Towards Multimodal Reasoning and Inference using Large Language Models

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics by Research

by

Suyash Vardhan Mathur
2019114006
suyash.mathur@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
July 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Towards Multimodal Reasoning and Inference using Large Language Models" by Suyash Vardhan Mathur, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Advisor: Prof. Manish Shrivastava

To Baba, Dadi, Nanaji and Nani

Acknowledgments

I would like to extend my sincere gratitude to my advisor, Prof. Manish Shrivastava, for his unwavering support, invaluable guidance, and insightful feedback throughout my research journey. From teaching the meaning of *K, Q, V* in *Transformer Architecture* during Intro to NLP to writing this thesis today, he has guided me through every step along this journey.

I am also grateful to my collaborator and mentor, Dr. Vivek Gupta, for his exceptional collaboration and for pushing me to strive towards truly impactful research. From brainstorming ideas to motivating me to work when I had no faith in myself, this journey wouldn't have been possible without him.

I am highly grateful to Prof. Dan Roth and Prof. Mohit Bansal for their invaluable guidance and mentorship throughout the course of my work on Multimodal Tables. Without their expertise and experience, it wouldn't have been possible to complete the project.

I am incredibly thankful to my collaborators, Jainit Sushil Bafna, Kunal Kartik, and Harshita Khandelwal, for helping me out with my research work. It would have been incredibly difficult to do as extensive work without their massive involvement and help. I would also like to thank my collaborator and dear friend, Akshett Rai Jindal, for his help with the SemEval tasks. His support and his ideas made it possible to submit the papers on time and achieve impressive results on the tasks.

I am also thankful to Prof. Ponnurangam Kumaraguru, Prof. Manish Gupta and Dr. Maneesh Singh for fruitful research collaborations and insights over various research projects through my college years.

I would also extend my thanks to my dear friends, Aditya Verma, Tushar Choudhary, Ayush Goyal, Rutvij Menavlikar, Chayan Kochar, CYK Sagar, Jaywant Patel, Charchit Gupta, Mihir Bani, Shubhransh Singhvi, Jayant Panwar, Arvapalli Akhilesh for their support throughout my years of engineering, whether academic or non-academic. Without these people, I wouldn't have made it to the end of my degree; even if I did, it would have been much less fun. A huge thanks to KV Aditya Srivatsa for being an awesome senior and guiding me from my first year until my thesis submission.

Finally, I would like to thank my father, Mr Anurag Mathur, my mother, Mrs Sonia Mathur, and my sister, Ms Ayushi Mathur, for their constant and unwavering faith in me. Without them, I would not have had the motivation to complete my research work in time and wouldn't be writing this document right now. I would also like to thank my Nanaji, who, being a Professor himself, encouraged me to pursue good research work and strive for the best. I wouldn't be anywhere without you! I am also incredibly grateful to my Babaji, Dadiji, and Naniji for their love and encouragement throughout all these years. I hope I will manage to make you all proud someday!

Abstract

Great strides have been made in Natural Language Processing (NLP) and Computer Vision (CV) in recent years. Large Language Models (LLMs), especially those of the parameter sizes of GPT-3.5 and GPT-4 have revolutionized tasks ranging from summarization to question answering, while Vision Transformers have enabled the development of highly efficient image segmentation, object detection, image synthesis models. However, there is still much work to be done in Multimodal space, involving the usage of both NLP and CV to process input/output involving both text and images.

In this dissertation, we work towards Multimodal Inference and Reasoning by LLMs and pursue research questions related to Multimodal Question Answering by such LLMs through three distinct problems: Multimodal Emotion-Cause Pair Extraction in Conversations, Question Answering using LLMs for Unconventional Reasoning, and using Multimodal Large Language Models (MLLMs) to perform Knowledge-aware Inference and Reasoning over Semi-Structured Multi-modal Tables.

We first explore Multimodality using one of the most fundamental NLP tasks – Emotion Analysis through Multimodal Emotion-Cause Pair Extraction in Conversation. We model the task as both an utterance-labelling and a sequence-labelling problem and experiment with different encoders to encode the visual, audio and textual modalities in the conversations. We conducted a comparative study that involved baselines using different encoders with an MLP, BiLSTMs, and those incorporating a BiLSTM+CRF layer.

Going further, we explore the task of Unconventional Reasoning using LLMs on questions involving lateral thinking, which requires looking at problems from an unconventional perspective and defying existing conceptions and notions. We experiment on the BrainTeaser Dataset using few-shot prompts, including explanations for reasoning in the examples for the model to understand the unconventional reasoning tasks better, improving over the zero-shot LLM baseline results.

Building upon the areas of Multimodality and using LLMs for reasoning, we propose the task of Knowledge-aware Question-Answering over Semi-structured Multi-modal tables and experiment with SOTA LLM and MLLM for solving the task. We create the MultimodalTabQA dataset, which consists of 35,111 questions over 16,941 tables recast from three existing tabular question-answering datasets. The dataset involves complex questions requiring handling multiple images as input, performing knowledge-aware entity disambiguation, understanding the semi-structured information represented and understanding the entities in the context of the table. We experiment with three different approaches to answering these questions and demonstrate the capabilities of SOTA LLMs on such new tasks.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Preamble . . . 1
1.2 Background . . . 2
1.2.1 Multimodal Image-Text Processing . . . 2
1.2.2 Large Language Models . . . 3
1.3 Scope of the thesis . . . 4
1.3.1 Multimodal Emotion-Cause Pair Extraction . . . 4
1.3.2 Unconventional Question Answering using LLMs . . . 5
1.3.3 Knowlege-aware reasoning over multimodal semi-structured tables . . . 5
1.4 Research Questions Addressed . . . 5
1.5 Thesis Layout . . . 6

2 Multimodal Emotion-Cause Pair Extraction . . . 8
2.1 Introduction . . . 8
2.2 Background . . . 9
2.2.1 Dataset . . . 9
2.2.2 Related Work . . . 10
2.3 Task Definition . . . 11
2.4 System Architecture . . . 12
2.4.1 Baseline I: Utterance labeling . . . 12
2.4.2 Baseline II: BiLSTM Architecture . . . 15
2.4.3 Baseline III: BiLSTM-CRF Architecture . . . 17
2.5 Experimental Setup . . . 18
2.6 Results and Analysis . . . 19
2.7 Conclusion . . . 21

3 Prompting GPT-3.5 for Unconventional Reasoning . . . 22
3.1 Introduction . . . 22
3.2 Background . . . 23
3.2.1 Dataset . . . 23
3.2.2 Related Works . . . 24
3.3 System Overview . . . 24
3.3.1 GPT-3.5 . . . 24
3.3.2 Few-shot prompting . . . 24
3.3.3 Experimental Setup . . . 25

3.4 Results and Analysis 25
3.5 Conclusion 28

4 Reasoning over Multi-modal Semi-structured Tables 30
4.1 Introduction 30
4.2 Related Work 31
4.2.1 Tables with additional modality 31
4.2.2 Knowledge-Based Visual Question Answering 31
4.2.3 Multimodal Entity Linking 32
4.2.4 VQA involving multiple images 32
4.3 Potential Challenges/Complexities in the Problem 32
4.3.1 Entity Disambiguation Problem 32
4.3.2 Logical Reasoning Questions 34
4.4 Problem Definition 37
4.5 Conclusion 37

5 MultiModalTabQA Dataset 39
5.1 Introduction 39
5.2 Datasets Recast 40
5.2.1 WikiSQL Dataset 40
5.2.2 WikiTableQuestions Dataset 41
5.2.3 FeTaQA 41
5.3 Recasting the Datasets 41
5.3.1 Step 1: Getting raw HTML corresponding to the table 42
5.3.2 Step 2: Getting WikiData entry corresponding to Wikipedia Link(s) 44
5.3.3 Step 3: Getting image from the Wikidata ID 44
5.3.4 Step 4: Filtering out the entities to be replaced 45
5.3.5 Step 5: Replacing the text with images 45
5.3.6 Step 6: Creating Explicit and Implicit Questions 45
5.3.7 Step 7: Pruning Table Images 46
5.4 Dataset 47
5.5 Dataset Validation 48

6 Using MLLMs for Inference and Reasoning on MultiModalTabQA 51
6.1 Introduction 51
6.2 Metrics Used for Evaluation 52
6.2.1 Short-Answer Data Sources 52
6.2.2 Long-Answer Data Sources 52
6.3 Baseline I: Partial Input 53
6.4 Baseline II: Image Captioning 54
6.5 Baseline III: Multimodal Table 56
6.6 Conclusion 57

7 Conclusions 59
7.1 Future Work 59

Bibliography 62

List of Figures

Figure Page

2.1 An example for the task (taken from [1]) 9
2.2 Emotion frequency in the dataset 10
2.3 Model Architecture 13

4.1 Visually similar ships requiring more context for disambiguation 33
4.2 Entities exhibiting meronymy leading to ambiguity 34
4.3 Player images needing more context for disambiguation 35
4.4 Wikipedia table for political parties in India 36

5.1 Example from WikiSQL dataset 40
5.2 Distribution of Question Types in WikiSQL 40
5.3 Example questions from the WikiTableQuestions Dataset 42
5.4 Distribution of topics in FeTaQA 43
5.5 Example questions in FeTaQA dataset 43
5.6 Example from the dataset 49

List of Tables

Table Page

2.1 Results for baselines on the ECAC dataset 19

3.1 Results of zero-shot and few-shot prompting on GPT-3.5 for the two BRAINTEASER subtasks. Ori = Original, Sem = Semantic, Con = Context. 28

5.1 Dataset Statistics 47
5.2 Image Statistics in the dataset 47
5.3 Question-Type in the dataset 48
5.4 No. of Multi/Single Multimodal column Reasoning Questions 48
5.5 Categories of the images in the dataset 50
5.6 Dataset Validation Statistics 50

6.1 Results of Text-only baseline on WikiSQL 53
6.2 Results of Text-only baseline on WikiTableQuestions 53
6.3 Results of Text-only baseline on FeTaQA 53
6.4 Entity Prediction Results for Captioning Baseline 55
6.5 Results of Captioning baseline on WikiSQL 55
6.6 Results of Captioning baseline on WikiTableQuestions 55
6.7 Results of Captioning baseline on FeTaQA 55
6.8 Results of Multimodal Table baseline on WikiSQL 56
6.9 Results of Multimodal Table baseline on WikiTableQuestions 56
6.10 Results of Multimodal Table baseline on FeTaQA 57

Chapter 1

Introduction

1.1 Preamble

In the quest to develop Artificial Intelligence (AI) systems that mimic human intelligence, the ability to understand and process information from multiple modalities is paramount. Humans, since their birth, have the ability to process multiple senses simultaneously to perceive the world around us, including vision, audition, olfaction, gustation, and taction. It is the combination of these senses that enables us to perform our day-to-day tasks as intelligent beings. Without any one of these senses, our understanding and perception of the world become significantly impaired. Similarly, for AI systems to achieve human-like intelligence, they must possess the capability to interpret and reason across different modalities.

In today's information age, the prevalence of multimodal data is ubiquitous. Social media platforms are teeming with posts that combine textual content with accompanying images or videos, while scientific datasets often contain textual descriptions alongside graphical representations. Further, various entertainment videos come with audio transcripts, and different knowledge sources have representative images for the entities they describe. This multimodal nature of data further necessitates AI systems that can effectively process and extract meaningful insights from heterogeneous sources.

To the end of creating intelligent models that can handle a large variety of reasoning/inference tasks, Large Language Models (LLMs) have emerged as powerful tools, achieving state-of-the-art performance for a wide range of natural language processing tasks [2], thanks to their ability to learn complex patterns and relationships from vast amounts of text data [3]. These models have demonstrated remarkable performance in tasks such as text generation, sentiment analysis, and machine translation. Furthermore, the advent of Multimodal LLMs (MLLMs) has extended their capabilities to incorporate information from diverse modalities, opening up new avenues for research in multimodal AI [4].

The integration of multimodal capabilities into LLMs enables these models to leverage not only textual information but also visual and auditory cues present in multimodal data. By harnessing the synergies between different modalities, multimodal LLMs offer a more comprehensive understanding of the underlying semantics and context within the data. This holistic approach to information processing mirrors human cognition's multimodal nature, bringing AI systems closer to achieving human-like intelligence.

This thesis is a step towards Multimodal Reasoning and Inference involving text and images, leveraging the capabilities of LLMs and Multimodal LLMs to perform emotion detection and reasoning/inference on Multimodal Data. We begin our probing towards Multimodal Semantic Understanding through the simplest task in traditional NLP – Emotion Detection. We experiment with the data for the task of Multimodal Emotion-Cause Pair Extraction in Conversation [5] on the *Emotion Cause in Friends (ECF)* dataset [1], proposing various baselines involving different encoders for text, audio and images and modelling it as both a sequence-level and an utterance-level task.

Next, we explore different prompting strategies to enhance the reasoning capabilities of large language models for unconventional reasoning tasks. We describe the experimental methodology, results, and insights gained from the experiments. This allows us to better understand how we can prompt LLMs for reasoning and inference, albeit only on text data.

Building upon these two works, we propose the problem of Knowledge-aware Reasoning and Inference over Multimodal Tables. We define the problem and create a dataset for the problem by recasting three different datasets for the task. We also make use of three different baselines involving Gemini-Pro-1.0 [6], a family of SOTA LLM and MLLM, to benchmark the performance of existing models on the new problem.

1.2 Background

1.2.1 Multimodal Image-Text Processing

Text and image processing have long been considered challenging areas in artificial intelligence, where achieving even a fraction of human performance seemed far-fetched.

One of the key breakthroughs in Natural Language Processing (NLP) was the development of the Transformer architecture [7]. The Transformer model relies on self-attention mechanisms to weigh the importance of different words in a sentence, allowing it to capture long-range dependencies and improve performance on tasks such as machine translation, text summarization, and question answering.

The Transformer architecture has been particularly successful in improving semantic understanding tasks, with models like BERT (Bidirectional Encoder Representations from Transformers) [8] achieving state-of-the-art results on a wide variety of NLP tasks. BERT is pre-trained on a large corpus of text data and fine-tuned on specific tasks, allowing it to generalize well to new tasks and datasets.

In the field of computer vision, the Transformer architecture has also had a significant impact. Models like the Vision Transformer (ViT) [9] have shown impressive performance on tasks such as image segmentation, object detection, and action recognition. ViT treats images as sequences of patches, which are then processed by a transformer model to extract features and make predictions. This approach has proven to be highly effective, outperforming traditional convolutional neural networks (CNNs) [10] on certain tasks.

The advancements in text and image processing have led to a growing interest in multimodal research, where models are trained to understand and generate both text and images. One of the earliest tasks in the multimodal domain is optical character recognition (OCR), which combines computer vision techniques with textual information. The MNIST dataset [11], which consists of handwritten digits, is an example of a dataset that involves a multimodal problem, as it requires both image processing and understanding of the textual information (the digits).

Even in the pre-transformer era, researchers were exploring Visual Question Answering [12], making use of fusion-based methods for various tasks. Over time, a multitude of image-text tasks emerged, some of the prominent ones outlined below: [4]:

1. **Image Captioning:** Generating a textual description of images.

2. **Visual Question Answering:** Answering questions based on the content of an image.

3. **Image Retrieval:** Finding images relevant to a text query, making use of joint image-text representations.

4. **Image-Text Sentiment Analysis:** Understanding the emotional tone conveyed through both image and text.

5. **Multimodal Summarization:** Generating summaries that incorporate both visual and textual information.

Out of these popular tasks, we primarily make explorations related to Image-Text Sentiment Analysis [13] and Visual Question Answering [14]. To some extent, we also involve Image Retrieval in one of the problems tackled in this thesis.

1.2.2 Large Language Models

While LSTMs [15] and ELMo [16] previously held the status of state-of-the-art architectures in NLP, the breakthrough of the Transformer architecture in 2017 [7] marked a significant advancement. Unlike LSTMs and RNNs, which struggled with handling long-term dependencies and longer sequence lengths, the Transformer architecture introduced self-attention mechanisms and achieved parallelization, greatly enhancing its efficiency in processing long sequences. This architecture led to the creation of the first generation of LLMs, including BERT [17], GPT [18], among many other models. Soon, there was a rapid increase in the number of parameters and training data size, leading to models with hundreds of billions of parameters, which pushed the boundaries of what was previously achievable in natural language processing.

Further, due to being trained on large-scale data, these models could perform even zero-shot tasks and, given some few-shot examples, could even generalize to entirely new tasks. Such emergent abilities of LLMs, including in-context learning [19], instruction following, and Chain of Thought (CoT) reasoning [20] have brought significant attention to them. Several research areas make use of LLMs as agents for performing various human-like tasks through few-shot prompts and CoT reasoning. Current state-of-the-art LLMs include GPT-3.5 [21], GPT-4 [2], LLaMa 3 [22], Gemini [6] among various other models.

However, until recently, the capabilities of these models were limited to text. Inspired by the success of pre-trained NLP models, multimodal Transformers were developed that could process cross-modality inputs like text, image, audio, and point cloud. Some of these Vision-Language Models include CLIP [23], Visual BERT [24], BLIP [25], Flamingo [26], etc., which were large-parameter pre-trained models trained on large-scale cross-modal datasets comprising images and text.

Recently, these models have further scaled to instruction-tuned and conversational models, leading to the creation of MLLMs, which leverage the powers of LLMs like GPT-3.5, FLAN [27], etc. and extend them to tasks across multiple modalities. Some of the recent MLLMs include GPT-4V [2], Gemini-Vision [6], QwenVLM [28], CogAgent [29], etc.

In our work, we involve experiments with GPT-3.5 and the Gemini-1.0-Pro family of models due to accessibility and resource constraints with other LLMs/MLLMs.

1.3 Scope of the thesis

This section briefly describes the three tasks mentioned earlier that are tackled as a part of this dissertation. We then outline the specific directions explored in this thesis and our contributions to these areas.

1.3.1 Multimodal Emotion-Cause Pair Extraction

Emotion analysis [30] in the field of Natural Language Processing (NLP) has undergone significant development, transitioning from its initial focus on discerning emotions in news articles to its current emphasis on recognizing emotions within conversational contexts [31]. Recently, there has been a growing interest in analysing emotional causes, marking a notable advancement in this domain [32].

The Multimodal Emotion Cause Pair Extraction (MC-ECPE) task [5] is a step in this direction, which involves multimodal utterances in a conversation (audio, video and text) and not only requires identifying the emotions expressed within these utterances but also the utterances which are the cause for those emotions. We use the Emotion-Cause-in-Friends dataset [1], derived from the popular television series Friends, which provides annotated transcripts and video clips.

We experiment with different encoders for the individual modalities (text, audio, video) and various architectures (BiLSTM, BiLSTM+CRF, MLP) to model the task.

1.3.2 Unconventional Question Answering using LLMs

This thesis explores the task of solving brain teasers that require lateral thinking, known as the BRAINTEASER task [33]. Lateral thinking involves solving problems through an indirect and creative approach, often diverging from traditional logical reasoning. The thesis focuses on leveraging Large Language Models (LLMs) for this task, specifically through few-shot prompting, which allows the model to perform well even with limited training examples. The BRAINTEASER [34] dataset contains puzzles that challenge conventional thinking and require unique perspectives to solve. The dataset includes Sentence Puzzles, which involve unconventional interpretations of sentences, and Word Puzzles, which require reimagining the meaning of words. This task aims to bridge the gap between vertical (logical) and lateral (creative) thinking in NLP models, offering a new perspective on problem-solving capabilities.

1.3.3 Knowlege-aware reasoning over multimodal semi-structured tables

This thesis proposes and explores the novel problem of Question Answering over Multimodal Tables, a task that involves reasoning and inference over tables that contain both textual and visual information. While Natural Language Processing (NLP) research has extensively studied reasoning over text-only tables [35], the inclusion of images in real-world tables poses new challenges that have not been adequately addressed. This study builds upon previous work on Multimodal Emotion-Cause Pair Extraction and Prompting Large Language Models (LLMs) for Inference and Reasoning to propose the problem of exploring Multimodality in Tables using Multimodal LLMs.

The task involves answering questions over tables from Wikipedia, where certain entities are represented using their images alongside textual information. This requires not only understanding the textual content of the table but also identifying and linking the images to their corresponding entities. Additionally, answering questions may involve reasoning over multiple images, understanding the entity in the context of the table, and utilizing real-world factual knowledge apart from complex logical/numerical reasoning.

The motivation behind this task stems from the need to enhance NLP models' ability to reason over real-world data, which often includes multimodal information. By addressing this task, we aim to improve the understanding and utilization of multimodal data in NLP models, particularly focusing on handling semi-structured table information, entity linking, reasoning over multiple images, and leveraging real-world knowledge.

1.4 Research Questions Addressed

This thesis explores the following research questions:

RQ1 *To develop methodologies for performing emotion-cause detection in Multimodal Conversations and compare methodologies that model it on an utterance-level and sequence-level.* Multimodal Conversations can be modelled at an utterance level, where each spoken dialogue is considered independent of the surrounding contextual utterances. On the other hand, we can also take the utterances and even emotions in the surrounding dialogues into account while modelling the Emotion Prediction task. Thus, we explore whether the latter leads to any benefit, using Bi-LSTMs and CRFs for the exploration. We also experiment with different SOTA text, image and audio encoders to determine the best-performing encoder combination for the task.

RQ2 *To explore prompting strategies for LLMs such that they can perform unconventional reasoning.* While LLMs have become very popular for various inference and reasoning tasks, they still find it challenging to perform lateral reasoning, which involves puzzles needing *out of the box thinking*. We prompt GPT-3.5 to reason over the questions through constructed prompts that utilize few-shot examples, detailed task explanations and explanations of the reasoning behind the few-shot answers. We also compare results with different number of provided few-shot examples to understand the effectiveness of the approach.

RQ3 *Whether existing models can parse the semi-structured information containing both images and texts.* We propose the problem of reasoning over semi-structured multimodal tables. One of the most crucial aspects of the problem is that the model needs to parse the semi-structured nature of the table and perform reasoning across rows and columns to answer complex questions.

RQ4 *Whether models can disambiguate entities from their images in context of the table.* Our proposed task of reasoning and inference over tables comprises Wikipedia entity images. In order to perform the task of reasoning and question answering, disambiguating these entities becomes essential. While these entities can be challenging to disambiguate only through the image, the context of the table might greatly help, and we explore this through our problem.

RQ5 *How can models handle multiple images provided in the form of a table?* While models like Flamingo take arbitrary sequences of image and text as input, even they might find it extremely difficult to reason over multiple images simultaneously. On the other hand, models like BLIP-2, LLaVa, etc. process one image at a time. This makes the task of reasoning over multiple images in a semi-structured format extremely hard.

1.5 Thesis Layout

C1 This chapter introduces the field of Multimodality and LLMs, where we define the scope of the investigations and experiments conducted towards Multimodal Reasoning with LLMs. We enlist the specific research problems addressed in this thesis, the motivation behind them, and a quick summary of our approaches that will follow in later chapters.

C2 In this chapter, we introduce the problem of Multimodal Emotion-Cause Pair Detection and then discuss the background of the task and dataset that we use. We describe the different encoders we experiment with, including BERT, RoBERTa-Large, EmotionRoBERTa, MViTv2-small, WavLM and Wav2Vec2-Large. We also describe the different architectures we use, including an MLP-based model, a stacked BiLSTM-based model and a stacked BiLSTM+CRF-based model.

C3 In this chapter, we introduce the problem of Unconventional Question-Answering involving Lateral reasoning and discuss the motivation and background of the task and the BrainTeaser dataset that we use. We describe the experimental setup involving few-shot prompts for the task and the structure of the prompts we use.

C4 In this chapter, we introduce the new problem of Reasoning over multimodal Semi-structured Tables and discuss the work in different fields of Multimodality that the problem derives from. We provide motivation for the problem by examining the complexities involved in the task and give a formal definition for the task.

C5 In this chapter, we discuss the MultimodalTabQA dataset that we create for the problem of reasoning over Multimodal Tables. We discuss the datasets we recast to create MultimodalTabQA and describe our methodology for recasting these datasets. We also provide an analysis of the created dataset.

C6 In this chapter, we describe three approaches to using LLMs/MLLMs for reasoning on our created dataset. We use the publicly available Gemini-1.0 Pro family of models to evaluate these baselines across the test set of our dataset.

C7 We conclude our thesis with this chapter, summarizing our analysis, experiments and the insights gained. We also describe the directions in which this work can be further extended.

Chapter 2

Multimodal Emotion-Cause Pair Extraction

This chapter is adapted from the publication *"LastResort at SemEval-2024 Task 3: Exploring Multimodal Emotion Cause Pair Extraction as Sequence Labelling Task"* accepted at the 18th International Workshop on Semantic Evaluation (SemEval-2024). This work is a joint effort with: Akshett Rai Jindal (IIIT Hyderabad), Hardik Mittal (IIIT Hyderabad) and Prof. Manish Shrivastava (Prof, IIIT Hyderabad).

As the first step towards exploring Reasoning over Multimodal Information, we explore the most fundamental field of NLP in the context of Multimodal Information – Emotion Detection. In this chapter, we describe the task of Multimodal Emotion-Cause Pair extraction and the architecture that we use for performing the task, modelling it as an utterance-level as well as a sequence-level task.

2.1 Introduction

Emotion Analysis is one of the fundamental and earliest sub-fields of NLP that focus on identifying and categorising emotions expressed in text. Earlier, research in this domain focused on Emotion Detection in news articles and headlines [36, 37]. However, later, Emotion Recognition in Conversation gained popularity due to the widespread availability of public conversation data [31]. Recently, the task of emotion cause analysis has gained traction, which tries to identify the causes behind certain emotions [32]. This has widespread applications, such as building chatbots that can identify the user's emotions and even the cause behind the emotions to perform certain actions [38]. For instance, companies can locate causes behind dissatisfaction in customer interactions and take appropriate measures [39], AI-driven therapeutic insights can be gained using such models [40], social media content moderation can be better done [41], work management and team management by companies can be improved [42].

In this work, we analyse the problem of Multimodal Emotion Cause Pair Extraction [5], where given a set of utterances in a conversation, we must identify the following:

1. Emotion of every utterance (if any). These emotions can be one of Ekman's six basic emotions [43] – anger, disgust, happiness, sadness, fear, and surprise.

2. Cause of these emotions, which is considered as the utterance that explicitly expresses an event or argument that is highly linked to the corresponding emotion.

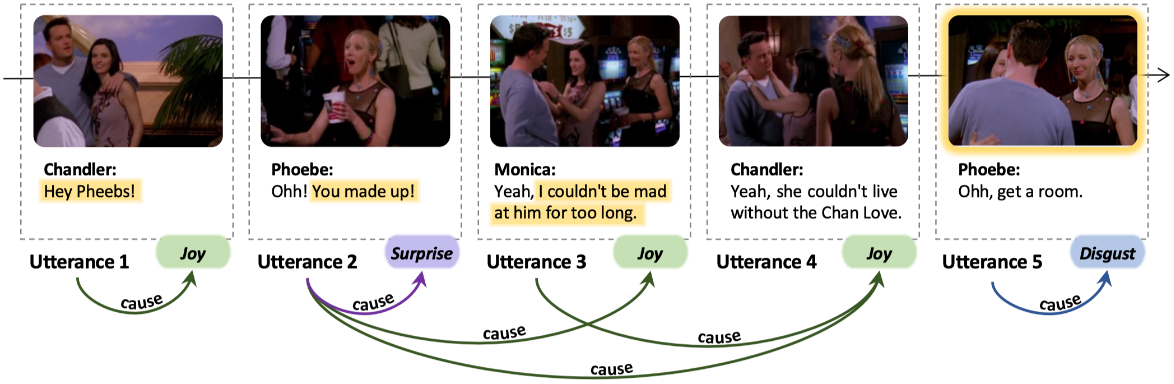


Figure 2.1: An example for the task (taken from [1])

Our proposed system tackles the task in a 3-step fashion – (a) First, we train a model to identify the emotions expressed in individual utterances in a conversation. (b) Next, we train a model to identify whether an utterance can be a cause of an emotion expressed in another/same utterance (candidate causes). (c) Finally, we train a model to pair emotion-utterances with their causes among the possible candidate causes. For both the (a) and (b) models we experiment with 3 basic architectures – (i) a simple Neural Network to determine the class of emotion (N-class classifier) and another Neural Network to identify whether the utterance is a candidate cause or not (binary classification). (ii) A BiLSTM [44] architecture that accounts for the surrounding context of the conversation while doing the N-class and binary-classification. (iii) A BiLSTM-CRF [45] architecture, which accounts for the surrounding emotions as well while doing the N-class classification. We also experiment with different encoders for the three modalities.

2.2 Background

2.2.1 Dataset

The dataset used for this problem is **Emotion-Cause-in-Friends** prepared by [1] specifically for this task. It has been prepared using conversations from the famous 1994 sitcom *Friends* as the source. This dataset contains 1,344 conversations made up of a total of 13,509 utterances, each conversation containing an average of 10 utterances. For each utterance, the dataset has an annotated transcript (covering text modality) and the corresponding video clip (covering visual and auditory modalities) from the show.

Each utterance is annotated with the emotion depicted by it, which is one of: `anger`, `disgust`, `fear`, `joy`, `neutral`, `sadness` and `surprise`. The dataset is highly skewed in terms of the frequency of different emotions in the dataset (see Fig. 2.2). Further, the emotion-causes pairs for all the non-`neutral` utterances are provided in the dataset in a separate list.

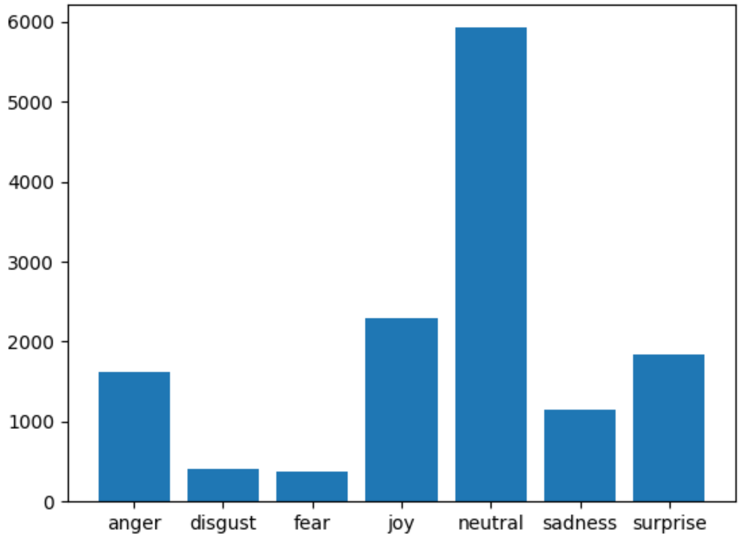


Figure 2.2: Emotion frequency in the dataset

Two such conversations can be found in Listing 1. In these samples, we can see the conversation ID (`conversation_ID`). It contains utterances under the key `conversation`. Each utterance has its ID (`utterance_ID`), its transcript (`text`), the speaker name (`speaker`), the emotion depicted and the corresponding video clip name (`video_name`) in dataset which corresponds to this utterance. We can also find the emotion-cause pairs under the key `emotion-cause_pairs`. For the 3rd utterance in (`conversation_ID`) 2, we can see that the emotion is `surprise` and it has two cause utterances - Utterance 1 and 3. We can identify the emotion for the third utterance to be of `surprise` because of the *?!* characters in the text, the surprised tone of the speaker's voice in the audio and also their expression in the video clearly showing surprise.

The task MC-ECPE expects the model to take a list of such conversations and predict the emotion and emotion-cause pairs labels.

2.2.2 Related Work

A lot of work has been done in the field of emotion analysis in textual settings. However, for many years, the main focus of researchers working in textual emotional analysis was only on emotion recognition, and no work was done to also identify the cause of that emotion until recent years. The research in the field of **emotion cause analysis** started with [36, 37], who worked on extracting potential causes given the emotions. However, these studies were still being conducted on texts such as news articles or micro-blogs. [46] extended this work to textual dialogues for the first time.

```
{ "conversation_ID": 37,
  "conversation": [ { "utterance_ID": 1,
      "text": "So , you and Angela , huh ?",
      "speaker": "Joey",
      "emotion": "neutral",
      "video_name": "dia37utt1.mp4" },
    { "utterance_ID": 2,
      "text": "Yep . Pretty much .",
      "speaker": "Bob",
      "emotion": "joy",
      "video_name": "dia37utt2.mp4" } ],
  "emotion-cause_pairs": [
    [ "2_joy", "1" ],
    [ "2_joy", "2" ]
  ] }
```

```
{   "conversation_ID": 2,
    "conversation": [ { "utterance_ID": 1,
            "text": "I do not want to be single , okay ? I
                just wanna be married again !",
            "speaker": "Ross",
            "emotion": "sadness",
            "video_name": "dia42utt3.mp4"},
        {   "utterance_ID": 2,
            "text": "And I just want a million dollars !",
            "speaker": "Chandler",
            "emotion": "neutral",
            "video_name": "dia42utt4.mp4"},
        {   "utterance_ID": 3,
            "text": "Rachel ? !",
            "speaker": "Monica",
            "emotion": "surprise",
            "video_name": "dia42utt3.mp4"} ],
    "emotion-cause_pairs": [
        [ "1_sadness", "1" ],
        [ "3_surprise", "3" ]
    ]}
```

Listing 1: Two sample conversations for the Multimodal Emotion-Cause Pairs task

Soon, work began on extracting not only the emotion but also the cause of that extracted emotion. People employed mainly two approaches for emotion cause analysis:

1. Extracting the potential causes given an emotion [47, 48, 49]

2. Extracting the emotion-cause pairs jointly [50, 51, 52].

[46] was the first to introduce the task of extracting emotion-cause in conversations, but their focus was also only on the textual dialogues. However, in our natural way of conversation, we rely on things like facial expressions and voice intonations to determine the emotion of the speaker. We also rely on auditory and visual scenes to determine the cause of the speaker's emotions. Hence, it is clear that identifying speaker's emotions involves not just their utterances, but also their visual expressions and intonations. Thus, the task of identifying Emotion-Cause Pairs is a multimodal task requiring visual and audio modalities apart from just utterance information. [53, 54, 55, 56] worked in the field of multimodal emotion analysis in conversations, but they did not consider the emotion causes.

The task of MC-ECPE was first worked on by [5].

2.3 Task Definition

As described above, the goal of the task is to jointly extract the emotions and corresponding causes in pairs. Say, we are given a conversation $D = [U_1, ..., U_i, ..., U_{|D|}]$, such that $U_i = [t_i, a_i, v_i]$ where

t_i represents the text modality, a_i represents the audio modality and v_i represents the video modality. Then, the goal of the task is to extract a set of emotion-cause pairs as $P = \{..., (U^e, U^c), ...\}$ where U^e denotes an emotion utterance while U^c denotes a cause utterance.

An example utterance from the dataset is shown in Fig. 2.1

2.4 System Architecture

In order to perform the task of detecting all emotion-cause pairs in conversation, we perform the task step-wise involving three steps. Given a conversation D as described in Section 2.2, the three tasks are:

1. **Candidate Emotion Identification:** This step involves finding a set of Candidate Emotion Utterances $\hat{U}^e$, which are utterances that exhibit an emotion (i.e. don't belong to the Neutral class of emotions). $\hat{U}^e_i$ comprises an utterance U^e_j along with the emotion E_j predicted for it out of the 6 emotion classes (N-class classification step).

2. **Candidate Cause Identification:** This step involves finding a set of Candidate Emotion Cause Utterances $\hat{U}^c$, which are utterances that exhibit a reason for a particular emotion and so can be paired with a different Candidate Emotion $\hat{U}^e$ identified in Step 1. $\hat{U}^c_i$ comprises an utterance U^c_j if it is predicted to possibly be a cause for another Candidate Emotion (binary classification step).

3. **Emotion-Cause pairing:** This stop comprises forming all pairs $\{..., (\hat{U}^e_i, \hat{U}^c_j), ...\}$ such that $\hat{U}^c_j$ is a cause-utterance behind the emotion-utterance $\hat{U}^e_i$. This step involves forming all possible pairs of Candidate Emotion Utterances (identified in Step 1) and Candidate Cause Utterances (identified in Step 2). Thus, this is a binary classification step for all the pairs, classifying the pair $(\hat{U}^e_i, \hat{U}^c_j)$ as a valid or invalid pair.

We propose three baselines, which are illustrated in Fig. 2.3 and are detailed below:

2.4.1 Baseline I: Utterance labeling

Our baseline model treats the problem as a simple **utterance labelling task**. We use pre-trained text, audio, and image encoders to encode the individual modalities and use these to train three models that can identify the emotions in the utterances, the candidate cause utterances, and finally identify valid emotion and cause utterance(s) pairs.

- **Text Encoding:** For encoding the transcription of each utterance, we use pre-trained BERT [57] embeddings as the baseline embeddings. BERT's effectiveness stems from its utilisation of the transformer architecture, which incorporates self-attention mechanisms. This architecture enables BERT to capture contextual information bidirectionally, facilitating a nuanced understanding of language semantics. Consequently, BERT embeddings serve as a robust foundation for our analysis, enabling comprehensive representation of utterances in our NLP tasks.

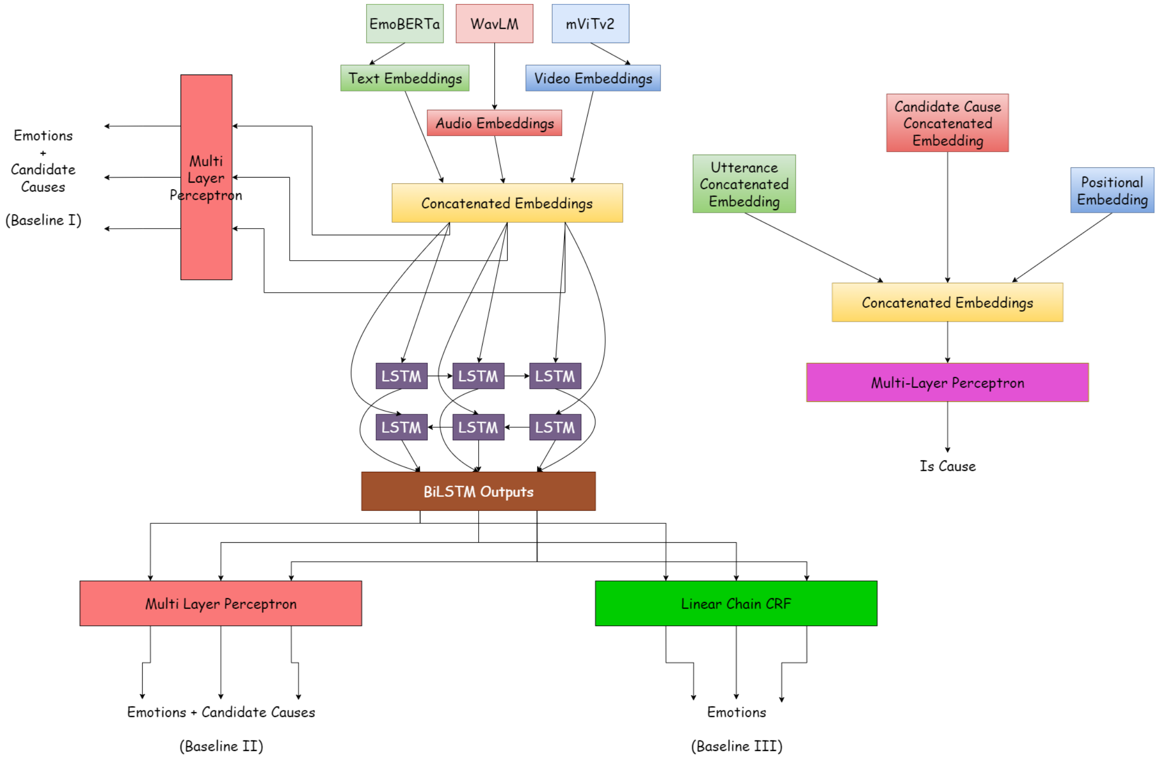


Figure 2.3: Model Architecture

Additionally, we finetune DeBERTa-Base [58] on the training data for our experiments. DeBERTa distinguishes itself through the incorporation of a disentangled attention mechanism, along with an enhanced masked encoder, which collectively augments its performance across diverse NLP tasks. Unlike BERT, DeBERTa's disentangled attention mechanism enables attention heads to focus independently on specific linguistic properties, enhancing its ability to capture intricate language structures. Additionally, the enhanced masked encoder further refines DeBERTa's contextual understanding, culminating in superior performance compared to BERT across a spectrum of NLP benchmarks.

Finally, we also tried RoBERTa-Large and [59] pre-trained EmotionRoBERTa-Base[1] which is publicly available RoBERTa-base model finetuned on the Go Emotions dataset [60]. RoBERTa, an improvement over BERT, implements dynamic masking during pre-training, increases training data and steps, and removes the next sentence prediction (NSP) task, resulting in enhanced performance across various NLP tasks. EmotionRoBERTa-Base offers specialised pre-training for emotion analysis tasks, thereby enriching our comparative analysis with a model tailored for emotion detection. For every text encoder, we perform mean-pooling of the word embeddings to get the textual representation of the utterance.

[1] https://huggingface.co/SamLowe/roberta-base-go_emotions

- **Video Encodings:** For encoding the videos, we sampled 16 equally spaced frames from the video and mean-pooled the embeddings for the 16 frames. For encoding these 16 images, we used MViTv2-small [61] encoder, which achieves state-of-the-art performance on the Kinetics video detection task [62], which makes it an obvious choice for recognising activities happening in the conversations relevant for emotion/cause detection. Traditional vision transformers, limited to processing images at a single resolution, might overlook crucial details for emotion recognition.

  MViT addresses this limitation through its powerful multi-stage architecture. Each stage meticulously analyses the image at a dedicated resolution and channel capacity. High-resolution stages, equipped with lower channel capacity, excel at capturing fine-grained details like facial expressions, which are vital indicators of specific emotions (e.g., furrowed brows for anger, upturned lips for happiness). Conversely, lower-resolution stages with higher channel capacity prioritise capturing broader emotional cues conveyed through posture and body language within the image.

  This multi-scale approach aligns perfectly with a scientific understanding of emotion recognition, where both subtle facial features and body language are paramount in conveying emotional states. By effectively capturing information across these distinct scales through MViT, we aimed to encode a richer representation of the image data within each video frame. This comprehensive encoding, encompassing both high-resolution details and lower-resolution contextual cues, can significantly contribute to the overall task of accurately detecting emotions and their causes within the multimodal framework.

- **Audio Encodings:** We used WavLM [63] for generating audio embeddings, which is trained on extensive audio data using masked speech representation and denoising in pre-training, making it suitable for various downstream speech tasks. WavLM excels in this role due to its multifaceted learning approach during pre-training on a massive speech dataset. It goes beyond just understanding the spoken content, also learning to identify the speaker and other characteristics embedded in the audio, like emotional tone. This rich embedding tackles various speech-processing tasks effectively.

  We also try Wav2Vec2-Large [64], which is trained by masking speech input in latent space and solving a contrastive task defined over a quantisation of the latent representations which are jointly learned. Unlike traditional methods requiring vast amounts of labelled data, Wav2Vec2-Large excels at self-supervised learning. It leverages unlabeled speech data by masking speech input in latent space and solving a contrastive task defined over a quantisation of the latent representations, which are jointly learned. This allows the model to identify patterns and relationships within the audio itself, building a strong understanding of speech even without explicit labels. Consequently, Wav2Vec2-Large requires significantly less labelled training data compared to conventional methods, making it efficient and adaptable to various speech variations. This efficiency, combined with its ability to achieve state-of-the-art performance in speech recognition tasks, even with less data, makes Wav2Vec2-Large a compelling choice for our experiments.

The model architecture is a combination of **three steps**, each of which is described below:

Step 1 – Emotion Classification

First, we concatenate the text, audio and video embeddings from the respective encoders and pass these concatenated embeddings into a dense layer, on which a Softmax function is applied to get the probability distribution over 7 classes (6 emotions and one neutral class). Due to a skewed distribution of the emotion labels in the dataset, we make use of **weighted Cross Entropy loss** to train the model, where the weights are taken as inverse of the frequency of the labels in the training dataset. It assigns weights to different classes based on their frequency in the training data. Classes with fewer examples (minority classes) are assigned higher weights, making their contribution to the overall loss function more significant. This forces the model to pay greater attention to these classes during training, improving its ability to learn and predict them accurately.

Step 2 – Candidate Cause Identification

For identifying the candidate cause, we similarly pass concatenated embeddings through a dense layer with a Sigmoid function, which predicts the probability of whether the utterance is a candidate cause or not. Binary Cross Entropy Loss is used to train the model.

Step 3 – Emotion-Cause pairing

For pairing the emotion utterances with the candidate causes, we concatenate the representations for the emotion utterance and the cause utterance, with a distance embedding. This distance embedding is generated by giving positional embedding to each utterance, sampled from a Normal Distribution. This representation is passed through a dense layer with a Sigmoid function, which learns to predict the probability of the emotion-cause utterance pair being a valid emotion-cause pair or not for the given conversation, trained using Binary Cross Entropy Loss.

2.4.2 Baseline II: BiLSTM Architecture

The BiLSTM architecture is inspired by the work in [5]. Baseline I architecture ignores surrounding utterances when classifying emotions and their causes in an utterance. However, emotions in an utterance are highly contextual (e.g., sarcasm, irony, subtle tone shifts) and require understanding the broader conversation and speaker relationships. Therefore, in Baseline II we consider the surrounding utterances while performing emotion and cause classification for a particular utterance.

Bidirectional Long Short-Term Memory networks (BiLSTMs) are a well-established approach for dealing with sequential data, where the order of information holds significance. Their strength lies in capturing long-term dependencies within sequences, making them ideal for tasks that require understanding context. BiLSTMs achieve this by processing data in both forward and backward directions.

This allows them to analyse how past and future elements in a sequence relate to the present element, providing a more comprehensive understanding of the context.

In the context of emotion and cause classification, this improved context awareness is crucial. The emotional state and cause behind an utterance can be heavily influenced by the flow of conversation. By considering both preceding and following utterances, BiLSTMs can more accurately identify the emotions expressed and the factors triggering them.

However, a single layer of BiLSTM might not always be sufficient to capture complex contextual relationships, especially in lengthy conversations. Therefore, we use stacked BiLSTMs, where each layer builds upon the previous one to extract progressively more intricate features and relationships within the conversation. This leads to a richer representation and improved learning of long-term dependencies, which is crucial for understanding how distant utterances influence emotions. Additionally, stacked BiLSTMs achieve gradual abstraction, with lower layers capturing low-level details and higher layers learning more abstract emotional flows within the conversation. This combination of enhanced representational power, improved dependency learning, and hierarchical abstraction allows stacked BiLSTMs to outperform single BiLSTM layers in emotion and cause classification tasks.

Thus, the stacked-BiLSTM architecture models the problem as a **Sequence Labeling task**. We use the best encoders in the Baseline I architecture to generate the embeddings in this architecture.

Step 1 – Emotion Classification

Similar to the Baseline Model, we concatenate the embeddings of the three modalities and pass them to a stacked BiLSTM. On top of the BiLSTM outputs, we apply a 7-class classifier to obtain the emotion category distribution. Similar to Baseline I, weighted cross-entropy loss is used.

Step 2 – Candidate Cause Identification

For Candidate Cause prediction, similarly, the concatenated embeddings are passed through a BiLSTM, on top of which a binary classifier is applied.

Step 3 – Emotion-Cause Pairing

The Emotion-Cause pairing model remains the same in this architecture as the Baseline I model.

In this architecture, BiLSTM provides the advantages of bidirectional and longer contexts, which should help better understand the emotions present in utterances. This is because, in a conversation, it is possible that the emotions are not just dependent on the current utterance but on surrounding multimodal utterances as well.

2.4.3 Baseline III: BiLSTM-CRF Architecture

In the BiLSTM model, each classification decision was conditionally independent. Thus, it predicted labels for each element in the sequence without considering the labels of its neighbours. This limitation can be overcome by employing Linear-chain Conditional Random Fields (CRFs). CRFs are models specifically designed for structured data where one output influences its neighbouring outputs. They excel at modelling these relationships by learning transition probabilities between labels. This is particularly beneficial for sequence labelling tasks like emotion prediction in text, where the emotion of an utterance is often influenced by the emotions in the preceding ones. For example, a happy utterance is more likely to be followed by another happy utterance. By incorporating a CRF layer on top of the BiLSTM, we can leverage the strengths of both models: the BiLSTM's ability to capture sequential features and the CRF's capability to model label dependencies.

Linear-chain CRFs have been extensively used with BiLSTMs for sequence labelling [65]. This could be useful for the emotion-cause prediction task as well because the emotion of one utterance is generally influenced by the emotions in its previous utterances.

Step 1 – Emotion Classification

For this architecture, we add a CRF layer on top of the BiLSTM layers, and make use of the CRF-loss to train the model instead of Cross-Entropy loss as in the previous architectures. This loss models the transitions between the labels in the architecture, modelling the task as a more complex sequence labelling task. Unlike Cross-Entropy loss, which focuses on the likelihood of individual labels, CRF loss considers the entire sequence of labels and their dependencies. Thus, while the BiLSTM layer learns more about the language and emotions expressed through the language, the CRF layer tries to learn about the relations between the emotions.

Step 2 – Candidate Cause Identification

For Candidate Cause prediction, the architecture remains the same as in Baseline II. This is because the transitions between cause labels (being the cause of an emotion in an utterance or not) do not make intuitive sense, and using BiLSTMs to capture surrounding context from other utterances is what seems more appropriate.

Step 3 – Emotion-Cause Pairing

The Emotion-Cause pairing model remains the same in this architecture as the Baseline I & II models.

2.5 Experimental Setup

We perform a random shuffle and use a 90-10% split for the train-validation split. The test set was provided by the authors, but its gold labels have not been made public.

The experiments involving Baseline II and III use *EmotionRoBERTa + WavLM + MViTv2* configuration. All the experiments involve applying a dropout of 0.3 on the audio, visual and textual embeddings before they are passed on to the main architectures. The BiLSTM for emotion detection consists of 4 stacked layers, while the one for candidate cause identification contains three stacked layers. The dropout between the stacked layers of the BiLSTM is kept at 0.3 as well. We use AdamW optimiser for all three models and use a linear learning rate scheduler with warmup for training the models. The Emotion Classification model is trained for 60 epochs, the Candidate Cause Identification model is trained for 40 epochs, and the Emotion-Cause Pairing Model is trained for 40 epochs as well.

In order to train the Emotion-Cause pairing model, we create positive and negative pairs during training. However, while the number of positive pairs is of the order *N*, the number of negative pairs comes to the order of N^2, and thus we perform a random sampling of the negative pairs to keep the positive and negative samples in the ratio 1:5. This helps us to maintain balance between the positive and negative classes.

Evaluation Metrics

We evaluate the three steps separately as well, apart from benchmarking the performance for the final Emotion-Cause pairs:

Emotion Identification: We use Weighted Precision, Recall and F1-score for the distribution between the 7 classes (6 emotions and neutral class). These metrics are formally defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP denotes true positives, FP denotes false positives, and FN denotes false negatives.

Candidate Cause Identification: Similarly, we utilize Weighted Precision, Recall, and F1-score to evaluate predictions between the binary classes: *is_candidate_cause* and *not_candidate_cause*. The formal definitions of these metrics remain consistent with those mentioned above.

Emotion-Cause Pairing: In this stage, we generate positive and negative pairs and assess the classification between the two classes using Weighted Precision, Recall, and F1-score. The formal definitions of these metrics remain consistent with those mentioned above.

Emotion-Cause Pairs: The official metrics used for the final evaluation of this task are Weighted F1-score and Macro F1-score. These scores are calculated as follows:

$$\text{Weighted F1-score} = \frac{\sum_{i=1}^{n} w_i \times F1_i}{\sum_{i=1}^{n} w_i}, \quad \text{Macro F1-score} = \frac{\sum_{i=1}^{n} F1_i}{n}$$

where $F1_i$ denotes the F1-score for class i and w_i denotes the weight for class i.

2.6 Results and Analysis

Model Name	**Emotion Detection**			**Candidate Cause Detection**			**Emotion-Cause Pairing**			**Emotion-Cause Pairing (Eval.)**		**Leaderboard**	
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**wt. F1**	**Macro F1**	**wt. F1**	**Macro F1**
Baseline I													
BERT + WavLM + MViTv2	0.61	0.52	**0.55**	0.71	0.71	**0.71**	0.93	0.87	**0.89**	**0.26**	0.20	0.182	0.165
EmotionRoBERTa + WavLM + MViTv2	0.55	0.45	0.47	0.67	0.66	0.66	0.93	0.86	**0.89**	0.20	0.18	**0.187**	0.170
DeBERTa (finet.) + WavLM + MViTv2	0.44	0.36	0.38	0.60	0.60	0.60	0.92	0.85	0.87	0.10	0.10	0.094	0.094
RoBERTa-L + WavLM + MViTv2	0.59	0.47	0.49	0.66	0.65	0.66	0.93	0.86	0.88	0.21	0.19	0.180	0.165
EmotionRoBERTa + Wav2Vec2 + MViTv2	0.55	0.47	0.48	0.67	0.67	0.67	0.93	0.87	**0.89**	0.21	0.20	0.172	0.170
BiLSTM **(Baseline II)**	0.55	0.51	0.52	0.67	0.67	0.67	0.93	0.86	**0.89**	0.22	**0.21**	0.184	**0.179**
BiLSTM + CRF**(Baseline III)**	0.53	0.56	0.54	0.67	0.67	0.67	0.93	0.86	**0.89**	0.24	0.18	0.165	0.172

Table 2.1: Results for baselines on the ECAC dataset

The performance of the three Baselines can be seen in Table 2.1. During the Evaluation phase in our participation in SemEval task 2024, our best-ranked submission of Baseline II had Wt. F1 score of 0.1836 and Macro F1 score of 0.1759, ranking 8th on the leaderboard.

Baseline I

Among the encoders in Baseline I, *BERT + WavLM + MViTv2* configuration performs the best on the validation set, including the individual steps as well as the final emotion-cause pair predictions. However, on the leaderboard, *EmotionRoBERTa + WavLM + MViTv2* gives the best performance, although the difference in the leaderboard scores is marginal among the encoders. This observation might indicate that the test data is a bit different in nature from the training data.

The better performance of EmotionRoBERTa can be attributed to the fact that the model's weights have already been finetuned towards emotion-related tasks. Further, it seems that finetuning DeBERTa on the training data caused it to overfit, leading to worse performance than vanilla BERT/RoBERTa models. RoBERTa-L performed slightly worse than BERT and EmotionRoBERTa.

Finally, WavLM, being the newer architecture, as expected, performed better than Wav2Vec2. This is because WavLM is more robust than Wav2Vec2, and it is trained in a combination of supervised and self-supervised learning, making its performance much better.

Baseline II

In our experimental setup, we adopt the *EmotionRoBERTa + WavLM + MViTv2* configuration as the encoders for the Baseline II architecture. Surprisingly, despite expectations of potential performance improvements, the weighted F1 score on the leaderboard experiences a slight decrease, while the Macro F1 score shows a marginal increase. This discrepancy may be attributed to the unique characteristics of the dataset, wherein the average length of a conversation is minimal, typically around ten utterances. As a consequence, the contextual information available to the model within each utterance is severely limited. Consequently, the inclusion of additional context from previous utterances may not provide significant benefit and could potentially introduce noise into the model's predictions. This contextual mismatch could explain the observed results on the leaderboard.

Baseline III

In analysing the results, we note a substantial decrease in the weighted F1 score and a slight decline in the Macro F1 score compared to the Baseline I and II architectures. This observation aligns with the findings of Baseline II, suggesting that sequence labelling may not be the optimal approach for modelling the dataset. Moreover, given the limited number of utterances in conversations, it is probable that the transitions between labels required for Conditional Random Field (CRF) modelling are not adequately learned, resulting in diminished performance. This discrepancy underscores the challenges inherent in effectively capturing the nuanced transitions between labels within the dataset's context.

2.7 Conclusion

In conclusion, we observe that the utterance labelling systems perform as well as sequence labelling systems for this specific dataset. Further, we also see that encoders which are trained on other emotion-related tasks tend to perform better on similar emotion-related tasks. However, we also observed that all text encoders gave almost similar results, indicating that only fundamental language knowledge needs to be integrated into the LLMs for emotion-related tasks, to the extent that BERT, RoBERTa and DeBERTa all have similar performance.

We also clearly observe the improvements of WavLM over Wav2Vec2. This demonstrates that while Wav2vec2 excels at deciphering the spoken words, WavLM goes further. It can analyse the speaker's voice itself, capturing paralinguistic cues that reveal emotions. By understanding both what is said and how it's said, WavLM becomes a powerful tool for extracting emotion-cause pairs in conversations.

We also understand the potential benefits LSTMs and CRFs could provide, modelling the problem as a sequence labelling task rather than an utterance labelling task. However, we clearly observed the unique nature of our dataset, which instead made the context information a source of noise rather than helpful information. In future, it is possible to learn joint embeddings over the three modalities, which should provide better representations for each utterance [66]. Further, it can be experimented to utilise the speaker information for each utterance while creating utterance representations [67].

This task serves as an introduction to Multimodal tasks, primarily utilising classical methods like BiLSTMs and BiLSTM-CRFs while only using basic Transformer-based encoders. However, in the next chapter, we move towards using more modern generative LLMs for performing reasoning, and finally build upon the concepts of Multimodality from this chapter to introduce the task of Reasoning over Semi-structured Multimodal Tables using LLMs.

Through this work, we gain insights into various image, text, and audio encoders and explore combining embeddings from the three modalities to perform complex analysis tasks. This exploration provides the basis for further investigation in the Multimodality domain later in this thesis.

Chapter 3

Prompting GPT-3.5 for Unconventional Reasoning

This chapter is adapted from the publication *“DaVinci at SemEval-2024 Task 9: Few-shot prompting GPT-3.5 for Unconventional Reasoning”* accepted at the 18th International Workshop on Semantic Evaluation (SemEval-2024). This work is a joint effort with: Akshett Rai Jindal (IIIT Hyderabad), Hardik Mittal (IIIT Hyderabad) and Prof. Manish Shrivastava (Prof, IIIT Hyderabad).

While the previous chapter dealt with understanding Multimodal data, we made use of more traditional NLP techniques like LSTMs, CRFs and simple MLP on top of embeddings obtained from different encoders to perform the task. However, given the impressive performance of LLMs, especially of the large parameter models like GPT-3.5 and GPT-4 in solving various complex NLP tasks, we are highly interested in utilizing such models as a part of this thesis. In this chapter, we discuss using few-shot prompting on GPT-3.5 and defining prompt-structures to perform Question Answering and reasoning using LLMs, which forms the base for the work on prompting MLLMs for Multimodal Question Answering later in the thesis.

3.1 Introduction

The human brain consists of two hemispheres - left and right. Both of them are responsible for different kinds of thinking strategies. The left hemisphere is involved in vertical thinking, and the right hemisphere is involved in lateral thinking [68]. Vertical (linear, convergent, logical) thinking is a more sequential analytical process. In contrast, in Lateral (outside the box, divergent, creative) thinking, we look at the problem from a new point of view, ignoring the expected associations with items.

In the field of NLP, much research has been done around vertical thinking and significant progress has been made. The recent work around Large Language Models (LLMs) [57, 69] has achieved great performance in solving complex reasoning tasks [70, 71, 72]. This performance is consistent in both cases when no task examples have been provided to the model during inference (zero-shot) [73] and when the model is introduced with the task during inference time (few-shot) [74].

However, lateral thinking has been overlooked when training NLP models like LLMs. When creating datasets for various models, texts that involve lateral thinking are mostly considered noise and filtered out from the data because researchers want their models to perform better at traditional reasoning tasks and not get confused by lateral thinking.

The task BRAINTEASER [34, 33]) tries to bridge this gap that exists between vertical and lateral thinking for LLMs and other NLP models. They formulated a set of Multi-choice Question Answers containing puzzles that can be solved only using lateral thinking. The benchmark dataset contains two types of lateral thinking puzzles - Sentence Puzzles and Word Puzzles. This has been constructed by designing a data collection procedure that crawled relevant puzzles from many websites that were publicly available performing semi-automatic filtering of irrelevant questions.

3.2 Background

3.2.1 Dataset

The dataset being used in this task is BRAINTEASER [34]. It was prepared by scraping puzzles from various publicly available websites and then semi-automatically filtering them out. Then *semantic reconstruction* and *context reconstruction* techniques were used to create variants of each puzzle without affecting its out-of-the-box thinking style. This helped in preventing possible memorization by LLMs and the lack of consistency of the puzzles. Each puzzle comprises of a premise and a question based upon it with four options. The task is to select the correct answer from the four options.

The puzzles in this dataset can be divided into two categories:

- **Sentence Puzzles**: These are brain teasers where the puzzle-defying commonsense is centered on sentence snippets.

  For example, **Question**: *You are running so fast but you're not getting closer. Where are you?* **Answer**: *Treadmill.* **Explanation**: This is because while running on a treadmill, we stay put where we are. The key is understanding that running on a treadmill means you remain stationary despite the motion.

- **Word Puzzles**: These are brain teasers where the answer violates the default meaning of the word and focuses more on the letter composition.

  For example, **Question**: *How can you make "ten" out of "net"?* **Answer**: *Just flip it around.* **Explanation**: This is because if we consider the spelling of the word "ten" and we flip the letters of the word around, we get the word "net" which is what we want to make out of "ten".

The training data contains **507 Sentence Puzzles** and **396 Word Puzzles**. Each of these puzzles has 4 options to choose from and only one option is the correct answer. Each puzzle further is of 3 different types, depending upon how it was constructed:

- **Original Questions** are the questions that are scraped from various puzzle websites on the web.

- **Semantic Reconstruction Questions** are created by rephrasing the Original Questions without changing its answer, distractor or the premise.
- **Context Reconstruction Questions** keep the commonsense premise intact but changes both the question and answer to new situational context.

3.2.2 Related Works

With the recent success of LLMs in various NLP tasks, researchers have also started exploring their use for Multiple Choice Question Answering (MCQA) tasks [75, 76].

Researchers have also started employing the technique of **few-shot prompting** [77, 78, 79] for various tasks and it has shown improvements when compared with **zero-shot prompting**.

LLMs like GPT-3.5 have been trained on vast amounts of human-generated text. The main features around which such models are trained are **Pattern Recognition**, **Creative Reasoning** and **Wide Knowledge Range**.

Thus, we decided to employ few-shot prompting on LLMs for this task.

3.3 System Overview

Our architecture uses GPT-3.5 ([21]) (specifically *gpt-3.5-turbo*) with few-shot prompting to answer the question.

3.3.1 GPT-3.5

In NLP, the architecture of Generative Pre-trained Transformer (GPT) 3.5 (GPT-3.5) stands as a significant advancement, which is the culmination of iterative improvement over its predecessors. The architecture of the model is based upon the Transformer model [80], which uses self-attention to enhance performance over the prior sequential models. GPT-3.5 scales this Transformer architecture to over hundereds of billions of parameters, which have been trained by exposing and training the model on hundreds of billions of tokens.

In particular, due to the autoregressive nature of GPT-3.5 and due to being trained on extremely large data, it has enough knowledge about the language and the real world to perform tasks in a Zero-shot setting [73]. This Zero-shot setting allows the model to understand and execute a task it hasn't been explicitly trained for. These capabilities have been reflected in GPT-3.5 being used in Summarization [81], Question Answering [82], Natural Language Inference [83], etc.

3.3.2 Few-shot prompting

While zero-shot prompting works well for simple tasks, tasks like BrainTeaser are a bit more complex in nature, and in such cases providing explicit instructions to the LLM about the nature of the task along with few examples of the task *(few shots)* becomes extremely helpful for the model [74]. Here, the few-shot technique involves providing GPT-3.5 some examples, allowing GPT-3.5 to generalize from the few examples, drawing on its large pre-trained knowledge about the language and the real-world.

Thus, 2 different sets of prompts are created for the task, one for the Sentence Puzzle Task and another for the Word Puzzle task, since the 2 tasks are fundamentally different and need different instructions and examples.

3.3.3 Experimental Setup

We provide 2-shot prompts to GPT-3.5 for our leaderboard submission. We also try out 5-shot prompt in the post evaluation phase to test if providing more examples helps the model perform better.

The prompt used for the Sentence Puzzle task is shown in Listing 2. As we can see, the prompt first details the task, and *IMPORTANT* keyword is used to express to GPT-3.5 that commonsense must not be used in the task, but instead it should look at meaning from an unconventional sense. Then, 2 examples are given, along with **reasoning** behind the answers too. This was important, as this gave the model more knowledge to be able to generalize the task from the examples. Further, the output format was clearly specified in the prompt so as to avoid getting extra information in the model output.

Similar prompt for Word Puzzle can be seen in Listing 3. The prompt clarifies that the structural aspect of the words should be focused on, emphasizing that unconventional meaning should be looked at. Then, 2 examples that exhibit structural aspect are given along with reasoning behind their answers as well as constraints for the output format.

3.4 Results and Analysis

The results are detailed in Table 3.1. We report Instance-based accuracy, which considers each question (original or reconstruction) separately, reporting accuracy on original puzzles and their semantic and context reconstructions. On the other hand, Group-based accuracy considers each original puzzle and its variants as a group, scoring 1 only when all three puzzles in the group are successfully solved; otherwise, the score is 0.

For comparison, we also list the zero-shot prompting results reported in [34]. As we can see, the two-shot performance on Word puzzle improved over the zero-shot setting for all the categories, while the same worsened in case of the Sentence puzzle.

This is because of the very nature of the two problems. Sentence puzzle involves deeper non-conventional semantic understanding of the question and the choices, which despite conveying reasoning behind the answers in the few-shot examples, cannot be generalized as easily with just 2 examples. On the other hand, the only tricky component of the Word Puzzle is that the structural aspect of certain words needs to be taken instead of the actual surface meaning of the said words. This can be much more easily generalized through just as few as two examples in the prompt. Further, adding the examples

```
You are given a question with multiple choices that you need to answer. The answer would only be one index of the multiple
↪  choices available. Such a question would involve  brain teaser questions where the puzzle defying commonsense is
↪  centered on sentence snippets.
IMPORTANT: It's crucial to analyze the question from an unconventional perspective, focusing on the literal or alternative
↪  meanings of the words used, rather than relying on common sense. You must not use commonsense, but look at meaning from
↪  a different perspective than what would commonly be done. For example,

Example 1:
Question: You are running so fast but you're not getting closer. Where are you?

Option 0: Country road.
Option 1: Treadmill.
Option 2: High way.
Option 3: None of above.

Answer: 1
Reason: This is because while running on a treadmill, we stay put where we are. The key is understanding that running on a
↪  treadmill means you remain stationary despite the motion. This is not valid for Country road or High way. Thus, the
↪  answer is 1 - Treadmill.

Example 2:
Question: From elementary school to collage, how many "first day of school" does the average person have in their lifetime?

Option 0: They technically only have one first day of school in their lifetime. That's the very first day they started
↪  attending school as a child.
Option 1: Average people have 4: elementary school, middle school, high school, and college.
Option 2: Average people have "first day of school" in each semester, so it will be more than 10!
Option 3: None of above.

Answer: 0
Reason: First day of school can only be one day in a person's lifetime. Here, it is important to understand that first day
↪  of middle school, high school, college won't be first day of school. Similarly, each semester's first day is not
↪  TECHNICALLY first day of school. This, the answer is 0 -  They technically only have one first day of school in their
↪  lifetime. That's the very first day they started attending school as a child. Thus, the key here is the term 'first day
↪  of school' technically refers to the very first day a person attends school, making all subsequent 'first days' at
↪  different educational levels irrelevant to the specific question.

Now, using these examples, answer the question below. It is IMPORTANT that you just provide the index of the answer in the
↪  response. DO NOT output the reason behind choosing the answer:

Question: In a small village, two farmers are working in their fields - a diligent farmer and a lazy farmer. The
↪  hardworking farmer is the son of the lazy farmer, but the lazy farmer is not the father of the hardworking farmer. Can
↪  you explain this unusual relationship?
Option 0: The lazy farmer is his mother.
Option 1: The lazy farmer is not a responsible father as he is lazy.
Option 2: The diligent farmer devoted himself to the farm and gradually forgot his father.
Option 3: None of above.

Answer:
```

Listing 2: Prompt for the Sentence Puzzle

in the Sentence puzzle that don't generalize very well for other questions in the testing set might have acted as noise for the model, which led to poorer performance.

```
You are given a question with multiple choices that you need to answer. The answer would only be one index of the multiple
↪  choices available. The question demands an unorthodox approach, focusing on the spellings or structural aspects of
↪  words, rather than their standard meanings. Your task is to choose the correct answer from the given multiple-choice
↪  options by analyzing the words in a literal or unconventional way.
IMPORTANT: It's crucial to analyze the question from an unconventional perspective, focusing on the spellings of certain
↪  words, rather than relying on common sense. You must not use commonsense, but look at meaning from a different
↪  perspective considering arrangement of the letters in certain words than what would commonly be done. For example,

Example 1:
Question: How can you make "ten" out of "net"?

Option 0: Just flip it around.
Option 1: Remove the letter "e".
Option 2: Move the letter "t" to the end.
Option 3: None of above.

Answer: 0
Reason: This is because if we consider the spelling of the word 'ten' and we flip the letters of the word 'ten' around, we
↪  get the word 'net', which is what we want to make out of 'ten'.  The answer focuses on the literal rearrangement of the
↪  letters, disregarding the typical meanings of the words. Thus, the answer is 0 - Just flip it around.

Example 2:
Question: What is the most fast city?

Option 0: Urban city.
Option 1: Inner city.
Option 2: Velocity.
Option 3: None of above.

Answer: 2
Reason: The term 'fast' in the question prompts an unconventional interpretation. All options contain the word "city", but
↪  "velocity" stands out as it directly relates to speed or 'fastness'. The question cleverly uses the term 'city' as a
↪  red herring, while the actual focus is on the concept of speed.

Now, using these examples, answer the question below. It is IMPORTANT that you just provide the index of the answer in the
↪  response. DO NOT output the reason behind choosing the answer:

Question: What sort of cheese is made in reverse?
Option 0: Cheddar cheese..
Option 1: Edam cheese.
Option 2: Blue cheese.
Option 3: None of above.

Answer:
```

Listing 3: Prompt for the Word Puzzle

We also note that using five-shot prompt instead of two-shot prompt hugely increases the performance. This is to be expected, as providing more examples would help the model generalize even better towards solving the task. This is specially true for Word Puzzle questions, where adding more examples allows the model to generalize the task much better.

Table 3.1: Results of zero-shot and few-shot prompting on GPT-3.5 for the two BrainTeaser subtasks. Ori = Original, Sem = Semantic, Con = Context.

Model	**Instance-based**			**Group-based**		**Overall**
	Original	**Semantic**	**Context**	**Ori & Sem**	**Ori & Sem & Con**	
Sentence puzzle						
GPT-3.5 (zero-shot)	60.7	59.3	**67.9**	50.7	39.7	**62.6**
GPT-3.5 (two-shot)	57.5	55.0	42.5	50.0	30.0	51.7
GPT-3.5 (five-shot)	**62.5**	**65.0**	55.0	**62.5**	**42.5**	60.8
Word puzzle						
GPT-3.5 (zero-shot)	56.1	52.4	51.8	43.90	29.3	53.5
GPT-3.5 (two-shot)	71.9	71.9	62.5	59.4	46.9	68.6
GPT-3.5 (five-shot)	**78.1**	**90.6**	**84.4**	**78.1**	**68.8**	**84.4**

However, in Sentence Puzzle we still notice a drop in the overall performance as compared to the zero-shot model. This is because of a drop in the performance of the context reconstruction questions, and a marginal increase in comparison to zero-shot in other types of questions. However, group based accuracy increases in five-shot, which might indicate that with five examples, the model is able to handle the variations in reconstructions better, albeit with performance of Contextual Reconstruction taking a dip. These observations are in line with the drop observed in two-shot prompt in comparison to the zero-shot prompt, highlighting the difficult nature of the task of Sentence Puzzle questions and the inability of the model to generalize using few Sentence Puzzle examples. However, we do note that the performance on Sentence Puzzle also does improve with additional examples between two-shot and five-shot prompting.

3.5 Conclusion

In conclusion, we explored the effectiveness of few-shot prompting for LLMs for complex and unconventional tasks. Further, it demonstrates that few-shot prompting is helpful only in scenarios where the examples convey enough information that can be better generalized, as the results worsened in the Sentence Puzzle while improved in the Word Puzzle.

In future, better prompting strategies like Chain of Thought prompting [84] can be utilized to improve the performance. Additionally, finetuning GPT-3.5 might also help in the task further. Also, increasing the number of training examples might help in further improving the model performance, as observed in the gains of performance in the five-shot prompt in comparison to the two-shot prompt.

Through this work, we get a clearer idea about how defining the reason behind certain question answers and even providing few-shot examples helps model in complex reasoning. We make use of these insights later in the thesis while reasoning over complex multimodal tables by prompting LLMs and MLLMs.

Chapter 4

Reasoning over Multi-modal Semi-structured Tables

This chapter is adapted from the upcoming publication *"Knowledge-Aware Reasoning over Multimodal Semi-structured Tables"* under review at an NLP venue. This work is a joint effort with: Jainit Sushil Bafna (IIIT Hyderabad), Kunal Kartik (IIT Guwahati), Harshita Khandelwal (UCLA), Prof. Manish Shrivastava (Prof, IIIT Hyderabad), Vivek Gupta (Postdoc, University of Pennsylvania), Prof. Mohit Bansal (Prof, University of North Carolina) and Prof. Dan Roth (Prof, University of Pennsylvania). This chapter incorporates the methodology and experiments outlined in the version of the paper dated 6th May 2024.

This chapter introduces the problem of Question Answering over Multimodal Tables. We discuss the prior work related to the problem, explore the motivation and challenges behind performing Multimodal Table Question Answering, and provide a formal formulation of the problem.

4.1 Introduction

Tables are essential tools for summarizing and conveying information efficiently across numerous fields. Although inference and reasoning over tables have been extensively explored within the realm of Natural Language Processing (NLP) [35], prior research has predominantly focused on text-only tables.

However, real-world tables frequently incorporate images, such as logos or flags, representing various entities like teams or countries. Additionally, in cases where the visual characteristics of entities are crucial, tables may include images corresponding to these entities. This aspect is vital for the inference and reasoning capabilities over such tables, yet it remains unexplored. In this thesis, we build upon the tasks of Multimodal Emotion-Cause Pair Extraction and Prompting LLMs for Inference and Reasoning to propose the task of exploring Multimodality in Tables, using Multimodal LLMs [85] to solve the task.

To this end, we introduce the task of Question-Answering over Multimodal Knowledge-aware Semi-structured tables. Our work involves Question-Answering over tables from Wikipedia, where Wikipedia entities belonging to various categories are represented using their images. Effectively, performing inference on such tables requires understanding what entity is represented by said image in the context of the table in addition to gaining simply a visual understanding of the image. Thus, Multimodal Entity

Linking [86] of the image in the context of the table is a crucial aspect of our methodology. Additionally, understanding the visual attributes of the entities is essential for answering queries that involve visual characteristics. Since the tables feature multiple images, our approach also incorporates elements of multi-image visual question-answering [87]. This also makes the problem interesting from the perspective of exploring and probing the capabilities of multimodal LLMs in the aspects of multimodal entity linking, reasoning over multiple images, using real-world factual knowledge, and understanding semi-structured information.

4.2 Related Work

Since the problem is novel, there is no prior work directly related to the problem in its entirety. However, the problem draws from various independent Multimodal problems and their aspects, combining them into one bigger problem. In this Section we look at the background works for each aspect of our problem independently, which provides a direction to how we approach it.

4.2.1 Tables with additional modality

While traditionally only homogeneous tables were considered for various NLP tasks, few recent works have explored using extra modalities as additional context for tabular question answering. [88] added additional context of paragraphs in addition to the information present in the table for reasoning and question answering. Similarly, [89, 90] created hybrid data by combining tabular and textual data from real financial reports to build the benchmark. Likewise, [91] explored the task of numerical reasoning over hybrid data comprising both textual and hierarchical table content.

Building upon hybrid data comprising tables and additional paragraph context, [92] adds the additional context of images over the paragraph and table information from the previous works. Similarly, [93] involves conversational question answering over tables, images and texts. Both these works propose models that involve combining the relevant extracted information from the individual three modalities for question answering. [94] propose a unified language representation for combining the image, text and table modalities for question answering. [95] propose an end-to-end prompting method and use In-context learning for performing Question Answering on the hybrid datasets.

However, these works still **don’t consider the table itself to be non-homogeneous**. Instead, they add some additional context to the homogeneous table, which still remains text-only.

4.2.2 Knowledge-Based Visual Question Answering

Recently, huge progress has been made towards Visual Question Answering [96], but entity-specific knowledge-aware visual question answering is relatively unexplored. While datasets like [97] and [98] require real-world knowledge for reasoning and performing Visual Question Answering, they don’t require much entity-specific finegrained knowledge for the question answering task. [99] introduced the task of knowledge-aware visual question answering over named entities, making used of Wikipedia/Wikidata entity images to create the dataset and the Knowledge Base over which retrieval-based approaches performed well for the task, which forms highly relevant to our task of performing tabular QA over Wikipedia entities.

[100] and [101] tackle similar task, requiring fine-grained knowledge about the entities provided in images for performing the task of Visual Question Answering. These also make use of Large Vision-Language Models like BLIP-2 [102], PaLI [103] to answer such questions.

4.2.3 Multimodal Entity Linking

Multimodal Entity Linking is a process that combines information from different modalities, such as text, images, or audio, to accurately identify and link entities mentioned in a given context to their corresponding entries in a knowledge base or database. [86] introduced the task of Multimodal Entity Linking under zero-shot setting using social media data. Several approaches were tried to Multimodal Entity Linking, like determining relations between image and text and performing disambiguation [104], combining text and image to perform the disambiguation [104], while others also implemented graph-matching based models [105].

Recently, [106] introduced the WikiDiverse datset, which involved Multimodal Entity Disambiguation for Wikipedia entities, which is very close to the entity linking/disambiguation aspect of our task, albeit in the context of a table. The analysis in this dataset also emphasizes that both image and text modalities are important in disambiguating the entity.

4.2.4 VQA involving multiple images

Visual Question Answering has made significant progress about understanding and responding about single images. However, real-world scenarios, including such Multimodal Tables, involve multiple related images that provide a richer context. [107] studies the task where answer needs to be mined from a pool of images, and proposes retrieval-based methods for doing the question answering. [108] proposes a dataset and model that focuses on question-answering over multiple images.

4.3 Potential Challenges/Complexities in the Problem

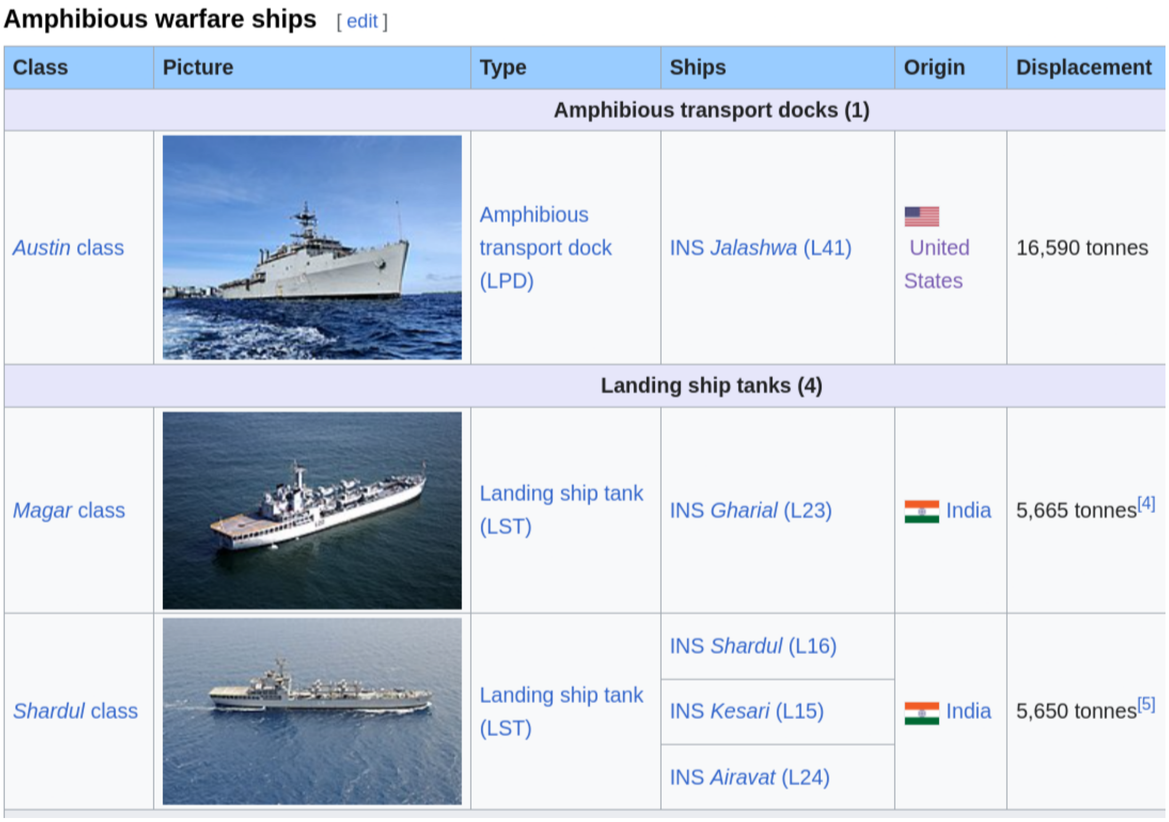
To provide some further motivation to the problem, we discuss some complexities and challenges that are involved in performing knowledge-aware reasoning over Multimodal tables:

4.3.1 Entity Disambiguation Problem

Since the entities are represented via images in the table, disambiguating the entities forms a big part of answering questions based on such tables. However, this entity disambiguation can be very complex in nature. Some cases below are illustrated to describe its intricacies:

Case 1: Using other columns entries from the entity's row

Often, it is difficult to disambiguate entities from the image alone – entire table context is needed. For example, in Fig. 4.1, just given the images it would be difficult to figure out the Class of the ship in question. However, given the context coming from Types and Origin, we can disambiguate it (assuming the Class column was absent and needed to be predicted). Thus, we would require the information provided about the entity in other columns to disambiguate the entity in question. This information in other columns can be text or image, and so would ideally require modelling that can handle any input type arbitrarily.



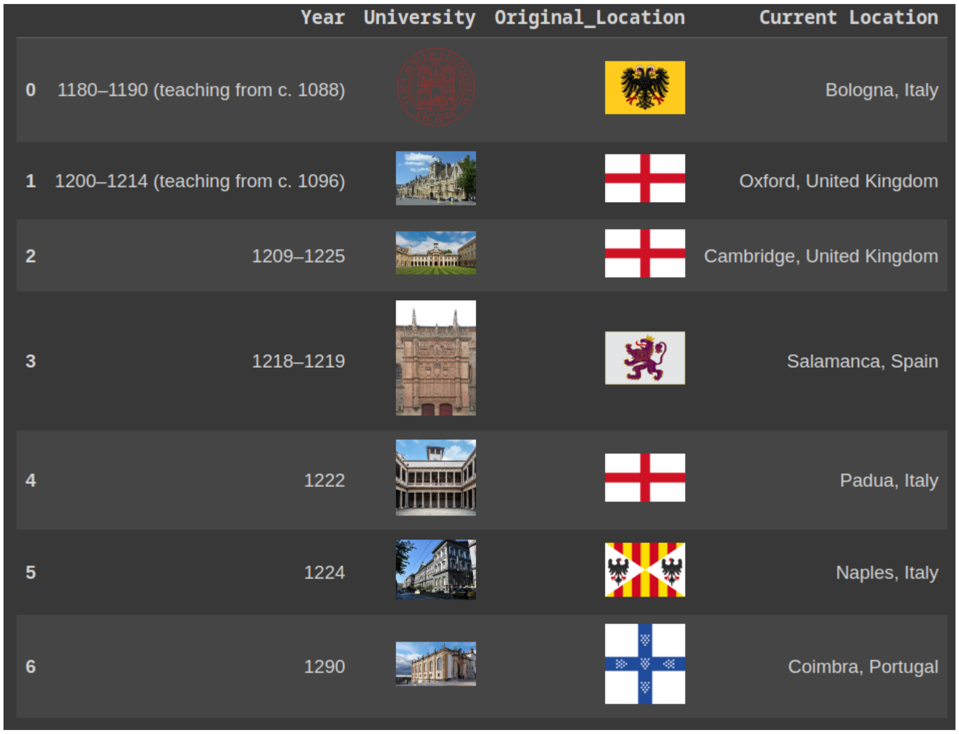
Amphibious warfare ships [edit]

Class	Picture	Type	Ships	Origin	Displacement
		Amphibious transport docks (1)			
Austin class		Amphibious transport dock (LPD)	INS *Jalashwa* (L41)	United States	16,590 tonnes
		Landing ship tanks (4)			
Magar class		Landing ship tank (LST)	INS *Gharial* (L23)	India	5,665 tonnes[4]
Shardul class		Landing ship tank (LST)	INS *Shardul* (L16) INS *Kesari* (L15) INS *Airavat* (L24)	India	5,650 tonnes[5]

Figure 4.1: Visually similar ships requiring more context for disambiguation

Case 2: Using information about the entity type from other rows

Context from other rows of the table is needed when the image used to represent the entity is one constituent of it, a phenomenon called as Meronymy. For example, in Fig. 4.2 the Balliol College from Oxford is being used as a representative of Oxford University rather than as an entity in itself. Thus, looking at some University logos in other rows and the *University* table header is necessary to understand that the Balliol College here is representing the Oxford University rather than just the college itself.



	Year	University	Original_Location	Current Location
0	1180–1190 (teaching from c. 1088)			Bologna, Italy
1	1200–1214 (teaching from c. 1096)			Oxford, United Kingdom
2	1209–1225			Cambridge, United Kingdom
3	1218–1219			Salamanca, Spain
4	1222			Padua, Italy
5	1224			Naples, Italy
6	1290			Coimbra, Portugal

Figure 4.2: Entities exhibiting meronymy leading to ambiguity

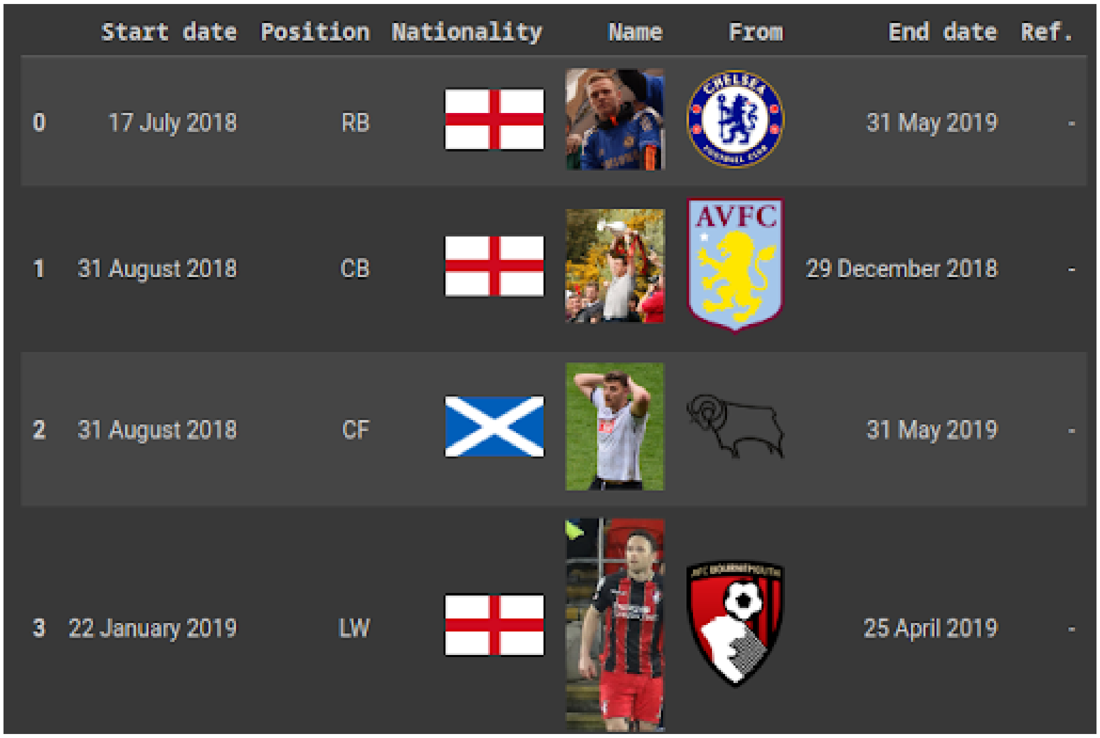
Case 3: Using entire table context to understand entity
In Fig. 4.3, when questioning about the nationality of a player, the answer should be English or Scottish rather than England or Scotland or Flag of England or Flag of Scotland. Here, understanding that the table is describing information about different players, and an image of a flag in the *Nationality* column would represent England/Scotland as countries rather than their respective flags is necessary.

Case 4: Table context helps identifying celebrities/humans
In Fig. 4.3, disambiguating the person depicted in a particular image (i.e. the task of facial recognition) from the image alone is not easy, especially when similar-looking people might exist. However, when we incorporate the surrounding context of the table, then the information about which position the player plays at, the nationality, their club and their tenure at the club provides significant extra context, which becomes crucial to identifying the players.

4.3.2 Logical Reasoning Questions

In case of abstractive questions, the task can involve logical reasoning as well, including the following types:

1. **Numerical Reasoning:** For example, in Fig. 4.4, we can ask a question like *By many more states does the lotus-symbol party rule compared to the book-symbol party?* Such questions might involve arithmetic operations, such as addition, subtraction, etc.



	Start date	Position	Nationality	Name	From	End date	Ref.
0	17 July 2018	RB				31 May 2019	-
1	31 August 2018	CB				29 December 2018	-
2	31 August 2018	CF				31 May 2019	-
3	22 January 2019	LW				25 April 2019	-

Figure 4.3: Player images needing more context for disambiguation

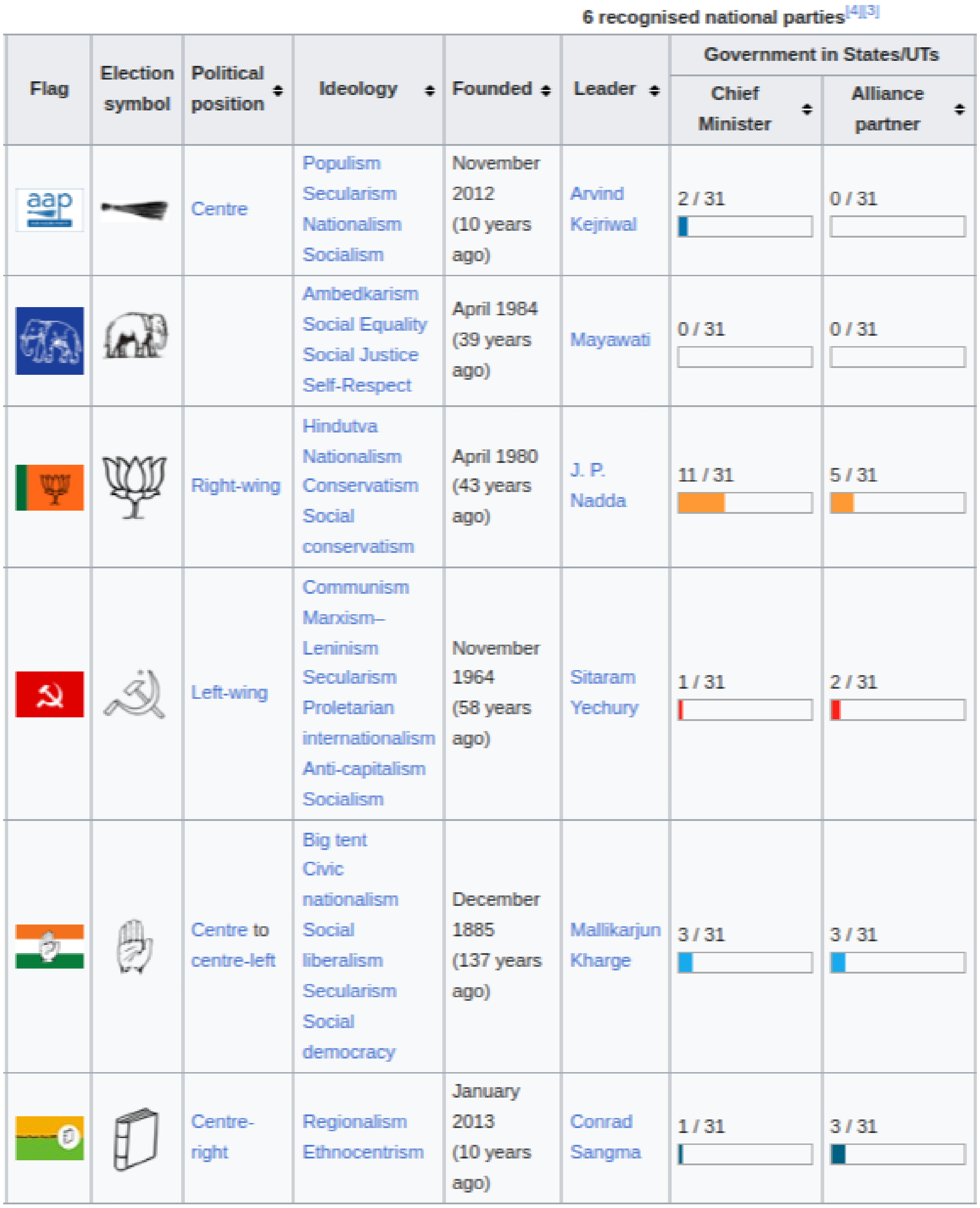
2. **Temporal Reasoning:** For example, in Fig. 4.3, we can ask a question like *How many days passed between the start dates of Todd Kane and Tommy Elphick?*. Such questions can even extend to other temporal aspects.

3. **Commonsense Reasoning:** In Fig. 4.4, we can ask a question like *Which political party symbol is relevant to cleaning?*, which would require commonsense reasoning that a broom is used for cleaning, which corresponds to the AAP symbol in the table.

4. **Entity Type:** We may want to predict and reason over what are the entity types in each table, based upon other column values. For example, in Fig. 4.3, the position and the football club images reveal that the players are football players, which requires understanding that someone joining a football club and playing on an RB position is a player.

5. **Multi-row Reasoning:** Questions can also involve multiple multimodal rows in the tables for reasoning and answering. For example, in Fig. 4.3, one can ask for the longest difference in the start date between two players who belonged to the same country but played for different clubs.

6. **Entity Reasoning**: The models should understand the textual part to the extent that it can understand that *Oxford University* implied that it is a kind of university. Further, abbreviations and entity resolutions need to be handled too, like BJP and Bharatiya Janta Party.



6 recognised national parties[4][3]

Flag	Election symbol	Political position	Ideology	Founded	Leader	Government in States/UTs	
						Chief Minister	Alliance partner
aap		Centre	Populism Secularism Nationalism Socialism	November 2012 (10 years ago)	Arvind Kejriwal	2 / 31	0 / 31
			Ambedkarism Social Equality Social Justice Self-Respect	April 1984 (39 years ago)	Mayawati	0 / 31	0 / 31
		Right-wing	Hindutva Nationalism Conservatism Social conservatism	April 1980 (43 years ago)	J. P. Nadda	11 / 31	5 / 31
		Left-wing	Communism Marxism–Leninism Secularism Proletarian internationalism Anti-capitalism Socialism	November 1964 (58 years ago)	Sitaram Yechury	1 / 31	2 / 31
		Centre to centre-left	Big tent Civic nationalism Social liberalism Secularism Social democracy	December 1885 (137 years ago)	Mallikarjun Kharge	3 / 31	3 / 31
		Centre-right	Regionalism Ethnocentrism	January 2013 (10 years ago)	Conrad Sangma	1 / 31	3 / 31

Figure 4.4: Wikipedia table for political parties in India

7. **Quantification:** Reasoning based on Quantifications like many/less and comparison should also be encoded. For example, the model should identify in Fig. 4.4 that BJP won in more states while Congress won in fewer states.

4.4 Problem Definition

Let T be a multimodal table consisting of m rows and n columns. Each cell in the table, denoted by $T_{i,j}$, can contain both textual and visual information. Formally, $T_{i,j}$ can be represented as a tuple $(t_{i,j}, I_{i,j})$, where $t_{i,j}$ represents the textual content of the cell and $I_{i,j}$ represents the image content of the cell.

Let Q be a set of questions, where each question q_k is associated with a specific table T. Each question can be represented as a textual query.

The goal of Multimodal Table Question Answering is to find the correct answer A_k to each question q_k based on the information provided in the corresponding table T. The answer A_k can be either a textual response or a numerical value, depending on the nature of the question.

Mathematically, given a table T and a question q_k, the Multimodal Table Question Answering task can be formulated as follows:

$$A_k = f(T, q_k)$$

where f is a function that maps the input table T and question q_k to the corresponding answer A_k.

The function f can be implemented using various techniques, such as natural language processing (NLP) models, computer vision models, and multimodal fusion techniques, to extract and integrate information from both textual and visual modalities in the table to generate the answer.

This formulation provides a formal framework for defining and solving the Multimodal Table Question Answering problem.

4.5 Conclusion

This chapter introduced the novel problem of Question Answering over Multimodal Tables, bridging the gap between traditional text-based table inference and the rich, multimodal landscape of real-world data. Drawing upon insights from various research domains, we highlighted the significance of incorporating images into tables and the unique challenges they pose for reasoning systems. Through a collaborative effort, we delineated the task's complexities, including entity disambiguation, logical reasoning, and the integration of textual and visual cues.

Expanding on the complexities identified, we undertake the development of a complex and diverse dataset specifically designed to address the multifaceted challenges of Multimodal Table Question Answering. This dataset, created by recasting diverse table question-answering datasets, incorporates nuances such as entity disambiguation, logical reasoning, and the fusion of textual and visual information. Through meticulous curation, we aim to provide a comprehensive benchmark for evaluating and advancing multimodal reasoning models. The subsequent chapter will offer insights into the dataset's construction methodology, statistics, and detailed analysis.

Chapter 5

MultiModalTabQA Dataset

This chapter is adapted from the upcoming publication *"Knowledge-Aware Reasoning over Multimodal Semi-structured Tables"* under review at an NLP venue. This work is a joint effort with: Jainit Sushil Bafna (IIIT Hyderabad), Kunal Kartik (IIT Guwahati), Harshita Khandelwal (UCLA), Prof. Manish Shrivastava (Prof, IIIT Hyderabad), Vivek Gupta (Postdoc, University of Pennsylvania), Prof. Mohit Bansal (Prof, University of North Carolina) and Prof. Dan Roth (Prof, University of Pennsylvania). This chapter incorporates the methodology and experiments outlined in the version of the paper dated 6th May 2024.

In this chapter, we outline the process used to construct the MultiModalTabQA dataset for our proposed problem of Multimodal Tabular Question Answering. We also present a detailed analysis and statistics about the dataset.

5.1 Introduction

As described in Chapter 4, we aim to create a dataset that incorporates the following challenges:

1. The entities should require the context of the table and external knowledge in addition to the information provided through just the image to be disambiguated.

2. The table should contain sufficient images and text entries. For this dataset, we keep the number of image-cells in a particular column between 30%-75% of the total number of cells.

3. The questions based on the table should require truly multimodal reasoning and should be hard to answer solely based on the text component of the table.

4. The dataset should require advanced reasoning such as Numerical reasoning, Temporal Reasoning, Multi-row Reasoning, and Commonsense Reasoning. Thus, the questions should not just require parsing and retrieving specific cells from the table but also complex reasoning over multiple cells.

To fulfil the requirement of external knowledge and image context for entity disambiguation, a good choice would be to recast textual tabular reasoning datasets containing many real-world entities as entries within the table into multimodal format. A perfect fit for such a dataset would be Tabular Datasets based on **Wikipedia Data**. To create a diverse dataset, we recast **three** existing Wikipedia Table Question Answering datasets' tables with images corresponding to some entities in the table.

5.2 Datasets Recast

5.2.1 WikiSQL Dataset

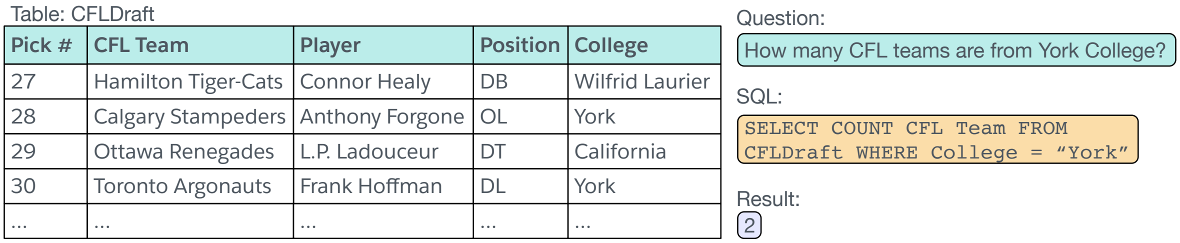


Table: CFLDraft

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Ottawa Renegades	L.P. Ladouceur	DT	California
30	Toronto Argonauts	Frank Hoffman	DL	York
...	...	...	...	...

Question:

How many CFL teams are from York College?

SQL:

```
SELECT COUNT CFL Team FROM
CFLDraft WHERE College = “York”
```

Result:

2

Figure 5.1: Example from WikiSQL dataset

The most basic requirement for a Multimodal Table Question-Answering model would be an ability to parse the entities in the table correctly and answer SQL-query-based questions from the multimodal table. To benchmark such abilities of Multimodal LLMs, we first recast the WikiSQL Dataset [109], which is a large dataset of 80,654 hand-annotated examples of questions and SQL queries distributed across 24,241 tables from Wikipedia. We can see an example from the dataset in Fig. 5.1

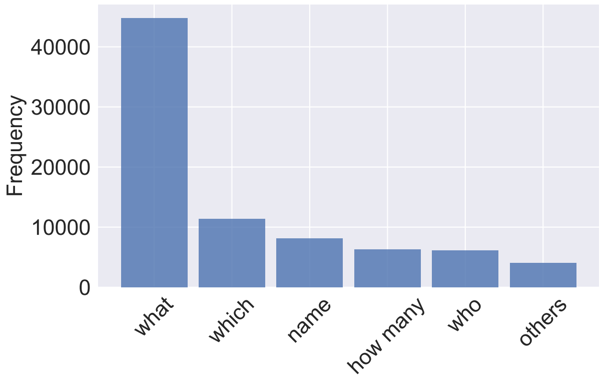


Figure 5.2: Distribution of Question Types in WikiSQL

Since this dataset involves SQL queries for answering each question, recasting this dataset provides us with the means to evaluate whether the MLLMs can parse the basic table structure and its contained entities correctly – which are the core aspects of a Multimodal Table-QA model. Further, since it involves Wikipedia tables belonging to various categories, it also captures entities in the images that are hard to disambiguate without the context of the entire table. The types of questions in the dataset are also diverse, as can be seen in Fig. 5.2, making the dataset suitable for recasting for testing the model's abilities to parse the **fundamental semi-structured nature of the table** and also **disambiguate the different entities in the dataset**. Since the answers are obtained by executing specific SQL queries on the table, this dataset is **short-answer** in nature.

5.2.2 WikiTableQuestions Dataset

Extending the complexity of our dataset beyond simple SQL-based question-answering, we recast the WikiTableQuestions dataset [110], which involves more complex reasoning over the table than simple SQL-based queries. We can see an example from the dataset in Fig. 5.3.

The dataset contains 22,033 questions on 2,108 tables. The dataset incorporates 3,929 unique column headers among 13,396 columns, exhibiting a diverse range of relationships in the table. Additionally, the dataset also requires complex reasoning, including temporal reasoning, numerical reasoning, counting, etc. Since it involves tables scraped from Wikipedia, it also incorporates the problem of entity disambiguation within the context of the table. This dataset is also **short-answer** in nature and is the source of such questions in MultimodalTabQA dataset which require **complex reasoning** over the table.

5.2.3 FeTaQA

While all the above datasets involve short-form answers, the FeTaQA dataset [111] comprises 10,330 **long-form answer** questions over 8,551 tables grounded in Wikipedia tables sourced from the ToTTo dataset. The dataset provides the cells used in reasoning for the answer. Additionally, it guarantees that the reasoning spans more than a single row or column, ensuring that the questions are complex, with an average of 6 cells used for reasoning in a question. We can see an example from the dataset in Fig. 5.5.

Further, the dataset contains tables from diverse sources, as seen in the dataset's topic distribution in Fig. 5.4. Using this dataset as one data source for recasting significantly improves the diversity and complexity of our dataset by involving complex and abstract long-form answer questions.

5.3 Recasting the Datasets

The primary step in converting a table from textual to multimodal form involves establishing a link between a textual mention of an entity and its corresponding image. Given that we are working primarily with Wikipedia datasets, where an entity's representative image can be obtained from either the Wikipedia Infobox or the Wikidata entry of the entity using just the page link, this task simplifies to finding the Wikipedia page link corresponding to the entity texts. Subsequently, we filter which entity

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x_1: *"Greece held its last Summer Olympics in which year?"*

y_1: $\{2004\}$

x_2: *"In which city's the first time with at least 20 nations?"*

y_2: {Paris}

x_3: *"Which years have the most participating countries?"*

y_3: $\{2008, 2012\}$

x_4: *"How many events were in Athens, Greece?"*

y_4: $\{2\}$

x_5: *"How many more participants were there in 1900 than in the first year?"*

y_5: $\{10\}$

Figure 5.3: Example questions from the WikiTableQuestions Dataset

texts should be converted to their representative images. The steps involved in recasting the datasets from textual to multimodal form are outlined below:

5.3.1 Step 1: Getting raw HTML corresponding to the table

To get the Wikipedia links corresponding to various entity texts, we need the raw HTML of the Wikipedia tables and parse the entity links using *BeautifulSoup*[1]. For the WikiTableQuestions data source, the released dataset provided all the original raw HTML files corresponding to the tables. However, such raw HTML was absent in other data sources and had to be scraped separately for WikiSQL and FeTaQA datasets.

In the case of both these datasets, the originally released dataset provides the Wikipedia page links from where the table was originally taken. We scraped the Wikipedia page version of these pages corresponding to their revision IDs from a date closest to the time when the respective datasets (ToTTo in the case of FeTaQA) were released so that the table content from the HTML could be as close to the

[1] https://beautiful-soup-4.readthedocs.io

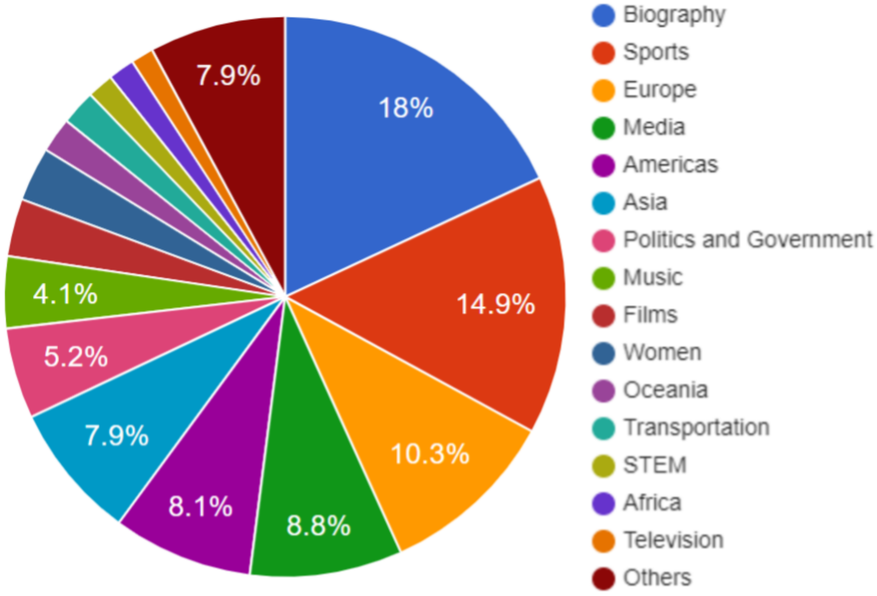
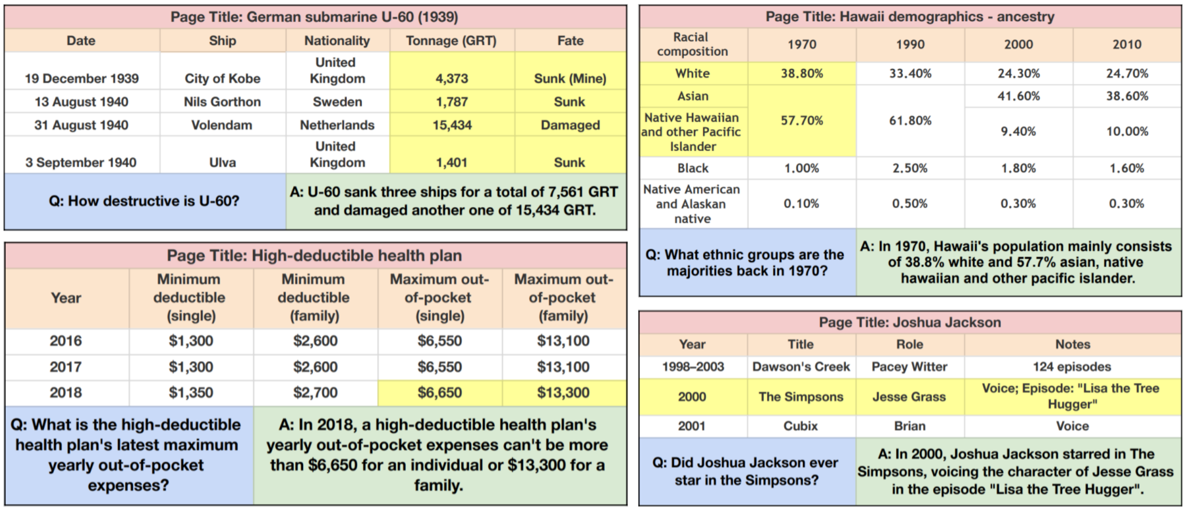


Figure 5.4: Distribution of topics in FeTaQA



Page Title: German submarine U-60 (1939)

Date	Ship	Nationality	Tonnage (GRT)	Fate
19 December 1939	City of Kobe	United Kingdom	4,373	Sunk (Mine)
13 August 1940	Nils Gorthon	Sweden	1,787	Sunk
31 August 1940	Volendam	Netherlands	15,434	Damaged
3 September 1940	Ulva	United Kingdom	1,401	Sunk

Q: How destructive is U-60?

A: U-60 sank three ships for a total of 7,561 GRT and damaged another one of 15,434 GRT.

Page Title: Hawaii demographics - ancestry

Racial composition	1970	1990	2000	2010
White	38.80%	33.40%	24.30%	24.70%
Asian	57.70%	61.80%	41.60%	38.60%
Native Hawaiian and other Pacific Islander			9.40%	10.00%
Black	1.00%	2.50%	1.80%	1.60%
Native American and Alaskan native	0.10%	0.50%	0.30%	0.30%

Q: What ethnic groups are the majorities back in 1970?

A: In 1970, Hawaii's population mainly consists of 38.8% white and 57.7% asian, native hawaiian and other pacific islander.

Page Title: High-deductible health plan

Year	Minimum deductible (single)	Minimum deductible (family)	Maximum out-of-pocket (single)	Maximum out-of-pocket (family)
2016	$1,300	$2,600	$6,550	$13,100
2017	$1,300	$2,600	$6,550	$13,100
2018	$1,350	$2,700	$6,650	$13,300

Q: What is the high-deductible health plan's latest maximum yearly out-of-pocket expenses?

A: In 2018, a high-deductible health plan's yearly out-of-pocket expenses can't be more than $6,650 for an individual or $13,300 for a family.

Page Title: Joshua Jackson

Year	Title	Role	Notes
1998–2003	Dawson's Creek	Pacey Witter	124 episodes
2000	The Simpsons	Jesse Grass	Voice; Episode: "Lisa the Tree Hugger"
2001	Cubix	Brian	Voice

Q: Did Joshua Jackson ever star in the Simpsons?

A: In 2000, Joshua Jackson starred in The Simpsons, voicing the character of Jesse Grass in the episode "Lisa the Tree Hugger".

Figure 5.5: Example questions in FeTaQA dataset

table given in the datasets. Next, among all the tables on a Wikipedia page, we choose the table with the highest Jaccard Coefficient and the table in the dataset as the corresponding raw HTML table. This computation is described below:

Let A be the set of bigrams extracted from the text of the candidate table scraped from the Wikipedia URL, and B be the set of bigrams from the table array provided in the dataset. The Jaccard coefficient for the similarity of these two sources based on overlapping bigrams is given by:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where:

- $|A|$ denotes the number of unique bigrams extracted from the text of the table scraped from the Wikipedia URL.

- $|B|$ denotes the number of unique bigrams in the table array provided in the dataset.

- $|A \cap B|$ denotes the number of overlapping bigrams between the table text scraped from the Wikipedia URL and the table array provided in the dataset.

- $|A \cup B|$ denotes the total number of unique bigrams from both sources.

This provides us with the raw HTML that we want corresponding to the tables provided in the WikiSQL and FeTaQA datasets.

5.3.2 Step 2: Getting WikiData entry corresponding to Wikipedia Link(s)

We extract Wikipedia links from the raw HTML of the tables and use them to find the corresponding WikiData IDs for the entities they represent. To accomplish this, we primarily utilize the Wikidata SPARQL query system. In cases where the Wikidata ID cannot be obtained through SPARQL queries, we resort to scraping the information from the corresponding Wikipedia pages.

5.3.3 Step 3: Getting image from the Wikidata ID

After retrieving the WikiData entry corresponding to the links, we proceed to fetch images using the WikiData information and infoboxes on Wikipedia. The image retrieval follows the preference order outlined in [99]: (i) Infobox image from the Wikipedia page; (ii) P18 “image” (roughly equivalent to the infobox image in Wikipedia articles); (iii) P154 “logo image”; (iv) P41 “flag image”; (v) P94 “coat of arms image”; (vi) P2425 “service ribbon image”;

5.3.4 Step 4: Filtering out the entities to be replaced

Initially, we considered creating the dataset by replacing all entity mentions with images scraped from their corresponding Wikipedia URLs. However, many images were too vague, depicting unrecognizable municipalities or roads/streets. The predominance of such images would have made the problem nearly unsolvable. To address this, we opted to scrape WikiData information, specifically the P31 "instance of" property (referred to as category, herewith) and the corresponding Wikipedia pageviews, to filter out entities that are replaced by images.

Our process involved manually annotating over 1,500 categories to determine which images should be included or excluded from the dataset. For specific categories, we set a page view threshold, ensuring that only those entities in such categories deemed "popular enough" to be recognizable from their images to a great extent were included in the dataset (e.g., certain cities, tourist attractions, etc.). Additionally, for categories like locations, we created collages from all images in the infobox of the entity's Wikipedia page since a single image is not a sufficient representation in such cases.

Furthermore, we re-scraped images for specific categories where logos, seals, or coats of arms were available and were better representations than the previously scraped images (e.g., town seals, company logos, tournament posters, etc.).

5.3.5 Step 5: Replacing the text with images

After finding the images that correspond to different links in the scraped HTML, we still need to replace the text with images in the dataset array(s) we already have. Due to differences in the nature of the dataset provided, we describe this step separately for each dataset:

- **WikiTableQuestions:** In this dataset, the raw HTMLs provided were used to create the final table arrays. Thus, we didn't need to map the links from HTML separately to the existing tables in the dataset. We simply replaced the text corresponding to the filtered Wikipedia links in the provided HTMLs with image IDs and generated the table arrays from them.

- **WikiSQL and FeTaQA:** Since the original raw HTMLs were not available in these two datasets, we create a mapping of $entity\_link \rightarrow [entity\_texts]$ from the raw HTMLs where all the $entity\_link$ have an image linked to them. Now, we prune out the entity_texts such that entity replacement doesn't happen for erratic strings, using heuristics like the percentage of non-alphabetic characters in the string, its length, etc. Now, we pass through the table provided in the dataset and replace all strings matching with *entity_texts* with the image corresponding to *entity_link*, prioritising replacement for longer words in *entity_texts*.

5.3.6 Step 6: Creating Explicit and Implicit Questions

We create the following types of questions in our dataset:

1. **Explicit Questions:** These are the questions in which an entity is explicitly mentioned in the question that has been replaced by an image in the original table.

2. **Answer-Mention Questions:** These are the questions in which an entity is explicitly mentioned in the answer that has been replaced by an image in the original table.

3. **Implicit Questions:** These are the questions in which the question or answer doesn't have an explicit mention of the entity, but the intermediate reasoning for the answer involves an entity that has been replaced by an image in the original table.

We must filter out the explicit, answer-mention and implicit questions from the provided questions to create the final dataset. For explicit and answer-mention questions, we include out all questions naively where any image-replaced entity text in the table occurs exactly in the question.

For implicit questions, we use evidence cells provided in WikiSQL and FeTaQA datasets, which provide the cells that have been used in the reasoning for getting the answer. However, these annotations were missing for WikiTableQuestions, so we hand-annotated the evidence cells for pruned questions in the dataset and, based on these annotations, chose implicit questions for our final dataset.

5.3.7 Step 7: Pruning Table Images

In order to make the problem interesting, we keep a lower limit of at least 30% images in at least one column of a table and an upper limit of 75% images in all columns of the table. Thus, we first prune out tables in which no column has at least 30% images. We also discard tables with more than 55 rows as they were too large for processing with different models.

Next, we remove excess images in the tables. For this, we prioritize keeping cells which contain explicit mention(s) for different questions/answer-mentions as images, keeping the priority order as (i) One explicit mention (question/answer-mention) cell from each question of the table (ii) Explicit mention cells (question/answer-mention) from all questions of the table (iii) At least one cell involving implicit reasoning for every question based on the table (iv) All cells involving implicit reasoning for any questions based on the table (v) Cells which are not involved in any reasoning/mention.

Now, we compute the number of image-cells that must be changed back to text to reduce the percentage of image-cells to 75%, and choose the tentative cells that can be converted back to text while prioritizing the cells described above. Now, out of the tentative cells which can be converted back to text, we randomly pick the required number of cells and replace the image tags in those cells with the original entity text.

After pruning the table images, we recheck the explicit/answer-mention and implicit questions for their respective mention/evidence cells and create our final dataset of explicit, answer-mention and implicit questions.

Lastly, we divide the recast dataset into training, development, and test sets using a split of 65%-15%-20%. To create the split, we perform random sampling over the questions in the dataset.

5.4 Dataset

After following the steps described previously, we create our MultimodalTabQA dataset v0.1 consisting of **35,111 questions over 16,941 tables**. The details of the tables and questions in the dataset are described in table 5.1. All the data sources in our dataset have almost the same average number of columns. In contrast, the average number of rows is slightly higher in WikiTableQuestions tables than in the other data sources. WikiTableQuestions has the smallest number of tables, while WikiSQL and FeTaQA have a much higher number. However, FeTaQA has very few questions per table, with almost a single question being asked on most tables. We note that WikiSQL has much more questions than FeTaQA and WikiTableQuestions.

Data Source	No. of Tables	No. of Questions	Avg. Rows	Avg. Columns
WikiSQL	9784	19645	13.95	6.25
WikiTableQuestions	1259	9175	18.23	6.30
FeTaQA	5898	6291	14.44	6.12

Table 5.1: Dataset Statistics

Table 5.2 details the statistics about the overall image replacements in the dataset. We notice that FeTaQA has a considerably lesser number of images than the other two data sources per table. WikiTableQuestions has a very high degree of multimodality in the tables despite the low number of tables. This is possibly because the HTMLs of WikiTableQuestions were readily available, aiding more entities in being replaced by images.

Data Source	Unique Images	Avg. Unique Images per Table	Avg. Images per Table
WikiSQL	35202	13.65	21.19
WikiTableQuestions	15387	17.51	25.66
FeTaQA	35683	10.36	17.43

Table 5.2: Image Statistics in the dataset

In Table 5.3, we can see the distribution of the different questions in the three data sources. We expect Answer-Mention Questions and Explicit Questions to be more challenging because in these questions, the entities mentioned in the question/answer are definitely replaced by images and so are NECESSARY to be disambiguated when performing the task. While WikiSQL and FeTaQA have a small number of Implicit questions, WikiTableQuestions has many more, indicating the complex reasoning involved in the dataset.

Data Source	Explicit Questions	Answer-Mention Questions	Implicit Questions
WikiSQL	12956	6374	315
WikiTableQuestions	3523	2773	2879
FeTaQA	2499	3180	612

Table 5.3: Question-Type in the dataset

Data Source	Single-Column	Multi-Column
WikiSQL	16192	3453
WikiTableQuestions	7737	971
FeTaQA	3933	2280

Table 5.4: No. of Multi/Single Multimodal column Reasoning Questions

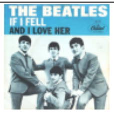
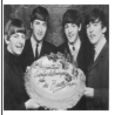
Table 5.4 shows the number of questions requiring single/multi-column multimodal reasoning. Handling multiple images of different categories (as they belong to different columns) and connecting them might be much more complex than multimodal reasoning over a single column. The high percentage of such questions in FeTaQA underscores the complexities of answering Free-form Questions. From Table 5.5, we can see the diverse sets of images present in the dataset, emphasizing the complexity of the problem.

5.5 Dataset Validation

Since we recast existing tables, we first need to verify whether the entity replacements are correct. We sample 250 tables from each data source (total 750 tables) and have 3 annotators score all unique $(image, original\_text)$ pair per table, verifying the correctness of the image replacements.

- Label 0 indicates that the image used for the entity is incorrect (e.g., the 2001 Championship logo for the 2004 Championship, an invalid image, or an incomprehensible image)
- Label 1 indicates the image represents the entity but is ambiguous for a human to identify (e.g., a generic stadium, an F1 racer with an obscured face, a current logo of a previously renamed company, or an unrecognizable town/city collage with generic buildings/images)
- Label 2 indicates the image clearly represents the entity (e.g., a company logo, a country's flag, or a recognizable monument)

We report the annotation results in Table 5.6.

	Week #	Theme	Song choice	Original artist	Order #	Result
0	Top 24 (12 Men)	1960s	" Daydream The Lovin' Spoonful "		11	Safe
1	Top 20 (10 Men)	1970s	" I just want to be your everything ANDY GIBB "		2	Safe
2	Top 16 (8 Men)	1980s	" Hallelujah "	Leonard Cohen	7	Safe
3	Top 12	Lennon–McCartney	"  "		4	Safe
4	Top 11		" THE BEATLES MICHELLE GIRL "		8	Safe
5	Top 10	Year They Were Born	" Fragile "	Sting	2	Bottom 3
6	Top 9		" "		4	Safe
7	Top 8	Inspirational Music	" "	Judy Garland	3	Safe
8	Top 7		" "		7	Safe
9	Top 6	Andrew Lloyd Webber	" Memory "	Elaine Paige	2	Safe
10	Top 5	Neil Diamond	" Forever in Blue Jeans " "September Morn"	Neil Diamond	1 6	Safe
11	Top 4	Rock and Roll Hall of Fame	" I Shot the Sheriff " " Mr. Tambourine Man "	Bob Marley Bob Dylan	3 7	Eliminated

Question What is the Song choice when The Beatles were the original artist, with an order # of 4?

Answer if i fell

Figure 5.6: Example from the dataset

Image Category	**Data Source**		
	WikiSQL	**WikiTableQuestions**	**FeTaQA**
Human	16,915	6,305	10,043
Location	4,518	3,082	3,779
Seals	738	356	478
Coat of Arms	703	460	779
Flags	1,149	831	1,158
Poster	751	455	5,446
Logo	4,572	2,380	5,628
Misc.	5,856	1,518	8,372

Table 5.5: Categories of the images in the dataset

Data Source	**No Agreement**	**Label 0**	**Label 1**	**Label 2**
FeTaQA	0.28% (14)	0.00% (0)	0.08% (4)	99.64% (5030)
WikiSQL	0.00% (0)	0.10% (6)	0.04% (2)	99.86% (5688)
WikiTableQuestions	0.43% (40)	0.26% (24)	0.13% (12)	99.19% (9282)

Table 5.6: Dataset Validation Statistics

Chapter 6

Using MLLMs for Inference and Reasoning on MultiModalTabQA

This chapter is adapted from the upcoming publication *"Knowledge-Aware Reasoning over Multimodal Semi-structured Tables"* under review at an NLP venue. This work is a joint effort with: Jainit Sushil Bafna (IIIT Hyderabad), Kunal Kartik (IIT Guwahati), Harshita Khandelwal (UCLA), Prof. Manish Shrivastava (Prof, IIIT Hyderabad), Vivek Gupta (Postdoc, University of Pennsylvania), Prof. Mohit Bansal (Prof, University of North Carolina) and Prof. Dan Roth (Prof, University of Pennsylvania). This chapter incorporates the methodology and experiments outlined in the version of the paper dated 6th May 2024.

This chapter describes the different experiments we perform on the MultimodalTabQA dataset. First, we detail various baselines we evaluate on the MultimodalTabQA dataset and explain the metrics we use to assess the model performance. Subsequently, we report preliminary results on different question types (explicit, answer-mention, and implicit) for the three data sources (WikiTableQuestions, WikiSQL, and FeTaQA) and discuss the key insights gained from these experiments.

6.1 Introduction

After creating the dataset for the task, we now evaluate different Multimodal LLM-based baselines on our dataset. Due to the lack of resources, we only explore zero-shot and few-shot approaches to prompting the LLMs and MLLMs involved in the baselines. We perform **preliminary experiments** with the following approaches:

1. **Partial Input Approach:** In this baseline, images are withheld, and only the table with replaced image tags alongside the question is given to the model. The model is thus prompted to generate an answer solely based on the textual input.

2. **Captioning Approach:** In this baseline, we replace each image tag with a caption/entity replacement for the particular image obtained from a Vision Language Model. Then, we prompt another LLM to answer the question based on the text-only table.

3. **Multimodal Approach:** In this approach, we provide the model with the entire table represented as a single image and prompt it to perform question-answering over the image itself.

6.2 Metrics Used for Evaluation

Since the answers involve real-world entities, we favor lexical matching metrics over semantic matching metrics. LLMs typically predict entities of the same or similar class in our entity-centric QA task, all of which would yield high score on semantic matching metrics (such as BertScore [112]), making them ineffective for our purposes.

6.2.1 Short-Answer Data Sources

For the short-form answer questions recast from WikiTableQuestions and WikiSQL dataset, we use the following metrics:

- **Exact Match** is used to report the percentage of predicted answers that were exactly matched with the gold answers.

- **Substring Match** is used to report the percentage of predicted answers which contain the respective gold answer as a substring. This is very relevant since the entity texts are replaced by images, which might lead to the answer being in a different form than what was written in the original table/corresponding gold answer.

- **F1-score** is computed between the tokens of the predicted answer and the gold answer. Since this is the Harmonic Mean of precision and recall on a token level, it also takes hallucination into account as a metric.

6.2.2 Long-Answer Data Sources

For the long-answer questions in the FeTaQA dataset, we use the following metrics:

- **SacreBLEU** is used, which is the BLEU score computed with a fixed set of parameters. This score uses a modified precision calculation over the n-grams in predicted and gold text, which penalizes over-reliance on a single n-gram. It also incorporates a brevity penalty to discourage overly short answers.

- **ROUGE - (1,2,L)** is used, which is a family of metrics which also uses N-grams for computation. ROUGE-1 and ROUGE-2 compute the recall of unigram and bigram in the predicted text and the reference texts, while ROUGE-L uses the Longest Common Sequence and is computed as an F1-score that combines recall (ratio of LCS length to reference length) and precision (ratio of LCS length to summary length).

- **BLEURT** is used, which is computed between the predicted and reference text in the form of a similarity score between encodings of predicted and reference text obtained through pre-trained Transformer models.

6.3 Baseline I: Partial Input

In this baseline, we exclude the images from the input, and only provide the textual table and question to the model for question answering. This baseline serves as a lower bound to other baselines involving images, as ideally, such models should perform better than those provided with no image.

For this experimental setup, we make use of Gemini-1.0 Pro, providing it with few-shot examples from the respective data source and asking it to answer the question. We don't fine-tune the model, opting instead to rely on few-shot examples and the task description for the model to generalize to new samples. The results for the different data sources are in Table 6.1, 6.2, 6.3.

Question Type	**Exact Match**	**Substring Match**	**F1-score**
Explicit	30.36%	35.80%	0.3656
Answer-Mention	19.06%	20.16%	0.3004
Implicit	61.90%	66.67%	0.6190

Table 6.1: Results of Text-only baseline on WikiSQL

Question Type	**Exact Match**	**Substring Match**	**F1-score**
Explicit	32.48%	35.18%	0.3485
Answer-Mention	18.38%	19.10%	0.2682
Implicit	38.72%	39.58%	0.3955

Table 6.2: Results of Text-only baseline on WikiTableQuestions

Question Type	**sacreBLEU**	**ROUGE1**	**ROUGE2**	**ROUGEL**	**BLEURT**
Explicit	25.85	0.5332	0.3433	0.4460	-0.3350
Answer-Mention	20.39	0.5226	0.3101	0.4304	-0.3539
Implicit	21.24	0.5122	0.3125	0.4301	-0.2095

Table 6.3: Results of Text-only baseline on FeTaQA

Some insights from these results

1. The relatively impressive performance of this baseline on different datasets highlights the **parametric real-world knowledge** that is encoded in LLMs today. Even when the entity texts were not provided, the entities could be guessed from the context of the table and the questions were answered.

2. As expected, the accuracy in Answer-Mention questions is much lower than the corresponding F1-score, as in these cases, the entire answer string is replaced by an image in the table, which leads to different forms of the same entity (like M.K. Gandhi instead of Gandhi). Such incorrect forms, despite being semantically correct, might get marked incorrect.

3. Answer-Mention questions seem more challenging than Explicit questions, probably because generating a correct answer is more complex than approximating it during intermediate steps and providing an answer for an entity with a textual mention in the table during subsequent steps.

4. The Implicit questions appear to be easier than the Explicit and Answer-Mention questions. This is perhaps because entity disambiguation, which is particularly hard, is not an integral part of those questions, making the performance better for them.

6.4 Baseline II: Image Captioning

In this baseline, we break down the problem into two individual steps:

1. **Entity Prediction:** We first predict the entity corresponding to the different image occurrences. In order to do this, we create an infobox-style table of the row in which the image to be captioned occurs. Similarly, we create infobox-style tables for a few other rows where the cells corresponding to the column of interest are textual only. Subsequently, we instruct MLLMs to use the context of these infoboxes along with the provided image to predict the original text corresponding to the image, using Chain of Thought prompting. We experiment with Gemini-1.0 Pro-Vision to run these experiments in a few-shot setting (without fine-tuning) and these entity prediction results are in Table 6.4.

2. **Question Answering:** Following the previous step, we have a Table T, question Q and a set of predicted entities corresponding to image tags in the table E. We use Chain of Thought prompting on the LLM to generate the answer to Q using T while considering separately provided E, which might be accurate or inaccurate. We also provide few-shot examples in the prompt but don't include the reasoning in those examples as it seemed to hinder model performance by restricting reasoning directions that the model explored. We use the Gemini-1.0 Pro model to run this step of the experiment with the few-shot methodology described above.

Data Source	Exact Match	Substring Match	F1-score
WikiSQL	45.18%	57.52%	0.5666
WikiTableQuestions	43.86%	53.06%	0.5432
FeTaQA	46.61%	59.16%	0.6043

Table 6.4: Entity Prediction Results for Captioning Baseline

Question Type	Exact Match	Substring Match	F1-score
Explicit	30.25%	35.42%	0.3530
Answer-Mention	22.20%	36.86%	0.3846
Implicit	52.38%	55.56%	0.5238

Table 6.5: Results of Captioning baseline on WikiSQL

Question Type	Exact Match	Substring Match	F1-score
Explicit	35.04%	44.11%	0.4111
Answer-Mention	22.52%	38.20%	0.4222
Implicit	29.51%	44.44%	0.3922

Table 6.6: Results of Captioning baseline on WikiTableQuestions

Question Type	sacreBLEU	ROUGE1	ROUGE2	ROUGEL	BLEURT
Explicit	20.10	0.4214	0.2552	0.3529	-0.7151
Answer-Mention	15.44	0.4175	0.2393	0.3469	-0.7811
Implicit	15.02	0.3457	0.1900	0.2918	-0.7422

Table 6.7: Results of Captioning baseline on FeTaQA

The results for this baseline are in Table 6.5, 6.6, 6.7. Some insights from these results are listed below:

1. In the case of the short-answer metrics, there is a clear improvement from the Text-only baseline, indicating that images form an integral part of the task.

2. The increase in Answer-Mention questions is much higher, which is expected since earlier, there was minimal context about the entity which was to correspond to the answer. While the context

about the entity is important for explicit questions too, it is much more important when the entity corresponds to the answer. This further highlights the importance of image data for the task.

3. However, for implicit questions, we observe some dip from Baseline I. This is possibly because the information about entities was not very relevant to these questions, as the multimodal cells are only used during intermediate reasoning in these questions. Thus, the additional image-entity information instead acted as noise during inference.

4. For FeTaQA, the metrics don't clearly outline the entity-retrieval challenges the previous baseline would've faced, and their fall might correlate to extra noisy information about entity predictions being provided. In future, we should evaluate this data source with better metrics that are more relevant to the task.

6.5 Baseline III: Multimodal Table

In this baseline, we create an image of the table containing all entity images embedded within it. This multimodal table-question input, comprising the table image and question, is directly provided to the model. We also include few-shot question-answer examples to explain the answer format to the model and evaluate it. For this baseline, we once more utilize pre-trained Gemini-1.0 Pro-Vision in our experimentation. The results for this baseline are in Table 6.8, 6.9, 6.10.

Question Type	Exact Match	Substring Match	F1-score
Explicit	20.10%	25.15%	0.2661
Answer-Mention	17.18%	21.02%	0.2840
Implicit	20.63%	36.50%	0.2222

Table 6.8: Results of Multimodal Table baseline on WikiSQL

Question Type	Exact Match	Substring Match	F1-score
Explicit	19.14%	29.56%	0.2541
Answer-Mention	16.63%	20.63%	0.2540
Implicit	25.64%	41.03%	0.3162

Table 6.9: Results of Multimodal Table baseline on WikiTableQuestions

Some insights from these results:

Question Type	sacreBLEU	ROUGE1	ROUGE2	ROUGEL	BLEURT
Explicit	6.026	0.2633	0.1576	0.2330	-0.7857
Answer-Mention	4.634	0.2688	0.1402	0.2327	-0.7857
Implicit	4.131	0.2170	0.1163	0.1874	-0.7346

Table 6.10: Results of Multimodal Table baseline on FeTaQA

1. This baseline performs worse than Baseline I for implicit questions, which might indicate that parsing semi-structured information directly from table images is still challenging for Multimodal LLMs.

2. Further, the performance on explicit questions is worse, indicating that entity disambiguation for multiple images contained within a single image is a hard problem for MLLMs. Further, since this is even worse than the text-only model, it might indicate that the parametric real-world knowledge of this model is not as good as the text-only model.

3. The performance on answer-mention questions seems similar to the Text-only baseline. This is possibly because the poor understanding of the table structure in some questions might have been countered by the availability of the precise answer-entity in visual form, helping the model.

6.6 Conclusion

These experiments demonstrate that the proposed task of Question-Answering over Multimodal Tables is a **challenging** task, with various aspects that make it hard. We also gain various insights about the nature of the MultimodalTabQA dataset and the LLMs/MLLMs used in our experiments. Our best-performing baseline was the Image Captioning Baseline, beating the other two by a huge margin. Its significant improvement over the Text-only baseline shows that the task and dataset are *truly multimodal* in nature, requiring images as an important component alongside the tabular information.

On the other hand, contrary to expectations, the Multimodal Table baseline performed the worst, highlighting the complex nature of the task. It also showed how difficult it is for MLLMs to handle multiple visual elements simultaneously in a semi-structured format and provides research direction towards models that can handle such complex multimodal information.

We believe this task and dataset can be a strong benchmark for evaluating the visual parsing and semantic understanding of Multimodal LLMs. In future, we plan to evaluate these baselines with a wider variety of models, like GPT-4V [2], QwenVLM [28], CogAgent [29], etc. Further, we also plan to create another baseline which uses models which accept arbitrary sequences of images and text like IDEFICS-2 [113], Flamingo [26], Mantis [108] and benchmark their performance on our dataset. While it is possible they won’t be able to process so many images well together, these experiments may provide us newer insights into the ever-evolving Multimodal space.

Chapter 7

Conclusions

In this thesis, we make advancements towards the understanding of different problems in Multimodality and how we can exploit LLMs to tackle those problems. We first explored the field of Multimodal Emotion Analysis, and performed experiments on the task of Multimodal Emotion Cause Pair Extraction. We made use of fusion-based techniques, concatenating the audio, video and text embeddings and used architectures like BiLSTMs, CRFs and even a simple MLP to benchmark the performance of the fused embedings on the task. We further explored prompting LLMs for complex and unconventional reasoning, making use of few-shot and reasoning-based prompts to make the LLM better at performing the task.

Building upon these works, we proposed the task of Knowledge-aware Question Answering over Multimodal Semi-structured Tables and describe the motivation for exploring the problem. We contribute a large-scale dataset with a diverse set of questions and tables for Multimodal Tabular Question Answering, being the first dataset for the task. We also evaluated Gemini, a SOTA MLLM through three different strategies for the task, and gain various insights from it.

7.1 Future Work

Creating the dataset and benchmarking MLLM through different stratagies led to some interesting results on the task. However, our work has some shortcomings, which can be addressed as future work:

- More datasets based on Wikipedia tables can be recasted into the MultimodalTabQA dataset. For example, HybridQA [88], which contains additional paragraphs apart from just the table context might be interesting to benchmark as an additional-context multimodal table problem. Another dataset, Open-WikiTable [114], which involves Open Domain Question Answering over Table can be repurposed for the task, making the task more complex and interesting.

- While the questions in our dataset require multimodal reasoning, none of them require a visual understanding of the images in the table. In future, we should explore augmenting some existing questions to create visual questions, which instead of explicitly mentioning an entity describe it through its visual attributes in the question.

- We should conduct experiments with Multimodal Models that can take arbitrary sequences of image and text as input, like IDEFICS-2 [113], Flamingo [26], Mantis [108] and benchmark their performance on our dataset. While it is possible they won't be able to process so many images well together, the insights might be interesting.
- We could only use Gemini due to resource constraints for the different types of models. In future, we would extend the experiments to include models like GPT-4V [2], QwenVLM [28], CogAgent [29].
- While we do provide the results on FeTaQA using some Natural Language Generation metrics, these metrics are not very insightful in context of our task. We should try out other metrics which are more relevant to our task for free-form questions.

Related Publications

1. **Suyash Vardhan Mathur***, Akshett Rai Jindal*, Hardik Mittal and Manish Shrivastava. LastResort at SemEval-2024 Task 3: Exploring Multimodal Emotion Cause Pair Extraction as Sequence Labelling Task. *Accepted at 18th International Workshop on Semantic Evaluation (SemEval-2024)*

2. **Suyash Vardhan Mathur***, Akshett Rai Jindal* and Manish Shrivastava. DaVinci at SemEval-2024 Task 9: Few-shot prompting GPT-3.5 for Unconventional Reasoning. *Accepted at 18th International Workshop on Semantic Evaluation (SemEval-2024)*

3. **Suyash Vardhan Mathur**, Jainit Sushil Bafna*, Kunal Kartik*, Harshita Khandelwal*, Manish Shrivastava, Vivek Gupta, Mohit Bansal and Dan Roth. Knowledge-Aware Reasoning over Multimodal Semi-structured Tables. *Under review at an NLP venue*

Bibliography

[1] Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844, 2023.

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

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.

[5] Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

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

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al.

An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] 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 (ICCV)*, December 2015.

[13] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171, 2023.

[14] Ana Cláudia Akemi Matsuki de Faria, Felype de Castro Bastos, José Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. Visual question answering: A survey on techniques and common trends in recent literature, 2023.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[16] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *ArXiv*, abs/1802.05365, 2018.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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.

[18] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[19] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[21] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[26] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[27] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[28] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[29] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023.

[30] Flor Miriam Plaza del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in nlp: Trends, gaps and roadmap for future directions, 2024.

[31] Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *arXiv preprint arXiv:1707.06996*, 2017.

[32] Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, 2019.

[33] Yifan Jiang, Filip Ilievski, and Kaixin Ma. Semeval-2024 task 9: Brainteaser: A novel task defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1996–2010, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[34] Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. BRAINTEASER: Lateral thinking puzzles for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore, December 2023. Association for Computational Linguistics.

[35] Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. A survey on table question answering: Recent advances. In Maosong Sun, Guilin Qi, Kang Liu, Jiadong Ren, Bin Xu, Yansong Feng, Yongbin Liu, and Yubo Chen, editors, *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy*, pages 174–186, Singapore, 2022. Springer Nature Singapore.

[36] Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. Towards building a social emotion detection system for online news. *Future Generation Computer Systems*, 37:438–448, 2014. Special Section: Innovative Methods and Algorithms for Advanced Data-Intensive Computing Special Section: Semantics, Intelligent processing and services for big data Special Section: Advances in Data-Intensive Modelling and Simulation Special Section: Hybrid Intelligence for Growing Internet and its Applications.

[37] Muhammad Abdul-Mageed and Lyle Ungar. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[38] Endang Wahyu Pamungkas. Emotionally-aware chatbots: A survey, 2019.

[39] Jeewoo Yun and Jungkun Park. The effects of chatbot service recovery with emotion words on customer satisfaction, repurchase intention, and positive word-of-mouth. *Frontiers in psychology*, 13:922503, 2022.

[40] Simon D'Alfonso. Ai in mental health. *Current Opinion in Psychology*, 36:112–117, 2020.

[41] Ramit Sawhney, Harshit Joshi, Alicia Nobles, and Rajiv Ratn Shah. Towards emotion-and time-aware classification of tweets to assist human moderation for suicide prevention. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 609–620, 2021.

[42] Ivo Benke, Michael Thomas Knierim, and Alexander Maedche. Chatbot-based emotion management for distributed teams: A participatory design study. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–30, 2020.

[43] Paul Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[44] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.

[45] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[46] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Romila Ghosh, Niyati Chhaya, Alexander F. Gelbukh, and Rada Mihalcea. Recognizing emotion cause in conversations. *CoRR*, abs/2012.11820, 2020.

[47] Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. A text-driven rule-based system for emotion cause detection. In Diana Inkpen and Carlo Strapparava, editors, *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA, June 2010. Association for Computational Linguistics.

[48] Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. Emotion cause detection with linguistic constructions. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[49] Lin Gui, Ruifeng Xu, Qin Lu, Dongyin Wu, and Yu Zhou. Emotion cause extraction, a challenging task with corpus construction. In Yuming Li, Guoxiong Xiang, Hongfei Lin, and Mingwen Wang, editors, *Social Media Processing*, pages 98–109, Singapore, 2016b. Springer Singapore.

[50] Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy, July 2019. Association for Computational Linguistics.

[51] Zixiang Ding, Rui Xia, and Jianfei Yu. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In Dan Jurafsky, Joyce Chai, Natalie

Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170, Online, July 2020. Association for Computational Linguistics.

[52] Penghui Wei, Jiahao Zhao, and Wenji Mao. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online, July 2020. Association for Computational Linguistics.

[53] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008.

[54] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[55] Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. Ecpec: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1754–1765, 2022.

[56] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics.

[57] 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.

[58] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[59] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[60] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics.

[61] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[62] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[63] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pretraining for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[64] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[65] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging, 2015.

[66] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, June 2023.

[67] Xingwei Liang, You Zou, and Ruifeng Xu. Si-lstm: Speaker hybrid long-short term memory and cross modal attention for emotion recognition in conversation. *arXiv preprint arXiv:2305.03506*, 2023.

[68] Shlomo Waks. Lateral thinking and technology education. *Journal of Science Education and Technology*, 6:245–255, 1997.

[69] OpenAI. Chatgpt: A language model by openai. `https://www.openai.com`, 2022.

[70] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[71] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[72] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and

Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[73] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022.

[74] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[75] Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*, 2022.

[76] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*, 2023.

[77] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[78] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg,

S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[81] Sengjie Liu and Christopher G. Healey. Abstractive summarization of large document collections using gpt, 2023.

[82] Hossein Bahak, Farzaneh Taheri, Zahra Zojaji, and Arefeh Kazemi. Evaluating chatgpt as a question answering system: A comprehensive analysis and comparison with existing models, 2023.

[83] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, 2023.

[84] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada, July 2023. Association for Computational Linguistics.

[85] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.

[86] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity disambiguation for noisy social media posts. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[87] Abhirama Penamakuri, Manish Gupta, Mithun Gupta, and Anand Mishra. Answer mining from a pool of images: Towards retrieval-based visual question answering. pages 1312–1321, 08 2023.

[88] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online, November 2020. Association for Computational Linguistics.

[89] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3277–3287, Online, August 2021. Association for Computational Linguistics.

[90] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[91] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[92] 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*, abs/2104.06039, 2021.

[93] Yongqi Li, Wenjie Li, and Liqiang Nie. MMCoQA: Conversational question answering over text, tables, and images. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4220–4231, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[94] Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, and Yongbin Li. Unified language representation for question answering over text, tables, and images. *arXiv preprint arXiv:2306.16762*, 2023.

[95] Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. Mmhqa-icl: Multimodal in-context learning for hybrid question answering over text, tables and images, 2023.

[96] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163, 07 2016.

[97] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. pages 3190–3199, 06 2019.

[98] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022.

[99] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022.

[100] T. Mensink, J. Uijlings, L. Castrejon, A. Goel, F. Cadar, H. Zhou, F. Sha, A. Araujo, and V. Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3090–3101, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society.

[101] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *ArXiv*, abs/2302.11713, 2023.

[102] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

[103] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel M. Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. *ArXiv*, abs/2209.06794, 2022.

[104] Li Zhang, Zhixu Li, and Qiang Yang. Attention-based multimodal entity linking with high-quality images. In Christian S. Jensen, Ee-Peng Lim, De-Nian Yang, Wang-Chien Lee, Vincent S. Tseng, Vana Kalogeraki, Jen-Wei Huang, and Chih-Ya Shen, editors, *Database Systems for Advanced Applications*, pages 533–548, Cham, 2021. Springer International Publishing.

[105] Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. Multimodal entity linking: A new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 993–1001, New York, NY, USA, 2021. Association for Computing Machinery.

[106] Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[107] Abhirama Subramanyam Penamakuri, Manish Gupta, Mithun Das Gupta, and Anand Mishra. Answer mining from a pool of images: towards retrieval-based visual question answering. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[108] Tiger AI Lab. Mantis: A Platform for Accelerating Transformer-based Machine Learning Workflows. `https://tiger-ai-lab.github.io/Blog/mantis`, 2024. Accessed: April 30, 2024.

[109] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*, 2017.

[110] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics.

[111] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022.

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

[113] Hugging Face. idefics: How hugging face's new platform can help you train better language models. `https://huggingface.co/blog/idefics`. Accessed: 2024-04-30.

[114] Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. Open-WikiTable : Dataset for open domain question answering with complex reasoning over table. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8285–8297, Toronto, Canada, July 2023. Association for Computational Linguistics.