algorithm 1** Process and Analyze Images using Vertex AI's Gemini Model

**Initialize** configuration parameters
**Load** qa_pairs JSON file
**Set** Block Size and Rate Limiting Parameters
**procedure** ProcessImage(image_url, question, api_key)
    Initialize Vertex AI client
    Send {image_url, question} to Gemini API
    **Extract** and **Return** answer from response
**end procedure**
**procedure** AnalyzeAndSave(start_index, end_index)
    **For each item** in qa_pairs from {start_index} to {end_index}
    Retrieve image URL
    Get question string
    Process image and question using ProcessImage
    Store result in list
    Delay next request by {TIME_BETWEEN_REQUESTS}
    **Save** results to JSON file
**end procedure**
**Execute** analyze and save in blocks:
**for** block_num **in range** (0, total_images, BLOCK_SIZE) **do**
    Compute start_index and end_index for block
    AnalyzeAndSave(start_index, end_index)
**end for**

***process_ image* Function**:

- Initializes the Vertex AI client using the provided project ID and location.
- Constructs a request to the Gemini API using the image retrieved from the storage and the question text.
- Sends the constructed request and receives a response from the Gemini API.
- Extracts whether the response contains the word "yes" to determine the answer, and returns the result accordingly.

***analyze_ and_ save* Function**:

- Opens the JSON file that contains the QA pairs.
- Iterates over a specified range of QA pairs.
- For each QA pair:
  - Constructs the image URL for the Google Storage bucket.
  - Calls the *process_ image* function to obtain the answer.
  - Adds the retrieved answer to the QA pair.
  - Appends the processed QA pair to a results list.
  - Sleeps for a specified time to ensure compliance with rate limits.
- Saves the results for the current block in a new JSON file.

**Main execution**

- Reads the JSON file containing all QA pairs.
- Calculates the total number of images to be processed.
- Divides the processing tasks into blocks of a manageable size.
- For each block:
  - Computes the starting and ending indices for the block.
  - Calls the *analyze_ and_ save* function to process and save the results for the current block.

| **Metric** | Value |
|---|---|
| **Total Questions** | 13,350 |
| **Total Correct Gemini Answers** | 8,128 |
| **Total Wrong Gemini Answers** | 5,222 |
| **Accuracy** | 60.88% |

Table 4.1: Performance Metrics for Gemini-1.5-Pro

Now, this does not mean that the Gemini-1.5-pro understands 60% of the charts. All it means that it correctly answered 60% of the questions correctly. One pass through all of this 1000 images costs about $40 so we could not do this for GPT-4o and Claude 3 Opus. But comparing them was important.

So we decided to use a smaller subset of 100 images to do the same. We selected images whose index ends in 1 for this purpose, although any possible subset could be used. We selected these indices hoping to get more variation in image types and appearance. We wrote python scripts to do this for Gemini, OpenAI and Claude API's. Again, the same google cloude storage bucket was used to pass the images.

| **Metric** | Gemini-1.5-pro | Gpt4o | Claude 3 Opus |
|---|---|---|---|
| **Total Questions** | 1,342 | 1,342 | 1,342 |
| **Total Correct Answers** | 863 | 886 | 733 |
| **Total Wrong Answers** | 479 | 456 | 609 |
| **Accuracy (%)** | 64.31% | 66.02% | 54.61% |

Table 4.2: Comparison of Performance Metrics Between Gemini-1.5-pro and Gpt4o Models

As we can see, Gpt4o is the best performer, followed by Gemini-1.5-pro which is slightly behind and then Claude 3 Opus, which is much worse when compared to these two models. One thing to keep in mind is that sometimes Gpt4o gave very elaborate answers. Some examples of it are given below:[1]

**Question:** "Does Sandy Brown has maximum area under the curve?"
**Response:** "To determine which color has the maximum area under the curve from the data points in the scatter plot, we can observe the general trend of the points for each color. **Dark Magenta**: Points generally increase but seem to cap around 90. **Sandy Brown**: Points clearly have a steeper upward trend, reaching above 90. **Indian Red**: Points have a more gradual increase relative to the other two and seem to cluster in the middle range.From visual observation: The **Sandy Brown** points show a steeper and consistent upward trend compared to both Dark Magenta and Indian Red. The **Dark Magenta** points, although starting with a high upward trend, do not maintain as high

[1] These responses were extracted from the API response.

values across the entire x-axis as Sandy Brown. The **Indian Red** points have the least steep trend.Therefore, **Sandy Brown** has the maximum area under the curve based on the visible trends in the scatter plot."

**Question:** "Is Light Sky Blue the roughest?"
**Response:** 'The "roughness" of the Chartreuse series of data points isn't immediately clear from the graph, as "roughness" could refer to several characteristics such as variability or distribution. From the given graph, the Chartreuse points appear relatively consistent without a lot of fluctuation in their position, which suggests that they might not be the roughest in terms of variability. If "roughness" refers to variability, then the Light Sky Blue data series, which shows a small downward trend, might be considered rougher since it does exhibit some variation. However, for statistical clarity, further analysis like calculating the standard deviation or visualizing a pattern in the changes is required."

We've used a simple logic to extract answer. Since, all the questions are of "yes" or "no" type, we convert the LLMs response to lowercase and check whether the first word is "yes" or "no" and extract the answer. This works extremely well for most part except for such cases, which are few. In such cases (about 5% of all the responses), we use the default answer, which is set to "no". So, there might be some variance in the accuracy of Gpt4o.

## 4.2 Need for Manual Analysis

While this initial automated testing with the FigureQA dataset provided quantitative metrics for evaluating the performance of the selected LLMs, we know that relying solely on binary questions does not offer a comprehensive assessment of the model's true comprehension abilities. The binary nature of the FigureQA questions introduces a significant limitation: the susceptibility to random guessing. Models can achieve approximately 50% accuracy by making random choices without genuinely understanding the content of the figure/chart. This might lead to misleading performance evaluations, as the models might appear to perform well despite not understanding the underlying charts at all.

Binary questions, while convenient for initial testing, lack the depth and complexity necessary to evaluate a model's reasoning capabilities fully. They oversimplify the issue by restricting the options to a binary "yes" or "no," which fails to account for the complex relationships that must be understood in visual data. This oversimplification can mask underlying flaws in the models' reasoning processes and provide an inflated sense of their true capabilities.

To address this limitation, we moved beyond automated binary questioning and incorporated manual analysis as a crucial step in our methodology. This involved developing custom, non-binary questions aimed at probing deeper into the visual reasoning abilities of the models. By using questions that require more elaborate responses, we aim to provide a more complete and accurate picture of whether the models truly understand the underlying charts rather than merely guessing the answers.

In the subsequent sections, we will describe the process and findings of our manual analysis, which was essential to gauge the LLMs' genuine interpretative and reasoning skills. This approach ensures a more robust evaluation of whether the models can understand and extract meaningful insights from complex visual data, thereby overcoming the limitations posed by binary questioning in the FigureQA dataset.

## 4.3 Manual Analysis

For the manual analysis, we've selected 20 random charts for each of chart type in the FigureQA dataset, namely,

1. Vertical Bar Charts
2. Horizontal Bar Charts
3. Line Charts
4. Pie Charts

We've introduced new questions for each chart type that are useful to evaluate the level of understanding a model has of a chart. For each of the chart, we created a group of questions whose answers we thought would be a better indicator of LLMs understanding of those charts. Table 4.3 shows the types of questions added for each of the chart type. A subset of these questions was selected to evaluate the LLM for that particular chart.

Table 4.3: Questions used for the manual analysis, modifications on top of FigureQA dataset

| **Chart Name** | **Question Types(templates)** |
|---|---|
| Vertical Bar Chart | • How many bars are there?<br>• What are their colors?<br>• Which color has the maximum/minimum value?<br>• Is the bar with color X greater/larger than the bar with color Y?<br>• Is the value for the bar with color X the same as that of the bar with color Y?<br>• What is the value for the bar with color X?<br>• Which color bars are greater/larger than the bar with color X?<br>• Are there X bars? If yes, which color bar has the height that is in the middle?<br>• How many colors have *some word* in their name?<br>• Do bars with color X and color Y have the same value?<br>• How many shades of color X are there in the chart? |
| Horizontal Bar Graph | • How many bars are there?<br>• What are their colors?<br>• Which color has the maximum/minimum value?<br>• What is the value for the bar with color X?<br>• How many colors have *some word* in their name?<br>• Is the value for the bar with color X the same as that of the bar with color Y?<br>• What are the values of bars with color X and color Y respectively?<br>• Is there a bar with value 'X'? If yes, what is the color of that bar?<br>• Which bar is bigger among color X and color Y?<br>• Which color bars are bigger than the bar with color X? |

*Continued on next page*

Table 4.3 – *Continued from previous page*

| **Chart Name** | **Question Types(templates)** |
| --- | --- |
| Line Charts | • How many lines are there?<br>• What are their colors?<br>• What is the maximum/minimum value on the X-axis/Y-axis?<br>• Do lines with color X and color Y intersect?<br>• How many dotted/non-dotted lines are there?<br>• Do any of the lines intersect?<br>• Is there a straight line/point where all the lines intersect?<br>• Do any of the dotted and non-dotted lines intersect?<br>• How many intersection points are there?<br>• Does the non-dotted line intersect all the dotted lines? (Not necessarily at the same point)<br>• What is the maximum number of lines that intersect at a single point?<br>• Are there any straight dotted lines?<br>• Is there an intersection point in the chart? |
| Pie Chart | • How many pies are there?<br>• What are their colors?<br>• Which color pie has the most area?<br>• Does the lower half of the pie look like a pyramid?<br>• Which colors surround the pie with color X? |

**Evaluation Process:**

For each LLM, for each of the chart, we uploaded the image and then asked it all the questions in one go. All the answers were checked against my answers for the same questions. This, along with the inter-LLM comparison, will inherently compare the LLMs performances with a human baseline. We've also compared the models performances with and without **system prompts**.

**System Prompt**: The system prompt for the is the initial text or message that is provided by the user to the API in order to generate a response from the model. The system prompt can be thought of as the input or query that the model uses to generate its response. The quality and specificity of the system prompt can have a significant impact on the relevance and accuracy of the model's response. Therefore, it is important to provide a clear and concise system prompt that accurately conveys the user's intended message or question.

We used a simple system prompt given below, just like any normal user would use.

**Analyse the following chart carefully and answer the following questions correctly.**

### 4.3.1 Results of the manual analysis:

**Vertical Bar Charts:**

| | **Gemini 1.5 Pro** | **GPT-4o** | **Claude 3 Opus** |
|---|---|---|---|
| Images for which all questions were answered correctly without system prompt. | 55% | 55% | 15% |
| Images for which all questions were answered correctly with system prompt. | 60% | 70% | 15% |
| Questions answered without system prompt. | 86.2% | 91.1% | 50.9% |
| Questions answered correctly with system prompt. | 88.2% | 94.1% | 54.9% |
| Incorrect answers without system prompt. | 13.8% | 8.9% | 49.1% |
| Incorrect answers with system prompt. | 11.8% | 5.9% | 45.1% |

Table 4.4: Performance Comparison of Different Models for Vertical Bar Graphs

Key Observations:

1. Claude 3 Opus did not use the actual given color names for most of these charts. If it had some association of a color to a name, it used those.

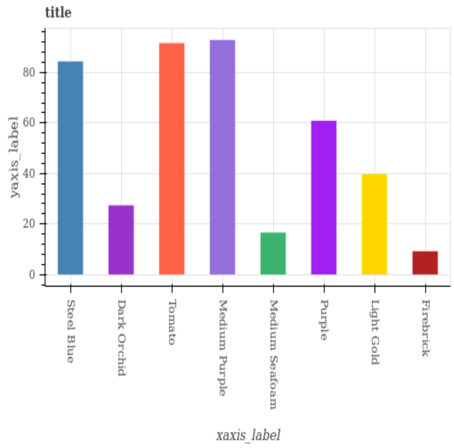



Figure 4.1: Example of a chart with close bars where all the models struggled.

2. All three models answered correctly when the heights of the bars were close, e.g.
3. In few of the charts, Claude 3 Opus hallucinated and stated things that are not present anywhere on the image, e.g.

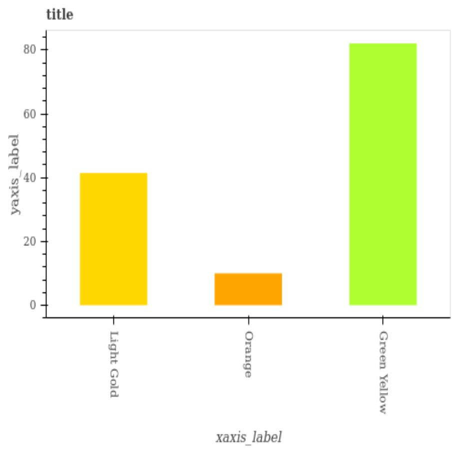



Figure 4.2: Claude 3 Opus hallucinated and found "South Africa" on this chart.

4. System prompt significantly improved the performance of Gemini-1.5-Pro and Gpt4o but did not have much effect on the performance of Claude 3 Opus.
5. From our limited testing, it seems that Gpt4o is the best model for tasks involving such charts.

**Horizontal Bar Charts:**

| | Gemini 1.5 Pro | GPT-4o | Claude 3 Opus |
|---|---|---|---|
| Images for which all questions were answered correctly without system prompt. | 55% | 50% | 25% |
| Images for which all questions were answered correctly with system prompt. | 70% | 55% | 30% |
| Questions answered without system prompt. | 86.7% | 84.6% | 64.2% |
| Questions answered correctly with system prompt. | 93.8% | 86.7% | 68.3% |
| Incorrect answers without system prompt. | 13.3% | 15.4% | 35.8% |
| Incorrect answers with system prompt. | 6.2% | 13.3% | 31.7% |

Table 4.5: Performance Comparison of Different Models for Horizontal Bar Graphs

1. Again, when the bar lengths are close, all of the models struggle in determining the larger/smaller bar, e.g.

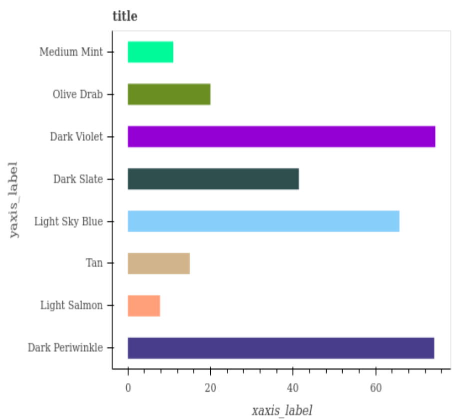



Figure 4.3: All three models incorrectly answered the color of the longest bar.

2. The system prompt significantly improved the performance of all three models, especially, Gemini-1.5-pro.

3. All the models are bad at determining the actual bar length. Gemini is comparitively the best for this, where as claude 3 almost always gets the length wrong, e.g.

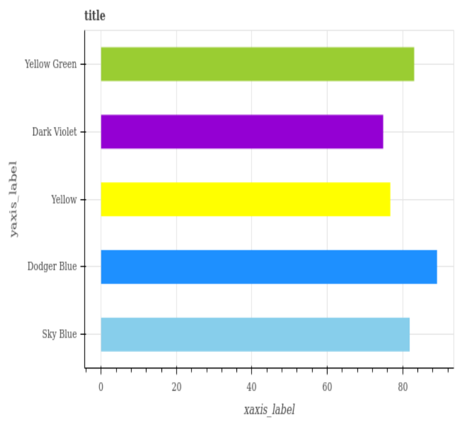



Figure 4.4: Claude 3 Opus and Gpt4o could not answer the correct length of yellow bar.

4. Claude did not use the given color names.
5. From our limited testing, it seems that Gemini-1.5-Pro is the best model for tasks involving such charts.

**Line Charts:**

| | **Gemini 1.5 Pro** | **GPT-4o** | **Claude 3 Opus** |
|---|---|---|---|
| Images for which all questions were answered correctly without system prompt. | 30% | 15% | 40% |
| Images for which all questions were answered correctly with system prompt. | 40% | 20% | 45% |
| Questions answered without system prompt. | 78.5% | 83.1% | 84.1% |
| Questions answered correctly with system prompt. | 84.1% | 84.1% | 88.7% |
| Incorrect answers without system prompt. | 21.5% | 16.9% | 12.2% |
| Incorrect answers with system prompt. | 15.9% | 15.9% | 11.3% |

Table 4.6: Performance Comparison of Different Models for Line Charts

1. Sometimes Gemini-1.5-Pro does not recognize the dotted lines at all. This always happened when there is a mixture of dotted and non-dotted lines, e.g.

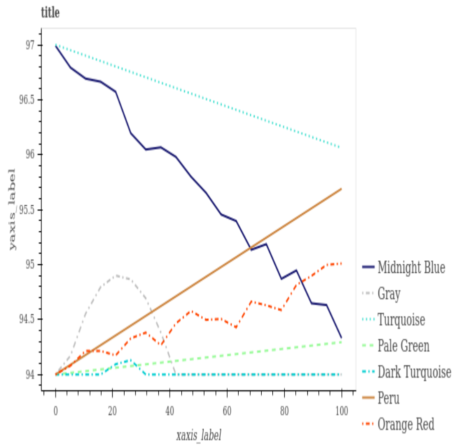



Figure 4.5: Gemini-1.5-Pro did not find any dotted line in this chart.

2. For the questions involving counting of lines/points, the models gave different answers when asked again 15% of the time, e.g.

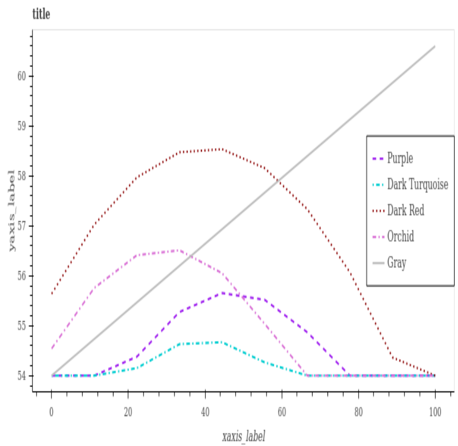



Figure 4.6: When asked again about the number of dotted lines, the models gave different answers.

3. All the models perform poorly on line charts, as compared to other type of the charts.
4. Surprisingly, Claude 3 Opus is the best performer in our limited testing.
5. The system prompt significantly improved the performance of Gemini-1.5-Pro.

**Pie Charts:**

| | **Gemini 1.5 Pro** | **GPT-4o** | **Claude 3 Opus** |
|---|---|---|---|
| Images for which all questions were answered correctly without system prompt. | 75% | 85% | 55% |
| Images for which all questions were answered correctly with system prompt. | 80% | 85% | 65% |
| Questions answered without system prompt. | 90.3% | 95.1% | 82.2% |
| Questions answered correctly with system prompt. | 91.9% | 95.1% | 85.4% |
| Incorrect answers without system prompt. | 9.7% | 4.9% | 17.8% |
| Incorrect answers with system prompt. | 8.1% | 4.9% | 14.6% |

Table 4.7: Performance Comparison of Different Models for Pie Charts

1. Most of the models performed much better on pie charts as compared to other chart types. Best performing model was Gpt4o.
2. System prompts did not help much in this as the accuracy was already pretty high.
3. We did not ask any questions about the area as we ourselves could not find it out and had no answers to it. Maybe that could be done in some future work.
4. For the most part, Claude 3 Opus uses its own color names and not the ones given in the chart.
5. Gpt4o and Claude 3 Opus again got answers to the questions that involved close things, e.g.

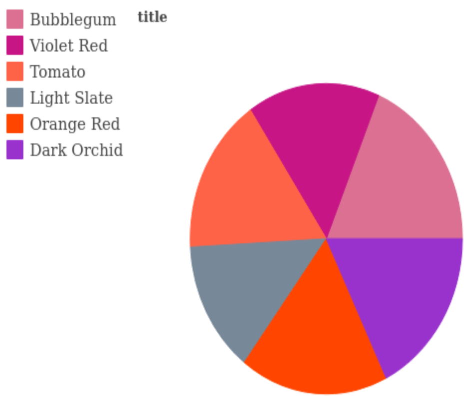



Figure 4.7: Only Gemini-1.5-pro answered this correctly.

**Overall Statistics:**

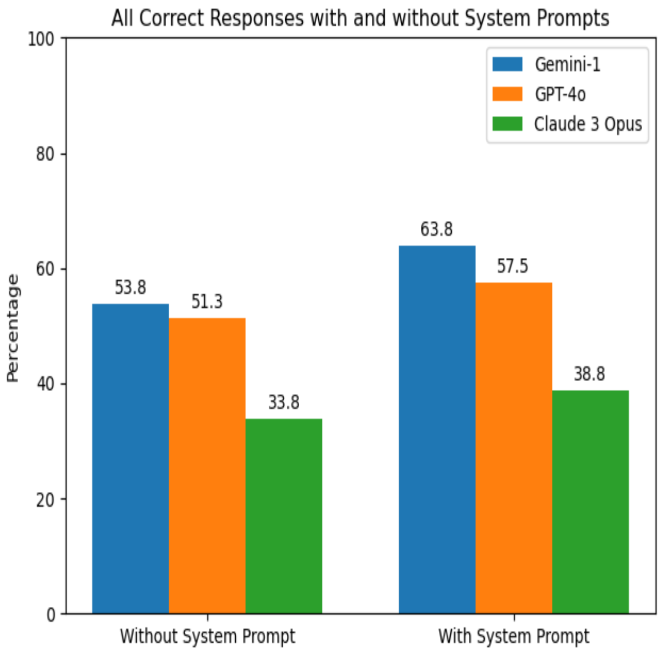



Figure 4.8: % of images with all questions answered correctly.

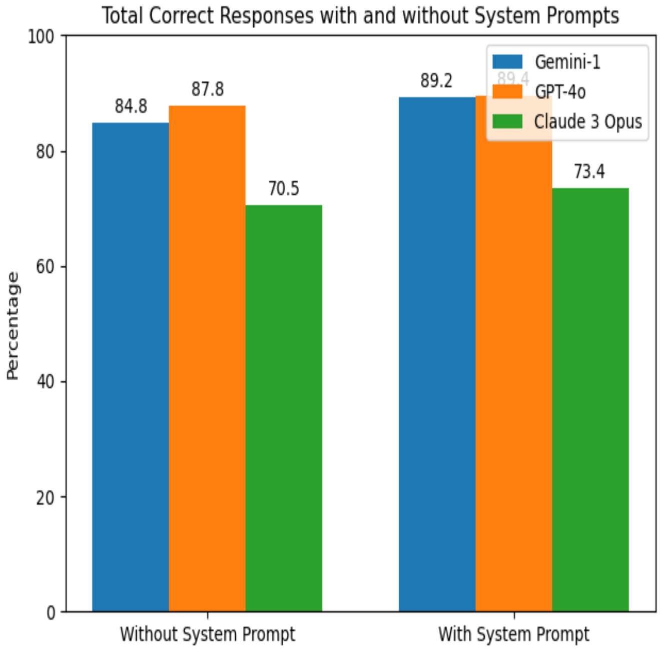



Figure 4.9: % of questions answered correctly.

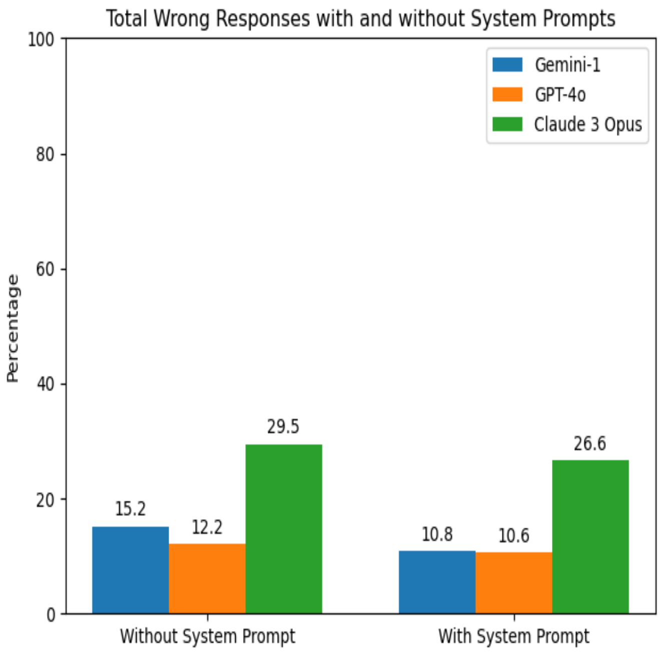



Figure 4.10: % of questions answered incorrectly.

# Chapter 5

# CONCLUSION

In this thesis, we explored the capabilities of LLMs in understanding and interpreting scientific charts. To achieve this, we decided to use three of the leading LLMs-OpenAI's GPT4o, Google's Gemini-1.5-Pro, and Anthropic's Claude 3 Opus for comparison on tasks involving scientific charts and figures.

We looked at some of the latest research in the field of CQA, highlighting their contributions but also pointed out some of their shortcomings, especially the use of binary questioning in datasets such as FigureQA. To address these limitations, we incorporated manual analysis with more nuanced, open-ended, non-binary questions. This approach allowed us to more accurately assess the genuine reasoning and interpretative abilities of these LLMs.

Using a combination of automated and manual analysis, we gained several key insights as follows:

1. **Performance Across Different Chart Types**
   - The performance of the LLMs varied heavily among different chart types. For example, Gpt4o had 85% accuracy on pie charts with system prompt and a mere 20% on line charts. This suggest that certain visualization formats may be much easier for machines to interpret than others.
   - All the models performed very poorly on the line charts. This might be because the presence of dotted lines which might be treated as some kind of noise by the models. Most of them got the questions related to the dotted lines wrong.
   - All the models struggled with identifying relationships between close boundaries and lengths of shapes.

2. **Impact of System Prompts**
   - In all the cases, use of system prompts improved the models performance. It's just that the improvement varied with the models and chart types.
   - The accuracy of Gemini-1.5-Pro improved significantly with the use of system prompts.
   - Conversely, for models like Claude 3 Opus, which already had high accuracy, the addition of system prompts did not have that much impact.
   - But the fact that in all the cases we had improvements highlights the importance of context and guidance in enhancing model outputs.

3. **Challenges and Limitations**

- The models struggled with certain complex visual elements and exhibited inconsistencies in answering repetitive questions.
- Due to financial and computational(rate-limits) constraints, we could not test the models on the entire FigureQA dataset. This emphasizes the need for more cost-effective and efficient access to the LLMs.
- We also need to create robust evaluation metrics and more sophisticated datasets that better capture the complexities of real-world scientific visualizations.

The results reinforce our beliefs on the capabilities of the LLMs to revolutionize how we interact with and extract knowledge from scientific figures. By advancing these capabilities, we can make the scientific information more accessible and interpretable. LLMs also have the potential to help automate complex data analysis tasks. While our findings show that LLMs have made promising strides in understanding basic chart elements and answering straightforward questions, there's still a long way to go. They struggle with moderately complex reasoning and interpretations.

# Chapter 6

# FUTURE WORK

There are several avenues for future work that can help us understand if the LLMs are ready for visualizations, especially scientific charts.

1. **Better datasets**
   - More diverse and comprehensive datasets should be created that contain a wide variety of charts and real-world scientific visualizations.
   - The datasets should contain a greater number of open-ended, contextually rich questions. This will show us whether the models actually comprehend the charts and allow us to assess their reasoning skills to some degree.
2. **Use of summarization for analysis**
   - We need to implement and test summarization techniques to assess the LLMs capabilities in generating accurate summaries of scientific charts and figures. This will help us understand how well the models can condense visual data into meaningful information.
   - We also need to develop a way to evaluate the quality of these summaries. It could be automated based on some criteria or we can have humans to do that job. But there has to be some standard that separates the good, the bad and the ugly.
3. **Standard Evaluation Metrics**
   - We need to find a way to standardize the evaluation of LLMs on scientific chart interpretation tasks. This will help us get comparable and consistent assesments accross different studies and models.
   - We need metrics that go beyond accuracy.
4. **Insight Discovery**

   Can LLMs identify trends, outliers, and relationships within the data visualized in the charts? We need to find a way to answer this question.
5. **Comparative Analysis**

   We need to test if the LLMs can compare and contrast multiple charts and draw connections between them. This will test their ability to handle a broader visual context.

# Bibliography


[AJP+23] Saleem Ahmed, Bhavin Jawade, Shubham Pandey, Srirangaraj Setlur, and Venu Govindaraju. *RealCQA: Scientific Chart Question Answering as a Test-Bed for First-Order Logic*, pages 66–83. 08 2023.

[AZG15] Rabah Al-Zaidy and C. Giles. Automatic extraction of data from bar charts. pages 1–4, 10 2015.

[CML+24] Victor Carbune, Hassan Mansoor, Fangyu Liu, Rahul Aralikatte, Gilles Baechler, Jindong Chen, and Abhanshu Sharma. Chart-based reasoning: Transferring capabilities from llms to vlms, 2024.

[Dat23] Databricks. What are large language models (llms)? `https://www.databricks.com/glossary/large-language-models-llm`, 2023. Accessed: 2024-06-10.

[JHvdM+16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. 12 2016.

[JJvdM16] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. 06 2016.

[KCPK18] Kushal Kafle, Scott D. Cohen, Brian L. Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018.

[KLL+22] Shankar Kanthara, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq R. Joty. Chart-to-text: A large-scale benchmark for chart summarization. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[KMA+18] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018.

[KSB13] Jagadish Kallimani, K. Srinivasa, and Eswara B. Extraction and interpretation of charts in technical documents. pages 382–387, 08 2013.

[KSK+16] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. volume 9908, pages 235–251, 10 2016.

[KSP+20] Kushal Kafle, Robik Shrestha, Brian Price, Scott Cohen, and Christopher Kanan. Answering questions about data visualizations using efficient bimodal fusion. pages 1487–1496, 03 2020.

[KSS+17] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017.

[LT23] Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs, 08 2023.

[PH17] Jorge Poco and Jeffrey Heer. Reverse-engineering visualizations: Recovering visual encodings from chart images. *Computer Graphics Forum*, 36, 2017.

[SCZ11] Stephanie Schwartz, Sandra Carberry, and Ingrid Zukerman. The automated understanding of simple bar charts. *Artif. Intell.*, 175:526–555, 02 2011.

[SKC+11] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. Revision: automated classification, analysis and redesign of chart images. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011.

# .1 APPENDIX

All the project files are available at the following GitHub repository: Major Project