OPEN ENDED QUESTION ANSWERING WITH CHARTS

SHANKAR KANTHARAJ

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO
SEPTEMBER 2022

Abstract

Charts are very popular to analyze data and convey important insights. People often analyze visualizations to answer open-ended questions that require explanatory answers. Answering such questions are often difficult and time-consuming as it requires a lot of cognitive and perceptual efforts. To address this challenge, we introduce a new task called OpenCQA, where the goal is to answer an open-ended question about a chart with descriptive texts. We present the annotation process and an in-depth analysis of our dataset. We implement and evaluate a set of baselines under three practical settings. Our analysis of the results show that the top performing models generally produce fluent and coherent text while they struggle to perform complex logical and arithmetic reasoning.

Acknowledgements

First of all, I would like to express my sincerest gratitude to my thesis supervisor, **Prof. Enamul Hoque Prince**, for his endless support from the very first day and throughout my masters degree. His valuable feedback and suggestions helped me improve the quality of this thesis work and submit it to a prestigious venue. I am very thankful to him for teaching and guiding me to conduct high-quality research in ML/NLP. It was a great honor for me to work with him for two years.

I would like to thank **Prof. Manos Papagelis** for being a committee member of my thesis.

I also want to thank **Prof. Shafiq Joty**, who is one of the co-authors of this work, for his great feedback and comments. I would like to thank **Do Xuan Long, Rixie Tiffany Ko Leong and Jia Qing Tan** for their collaboration and assistance in this work.

Finally, I would like to thank my parents for their encouragement and my sister and her family for supporting me during my time in this graduate program.

Preface

This thesis work was mainly done by **Shankar Kantharaj** under the supervision of **Prof. Enamul Hoque Prince**. This work is currently accepted for the EMNLP 2022 conference (First paper in list below). My role in this work was formulating the problem definition and conducting dataset construction through a user study, dataset post processing/analysis, model experimentation, output analysis and manuscript writing. My coauthors **Enamul Hoque & Shafiq Joty** provided guidance and assistance in manuscript writing. Coauthors (Do Xuan Long, Rixie Tiffany Ko Leong, Jia Qing Tan) assisted in dataset post processing/analysis, model experimentation, output analysis and manuscript writing. In the other paper [2] where I am listed as a first author, I assisted in dataset construction, model experimentation, metric development and manuscript writing.

1. **Shankar Kantharaj**, Do Xuan Long, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, Shafiq Joty "OpenCQA: Open-ended Question Answering with Charts", [**Accepted for 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)**]

2. **Shankar Kantharaj*** [1], Rixie Tiffany Ko Leong*, Xiang Lin*, Ahmed Masry*, Megh Thakkar*, Enamul Hoque, Shafiq Joty, "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization", accepted in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL'22) 2022.

[1] * Equal contribution. Listing order is based on the alphabetical ordering of author surnames

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

Using data visualizations such as bar charts and line charts to discover critical insights and explain them to others is at the heart of many decision making tasks [54]. Often, people explore such visualizations to answer high-level questions that involve reasoning and explanations. For example, Figure 1.1 shows an open-ended question which cannot be answered by a single word or phrase, rather it requires an explanatory answer. Answering such questions can be time consuming and mentally taxing as they require significant amounts of perceptual and cognitive efforts. Further, understanding and reasoning the visuals in a chart requires interpreting a combination of many different aspects of the chart to derive a high level overview, which can be very challenging. For the particular question in Figure 1.1, the user needs to find relevant marks (bars) in the given charts, compare their values and perform reasoning over them to generate an explanatory answer. Thus, the research question we address in this thesis is: can we build systems to automatically answer such open-ended questions about charts with descriptive texts?

Question
Compare the Democrats and Republicans views about providing health care to the population?
Answer
While 83% of Democrats say providing high - quality, affordable health care for all should be a top priority, a much smaller share of Republicans (48%) agree.

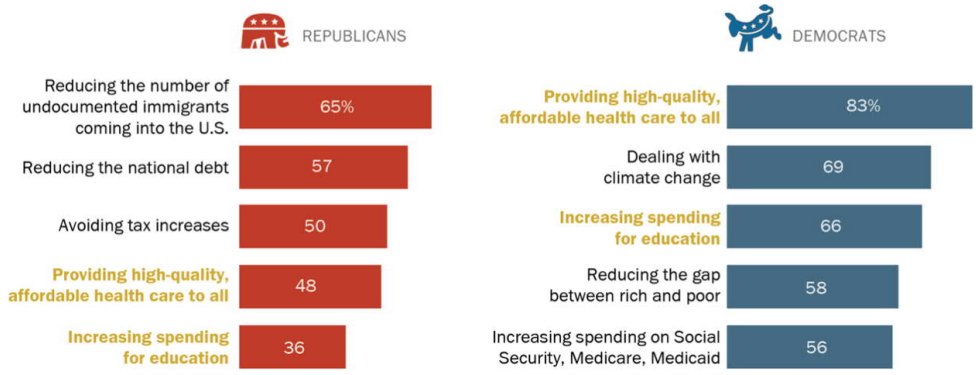


Figure 1.1: A sample open ended question-answer pair.

Chart Question Answering (CQA) is a task where the goal is to take a chart and a natural language question as input and generate the desired answer as output [31]. While CQA has received growing attentions in the last few years, existing datasets only focused on close-ended (factoid) questions where the answer is a word or phrase [36, 33, 8, 72] . These datasets typically use predefined templates to generate synthetic questions and answers to these questions come from a closed vocabulary (*e.g.,* 'yes', 'no', 'x-axis-label'). PlotQA [52] introduces some open-vocabulary questions that require

aggregation operations on the underlying chart data; the answer is still however a number or a word/phrase obtained from the chart. Kim et al. [39] attempt to automatically explain how the model computes the answer but only for close-ended ones. To our knowledge, we are the first to propose open ended question answering on charts as a computational task.

1.2 Our Approach

In this thesis, we introduce a novel task named, OpenCQA in which the system takes a chart and a question as input and is expected to produce a descriptive answer as output like the one in Figure 1.1. This task is multidisciplinary and challenging in nature as it involves natural language generation (NLG), information visualization and computer vision. It differs from data-to-text and reading comprehension, because unlike text or tables, charts serve a different communicative goal by creating visual representation of data. Readers can quickly notice important patterns, trends, and outliers from such visual representation which cannot be easily observed from a table of raw data [54]. For example, by looking at a line chart one can quickly discern an important trend whereas scatterplots may visually depict correlations and outliers. Existing NLG approaches for tables do not consider such chart features in the generation.

We have developed a benchmark dataset consisting of 7,724 human-written open-ended questions about a variety of real-world charts and the associated descriptive answers.

We formulate three practical task settings. In the first setting, a chart and the article containing the chart is provided as input, and the model generates an answer to an open-ended question. This setting poses extra challenge as articles often contain paragraphs that are irrelevant to the questions. To make the task more focused, the second setting provides only the relevant paragraph(s) to the chart. Hence we can measure the models' ability to answer a question without the effect of extra noise from irrelevant text. The third setting is more challenging as the related text is not provided and the model needs to generate an answer solely based on the chart. This is more relevant to real world scenarios where charts are not associated with any explanatory text.

To solve this problem, we adapted a number of state of the art multimodal,data2text,extractive and abstractive summarization models. Further details about these approaches are explained in Chapter 4. We evaluated these baselines using a number of automatic evaluation metrics, and analysed the model performance by chart types, ablation tests, naive baseline comparisons and human evaluations. Lastly we covered the errors and challenges that existing state-of-the-art models and metrics face on this dataset with suggested improvements. Details of all evaluations are given in Chapter 5.

1.3 Contributions

We provide the following contributions from this thesis:

- Since to our knowledge there is no existing task or datasets available for open-ended question answering on charts, we introduce this novel task and construct a relatively large benchmark dataset from real-world sources suitable for training large scale deep learning models;

- We adapt a set of state-of-the-art neural models that utilize multimodal, data2text and extractive/abstractive summarization methods to serve as strong baselines and a first step towards solving open-ended chart question answering;

- We identify a set of model evaluation metrics suitable for this new task and propose a new metric for chart question answering that can better identify the models ability to generate text relevant to the extracted data from the chart;

- We conduct a series of automatic and qualitative evaluations of our dataset and model results along with the current errors and challenges that state of the art models face on this proposed problem. Further we highlight the limitations of this work and identify future directions research can take to tackle this problem.

1.4 Outline

Before moving to the next chapter, we give a brief description regarding how the following chapters of this thesis are organised. We start the Chapter 2 with a literature review on

previous work in chart summarization, query focused summarization, data2text generation, chart question answering and end with the work done in visual question answering. Then in Chapter 3, we describe our chart question answering benchmark by detailing the dataset construction process with the step by step procedure guided by rich illustrations and explanatory figures. This chapter ends with an analysis of the dataset to understand the types and distribution of questions and charts available. Chapter 4 outlines the models adapted for the open-ended question answering task complete with their parameter settings with a brief description of the problem definition. This is followed up by Chapter 5 where we quickly highlight the automatic evaluation metrics used to measure model performance. Then we do a deep dive into the experimental results, looking at performance under many different settings complete with ablation tests and naive comparisons. We further define our newly proposed metrics for tackling the chart question answering problem and analyse the tradeoffs of different input settings on final performance along with the human evaluation of our baseline models. We finish this chapter off by understanding the errors and challenges current models and metrics face for this problem. We end this thesis in Chapter 6 where we discuss our final conclusions and limitations along with future directions of work to improve on this task. Ethical Considerations for this research are shared in the Appendix.

2 Literature Review

In this chapter we discuss some of the most relevant literature in chart summarisation, query focussed summarisation, data2text generation chart question answering and visual question answering, and how they are different from our proposed task of open ended question answering.

2.1 Chart Summarization

Chart summarization is a new area of research that is recently getting more attention. Chart summarization is the task of generating a succinct summary of the information available in the chart. Mittal et al. [53] adopts a planning-based architecture and used templates to describe charts with texts. These templates are based on the graphical structure and data mappings, a framework for identifying graphical elements and structure of the data from the graph. These templates are then further formatted to match target styles of surface realized text. They claim that their approach works well even in the absence of a detailed representation of the graph, but sometimes produces unconventional formats. It also

cannot create generalized template representations and their methods are limited to only describe how to read a chart without summarizing any insights from the chart. Demir et al. [18] present an approach to incorporate the salient features and intended message of an information graphic into a brief textual summary. To do this they use a combination of a visual extraction system from prior work for obtaining xml representations of input charts and a bayesian inference module for generating a high level message from the graphic. This xml and message input is given to a content identification module (CIM) to find important features in the graphic. They then pass these features to the Text Structuring and Aggregation Module (TSAM) and Sentence Ordering Module (SOM) to order features according to type of information and finally to the Sentence Generation Module (SGM) to get surface level sentences in natural language. They found that their approach performs well in selecting and organising most important information from the chart and conveying high level messages to users.

More recent work focuses on generating datasets for this task to train deep learning models. Chen et al. [10] attempted to investigate figure captioning where the goal is to automatically generate a natural language description of the figure. For this they proposed the FigCAP dataset based on FigureQA [36]. They also introduced a novel attention mechanism to enable the decoder to focus on specific figure labels and another to focus on relationships between figure labels. They incorporated these into an encoder-decoder

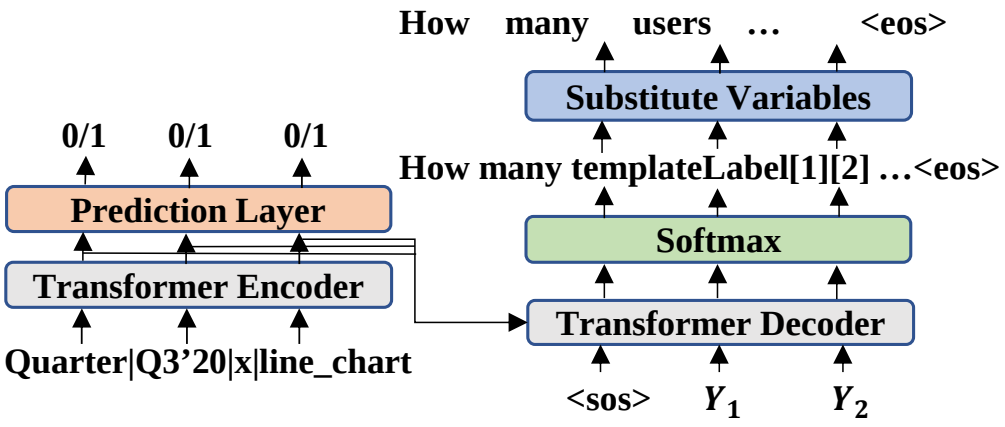


Figure 2.1: Enhanced Transformer Architecture [23].

architecture, where the encoder is a ResNet [27] used to extract features from figure, and the decoder is an LSTM [30] equipped with their attention methods. From experimental results they find that the models that use attention achieve better performance indicating the superiority of attention for figure captioning. They also find that the non attention approaches are better at generating high level descriptions since they capture high level general information of a figure better than attention methods. Obeid and Hoque [57] (Figure 2.1) adapted the enhanced transformer model [24] for chart summarization. They do this by including chart specific features important for summarization along with input record tuples. To identify ordering relationships in the chart, they leverage positional embeddings in the encoder. Then they finally substitute data facts in the reference with tokens to address hallucination issues [81, 59]. They found that their approach improves on

content selection [81] compared to enhanced transformer model and generated summaries were more informative, concise and coherent. Spreafico and Carenini [73] also tackle summarization of charts but constrain their problem to only deal with line charts. For the lack of existing datasets, they created a new dataset collected through crowdsourcing. They compared the effects of using an LSTM to two non neural approaches. The first alternative is the bi-gram model which calculates the statistical probability of generating any sequence of m words. The second alternative is the similarity model which implements a nearest neighbor approach to find the most similar input. Their results show that the neural approach beats non neural approaches in average f1 scores.

2.2 Query Focused Summarization

In this task, the goal is to generate a summary of the text document(s) based on a given query. For abstractive summarization, a key challenge is that there are only a few small datasets on specific domains such as news [17] debate [56] and meeting [89]. Given the lack of large training data, some prior methods focus on applying transfer learning techniques to first pretrain on a large generic summarization dataset and then finetune on a specific query focused summarization task. Baumel et al. [3] presented Relevance Sensitive Attention Query Focussed Summarisation (RSA-QFS) model for training sequence models to take relevance of query into account while generating summaries. For each input document and query pair, they compute the relevance of each sentence in the

document to the query through an attention based approach, and rescale predicted words at the decoder by the attention weights. They find that RSA-QFS improves the ROUGE scores [45] compared to previous abstractive models and achieves good results compared to extractive methods, but there is still much improvement needed in this area of research. OpenCQA is different from these tasks since answering open-ended questions requires more logical reasoning and inference. Further, charts are more difficult for deep learning models to interpret than text documents since a chart must be visually parsed to extract important trends and structure used for answering complex questions.

2.3 Data2text Generation

Data2text models generate a descriptive summary from a data table. Previous work has focused on specific domains such as sports [2, 80]. Barzilay and Lapata [2] propose a new learning framework for data2text generation. They argue that a content selection component is needed which determines the subset of information collected from multiple tables to include in generated summaries. For this they propose an efficient method to automatically learn content selection rules from databases by treating the content selection as a collective classification problem to identify relationships existing between table cells. Their experiments on American National Football League (http://www.nfl.com/scores) show improvements of 10% in F-score compared to models that do not consider content selection in input tables. Wiseman et al. [80] pointed out some of the challenges existing

in data2text generation. To overcome these challenges they proposed the BOXSCORE dataset (https://github.com/harvardnlp/boxscore-data) that was similar in token count to WEATHERGOV, but had longer target texts,larger vocabulary and posed more difficult content selection. Their dataset consisted of two types of articles, further categorising their dataset into two sub datasets. The first dataset called ROTOWIRE consists of medium length professional NBA game summaries intended for fantasy basketball fans interested in game statistics. The second dataset SBNATION contains summaries written by basketball fans for others, thus involving more informal structure. To solve this task they proposed three metrics to evaluate model performance. Content Selection (CS) measures how well the model can match the reference document in extraction of data records in generating summaries. Relation Extraction (RG) measures the factuality of generated summaries and Content Ordering (CO) measures the ordering of extracted data records. Their experiments on this task with a number of previous baselines show that existing model are fluent but still not comparable to human made documents. Other domains of focus were weather-forecast [65], recipe [84] and biography [42].

Most previous datasets were closed domain, causing bias in models urging the need for open-domain datasets. Parikh et al. [59] (Figure 2.2) present TOTTO dataset, an open-domain English table to text dataset built using Wikipedia tables. Each data sample consists of a table, set of highlighted cells that the model should use as reference for generation, and a short summary to train the model. Their experiments show that their BERT

Table Title: Gabriele Becker
Section Title: International Competitions
Table Description: None

Year	Competition	Venue	Position	Event	Notes
Representing Germany					
1992	World Junior Championships	Seoul, South Korea	10th (semis)	100 m	11.83
1993	European Junior Championships	San Sebastián, Spain	7th	100 m	11.74
			3rd	4x100 m relay	44.60
1994	World Junior Championships	Lisbon, Portugal	12th (semis)	100 m	11.66 (wind: +1.3 m/s)
			2nd	4x100 m relay	44.78
1995	World Championships	Gothenburg, Sweden	7th (q-finals)	100 m	11.54
			3rd	4x100 m relay	43.01

Original Text: After winning the German under-23 100 m title, she was selected to run at the 1995 World Championships in Athletics both individually and in the relay.
Text after Deletion: she at the 1995 World Championships in both individually and in the relay.
Text After Decontextualization: Gabriele Becker competed at the 1995 World Championships in both individually and in the relay.
Final Text: Gabriele Becker competed at the 1995 World Championships both individually and in the relay.

Figure 2.2: Example in the Totto dataset [59].

based encoder-decoder performs best in terms of BLEU and PARENT [21] scores. Their analysis of generated summaries indicate that current data2text generation models still suffer from hallucinations wherein the model generates data facts that cannot be entailed from the source. Current models also struggle with rare topics that have relatively few training samples and their numerical reasoning abilities such as counting rows/columns, comparing cells, min/max values are limited. They further purport the need for new evaluation metrics that can capture these conditional generation aspects. Mei et al. [51] propose an encoder-decoder model for jointly selecting content to generate and converting them into coherent summaries. The task is often formulated as two subproblems: content selection where the relevant records or cells in the input data table are selected, and surface

realization, where the natural language summaries are generated using the selected content. Their proposed model works by using a bidirectional LSTM to encode input records, then computing probability of an input record being relevant to target summaries and finally passing the selected encoded inputs to an LSTM decoder to get the surface realized text. They show state of the art results on the WEATHERGOV dataset in terms of F1 and BLEU scores and further improvements using k-nearest neighbor beam search. Lebret et al. [42] introduce a new dataset called WIKIBIO collected from English Wikipedia biography articles[2]. They propose a neural model to generate biographical summaries from fact tables. Evaluations show an improved performance of 15 BLEU compared to other models.

Most previous models give equal weight-age to the rows of the table when training a model, which may not be desirable for tasks such as question answering where only few rows are important for answering the question. Gong et al. [23] present a new transformer based model for data2text generation. In their proposed model they first train the transformer encoder separately by adding another binary layer to predict whether an input record should be used in the output. These binary predictions are later used to train the decoder to decide which records are important to use for generating text. Their approach shows state of the art results on the ROTOWIRE dataset in terms of BLEU

[2]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Biograph

scores, Content Selection and Content Ordering scores. Chen et al. [12] emphasizes on the importance of logical inference abilities of natural language generation models. For this they present the LOGICNLG dataset, an extension of TabFact dataset [11] paired with summaries that require logical and symbolic reasoning. They experiment on their new dataset by proposing the field-infusing and field-gating encoder, both using transformer based architectures. They further propose a coarse to fine generation scheme they use to train a GPT2 [62] and BERT [19] model. Their experiments show that the coarse to fine generation models are able to better generate summaries that require logical and symbolic reasoning compared to their field-infusing and field-gating models and other state of the art models. Chen et al. [13] performs similar research emphasizing the need for models that can perform logical inferencing on tables to generate high fidelity textual descriptions. They propose the LOGIC2TEXT dataset, where data tables are from WIKITABLES [5] and summaries are categorised under 7 common logic types and also converted to their respective logical forms. They trained common summarisation models under fully supervised and few shot settings and find that a GPT2 model performs best in terms of ROUGE scores. Unlike the task with data tables, our task involves understanding visual features of the charts and the natural language questions to perform reasoning in order to generate (or extract) texts as answers.

2.4 Chart Question Answering

Chart question answering (CQA) is very similar to visual question answering (VQA) and requires a model to perform reasoning on input charts for visual and reasoning based questions. Charts are more difficult to interpret than images since they have an inherent structuring to chart elements and often use a variety of styles to communicate information.

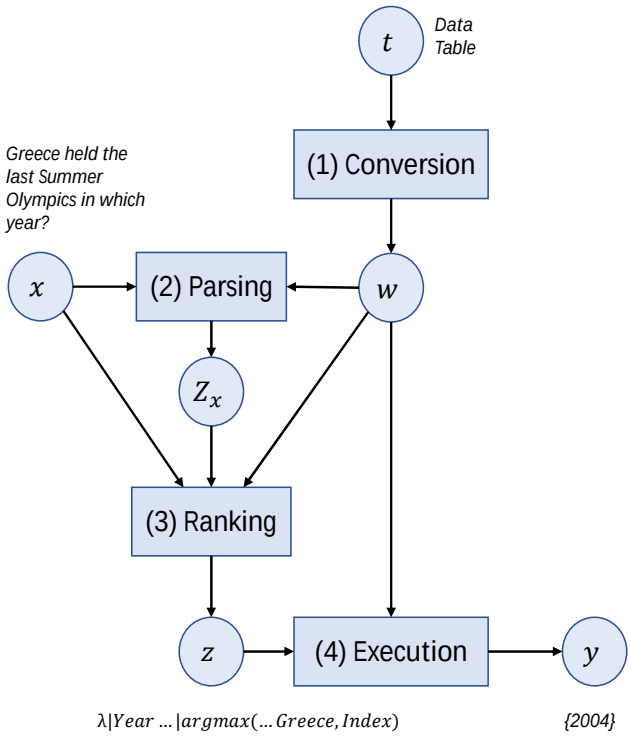


Figure 2.3: Compositional semantic parsing model [60]. (1) The table t is converted into a knowledge graph w; (2) using the knowledge graph the question x is parsed into candidate logical forms Z_x; (3) the highest scoring candidate z is chosen; (4) z is executed on w yielding the answer y.

Pasupat and Liang [60] (Figure 2.3) argue that previous work on semantic parsing for question answering tend to trade off between breadth of knowledge sources and depth of logical compositionality. They propose the task of answering complex questions on semi structured tables. For this they provide the WIKITABLEQUESTIONS dataset consisting of HTML tables from Wikipedia paired with question-answer pairs. They propose a logical form semantic parsing algorithm to handle the large space of logical forms created. Experimental results showed significantly improved performance compared to other models evaluated on this task.

Kim et al. [39] found in a study that users are often interested in visual aspects of charts such as colors, data values, specific labels etc. To answer such questions, they developed an automatic question answering system for charts that takes an input question and Vega-Lite chart [68], applies a series of natural language processing steps taking the help of Sempre [4] and produces a natural language answer with an explanation of how the answer was produced. Feedback on their system from users show that the generated explanations are very helpful to understand how answers are being generated. Nan et al. [55] argues that existing table question answering datasets lack questions that ask for reasoning and hierarchical structure from knowledge sources. They introduce the FeTaQA dataset consisting of Wikipedia table,question/answer triples. They propose a pipeline method based on semantic parsing QA systems and compare performance to

large pretrained language models. Their results show that end to end large scale pretrained models with simple table encoding strategies achieve higher scores than their proposed approaches.

Kafle et al. [34] explains that graphical visualizations are important for conveying numeric information and existing vision models fail to interpret simple bar charts. They propose the DVQA dataset to bridge this gap by improving a models ability to understand a chart and answer questions about information in the chart. They propose two baseline models that perform better than conventional Visual Question Answering(VQA) models. In their followup paper they propose PReFIL model [35] for improving the state of the art on chart question answering. PReFIL learns bimodal embeddings by fusing question and image features to answer given questions. Their proposed work significantly improved performance on DVQA and could additionally be used to construct a chart table for the given chart. Singh and Shekhar [71] propose STL-CQA model, a transformer based model that uses a combination of question encodings and localised chart feature extractions to answer questions. They further propose a new dataset LEAFQA++, an extension of LEAFQA [9] with more complex and balanced questions of different types. Their approach improves accuracy compare to other VQA methods on DVQA.

2.5 Visual Question Answering

Visual Question Answering involves answering a question regarding an input image. To solve this task many datasets and models have been recently proposed. Antol et al. [1] attempt to create a free-form visual question answering dataset consisting of roughly 0.25M images, 0.76M questions and 10M answers. These questions often target background or foreground details or can be more general in nature. As a result, more detailed understanding of the image and complex reasoning is required to generate suitable captions. They experiment with a number of baselines using MLP(Multi Layer Perceptron) and LSTM and measure performance in terms of accuracy. They further provide a breakdown of performance by question types to understand which question types are easier for models to answer. Their results show that models significantly underperform humans in these types of tasks and the questions can rarely be answered using a generic answer, prompting the need for more task specific architectures.

Due to the breadth in the range of visual question answering tasks, it is preferable to create large scale models that can perform well on all these tasks without requiring too much training data. Lu et al. [48] present ViLBERT model as a pretrained model that can perform visual question answering on a wide range of tasks by just fine tuning on them. Their model extends the BERT architecture by adapting it as a multi-modal two-stream model which can process both visual and textual inputs that are learned together using

co-attention transformer layers. Their model is pretrained using two tasks, masked multi-modal modelling which masks approximately 15% of both words and image region inputs and asks the model to reconstruct them, and multi-modal alignment prediction where the model is presented with an image-text pair and the model must tell if the image and text are related. They pretrain their model on the Conceptual Captions dataset [70] containing a collection of 3.3 million image-caption pairs crawled from the web. They fine-tune their model to a number of downstream visual & language tasks such as: (1) Visual Question Answering(VQA) [1] that relates to answering natural questions about images. (2) Visual Commonsense Reasoning(VCR)[87] which consists of two problems (a) visual question answering, (b) answer justification, both in the form of multiple choice problems. This dataset consists of 290k multiple choice question answer pairs from 110 movie scenes. (3) Grounding Referring Expressions which requires localising a region given a natural language reference. This task uses the RefCOCO+ dataset [38] consisting of 130,525 expressions referring to 96,654 distinct objects in 19,894 photographs of natural scenes. (4) Caption-Based Image Retrieval where a caption is used to retrieve the relevant image described in the caption. This task utilizes the Flickr30k dataset [85] consisting of 31,000 images from Flickr consisting of five captions for each image. From their experimental results they find that ViLBERT improves performance over single stream models under both pretrained and non pretrained model types. They further notice improvements in visiolinguistic representations and that finetuning on their model allows for powerful

transfer to downstream vision and language tasks.

To make the previous tasks more challenging Talmor et al. [76] attempt to address multimodal question answering by proposing the MULTIMODALQA dataset consisting of 29,918 question/answer pairs requiring cross modal reasoning. Their dataset slightly differs from previous work by including more complex, multi-hop questions that require reasoning over text, tables and images. They further propose ImplicitDecomp model for solving this task. This model is a RoBERTa-large model that takes a question as input and predicts one of 16 question types. This question type defines the order of operations to perform. Their experimental results show improvements in F1 scores compared to using single modality(i.e either text, image or table) for reasoning.

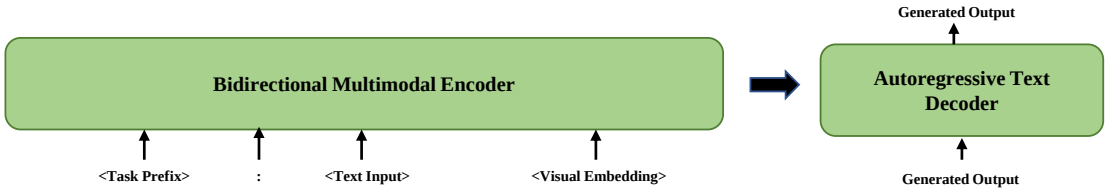


Figure 2.4: VL-T5 and VL-BART architectures [15]

Creating task specific architectures slow down the progress of visual question answering models. Cho et al. [15] (Figure 2.4) proposes getting rid of task specific architectures by using a unified framework that learns different tasks in a single architecture using multimodal conditional text generation as the language modelling objective. For this they propose two models i.e VL-T5 and VL-BART. VL-T5 is based on the T5 base model [63]

and VL-BART is based on BART base [43]. Both models first represent an image using n=36 object regions obtained using a Faster R-CNN[66] trained on Visual Genome [41] for object classification. These object regions and bounding box coordinates are encoded with a linear layer, while their image ids and region ids are encoded using BERT [20] and all features are summed. To embed the text, they prepend a prefix for each text input, specific to the task(ex. visual grounding : ...). This concatenated text input is tokenized and encoded and summed with BART/T5 positional embeddings. The ouput from the base models are passed to an autoregressive decoder to generate text. They gather images and captions from MS COCO [47], Visual Genome, VQA 2.0 [26] GQA balanced [32] and Visual7W [90] to get a pretraining dataset of 9.18M image-text pairs. They pretrain their dataset for a number of pretraining tasks such as Multimodal Lanugage Modeling, Visual Question Answering, Image-text matching, Visual Grounding(*i.e.,* given a description of an object region, predict the region id) and Grounded Captioning(*i.e.,* given a region id, predict the textual description of the region). They evaluate their model on a number of downstream tasks such as VQA, GQA, (NLVR)Natural Language Visual Reasoning [74](*i.e.,* determine if a statement is true about two images), RefCOCOg [49] (*i.e.,* localise an object described by natural language description), VCR and (Multi30k)Multimodal Machine Translation[22] (*i.e.,* translate English text to German for an associated image). Their results show that VL-T5 and VL-BART achieve comparable performance to state of the art multimodal models for a wide range of problems and superior performance on

visual question answering.

A major limitation of existing work is the lack of datasets having natural real world questions on charts. Masry et al. [50] addresses the drawback that current chart question answering datasets usually do not contain complex reasoning questions and are often template based questions and answers using a fixed vocabulary. Their work presents a large scale dataset consisting of 9.6k human written questions with 23.1k questions generated from chart summaries. They evaluate a number of transformer based multimodal models and propose a new model VisionTapas as an extension of Tapas [29] for this task. Their evaluation on datasets such as FigureQA, DVQA, PlotQA and their proposed ChartQA result in improved results with their VisionTapas model compared to previous state of the art models. They find that bounding boxes are very important to chart question answering tasks to identify hierarchical relationships existing within charts.

Although significant work has started on chart question answering, current models still lack the ability to perform inferencing and reasoning on complex question answering frameworks. Thus to bridge this gap we propose the task of open ended question answering on charts and curate a benchmark dataset so further research and exploration can take place.

3 Chart Question Answering Benchmark

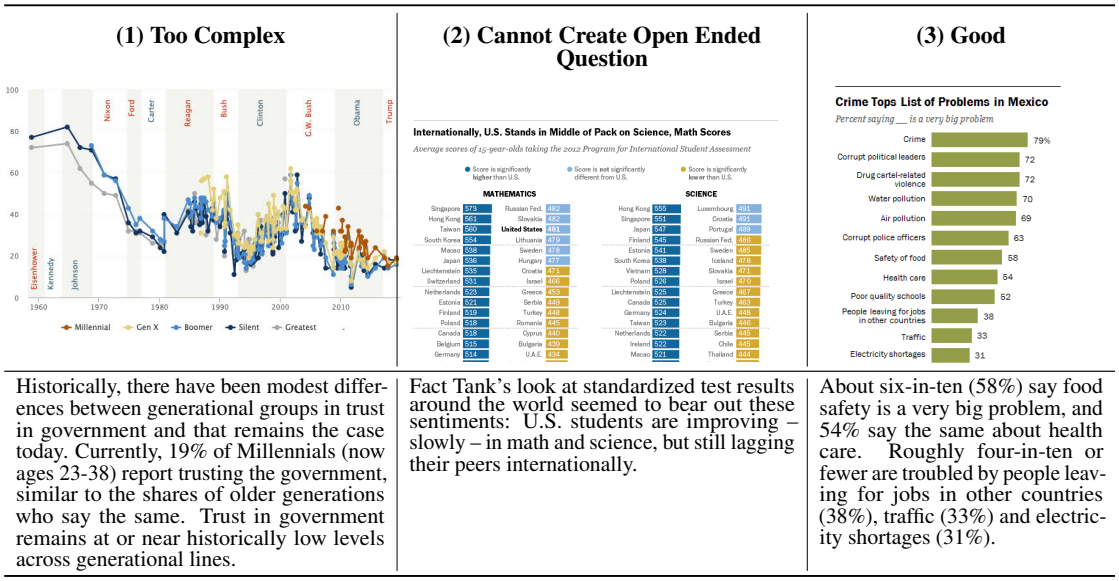
Our first contribution for this thesis is a novel dataset for the task of open ended question answering on charts. The process for dataset construction detailing how the data was collected and annotated is given in the following section.The second section is an in-depth analysis of our proposed dataset based on the chart types, topics, and question types.

3.1 Dataset Construction

This section discusses the steps involved for collection and annotation of data to curate an open ended question answering dataset. Since this is a new task, no datasets currently exist, motivating the need to create a large enough dataset suitable to train large scale deep learning models. We discuss the dataset collection procedure below followed by the data annotation procedure. Ethical Considerations are discussed in Section 7.1.

3.1.1 Data Collection

Building a dataset with open-ended questions and human-written descriptive answers is challenging because there are not many publicly available real-world sources with charts and related text Descriptions. After exhaustive search, we decided to use charts from Pew Research (pewresearch.org). Pew serves as a suitable source because the articles are written by professional writers covering opinions, market surveys, demographic trends and social issues. The articles are often accompanied by a variety of real-world charts and their summaries.



Historically, there have been modest differences between generational groups in trust in government and that remains the case today. Currently, 19% of Millennials (now ages 23-38) report trusting the government, similar to the shares of older generations who say the same. Trust in government remains at or near historically low levels across generational lines.

Fact Tank's look at standardized test results around the world seemed to bear out these sentiments: U.S. students are improving – slowly – in math and science, but still lagging their peers internationally.

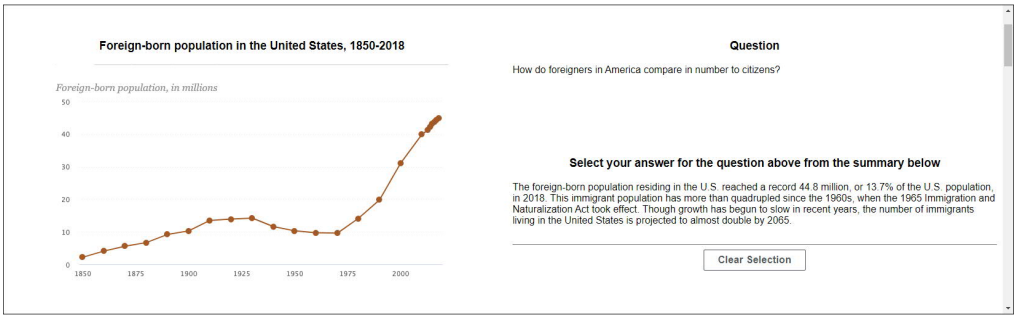
About six-in-ten (58%) say food safety is a very big problem, and 54% say the same about health care. Roughly four-in-ten or fewer are troubled by people leaving for jobs in other countries (38%), traffic (33%) and electricity shortages (31%).

Figure 3.1: Examples of charts in the Pew dataset: (1) too complex since the line charts are difficult to separate; (2) not suitable for creating an open-ended question since summary does not refer to information in the chart; and (3) a good sample consisting of a chart and the summary that refers to the chart information. We filter samples like (1) and (2) from our dataset.

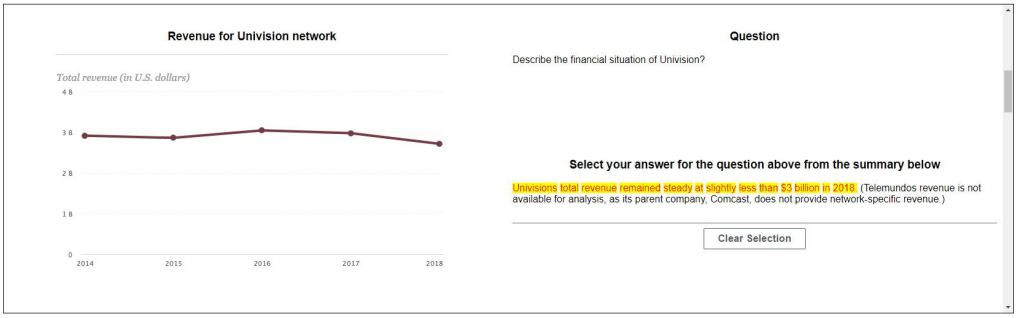
We collected 9,285 chart-summary-article triples scraped from nearly 4,000 articles as described in the Chart-to-text benchmark.[3] However, not all of the charts are suitable for creating open-ended questions. For example, some charts maybe too unconventional or too complex while a few others have poor resolution. Similarly, the text accompanying the chart may not discuss data values in the chart and instead refer to other external background facts. Hence, we manually went over all the charts to retain 7,724 samples that we deemed suitable for our study. In particular, we filtered out 1,019 samples as too complex and 542 as samples we cannot make an open-ended question (See Figure 3.1).

In the Chart-to-text benchmark, each sample consists of a chart, the title, text extracted using an OCR method and associated summary text and the original article. OCR text segments were extracted top to bottom left to right along with their rectangular bounding boxes given as $\{$x,y,len(x),len(y)$\}$. The extracted title from the chart image was taken as the final chart title if there was no associated `alt` text with the chart image. If the `alt` text (which gives a short chart description) was available, the longer of the two was taken by comparing it with the extracted title. The summary for each chart was obtained by finding the most relevant texts to the chart following an annotation process which is described in detail in [37].

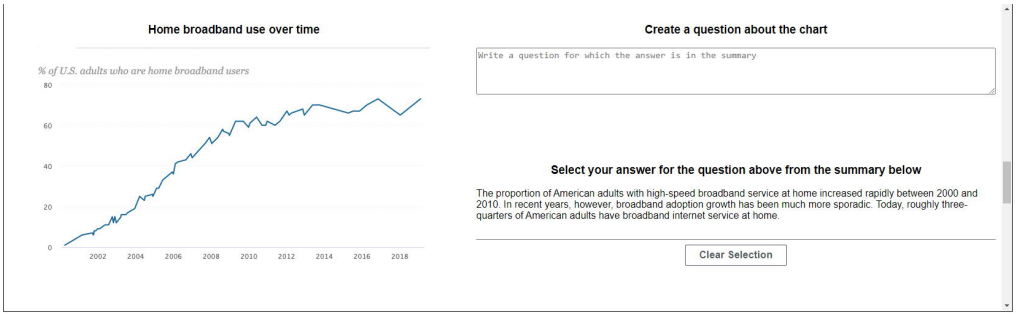
[3] More details on finding chart-summary-article triples are provided in [37].



(a) The interface shows a sample where the question is given and the crowdworker needs to highlight the corresponding answer from the summary.



(b) The interface shows the highlighted answer selected by the user for a given question.



(c) The interface shows an annotation task, where the annotator needs to formulate a new question and then subsequently select the answer from the summary paragraph.

Figure 3.2: Amazon Mechanical Turk study interface.

In each HIT (Human Intelligent Task), the worker answered three questions previously asked by another coworker for three separate charts respectively and also created three new question-answer pairs for three new charts (see Figure 3.2). The first round of 30 samples were created by the authors but were not included in the final dataset to remove any potential bias. Each subsequent round of samples were created by crowd workers who would answer existing question that were created in the previous round while also creating new question-answer pairs for the next round.

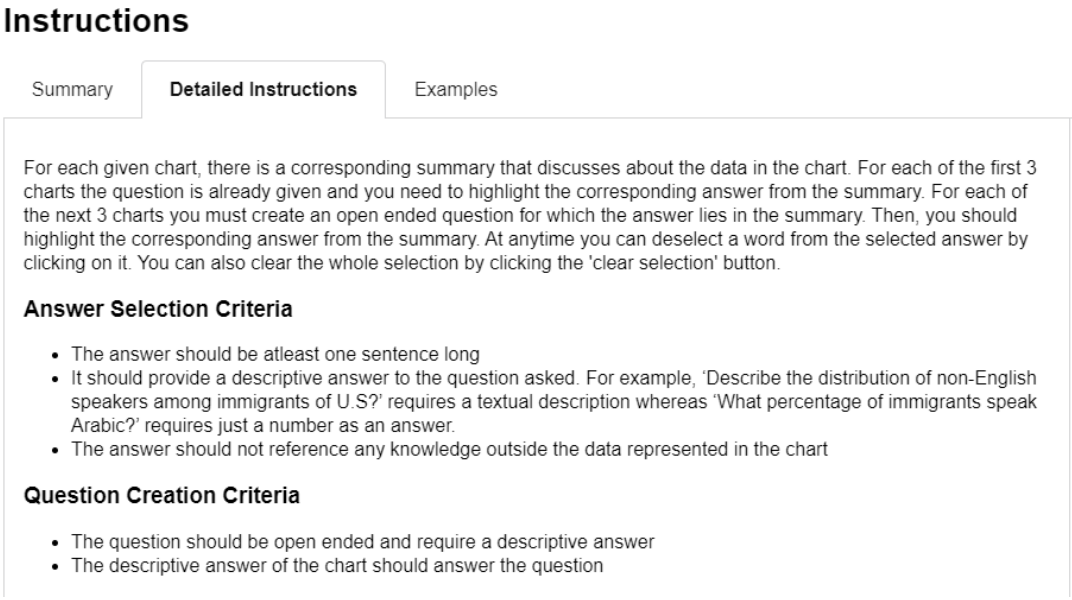
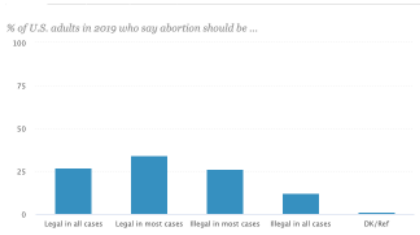


Figure 3.3: Amazon Mechanical Turk Instructions for Mturk workers.

Each participant was first provided with a consent form where the complete procedure of the study and data collection was explained. Once the participant provided consent, detailed instructions were shown on how to complete the task(see Fig. 3.3). In particular,



Question 2

What position do Americans hold on abortion?

Answer 2

Though abortion is a divisive issue, more than half of U.S. adults take a non-absolutist position, saying that in most – but not all – cases, abortion should be legal (34%) or illegal (26%). Fewer take the position that in all cases abortion should be either legal (27%) or illegal (12%).

Question 2

Do most Americans think abortion should be Legal?

Answer 2

Yes

(This is not a good example question because the answer is just a single word as opposed to descriptive text.)

Figure 3.4: Example of good and bad open ended question and answer pair on Amazon Mechanical Turk(The question on the left is a good example since it requires a descriptive answer and is open ended, and the question on the right is a bad example as it is not open ended).

participants were instructed to create open-ended questions that required a descriptive answer. An answer was considered to be descriptive if it was at least one sentence long and derivable from the chart(*i.e.,* it did not refer to any knowledge outside the data represented in the chart). This method of selecting the answer for the question from the summary is very controlled since the summary was written by a professional writer and these summaries usually had good linguistic style and grammatical structure. Participants were additionally shown sample question-answer pairs to explain how ideal questions looked like (see Figure 3.4).

To ensure quality, only participants with a hit approval rate greater than 95% and

over 5,000 approved hits were selected for pre-screening. In the pre-screening stage, participants completed a sample task that allowed us to assess their proficiency for this study. Those who successfully completed the task according to the given instructions were qualified for participating in the main study. The study protocol was approved by the ethics review board at York University with certificate ID e2021-317.

3.1.2 Data Annotation

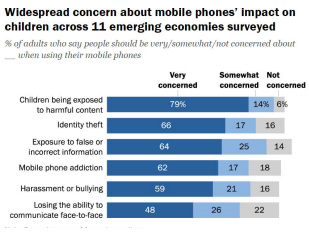
Question Validation & Editing	Disagreement Resolution	Decontextualization
Describe the concerns on mobile phones ~~addictions~~ among the surveyed countries?  	**Ans 1:** In addition to their concerns about the impact of mobile phones on children, majorities across the 11 countries surveyed also say people should also be very worried about issues such as identity theft (an 11-country median of 66% say people should be very concerned about this), exposure to false information (64%) addiction (62%) and harassment or bullying (59%) when using their mobile phones face-to-face due to mobile phone use (48%) Ans 2: the impact of mobile phones on children, majorities across the 11 countries surveyed also say people should also be very worried about issues such as identity theft (an 11-country median of 66 % say people should be very concerned about this), mobile phone addiction (62%) and harassment or bullying (59%) when using their mobile phones	In addition to their concerns about the impact of mobile phones on children, majorities across the 11 countries surveyed ~~also~~ say people should ~~also~~ be very worried about issues such as identity theft (an 11-country median of 66% say people should be very concerned about this), exposure to false information (64%), mobile phone addiction (62%) and harassment or bullying (59%) when using their mobile phones. Fewer are very concerned about the risk that people might lose the ability to communicate face-to-face due to mobile phone use (48%).

Table 3.1: Illustration of data annotation process (steps 2 - 4). Green text is newly added from the summary, ~~strikeout red~~ text is removed. Purple text is from the first worker, blue text is from the second worker, and brown represents texts in common between the participants.

We perform an annotation study on the collected chart data to create question-answer pairs. The steps we perform for data annotation are: (1) question-answer creation (2) question validation (3) disagreement resolution (4) decontextualization. Steps are described in Table 3.1.

Question-answer Creation We asked each crowdworker from Amazon Mechanical Turk to answer three existing questions (created by another crowdworker) for three separate charts respectively, and create three new question-answer pairs for three new charts. They were provided with the chart and the summary, and were asked to select portions of the text as an answer to the question. The selected segments can be noncontiguous. In this way, we collected two answers from different workers for each question, to verify answers and to remove any potential bias in answer selection.

Question Validation After collecting the question-answer (QA) pairs, this and the next two steps are performed by five internal annotators who are native speakers of English and have research background in summarization. Each QA pair is first examined by an annotator who first checks if the question is open-ended in nature, and edits the questions when the question is vague or incomplete, or not answerable from the charts. Then, the remaining annotators analyze the questions in terms of grammatical correctness and edit them as needed. Overall, the question was edited in 53% question-answer pairs. In this 22.7% cases were minor changes (less than 30% tokens changed), 15.5% cases were moderate changes (between 30% and 60% tokens changed) and 14.8% were major changes (over 60% tokens changed).

Disagreement Resolution As mentioned, we obtain two answers from the crowdworkers for each chart-question pair. To resolve any potential bias from one and/or

disagreement between the two answers, we build an annotation interface where an annotator can either choose one of the two answers, or select a new answer from the given summary. The annotator checks whether the answer contains irrelevant information to the question or any text that are not derivable from the chart (*e.g.,* background information). For 18.4% cases, the two answers matched exactly. For 68.2% samples, the two answers still had high overlaps (over 90% token matches); for another 10.1% the overlaps between the answers were moderate (between 30% and 90% token matches) and for the remaining 3.3%, the token matches between answers were less than 30%. While resolving the disagreements between crowdworkers, in 96% cases the annotators chose one of the two answers while for other 4% they selected a new answer from the summary.

Decontextualization In some cases, crowdworkers may have left out important information from the summary that is relevant to the question while in other cases, they may have included information that is not derivable from the chart. Thus, after selecting the most appropriate answer, annotators edit it further by adding tokens from the summary or removing tokens as necessary, which is taken as the extractive answer for the dataset. Also, if needed, they replace the occurrence of a pronoun with its antecedent (a proper noun) in cases where the entity is unknown, to put the answer in context, which is the *abstractive answer*.

3.2 Dataset Analysis

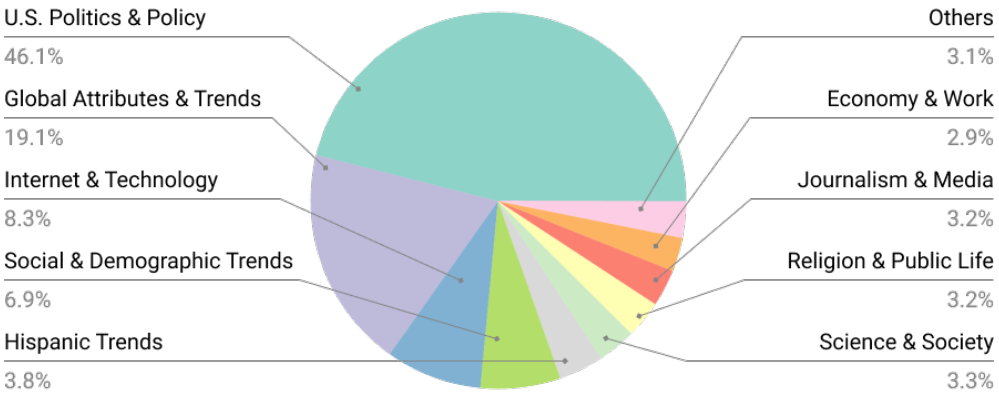
Statistics (on average)	
Tokens in article	1268.91
Tokens in summary	123.35
Tokens in title	17.94
Tokens in question	11.88
Tokens in abstractive answer	56.41
Tokens in extractive answer	56.21
Percentage of tokens extracted from the summary	52%
Percentage of tokens extracted from the article	7%

Figure 3.5: Dataset Statistics

In this section we analyse our dataset for basic linguistic statistics, question types and chart types. Figure 3.5 represents some basic linguistic statistics about the dataset. The questions and titles are generally short with both under 21 tokens on average. The percentage of tokens overlapping between the extractive answer and the article is 7% on average. Other characteristics of our dataset are:

Type	Simple	Complex
Bar	712	4,823
Line	234	1,667
Area	7	4
Scatter	0	42
Pie	235	0
Total	1,188	6,536

(a) Chart Types



(b) Distribution of topics.

Figure 3.6: (top) Chart Types, (bottom) Topic Distribution

• **Chart Types and Topics** Our dataset contains a variety of chart types (Figure 3.6). The most common is bar charts (71.7%), for both simple as well as stacked and group bar charts. Here, we consider a chart to be simple if the corresponding data table has only two columns. If the data table has more than two columns then we consider it to be a complex chart (e.g., stacked bar charts and line charts with multi-series). The next most common type is line charts (24.6%). Other types include area charts, scatter plots and pie charts.

The dataset also covers a diverse range of topics including politics, technology, society and media; about half of the charts cover *U.S. Politics & Policy* due to the nature of the dataset.

Type	**Example**	**%**
Identify	What are the current thoughts on direct democracy?	37%
Summarize	Explain the distribution of people who know a transgender person?	37%
Compare	Compare Americans and Germans views about the world economic leader?	20%
Discover	How do Americans' see the coronavirus statistics?	6%

Table 3.2: Example and distribution of question types among 100 randomly selected questions. The corresponding charts of these examples are shown in Figure 3.7

• **Question Types** We further analyze the question types using 100 randomly sampled question-answer pairs from our dataset. Table 3.2 shows the distribution of questions across four main types. Our categorization of questions are based on the specific analytical tasks with visualizations one would have to perform to answer the question [54]. The four categories are: (*i*) *Identify:* questions that require identifying a specific target (*e.g.,* a data item or a data attribute) from the chart and describing the characteristics of that target; (*ii*) *Compare:* questions that require comparisons between specified targets from

the chart; (*iii*) *Summarize:* questions that require summarizing the chart based on specified statistical analysis tasks (*e.g.,* describing data distribution, outliers(s), and trend(s)); and (*iv*) *Discover:* questions that require analyzing the whole chart to derive key insights through inference and reasoning. Unlike the *summarize* task there is no explicit analytical task that is specified in a *discover* type question.

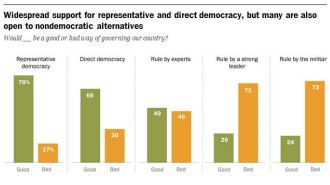
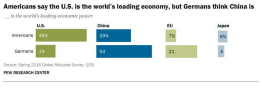
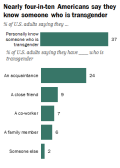
(1) Identify	**(2) Compare**	**(1) Summarise**	**(4) Discover**
Question: What are the current thoughts on direct democracy?	**Question:** Compare Americans and Germans views about the world economic leader?	**Question:** Explain the distribution of people who know a transgender person?	**Question:** How do Americans' see the coronavirus statistics?
			
A global median of 66% say direct democracy–in which citizens vote on major issues–would be a good way to govern.	As far as what share of Americans say they know a transgender person, 37% say they personally do, including 13% who say they have a close friend and 6% have a family member who is transgender (9% say they have a close friend and 6% have a family member who is transgender). About a quarter of Americans (24%) say they have an acquaintance who is transgender, while 7% say they have a transgender co-worker.	Americans and Germans diverge on who is the world 's leading economic power. Roughly half of Americans (49%) name the U.S. as the world 's economic leader, while only 19% of Germans say the same . Over half of Germans (53%) say China is the leading economy. A further 21% of Germans say the countries of the European Union are the world's top economic power, while only 7% of Americans say this.	The survey finds that a majority of Americans (60%) say the primary reason that the number of confirmed coronavirus cases is increasing is because there are more new infections, not just more testing for the disease . About four-in-ten (39%) say cases are rising primarily because more people are being tested than in previous months.

Figure 3.7: Examples question types in the Pew dataset: (1) Identify: The question refers to a specific target ('direct democracy') which represents a group of bars in the chart; (2) Compare: The question requires a comparison between two specific targets in the chart (e.g. between 'Americans' and 'Germans' which represent two groups of bars); (3) Summarize: The question asks to summarize the distribution statistic in the chart (i.e., 'people who know a transgender person'); (4) Discover: The question does not specify any specific statistical task but requires the inference and reasoning over the whole chart to derive key insights (e.g., majority of Americans say that the increase in coronavirus cases is due to more new infections, not just because of more testing).

From Table 3.2, we notice that people often ask to locate one or more portions of a chart (identify and compare) and then characterize them to answer the question. They may also ask for a descriptive summary of some trends or data distributions. In contrast, questions that require the user to focus on the whole chart (*e.g.,* discovery) are fewer. This

suggests that unlike the chart summarization problem which focuses on the whole chart, OpenCQA problem requires the model to identify relevant portions of the chart to answer the question.

4 Models for Open Ended Question Answering

In this chapter we explain about the models we adapted for solving this task. We first explain the problem setup in detail. Then we give a detailed description of the extractive and abstractive models used, along with their parameter settings throughout various experiments.

4.1 Problem Definition

For our OpenCQA problem, we consider three task settings as explained below:

- In the first setup, the model takes a chart and the article containing the chart as input and extracts an answer to an open-ended question. The data for this task can be represented as a tuple of 6 elements $\mathcal{D} = \{\langle C, T, M, Q, D, A\rangle_n\}_{n=1}^{N}$, where C, T, M, Q, D and A represent the chart image, title, metadata, question, document (article) text and answer text, respectively. The metadata $M = \langle C_{\text{label}}, C_{\text{bbox}}\rangle$ consists of the chart labels that are text segments extracted from the chart through OCR (*e.g.,* axis labels, data labels) and their respective bounding boxes.

- In the second setup, the chart summary is provided as input instead of the whole article. The dataset in this setup can be represented as $\mathcal{D} = \{\langle C, T, M, Q, S, A\rangle_n\}_{n=1}^{N}$, where S represents the chart summary.

- In the third setup, the chart summary is not accessible, and the model must rely only on the chart. This is a more difficult and interesting problem setting since real world charts often do not come with explanatory summaries. In this setting, for an input $I = \langle C, T, M, Q\rangle$, the model has to learn to generate an explanatory answer A.

For training the models, we use state-of-the-art *extractive* and *generative* QA models. The supervision of extractive models are *extractive* answers whereas for the generative models *abstractive* or edited (as described in Section 3.1.2) answers are used. Note that the third task setting applies to only generative models, whereas the first and second settings apply to both extractive and generative models. We describe the models below.

4.2 Extractive Models

We adopt two extractive models for the two problem setups where the models extract the answer to the question from the input summary or article.

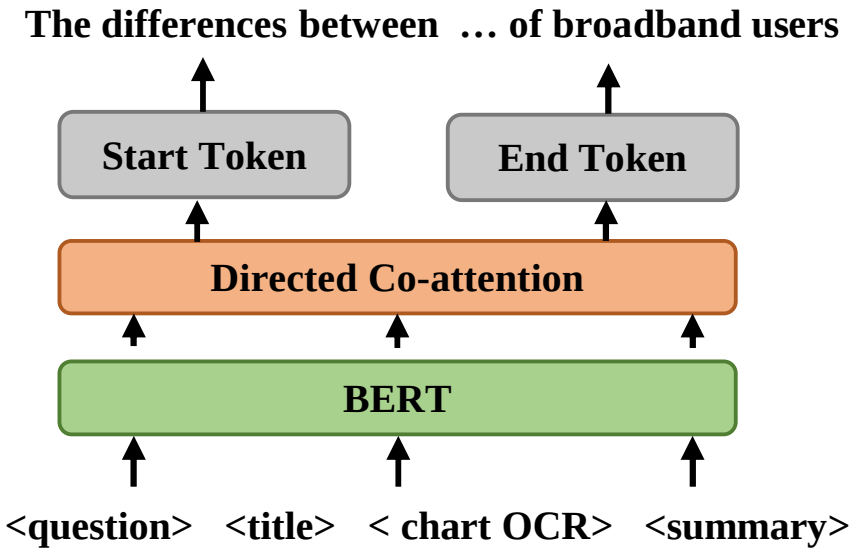


Figure 4.1: BERTQA model (when the summary is given as input).

BertQA [7] (Figure 4.1) is an extractive QA model that uses directed coattention layers [83] to improve the performance of the original BERT model [20]. In this approach, we first pass the question and the text (article or summary) through a BERT model to get the (self-attention based) representations. It then calculates the cross attention from the question to text and text to question and concatenates the resulting vectors to predict the start and end points of the answer span.

We fine-tune BERT-large (340M parameters, 24 layers) with a batch size of 4 for 2 epochs, with a max input sequence length of 512 tokens and output sequence length of 128 tokens.

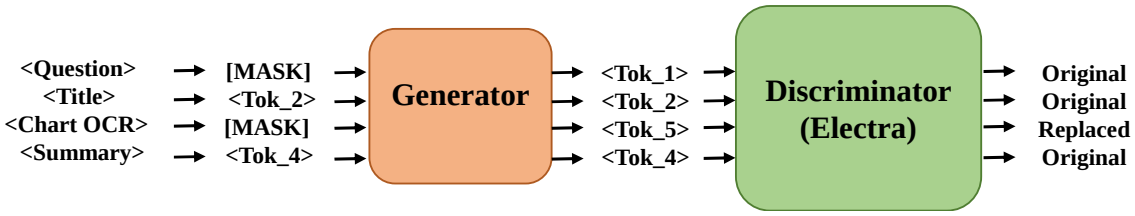


Figure 4.2: ELECTRA model (when the summary is given as input).

ELECTRA [16] (Figure 4.2) proposes a self-supervised representation learning method that emphasizes computational efficiency. In contrast to Masked Language Modeling (MLM), it uses Replaced Token Detection or RTD as the pretraining task. The training process is inspired by the training of Generative Adversarial Networks or GANs [25]. The training examples are first passed to a generator (typically a small MLM-based model) to replace some of the tokens in the input with other probable but incorrect tokens. ELECTRA (the discriminator) is then trained to distinguish between the "original" vs. "replaced" tokens in the input. This binary classification task is applied to every token, hence it requires fewer training examples compared to MLM training. ELECTRA achieves state-of-the-art results on SQuAD 2.0 [64]. We use the same experimental setup as with the BERTQA model.

We fine-tune Electra-Base for 25,000 iterations with a batch size of 4 and initial learning of 0.0001. The max sequence size of input is set to 2048. For inference, the batch size is set to 4 with search beam size of 20.

4.3 Abstractive Models

Below we describe five abstractive models which automatically generate an answer to the question as opposed to the extractive models which extract the answer from the input summary or article.

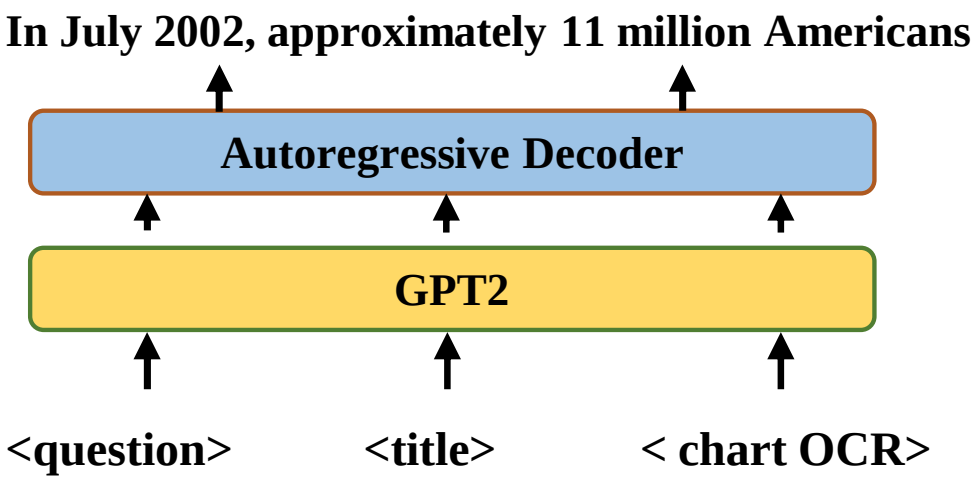


Figure 4.3: GPT2 (when the summary is not given as input).

GPT2 [62] (Figure 4.3) trains a transformer decoder on the unlabelled BooksCorpus dataset [91] using a conditional language modelling objective. Their pretrained model can be fine-tuned on downstream tasks such as textual entailment, similarity and question answering. We fine-tune GPT-2 under the three tasks where all the input elements in each task setting are concatenated as conditioning input to predict the answer.

We fine-tune GPT2 with a batch size of 6 for 5 epochs with a max input sequence length of 512 tokens and output sequence length of 128 tokens.

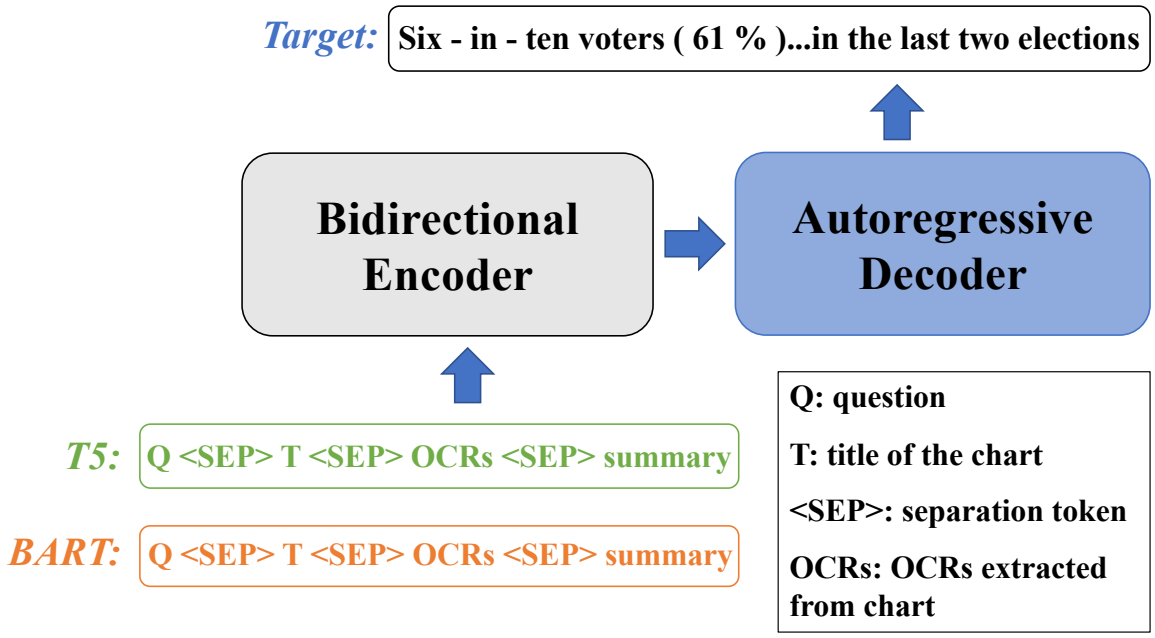


Figure 4.4: T5 and BART model (when the summary is given as input).

BART [44] (Figure 4.4) uses a standard encoder-decoder transformer architecture. Its pretraining task involves denoising, where text spans in the input text are replaced with a single mask token, and the decoder is tasked to predict the original input sequence. BART has been shown to achieve state-of-the-art performance on text generation tasks such as summarization. In each of our three task settings, we concatenate the corresponding inputs and feed into the model for fine-tuning.

We fine-tune BART-Base[4] (140M, 6-layers) with max input sequence length of 512 tokens, for 50K iterations with a batch size of 4 and evaluate after every 1,000 iterations on the validation set. The initial learning rate is set to 0.00005. For inference, we use the model with the lowest validation loss and decode with a beam size of 4.

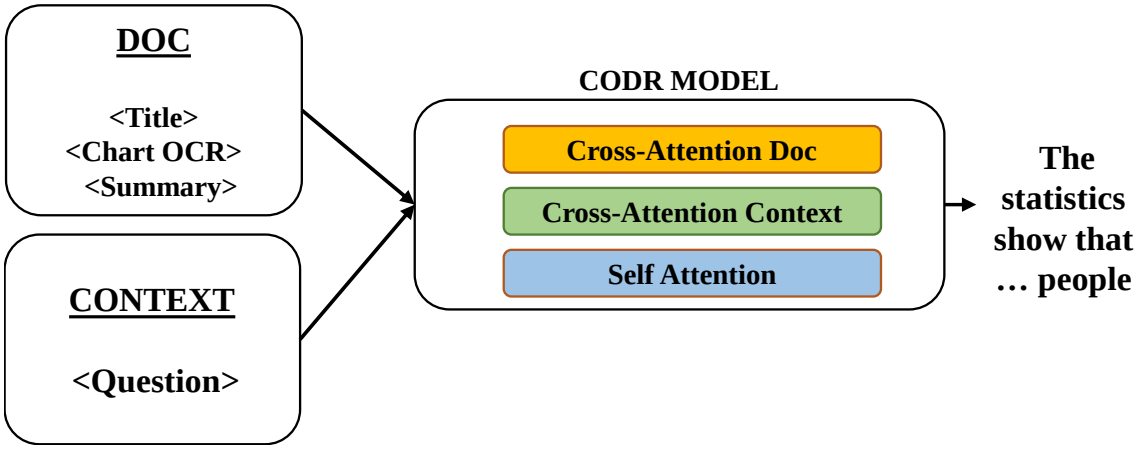


Figure 4.5: CODR model (when the summary is not given as input).

CODR [61] (Figure 4.5) proposes a document grounded generation task, where the model uses the information provided in a document to enhance text generation. In their setup, the context and source documents are concatenated and passed to a BART encoder to get a contextualized representation of the document. Then the same encoder is applied to the context alone and both representations are finally concatenated and passed to the BART decoder to generate the text. We fine-tune the pretrained BART encoder and

[4] huggingface.co/transformers

decoder following the approach of CODR for OpenCQA by taking the question as the context and concatenating the rest of input elements as the grounded document.

We fine-tune BART-large for 10 epochs and a batch size of 6 with max input sequence length of 512 tokens, output sequence length of 128 tokens,gradient accumulation set to 1 and the other parameters at default settings.

T5 Model [63] (Figure 4.4) is a unified encoder-decoder transformer model that converts language processing tasks into a text2text generation format. It is first pretrained with a 'fill-in-the-blank' denoising objective, where 15% of the input tokens are randomly dropped out. The spans of consecutive dropped-out tokens and dropped-out tokens that stand alone are then replaced by special sentinel tokens. Each sentinel token is assigned a token ID that is unique in the input sequence. The decoder then learns to predict those dropped-out tokens, delimited by the same input sentinel token plus a final sentinel token. We fine-tuned T5 on our tasks using the same input format as with BART.

Similar to BART, we fine-tune T5-Base[4] (220M, 12-layer Transformer as the encoder and decoder) with max input sequence length of 512 tokens, for 100k iterations with a batch size of 4 and an initial learning rate of 0.00001, evaluate after every 1,000 iterations on validation set, and use the model with best validation loss for testing. Inference is done with beam search with a beam size 4.

VLT5 Model [15] (Figure 2.4) is a T5-based framework that unifies the Vision-Language (VL) tasks as text generation conditioned on multimodal inputs. The input consists of both textual tokens and visual features of objects from the image extracted by Faster R-CNN [67]. The model is pre-trained on multiple multimodal tasks: language modeling, visual QA, visual grounding, image-text matching and grounded captioning. We fine-tuned VL-T5 on our OpenCQA generative task in the following manner. For the textual input, we use the same input format as T5. For the visual input, we extract the visual features of different marks in the chart image (*e.g.,* bars, lines) using Mask R-CNN [28] with Resnet-101 as its backbone.

We fine-tune the pretrained VLT5 (T5-base as its backbone) for 100 epoches with a learning rate of 0.0001 and max input sequence length of 512 tokens. Inference is done with beam search with a beam size of 4. Note that the original VL-T5 represents an input image with 36 object regions. However, since the number of bounding-box objects in charts varies from one to another, we either pad the extracted visual features (with zeros) or truncate them to maintain the length at 36.

5 Evaluation

This chapter covers all the evaluations and result analysis we performed as part of this research. We start by briefly discussing the automatic evaluation metrics used. We then do a deep dive into the experimental results and explain our findings. Next we experiment with model performance by chart type to better understand how different charts can influence model results. In addition, we conduct ablation tests and naive baseline comparisons to estimate the importance of various input components of our models and how useful they are to generating relevant answers. Further we propose a new metric that is more suitable for measuring the benefits of OCR text as model input. Finally we analyse the model performance by detailing a human evaluation of model outputs and summarizing the errors and challenges that current models and metrics face for solving this problem.

5.1 Automatic Evaluation

5.1.1 Evaluation Metrics

We utilized six measures for automatic evaluation which we explain in detail.

BLEU [58] was proposed as an automatic evaluation measure to compare generated summaries from deep learning models to actual reference summaries. It does so by measuring n-gram overlaps between the model generated text and the reference text, in a way that is similar to calculating precision, but also accounting for repeating tokens in the generated text. Due to this it is considered a *modified precision*. The implementation we use takes the geometric mean of BLEU1-4 and multiplies by the exponential of brevity penalty.

ROUGE [46], on the other hand, is a recall oriented metric and it measures n-gram overlaps between the model generated text and the reference as a percentage of n-grams in reference text. ROUGE is shown to correlate well with human judgement for text generation tasks. In our experimental results, we report ROUGE-F score.[5]

CIDEr [78] measures TF-IDF weighted n-gram overlaps between the model generated text and the reference.

[5]The rouge score is implemented by using the python package rouge-score, which is a Google re-implementation of the original Rouge-1.5.5 perl script.

BLEURT [69] is a model-based metric that measures the fluency of the generated sentence and the extent to which it conveys the meaning of the reference. We use BLEURT-base-128.

Content Selection or CS score [80] measures how well the generated text matches the gold answer in terms of selecting which records to generate. Since both the BLEURT and CS are calculated at the sentence-level, we average these scores over the whole test set.

BERTScore [88] which correlates better with human judgments and has been widely adopted recently for text generation evaluation. BERTScore computes token similarity between the candidate and reference using contextualized embeddings rather than relying on exact matches.

5.1.2 Main Results

As expected, the models perform better when only the relevant chart summary is given compared to when the entire article is provided (the first 2 setups in Table 5.1). The generative models have reduced performance due to the added texts in the article in setup 1. The extractive models also have a loss in performance in this setup mainly because they assume the answer comes from a single contiguous text segment whereas a gold answer can be constructed from multiple segments (Section 3.1.2). Specifically, these methods apply a sliding window over the tokens in the input sequence to find the start and end

Models	Type	ROUGE_-F (1/2/L) ↑	CS ↑	BLEURT ↑	CIDEr ↑	BERTScore ↑	BLEU ↑
Setup 1: With Article Provided							
BERTQA	Extractive	22.81/8.53/16.63	28.17%	-0.692	0.983	85.00	9.58
ELECTRA	Extractive	57.67/50.09/53.89	54.73 %	0.066	5.100	91.01	38.79
BART	Generative	50.68/38.95/45.55	54.29%	-0.017	4.121	91.08	23.91
T5	Generative	66.57/60.40/**63.46**	**68.24%**	**0.103**	**5.437**	**93.53**	**41.28**
VLT5	Generative	58.90/49.97/54.82	66.06%	0.076	5.227	92.34	40.06
GPT2	Generative	16.71/66.57/13.75	14.71%	-0.745	0.814	82.05	0.99
CODR	Generative	14.58/1.31/11.37	3.4%	-1.155	0.051	81.90	0.43
Setup 2: With Summary Provided							
BERTQA	Extractive	70.97/67.77/69.34	85.28%	0.297	7.177	94.65	**66.33**
ELECTRA	Extractive	76.24/74.33/**75.48**	**92.57%**	**0.378**	**7.823**	**96.06**	65.40
BART	Generative	66.41/61.86/64.41	68.71%	0.257	6.613	93.85	38.42
T5	Generative	75.77/73.51/74.51	81.73%	0.369	7.728	95.22	57.93
VLT5	Generative	75.07/72.35/73.86	85.36%	0.376	7.597	95.20	59.80
GPT2	Generative	14.29/11.13/13.42	22.00%	-0.782	1.725	49.18	12.68
CODR	Generative	13.81/0.76/10.25	2.480%	-1.039	0.038	81.87	0.31
Setup 3: Without Summary Provided							
BART	Generative	40.29/21.40/32.48	49.07%	-0.166	2.260	**89.69**	7.41
T5	Generative	41.12/22.09/32.97	52.30%	-0.173	2.357	89.59	9.28
VLT5	Generative	42.87/22.60/**33.29**	**54.47%**	**-0.134**	**2.447**	89.53	**14.73**
GPT2	Generative	28.55/11.26/22.46	32.00%	-0.493	1.314	85.05	4.89
CODR	Generative	14.67/1.05/10.90	4.14%	-1.170	0.053	81.86	0.32

Table 5.1: Evaluation results for different models on the test set. ↑ means higher is better, ↓ means lower is better. All models use the OCR-extracted data and chart title in the input. 70:15:15 split is used for train:test:val.

tokens of the answer span. However, if there are irrelevant tokens in the span, a window may miss portions of the answer that fall outside the maximum sequence length. Also, during training, the model rejects those instances where the answer span is not contained wholly inside the input sequence. The problem is worse with BERTQA because it uses a sequence length of 512 which is much shorter than the article size (1268 tokens on average), whereas ELECTRA uses a sequence length of 2048 tokens, reducing this effect.

We notice that the models perform better when the summary is provided compared to when only the chart is provided (setup 2 vs. 3 in Table 5.1). This is expected because

when the summary is not given, the model does not have any textual reference to generate the answer to a given question. The extractive models BERTQA and ELECTRA achieve high ROUGE-L since they directly extract the text from the given summary to answer the question. ELECTRA performs better than BERTQA on all metrics except BLEU showing the benefit of the generator-discriminator based cascaded training.

When the summary is not provided, VLT5 achieves the best results across most measures, likely because it uses both visual and textual features to generate answers. On the contrary, CODR performs poorly possibly because it was originally intended for discourse generation whereas chart QA is quite a different task. GPT-2 performs better than CODR but it has lower ROUGE-L and CS scores compared to BART and T5.

We conduct several additional experiments to understand: (*i*) how performance may vary with respect to different chart types; (*ii*) how including the bounding box information as input impacts the performance; (iii) how naive baselines that either randomly select the answer from the input text or only take the question/summary as input perform; and (*iv*) how much the OCR-extracted text contributes to the output. Details of these experiments are as follows.

5.1.3 Performance by Chart Types

We analyze the performance by chart types in Table 5.2. We observe that our best performing models perform better at summarizing simple (pie charts) and frequent chart

VLT5 (Without Summary)					
	Bar	**Line**	**Pie**	**Area**	**Scatter**
ROUGE ↑	43.75/23.41/33.96	39.86/19.82/30.64	48.18/26.12/39.12	100.0/100.0/100.0	31.89/9.21/19.33
BLEURT ↑	-0.134	-0.150	-0.020	0.800	-0.329
CIDEr ↑	2.576	2.058	2.912	0.0	1.181
BERTScore ↑	89.60	89.21	90.86	98.79	86.35
VLT5 (With Summary)					
	Bar	**Line**	**Pie**	**Area**	**Scatter**
ROUGE ↑	75.61/72.9/74.38	73.19/70.46/71.93	80.1/77.57/79.48	100.0/100.0/100.0	67.46/66.83/67.46
BLEURT ↑	0.381	0.353	0.457	0.800	0.302
CIDEr ↑	7.657	7.414	7.920	0.0	7.378
BERTScore ↑	95.28	94.87	96.24	98.79	93.47

Table 5.2: Results of VLT5 (without summary) and VLT5 (with summary) on our test set *w.r.t.* chart types.

types (bar charts), whereas the performance generally decreases for complex and less frequent charts (*e.g.*,scatter plots).

5.1.4 Ablation Test

In table 5.3 we show the results of an ablation test, where each model is given the question and OCR-extracted chart data as input. Here, models with ”Bboxes” prefix are given the bounding boxes C_{bbox} of text segments in the chart, and the other models are repeated here from Table 5.1 for comparison. We observes that the inclusion of the bounding box information does not necessarily improves the performance. One possible reason could be that bounding box extraction of the OCR model is not always perfect as the model sometimes merges or splits the bounding boxes incorrectly.

Models	Type	ROUGE_-F (1/2/L) ↑	CS ↑	BLEURT ↑	CIDEr ↑	BERTScore ↑	BLEU ↑
Setup 1:With Article Provided							
Bboxes-BERTQA	Extractive	23.95/9.17/16.69	32.71%	-0.736	1.080	85.10	9.35
Bboxes-ELECTRA	Extractive	52.12/43.07/47.49	49.74 %	-0.051	4.379	89.97	30.77
Bboxes-BART	Generative	45.07/31.93/39.55	48.48%	-0.108	3.411	90.20	16.78
Bboxes-T5	Generative	66.21/60.15/63.40	67.49%	0.076	5.414	**93.55**	**41.58**
BBoxes-VLT5	Generative	55.91/46.53/51.76	62.05%	0.048	4.932	91.83	37.94
Bboxes-GPT2	Generative	16.78/6.89/14.08	15.45%	-0.757	0.85	81.12	1.36
Bboxes-CODR	Generative	13.67/0.97/10.88	1.73%	-1.026	0.035	82.40	0.00
BERTQA	Extractive	22.81/8.53/16.63	28.17%	-0.692	0.983	85.00	9.58
ELECTRA	Extractive	57.67/50.09/53.89	54.73 %	0.066	5.100	91.01	38.79
BART	Generative	50.68/38.95/45.55	54.29%	-0.017	4.121	91.08	23.91
T5	Generative	66.57/60.40/**63.46**	**68.24%**	**0.103**	**5.437**	93.53	41.28
VLT5	Generative	58.90/49.97/54.82	66.06%	0.076	5.227	92.34	40.06
GPT2	Generative	16.71/66.57/13.75	14.71%	-0.745	0.814	82.05	0.99
CODR	Generative	14.58/1.31/11.37	3.4%	-1.155	0.051	81.90	0.43
Setup 2:With Summary Provided							
Bboxes-BERTQA	Extractive	71.76/68.56/70.13	85.25%	0.309	7.216	94.73	**67.24**
Bboxes-ELECTRA	Extractive	75.87/73.78/75.09	91.93%	0.367	7.766	95.98	65.17
Bboxes-BART	Generative	63.11/57.22/60.44	65.93%	0.205	6.090	93.23	36.12
Bboxes-T5	Generative	76.54/74.24/75.35	84.21%	0.377	7.780	95.36	60.26
BBoxes-VLT5	Generative	73.91/71.07/72.67	80.92%	0.341	7.470	94.98	58.25
Bboxes-GPT2	Generative	18.21/15.41/17.47	25.73%	-0.686	1.955	43.28	16.06
Bboxes-CODR	Generative	12.16/0.73/9.31	2.040%	-1.283	0.046	80.17	0.19
BERTQA	Extractive	70.97/67.77/69.34	85.28%	0.297	7.177	94.65	66.33
ELECTRA	Extractive	76.24/74.33/**75.48**	**92.57%**	**0.378**	**7.823**	**96.06**	65.40
BART	Generative	66.41/61.86/64.41	68.71%	0.257	6.613	93.85	38.42
T5	Generative	75.77/73.51/74.51	81.73%	0.369	7.728	95.22	57.93
VLT5	Generative	75.07/72.35/73.86	85.36%	0.376	7.597	95.20	59.80
GPT2	Generative	14.29/11.13/13.42	22.00%	-0.782	1.725	49.18	12.68
CODR	Generative	13.81/0.76/10.25	2.480%	-1.039	0.038	81.87	0.31
Setup 3:Without Summary Provided							
Bboxes-BART	Generative	39.79/21.24/31.94	49.12%	-0.171	2.254	89.63	7.40
BBoxes-T5	Generative	40.88/21.86/32.96	51.12%	-0.181	2.339	89.56	9.09
BBoxes-VLT5	Generative	42.62/22.18/32.93	53.26%	**-0.132**	2.382	89.52	14.01
Bboxes-GPT2	Generative	28.57/10.91/22.24	31.30%	-0.459	1.246	84.50	3.92
Bboxes-CODR	Generative	14.18/1.19/10.22	4.43%	-1.053	0.053	81.77	0.22
BART	Generative	40.29/21.40/32.48	49.07%	-0.166	2.260	**89.69**	7.41
T5	Generative	41.12/22.09/32.97	52.30%	-0.173	2.357	89.59	9.28
VLT5	Generative	42.87/22.60/**33.29**	**54.47%**	-0.134	**2.447**	89.53	**14.73**
GPT2	Generative	28.55/11.26/22.46	32.00%	-0.493	1.314	85.05	4.89
CODR	Generative	14.67/1.05/10.90	4.14%	-1.170	0.053	81.86	0.32

Table 5.3: Ablation study results for understanding the effect of including the OCR extracted bounding boxes as input. All models use OCR-extracted data as input. "Bboxes-" models use bounding box information of chart elements.

5.1.5 Evaluation with Naive Baselines

In order to better understand the dataset and how challenging the task is, we analyze the results of fine-tuning and testing T5 and VLT5, our best performing models with only

Models	ROUGE_-F (1/2/L) ↑	CS ↑	BLEURT ↑	CIDEr ↑	BERTScore ↑	BLEU ↑
T5-Only Question	31.57/13.19/25.05	31.66%	-0.396	1.401	87.94	5.34
T5-Only Summary	67.56/63.78/65.47	**82.22%**	0.199	**6.856**	93.94	49.18
VLT5-Only Question	29.56/11.56/22.43	32.98%	-0.363	1.322	87.74	6.34
VLT5-Only Summary	67.31/63.54/**65.79**	77.47%	**0.238**	6.803	**94.07**	**52.71**
Random Summary Baseline	54.48/46.22/49.07	62.28%	-0.015	4.997	90.89	45.69
Random Article Baseline	23.95/8.81/17.12	19.66%	-0.641	0.960	84.36	5.50

Table 5.4: Results for T5 and VLT5 models with only the question or the summary as input on OpenCQA test set and a baseline that randomly selects 52% and 7% tokens for the with summary and with article setups respectively.

the question Q or *the summary* S as input (see Table 5.4) In addition, we choose a naive baseline which randomly selects the answer from the given input text. This baseline spans 52% tokens of the input when only the summary S is given whereas it spans 7% tokens of the input when the article D is given (based on the statistics of tokens overlapping between the summary or article and the gold answer in our dataset(see Figure 3.5)). As expected, we can see that both models struggle to answer the question when only the question is given as input. However, they perform better with only summary as input as the model has access to the summary containing the answer. In contrast, the performance drops when we use the baseline that randomly selects the answer, especially when the entire article is given as input as it becomes more difficult to randomly guess the answer.

5.1.6 The Effect of the OCR-text

We further analyze how the model extracts the answer from different input sources (e.g., input summary vs. OCR-generated text). For this purpose, we consider three new metrics

Models	CEF-Precision ↑	CEF-Recall ↑	CEF-Coverage ↑
Test	49.32	25.89%	47.34
	Setup 1:With Article Provided		
Bboxes-VLT5	38.05%	21.16%	42.18%
VLT5	42.32%	22.57%	43.11%
Bboxes-T5	**42.82%**	22.63%	43.60%
T5	42.50%	**24.31%**	**45.30%**
	Setup 2:With Summary Provided		
Bboxes-VLT5	42.94%	25.18%	46.71%
VLT5	**43.48%**	25.15%	46.96%
Bboxes-T5	42.53%	25.78%	47.74%
T5	42.04%	**26.96%**	**49.06%**
	Setup 3:Without Summary Provided		
Bboxes-VLT5	48.70%	22.46%	43.54%
VLT5	50.18%	23.09%	**44.08%**
Bboxes-T5	60.71%	20.36%	38.50%
T5	**61.89%**	**21.24%**	40.32%

Table 5.5: Chart Extraction Factor (CEF) scores for VLT5 and T5 across all problem setups. "Bboxes-" models use bounding box information of chart elements.

for Chart Extraction Factor (CEF) scores as follow:

CEF-Precision: Shows how much fraction of generated tokens in the answer (excluding stopwords) A is coming from OCR-extracted text segments C_{label}. This will inform how much of the generated answer is taken from the OCR-extracted data of the chart.

CEF-Coverage: Represents how much fraction of OCR-extracted text segments C_{label} contribute to the generated answer A. An OCR-extracted text segment contributes to the answer if any generated tokens in the answer A appear in the segment. This will give a sense of how spread out the chart data extraction is. For more focussed tasks like *Identify*

question types, we would want the CEF-Coverage to be lower so it is more focussed in a particular region of the chart. For a higher level task like *Discover*, we would want the CEF-Coverage to be higher so it will get high level insights by looking at the chart overall.

CEF-Recall: Shows how much fraction of OCR tokens in C_{label} is recalled in answer A.

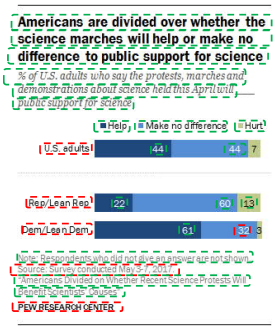
Model Input		Generated Answer
Question: What are the partisans views on whether the marches will increase public support for science? 	**OCR extracted data:** Americans are divided orever whether the | science marches will help or make no | difference topublic support for science | % of USS adults who say the protests, marches Aprilwill and | demonstrations public about science held this Aprilwill | public supportfor science | Help Make no d ifference Hurt | U.S ad ults | 44 | 44 | Rep/Lean Rep | 22 | 60 | 13 | Dem/Lean Dem | 61 | 32 | Note: Respondents who did notgive 2017 an answerare.not.shown. | "Americans Source Survey Divided conducted May3-7 notgive Recent 2017 | "Americans Survey Divided on Whether Recent Science Protests Will | Benefit Scientists' Ca uses | PEW RESEARCH CENTER **Summary:** Americans are closely divided on the issue of whether the marches will increase public support for science: 44% think they will help and 44% think they will make no difference. Another 7% believe they will hurt public support . Most Democrats and Democratic leaners (61%) believe the marches and demonstrations held in April will benefit public support for science. By contrast, just 22% of Republicans and independents who lean to the GOP say the marches will help drive public support for science, while six-in-ten (60%) of this group believes the marches will have no impact on public support and 13% say the marches will hurt public support.	Americans are closely divided on the issue of whether the marches will increase public support for science: 44% think they will help and 44% think they will make no difference. Another 7% believe they will hurt public support. Most Democrats and Democratic leaners (61%) believe the marches and demonstrations held in April will benefit public support for science. By contrast, just 22% of Republicans and independents who lean to the GOP say the marches will help drive public support for science , while six-in-ten (60%) of this group believes the marches will have no impact on public support and 13% say the marches will.

Table 5.6: Illustration of chart extraction scoring.Green text are answer tokens generated from OCR text, Blue text are answer tokens not found in OCR text but generated from summary, Red text highlights the remaining OCR-extracted text. All text sequences in the chart with bounding boxes are extracted by the OCR model and *Green* bounding boxes are used in the generated answer. CEF(Recall): **43.48**; CEF(Coverage): **86.36**; CEF(Precision): **54.54**.

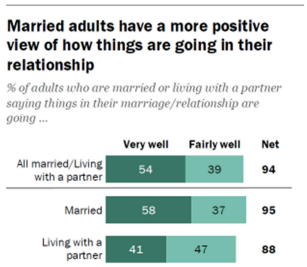
The results on our top performing models are shown in Table 5.5. We observe that the CEF-Precision score decreases across all models when the summary or the article is given as input compared to the setup where only the OCR text is given. This suggests that the model utilizes less information from the extracted OCR text when it has access to summary or article. In contrast, CEF-Coverage scores tend to increase when the

summary is provided in comparison to without summary as the answer tends to cover more OCR text segments if the summary is present. The CEF-recall scores increase when the summary is provided since many tokens in the summary overlap with the OCR extracted text. Table 5.6 illustrates an example of how OCR-extracted texts and the input summary contributed to the answer.

5.1.7 Additional Sample Outputs

Additional sample outputs from the best performing model are shown in Figure 5.1.

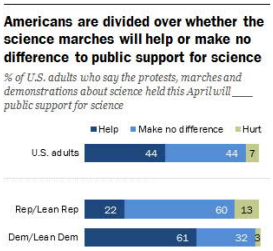
Question: How do married adults conceive their relationship satisfaction in comparison to adults living with a partner?



VLT5: Compared with adults who are married or living with a partner, many are more positive about how things are going in their relationship. Roughly nine-in-ten say that their marriage or relationship is going very or fairly well, compared with 37% of adults who are living with a partner.

Gold Answer: Married adults are more likely than those who are living with a partner to say things are going very well in their relationship (58% vs. 41%) They also express higher levels of satisfaction with specific aspects of their relationship, including the way household chores are divided between them and their spouse or partner how well their spouse or partner balances work and personal life, how well they and their spouse or partner communicate, and their spouse's or partner's approach to parenting.

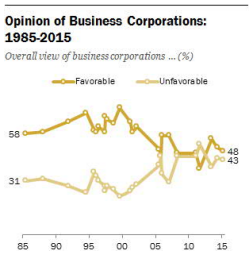
Question: What are the partisans views on whether the marches will increase public support for science?



VLT5: Democrats and Democratic-leaning independents are divided: 60% say the marches will increase public support for science, while just 32% say the demonstrations will not hurt the public.

Gold Answer: Most Democrats and Democratic leaners (61%) believe the marches and demonstrations held in April will benefit public support for science. By contrast, just 22% of Republicans and independents who lean to the GOP say the marches will help drive public support for science, while six-in-ten (60%) of this group believes the marches will have no impact on public support and 13% say the marches will hurt public support.

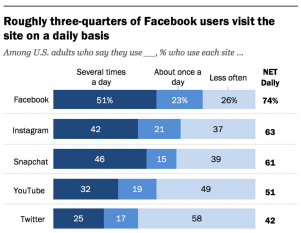
Question: What are the public opinions about business corporations



VLT5: Currently, 48% express a favorable view of business corporations, while 43% express an unfavorable view.

Gold Answer: 48% hold a favorable view of business corporations, compared with 43% who hold an unfavorable view.

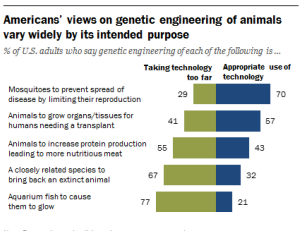
Question: Describe the trend of Facebook users usage of the site on daily basis?



VLT5: About three-quarters of Facebook users (71%) say they visit the site several times a day, or several times a day. The roughly one-third (26%) say they visit the site at least once a day, or several times a day.

Gold Answer: Roughly three-quarters of Facebook users (74%) visit the site daily, including about half who do so several times a day.

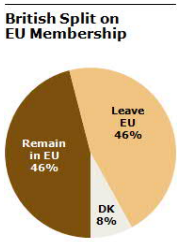
Question: What are views on genetic engineering of mosquitos?



VLT5: Two-thirds or more of Americans (70%) say genetic engineering of mosquitos is appropriate to prevent the spread of disease by limiting their reproduction. But a majority of Americans (55%).

Gold Answer: Seven -in-ten Americans (70 %) believe that genetically engineering mosquitoes to prevent their reproduction and therefore the spread of some mosquito-borne diseases would be an appropriate use of technology, while about three-in-ten (29%) see the use of genetic engineering for this purpose as taking technology too far.

Question: Compare the British approval and rejection levels about being a member of the EU.



VLT5: As the 2014-10 EU elections, Britain's approval rating is only slightly lower than the 49% who approved of becoming a member while only 8% said they would leave.

Gold Answer: The British, remain evenly divided on leaving the EU: 46% say stay and 46% say go.

Figure 5.1: Sample outputs from the best performing model VLT5 (without summary).

5.2 Human Evaluation

This section covers the human evaluation we perform on the outputs of the trained models, detailing the explanation of each measure used.

Automatic evaluation methods are often insufficient to evaluate model outputs as they may not take into account some important information or may not be suitable to measure some important aspects of textual quality. To further assess the quality of the answers generated by the models, we performed a human study on 150 randomly sampled charts from our OpenCQA dataset with three internal annotators who are native speakers of English. For comparison, we chose the outputs from the best performing generative model (VLT5) on the with and without summary settings, which we denote as VLT5-S and VLT5, respectively. We also provide the gold answer (as control) and randomise the order of all three answers to remove biases. For these three criteria, annotators provided a 5-point Likert scale rating from 1 (the worst) to 5 (the best). The annotators assess each answer based on three criteria adapted from [86]:

Factual correctness is a criteria we use to check if the summary contains data facts that are consistent with the information in the chart. It is calculated by calculating the number of factual tokens in the generated summary that is consistent with the data represented in the chart. A score of 1 indicates the summary has none or very little factual correctness with respect to the chart, and a score of 5 indicates it is nearly/perfectly aligned factually with the chart.

Relevance is another human evaluation criteria we use to check if the given summary is relevant to the given question and provides a suitable answer to the given question. It is

calculated by measuring the proportion of the summary that is relevant to the question. A score of 1 indicates the whole summary is either mostly or completely irrelevant, and a score of 5 indicates all sentences are nearly or completely relevant to the question.

Fluency is a criteria where we would like to measure the extent to which the generated text is deficient of formatting issues, capitalization errors or ungrammatical sentences (*e.g.,* fragments, missing components) which would otherwise make the text difficult to read. A score of 1 would indicate the whole summary is completely lacking fluency and grammar, but a score of 5 indicates it is very fluent with good grammar in place.

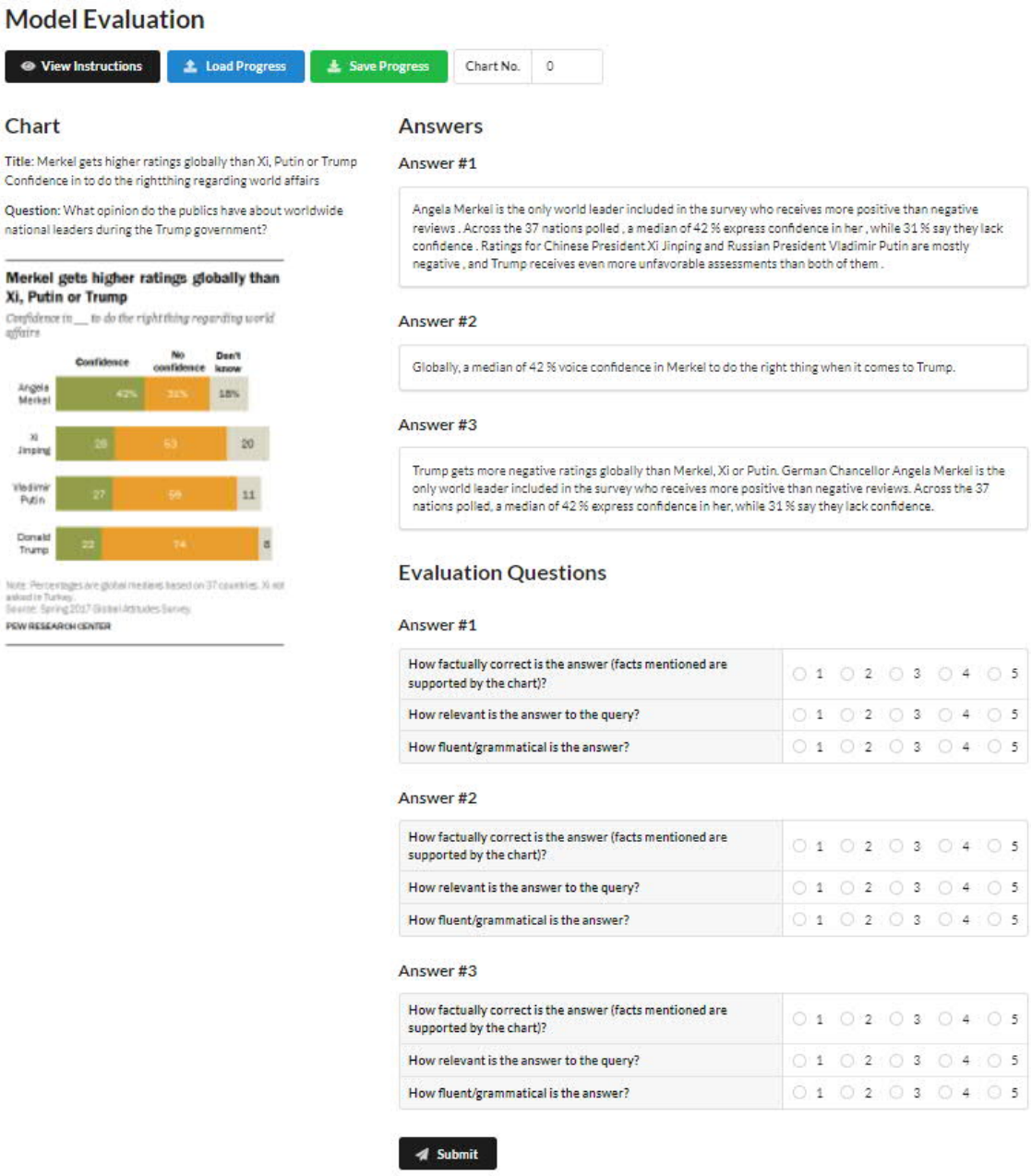
We further propose a new *recall-oriented* criteria called **Content Coverage** which measures the extent to which the facts from the gold answer are covered by the generated text, ranging from 1 (0-20%) to 5 (80-100%). For this, the order of outputs of the two models were randomized to counterbalance any potential ordering effect. Each comparison was performed by three annotators. To measure inter-annotator reliability, we computed Krippendorff's alpha [40] and found a good overall agreement with an alpha coefficient of 0.806.

	Factual	Relevance	Fluency	Content Cov.	CS
VLT5	44.89%	86.67%	92.45%	41.11%	53.19%
VLT5-S	82.45%	88.67%	98.45%	85.12%	81.15%
Gold	96.45%	96.00%	98.89%	100.00%	100.00%
Krippendorff's α	0.75	0.76	0.83	0.79	

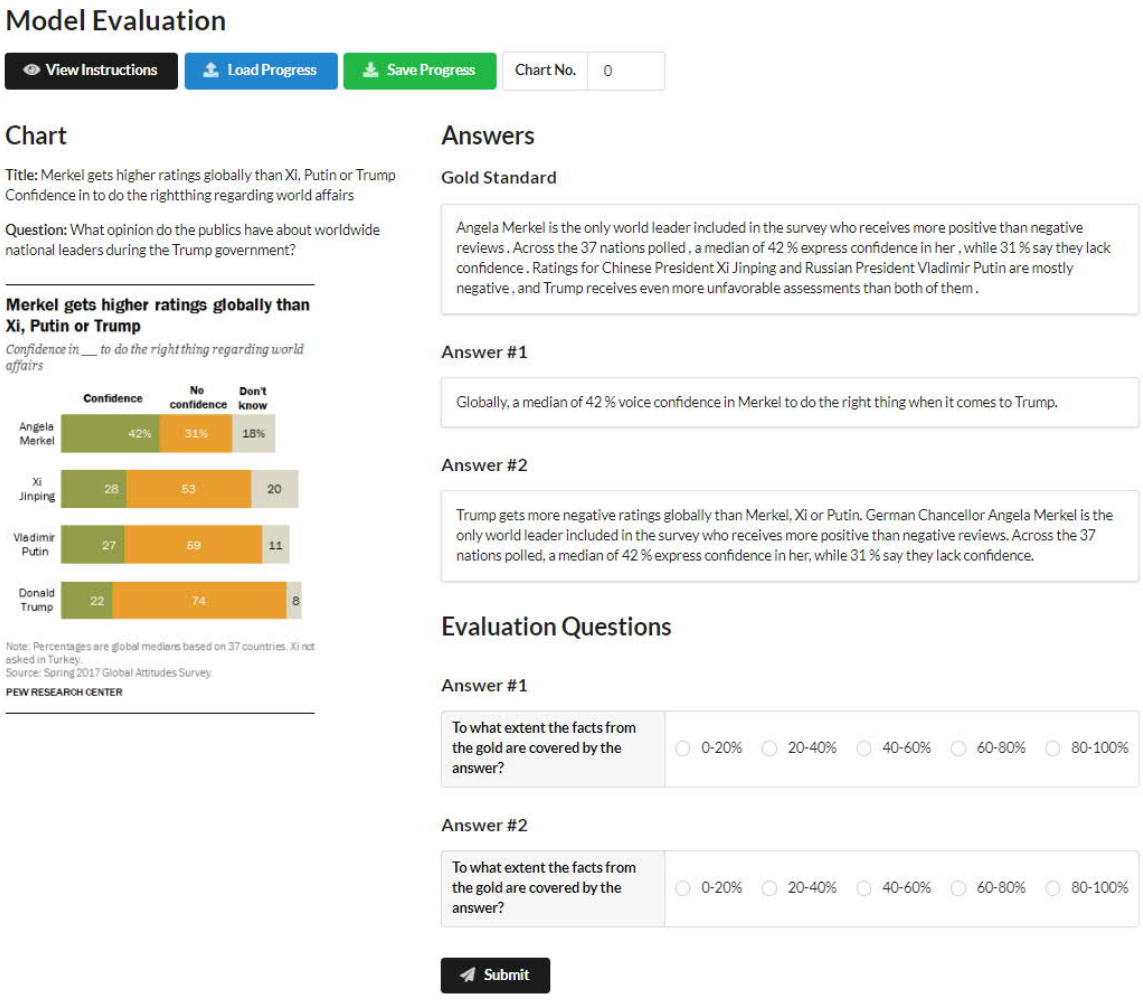
Table 5.7: Human evaluation results for comparing between the outputs of VLT5 (without summary), VLT5-S (with summary) and the gold answer. The last row shows the agreements.

Table 5.7 shows the results of human evaluation, where for each model we show the percentage of answers that received the nearly perfect/perfect ratings (*i.e.,* 4 and 5). The gold answers receive high ratings across the first three measures which attests the quality of our dataset. VLT5-S achieves higher ratings across all measures than VLT5 as it has access to the summary. The VLT5 model also received impressive ratings on Relevance and Fluency while it has more room for improvement on Factual Correctness and Content Coverage, especially when the summary is not given. Lastly, we see that the automatic and the human evaluation measures for content coverage, *i.e.,* the CS score and Content Coverage, both of which are recall oriented, tend to be correlated.

The user interface for the human evaluation annotation task of comparing among generated answers by different models are given in Figure 5.2.



(a) The interface for human evaluation that shows the gold answer as well as the outputs from VLT5 (VLT5 without summary) and VLT5-S(VLT5 with summary) in random order. The factuality, relevance and fluency measures are rated on a 5-point likert scale.



(b) The interface for human evaluation on the content coverage measure. The human compares between the gold answer and the outputs from VLT5 (VLT5 without summary) and VLT5-S(VLT5 with summary) presented in random order.

Figure 5.2: Human evaluation interfaces.

5.3 Error Analysis and Challenges

Here we analyse the outputs for errors and current challenges with current state of the art models on our proposed task.

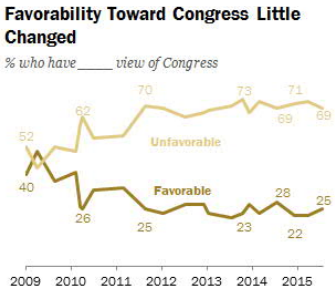
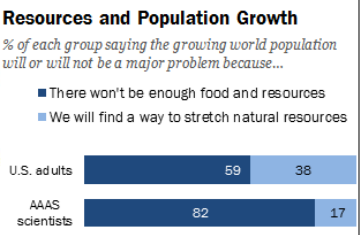
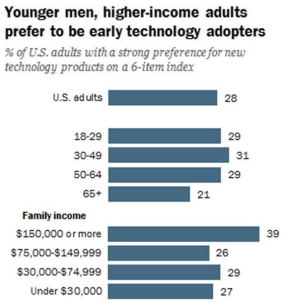
Q1: Describe the change in favorability towards Congress. 	**Q2:** Describe the proportions of U.S. adults and AAAS scientists that think population growth won't be a problem. 	**Q3:** How is the relation between family income and early technology adoption? 
Gold: Views of Congress continue to hover near record lows: Just 25% of Americans view Congress favorably.	**Gold:** Just 17% of AAAS scientists and 38% of Americans said population growth won't be a problem because we will find a way to stretch natural resources.	**Gold:** Adults with family incomes of at least $150,000 annually tend to express a stronger preference for early technology adoption on this summary index (39% do so) compared with those at middle and lower income levels
VLT5: 25% express a favorable opinion of Congress, while 69% view it unfavorably. But since April 2014, the share of Americans with a positive view of Congress has fallen off to 25%.	**VLT5:** Fully 59% of AAAS scientists say population growth will n't be a major problem because there will not be enough food and resources in the world, and 17% say we will find a way to stretch natural resources.	**VLT5:** Younger men (39%) are more likely than older adults (26%) to express a strong preference for being early technology adopters.
Bboxes-VLT5: A majority of the public (69%) has a favorable opinion of Congress, while 23% view it unfavorably. In July 2014, the balance of opinion was just the opposite: 47% expressed a favorable opinion of Congress, while 25% held an unfavorable view.	**Bboxes-VLT5:** Two-thirds of AAAS scientists say population growth will n't be a major problem because of scarce resources and other resources. At the same time, about six-in-ten of both groups say that the U.S. will find a way to stretch natural resources and hope that we will find a way to stretch these resources	**Bboxes-VLT5:** Older people in the U.S. tend to have higher levels of education and skills, rather than being early adopters. Many of those with higher incomes prefer early technology adoption; in fact, lower-income adults lag behind younger men in this regard.

Figure 5.3: Sample outputs from the Pew test set on VLT5 model (without summary provided)."Bboxes-" models use bounding box information of chart elements. Red indicates factual errors.

We have manually analyzed the outputs from our best performing model to explain the key challenges that existing models face. Overall, the outputs of VLT5 model are often fluent and grammatically correct as evident from Figure 5.3. Also, this model rarely hallucinates and is often true to the source context. However, there are still several areas

for improvement which we discuss below.

Factual Correctness and relevancy In Figure 5.3, both variants of VLT5 fail to associate some reported data values (highlighted in red) with their corresponding axis labels. In (Q3 in Figure 5.3), the VLT5 model outputs some facts about the relation between 'age' and 'early technology adoption' but the question asked about the relation with 'family income', not 'age'. To address the factual errors and irrelevant facts, future models could create better input representations including semantic graph representations [77] by extracting and exploiting relationships among the question, data values, chart elements (*e.g.,* axis labels) and their visual features (*e.g.,* colors).

Mathematical & Logical Reasoning Often gold answers include numerical reasoning to provide rich explanations. For example, in Figure 5.3(Q2), the gold answer provides a sense of limited proportion by mentioning "Just 17% of AAAS...", which is not captured in the generated outputs. Future models may improve such numerical reasoning as an intermediate step before surface realization [14]. In another case (Q3 in Figure 5.3), the gold answer considers that incomes below 150k are middle and lower income levels which is not captured by models. Future work could infer such rich semantics with the help of external knowledge sources and inference based modelling [6, 75].

Rare Topics We also observe that the VLT5 model tends to produce less fluent and factually correct answers when the sample comes from a rare topic (*e.g.,* Bboxes-VLT5 in Q2 in Figure 5.3). Future work could focus on transfer learning and building more diverse datasets to handle such variations.

Automatic Evaluation It has been acknowledged that metrics like BLEU do not capture many data-to-text generation issues [59]. This is especially the case for open-ended questions where correct answers can be expressed in multiple possible ways. While we evaluated the models through several metrics, we urge for further research on building better automatic metrics that capture model performance in terms of factual correctness, relevancy and reasoning aspects while also correlating strongly with human judgements.

6 Conclusions and Future Work

In this chapter we provide our final conclusions from this research. We follow up with a brief summary of all our contributions in this work. Lastly we discuss future directions for this work covering the limitations of our dataset and modelling approaches.

6.1 Conclusions

In this thesis we proposed a new benchmark, OpenCQA, for answering open-ended questions about charts. These open-ended questions generally identify high level insights about the whole chart or focused insights on part of the chart, summarise the information in the chart or compare between different parts of the chart. We also applied several state-of-the-art baselines and metrics and further introduced a new metric 'Chart Extraction Factor(CEF)' that does better at evaluating the models ability to extract information from the chart. Our evaluations suggest that current state of the art generative models can produce fluent text but still struggle to produce relevant and factually correct statements with numerical and logical reasoning. Further, the existing metrics are insufficient for accurately measuring model performance on open-ended question answering tasks while

measuring logical and numerical reasoning and inference abilities thus new metrics are required. We hope that OpenCQA will serve as a useful research benchmark for model and metric development and motivate other researchers to explore this new task.

6.2 Summary of Contributions

To summarise our overall contributions, we introduced a novel task for open ended question answering on charts by constructing a relatively large benchmark dataset from real world sources. We explained in detail the process of collecting real world data from users and filtering user annotations to produce a high quality benchmark. We then applied a set of state-of-the-art neural models to serve as strong baselines for performing this task, by using OCR based approaches for extracting text from charts and adapting models to use the OCR-extracted text and other inputs for answering questions. To evaluate these models we applied a set of automatic evaluation metrics and proposed some new metrics to identify the performance of current models on this task.

We then conducted extensive analysis on our dataset by understanding the topics, statistics, chart and question types of our dataset. Further we extensively analysed the scope of model performance on our task by understanding performance by chart types, conducting ablation tests to know the benefits of each component of the model input, comparisons with naive baselines to understand how much better our models perform than simple heuristics, human evaluations of model results to get expert judgements on model

performance and lastly an overview of the errors and challenges that existing models face based on a qualitative analysis of the model results.

6.3 Future Work

A limitation of our dataset is that we were limited to Pew research (pewresearch.org) as a source due to the nature of our task and thus could not consider other sources. Future work could expand on our dataset when suitable sources become available. Such a dataset should contain rich reasoning questions that require the model to interpret the chart and make inferences to answer questions. This will push the state of the art further by enabling models to reason and infer high level information, which is necessary for models to reach human level performance.

We did not consider long-range sequence models such as *linformer* [79] or recently proposed *memorized transformer* [82] for modeling long sequences. These models will be helpful for taking in inputs that contain multiple concatenated components that make the input sequence length very long, allowing the model to learn from more input sources for answering complex questions. A similar problem exists with current state-of-the-art extractive models, which are limited to extracting contiguous text sequences from given text documents, which do not work well for open-ended question answering tasks.

Future work can also focus on developing metrics that can measure numerical and

logical accuracy of models. Current evaluation metrics cannot identify this, and thus may give false impressions of good model performance. Models that can visually parse the meaning of the chart can allow for better chart representations than just using OCR-extracted chart data. This would require using a pretrained chart encoder trained on a large dataset of charts. This a challenging task due to the lack of large real world chart datasets.

Models for interpreting charts require more development. Due to the lack of large chart datasets, it has not yet been possible to train large scale models that can easily adapt to a variety of tasks with minimal training. By expanding the data sources we have and coming up with new ways to train a model to understand the high level hierarchical relationships that are visually salient in charts, we can improve chart understanding abilities allowing models to perform a number of chart related tasks. Lastly, our task setup is limited since we only have access to the automatically extracted OCR data which is often noisy. Future methods can focus on improving OCR extraction for this task to improve the input of the model thus boosting model performance.

Bibliography

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Regina Barzilay and Mirella Lapata. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 331–338, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/H05-1042`.

[3] Tal Baumel, Matan Eyal, and Michael Elhadad. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*, 2018.

[4] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on Freebase from question-answer pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[5] Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Methods for exploring and mining tables on wikipedia. In *Proceedings of the ACM SIGKDD workshop on interactive data exploration and analytics*, pages 18–26, 2013.

[6] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075`.

[7] Ankit Chadha and Rewa Sood. Bertqa–attention on steroids. *arXiv preprint arXiv:1912.10435*, 2019.

[8] R. Chaudhry, S. Shekhar, U. Gupta, P. Maneriker, P. Bansal, and A. Joshi. Leaf-qa: Locate, encode attend for figure question answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3501–3510, 2020. doi: 10.1109/WACV45572.2020.9093269.

[9] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521, 2020.

[10] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan A. Rossi, and Razvan C. Bunescu. Figure captioning with reasoning and sequence-level training. *CoRR*, abs/1906.02850, 2019. URL `http://arxiv.org/abs/1906.02850`.

[11] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

[12] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. Logical natural language generation from open-domain tables. *arXiv preprint arXiv:2004.10404*, 2020.

[13] Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. Logic2Text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2096–2111, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.190. URL `https://aclanthology.org/2020.findings-emnlp.190`.

[14] Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. Logic2text: High-fidelity natural language generation from logical forms. *arXiv preprint arXiv:2004.14579*, 2020.

[15] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[16] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[17] Hoa Trang Dang. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12, 2005.

[18] Seniz Demir, Sandra Carberry, and Kathleen F. McCoy. Summarizing information graphics textually. *Computational Linguistics*, 38(3):527–574, 2012. doi: 10.1162/COLI_a_00091. URL https://www.aclweb.org/anthology/J12-3004.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[21] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*, 2019.

[22] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL https://aclanthology.org/W16-3210.

[23] Li Gong, Josep Crego, and Jean Senellart. Enhanced transformer model for data-to-text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 148–156, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5615. URL https://aclanthology.org/D19-5615.

[24] Li Gong, Josep M Crego, and Jean Senellart. Enhanced transformer model for data-to-text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 148–156, 2019.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.

[29] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.398. URL https://aclanthology.org/2020.acl-main.398.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[31] Enamul Hoque, Kavehzadeh Parsa, and Ahmed Masry. Chart question answering: State of the art and future directions. *Journal of Computer Graphics Forum (Proc. EuroVis)*, 41(3):1–18, 2022.

[32] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[33] Kushal Kafle, Scott Cohen, Brian L. Price, and Christopher Kanan. DVQA: understanding data visualizations via question answering. *CoRR*, abs/1801.08163, 2018. URL http://arxiv.org/abs/1801.08163.

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

[35] Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. Answering questions about data visualizations using efficient bimodal fusion. In

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1498–1507, 2020.

[36] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

[37] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[38] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

[39] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[40] Klaus Krippendorff. Computing krippendorff's alpha-reliability. *Computing*, 1, 2011.

[41] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL https://arxiv.org/abs/1602.07332.

[42] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1128. URL https://www.aclweb.org/anthology/D16-1128.

[43] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[44] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

[45] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[46] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[48] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[49] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[50] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 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, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.177.

[51] Hongyuan Mei, TTI UChicago, Mohit Bansal, and Matthew R Walter. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *Proceedings of NAACL-HLT*, pages 720–730, 2016.

[52] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[53] Vibhu O. Mittal, Johanna D. Moore, Giuseppe Carenini, and Steven Roth. Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24(3):431–467, 1998. URL https://www.aclweb.org/anthology/J98-3004.

[54] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[55] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, et al. Fetaqa: Free-form table question answering. *arXiv preprint arXiv:2104.00369*, 2021.

[56] Preksha Nema, Mitesh M Khapra, Anirban Laha, and Balaraman Ravindran. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, 2017.

[57] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.inlg-1.20.

[58] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[59] Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*, 2020.

[60] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.

[61] Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. Focused attention improves document grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021.

[62] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *Open-AI Blog*, 2019.

[63] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[64] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018.

[65] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2): 137–169, 2005.

[66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[67] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[68] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics*, 23(1):341–350, 2016.

[69] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

[70] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

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

[72] Hrituraj Singh and Sumit Shekhar. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.264. URL https://www.aclweb.org/anthology/2020.emnlp-main.264.

[73] Andrea Spreafico and Giuseppe Carenini. Neural data-driven captioning of time-series line charts. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375351. doi: 10.1145/3399715.3399829. URL https://doi.org/10.1145/3399715.3399829.

[74] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

[75] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

[76] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.

[77] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3233–3241, 2017.

[78] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[79] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[80] Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1239. URL https://aclanthology.org/D17-1239.

[81] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.

[82] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022.

[83] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*, 2016.

[84] Zichao Yang, Phil Blunsom, Chris Dyer, and Wang Ling. Reference-aware language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1850–1859, 2017.

[85] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006.

[86] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*, 2021.

[87] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[88] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[89] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. Qmsum: A new benchmark for query-based multi-domain meeting summarization, 2021.

[90] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.

[91] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

7 Appendix

7.1 Ethical Considerations

During the dataset collection and annotation process, we had many ethical issues to take into consideration. To respect the intellectual property of the chart publishers, we only used publicly available charts from resources that provide publication rights of downloaded content for academic purposes. According to the terms and conditions for Pew,[6] users are allowed to use the content as long as they are attributed to the Center or are not attributed to a different party.

To fairly compensate the Mechanical Turk annotators, we compensated the annotators based on the minimum wage in the US (7.25 US$ per hour), which is 1 US$ for each task. Additionally, to protect the privacy of these annotators, all of their annotations were anonymized. To ensure the reproducibility of our experimental results, we have provided the hyperparameter settings in Chapter 4 .

We foresee one possible misuse of our models that is to spread misinformation. Currently, our model outputs tend to appear fluent but contain some factual errors, as detailed in Section 5.3. Hence, if such model outputs are published without being corrected, it may mislead and misinform the general public.

[6] https://www.pewresearch.org/about/terms-and-conditions/