M^2-TabFact: Multi-Document Multi-Modal Fact Verification With Visual and Textual Representations of Tabular Data

Mingyang Zhou[1], **Lingyu Zhang**[1], **Sophia Horng**[1],
Maximillian Chen[1], **Kung-Hsiang Huang**[2], **Shih-Fu Chang**[1],
[1]Columbia University, [2]Salesforce AI Research
mz2974@columbia.edu, lz2814@columbia.edu, h4230@columbia.edu,
maxchen@cs.columbia.edu, kh.huang@salesforce.com, shih.fu.chang@columbia.edu

Abstract

Tabular data is used to store information in many real-world systems ranging from finance to healthcare. However, such structured data is often communicated to humans in visually interpretable formats (e.g. charts and textual paragraphs), making it imperative that fact-checking models should be able to reason over multiple pieces of structured evidence presented across different modalities. In this paper, we propose Multi-Document Multi-Modal Table-based Fact Verification (M^2-TabFact), a challenging fact verification task that requires jointly reasoning over visual and textual representations of structured data. We design an automatic data generation pipeline that converts existing tabular data into descriptive visual and textual evidence. We then use Large Language Models to generate complex claims that depend on multi-document, multi-modal evidence. In total, we create 8,856 pairs of complex claims and multi-modal evidence through this procedure and systematically evaluate M^2-TabFact with a set of strong vision-language models (VLM). We find that existing VLMs have large gaps in fact verification performance compared to humans. Moreover, we find that they are imbalanced when it comes to their ability to handle reason about different modalities, and currently struggle to reason about information extracted from multiple documents.

1 Introduction

Structured data are widely used to organize and present information in various settings ranging from web pages to spreadsheets and infographics. In an age where misinformation and hallucinated text generation continue to spread rapidly on the internet (Gao et al., 2021), building autonomous systems that can verify factual claims against structured data will lead to the reduction of misinformation and a safer experience on the internet.

Recently, several benchmarks have been proposed to evaluate automatic fact verification systems' ability to reason over structured data (Chen et al., 2020; Wang et al., 2021). However, two main limitations still remain. First, structured data in documents are commonly presented in complex but interpretable formats (e.g., visual chart plots or natural language summaries) rather than simple tables. Second, existing table-based fact-checking systems built for these benchmarks simply assume that all of the evidence is contained in a single document or source. This is different from real-world scenarios, where human fact-checkers typically need to review multiple structured data evidence sources to evaluate the truthfulness of a complex claim.

To address the above limitations, we propose Multi-Document Multi-Modal Table-based Fact-Checking (M^2-TabFact), a benchmark task which requires table-based fact verification systems to reason about information from multiple sources of structured data represented in multiple modalities. An example instance from our corpus is given in Figure 1. Given a text hypothesis, the task is to verify the truthfulness of this claim against a chart and a text paragraph converted from two associated structured tables. Solving this task entails decomposing the claim into two simpler pieces of information to retrieve, identifying the evidence from each modality necessary to retrieve the corresponding information, and then finally combining the information obtained from the two pieces of evidence to make a final decision on whether the claim is factual.

M^2-TabFact is constructed through an automatic pipeline involving four high-level steps. (1) *Evidence Table Collection:* we split a table into two sub-tables to construct two plausibly related source tables for multi-document, multi-modal evidence creation. (2) *Multi-hop Claim Creation:* we sample various data points from each source table and generate multi-hop claims that require multiple rea-

Findings of the Association for Computational Linguistics: ACL 2025, pages 26239–26256

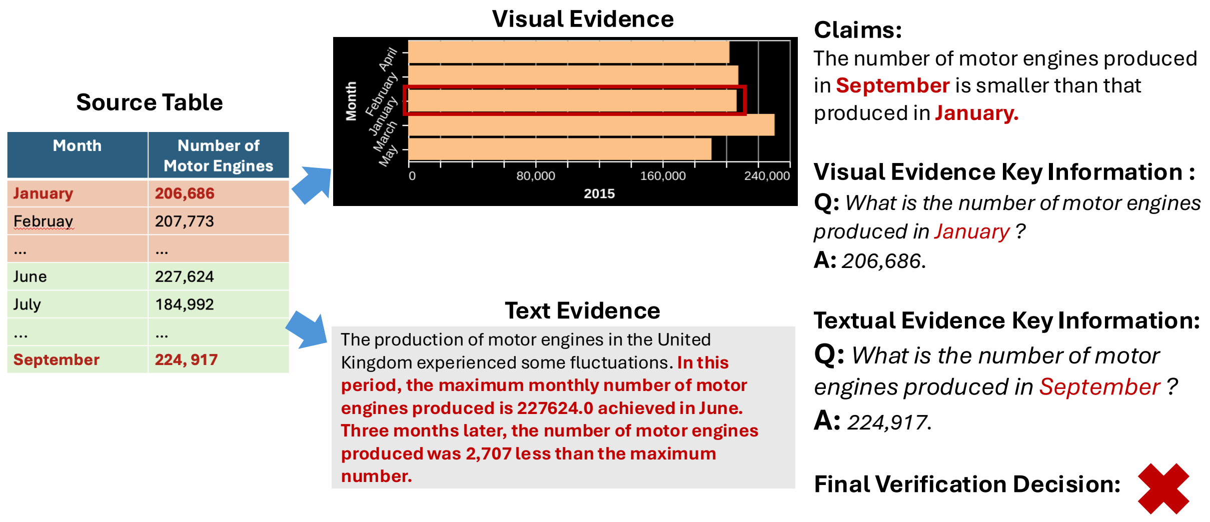


Figure 1: An example of M^2-TabFact where the claim verification requires: 1) parsing the key information of the claim; 2) verifying the key information against each piece of uni-modal evidence. 3) combining the verified key information from uni-modal evidence to assess the claim's truthfulness. The key information is highlighted in red.

soning operations between the sampled data points via pre-defined templates. (3) *Multi-modal Evidence Creation:* we convert one source sub-table into a chart using Data Visualization tools and the other sub-table into a text summary. (4) *Paraphrasing:* we prompt Large Language Models (LLM) to paraphrase the template-based text claim and text evidence into more diverse and fluent language.

To verify the unique challenges presented by our new dataset, we empirically evaluate a set of strong Vision and Language Models (VLMs) on M^2-TabFact, and compare their evaluations to human-level performance. We find that our dataset poses a great challenge to existing VLMs. The strongest model evaluated only achieves slightly less than 60% accuracy, significantly lagging behind human-level performance 88%. Hence, M^2-TabFactis a challenging problem and will stimulate progress on fact-checking against multi-modal structured data.

The contributions of our paper are as follows:

- We introduce M^2-TabFact, a large-scale fact-checking dataset consisting of 8,856 claim and evidence pairs constructed from multi-source, multi-modal tabular data.
- We propose an automatic pipeline to construct this dataset at scale.
- We systematically analyze the limitations of existing SOTA Vision and Language Models on this task and suggest future directions.

2 Related Work

2.1 Evidence-based Fact Checking

The task of predicting the truthfulness of a supposedly factual claim against evidence has been widely explored in the natural language processing research community. The majority of existing evidence-based fact-checking work focuses on text-based evidence (Thorne et al., 2018; Jiang et al., 2020; Augenstein et al., 2019; Kotonya and Toni, 2020; Wadden et al., 2020; Saakyan et al., 2021). As much information on the internet is disseminated in other modalities (e.g. infographics), there is naturally a growing interest in developing automated fact-checking (AFC) systems that can process evidence in other modalities such as images (Boididou et al., 2015; Fung et al., 2021; Jindal et al., 2020; Nakamura et al., 2019; Raj and Meel, 2021) and videos (Micallef et al., 2022; Papadopoulou et al., 2019; Rayar et al., 2022).

This has led to the development of fact-checking benchmarks that require grounding on multi-modal evidence (Mishra et al., 2022; Nielsen and McConville, 2022; Yao et al., 2023). While most of these multi-modal fact-checking benchmarks sourced their documents from news and social me-

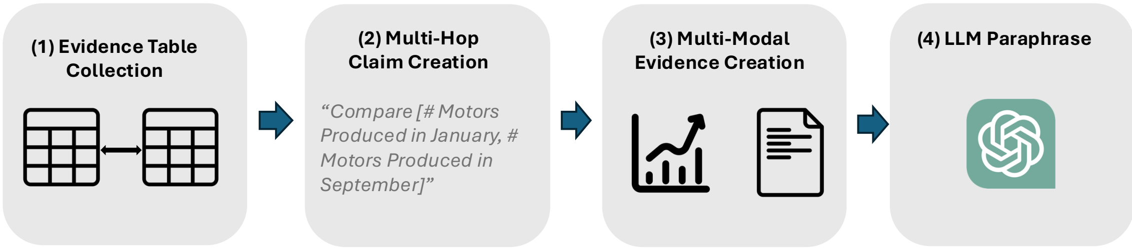


Figure 2: An overview of the automatic data generation pipeline for M^2-TabFact that involves four steps.

dia posts, in our work we explore multi-modal formats of structured data as the evidence source which is more often used in scientific papers, technical reports, and infographics.

2.2 Structured Data Fact Checking

Chen et al. (2020) proposed the first table-based fact-checking benchmark by collecting tables from Wikipedia as evidence and asking crowd workers to create contrasting claims where one does and does not contradict the source. Wang et al. (2021) extracted table evidence from scientific articles and created claims based on the sentences in the article that describes those same tables. There is also another line of research that collects fact-checking datasets from charts, a commonly used visual representation of tabular data. ChartFC (Akhtar et al., 2023a) extends TabFact dataset (Chen et al., 2020) by converting a subset of the table evidence into bar charts via visualization libraries. Following this work, ChartCheck (Akhtar et al., 2023b) collects real-world charts from the internet extending the coverage of more chart types. These existing efforts are still limited to a single document and a single modality setting. In contrast, the complex fact-checking process conducted by humans in real-world applications usually requires checking multiple different resources such as figures, tables and articles in a research paper. To address this need for more realistic fact-checking procedure, we construct the first multi-modal multi-hop fact-checking dataset grounded on tabular evidence.

3 M^2-TabFact: Data Creation

In this section, we introduce our automatic pipeline to systematically create a challenging, large-scale multi-modal fact verification dataset pairing textual claims with multi-modal structured data. Figure 2 provides an overview of the data creation process. As introduced in Section 1, the whole process is broken down into four high-level steps: (1) Evidence Table Collection; (2) Multi-Hop Claim Creation; (3) Multi-Modal Evidence Creation (4) LLM Paraphrasing. We introduce details of each step in the following section.

3.1 Evidence Table Collection

Our goal is to create a diverse multi-modal fact verification dataset from real-world tabular data, where the tabular data is suitable to be converted to both a chart plot and text summary. Therefore, we collect our tabular data resources from existing Chart Captioning or Chart Summary datasets that provide paired tabular data annotations. Our seed chart datasets include Vistext (Tang et al., 2023) and Chart-to-Text (Kantharaj et al., 2022). The tables for both datasets are crawled from Statista.com and cover a diverse set of topics including technology, trade, retail, and sports. They also cover a diverse set of chart types including: pie chart, line chart, bar chart, and area chart. We filter out tables that contain crushed values and are left with 8,856 source tables.

After collecting our source evidence tables, we need to create claims that depend on two pieces of evidence sources. In order to obtain two plausibly related tables, we split each original source table into two sub-tables. Each table is composed of two parts: the data and the title. For the data, we perform a column-wise or row-wise splitting strategy from the middle point of the table to ensure the information contained in the two sub-tables is relatively balanced. While the original title can be directly inherited by the sub-tables in the majority of the time, there are several cases where the sub-table titles should be adjusted accordingly. For example, titles that cover time-range of the table content (e.g. *"The average Boston Celtics ticket price from 2010 to 2020"*) or titles that cover all the categorical values of the row header or column header (e.g *"The total number of bilingual speakers in England, France, and Spain in 2024."*) need to be adjusted based on the time range or the categorical values found in the sub-tables data. To address this issue, we build a classifier to classify whether the title of the sub-tables needs to be changed from the original title and then map the title to the adjusted version using an LLM. We summarize the process of editing sub-table titles in Appendix A.1.

3.2 Multi-Hop Claim Creation

After we split source tables into two sub-tables, we adopt a template-based method to create claims that require information from both sub-tables to verify. We first parse the sub-tables into a more manipulable format. We extract key-value pairs and their units from the table. An example of this is shown in Fig 3.

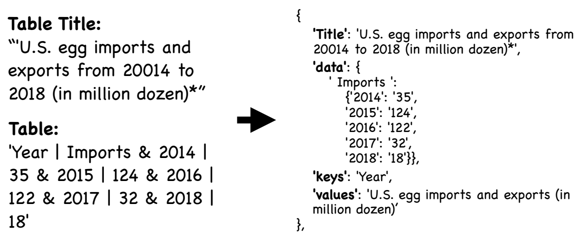


Figure 3: Parsing tables into dictionaries. We convert columns and rows of the table into key-value pairs.

To create a diverse set of claims, we designed 3 groups of multi-hop reasoning composition types: **Compare**, **Coref**, and **Math**. Each consists of a set of diverse templates using the key-value pairs extracted from the tables.

Compare requires comparing values from two individual sub-tables. We randomly sample a key-value pair from each of the sub-tables and create a true claim based on the relative percentage of the values (Table 1 row 1). Negative claims are simply reversing the relationship between values.

Coref requires identifying entities that are referenced across sub-tables. We filter out pairs of sub-tables that contain the same entities (e.g. Table1 and Table2 both contain the year "2020"). We create a claim by randomly sampling key-value pairs from both tables under the same entity. Negative claims are created by swapping values with randomly sampled values from the table.

Math requires performing algorithmic operations. We created 10 templates involving different mathematical operations, including the sum of all entries, sum of two entries, maximum, median, mean etc. Negative claims are generated by calculating results with randomly sampled entries or with missing or additional entries.

All multi-hop claims can be decomposed into two sub-claims, each requiring information from only one of the sub-tables. We prompt an LLM to convert these sub-claims into QAs.

Detailed templates will be released with our code.

Composition type	Template
Compare	`The {unit} of {k2} is {x}% larger than {k1}`
Coref	`The {entity_type} that {k1} has {v1} in {unit}, {k2} has {v2} in {unit}`
Math	`The total {unit} of all {x_axis}s in {chart_title} and {table_title} is {sum}`

Table 1: Templates for claim creation

3.3 Multi-Modal Evidence Creation

Visual Evidence To create the visual evidence, we convert one sub-table to a chart plot via existing data visualization tools. For tables from Vistext, as they provide the metadata to plot the chart from the original table with the Vega-Lite visualization library (Satyanarayan et al., 2017), we simply replace the original table data with the sub-table data in the meta-file to create the chart plot with the same tool. For tables from Statista, we use the Matplotlib library to plot the sub-tables into the same chart type as the chart type of the original table, also applying one of the 24 visual themes provided by the library. Details can be found in Appendix A.2.

Textual Evidence For the other sub-table, we convert it into a natural language summary as textual evidence. Learning from the human-annotated chart summary from Vistext and Chart-to-Text, we create a set of templates to compile key summary statistics such as variation trends over time, min, max, and mean values from the given sub-table. Besides this general key information, the text summary must also capture the sub-table's sampled data information used to create the multi-hop claim (e.g, in Figure 1, the text summary needs to capture the number of motor engines produced in September — 224,917). However, simply mentioning this sampled data point in the textual summary will make it fairly easy for the model to detect. Thus, we create templates to present a numerical connection of the sampled data information to one of the general key summary statistics captured from the sub-table (e.g the number of produced motor engines in September is 2,707 less than the maximum number), which forces the model to perform multi-hop reasoning in order to accurately identify the sampled data information. Additional details of textual evidence generation are included in Appendix A.3

3.4 LLM Paraphrasing

Finally, we leverage a highly capable LLM to improve the language diversity and fluency of the claim and textual evidence which were automatically generated from the pre-defined templates. The prompts we used for paraphrasing are presented in Appendix A.4. The prompt requests a rewritten version of a given sentence that is more natural with corrected grammar, while preserving the original content. We include a few examples in the prompt as context.

3.5 Quality Control

After we synthesize the dataset, we conduct human evaluation on a small subset of the generated data to understand the quality. We randomly check 100 examples of the generated data and verify three things: (1) is the claim verifiable by the two pieces of evidence; (2) does the verification require key information from both evidence modalities. (3) does the generated multi-modal evidences have factual inconsistency with the original table evidences. In our final version of the data, 92% of claims are verifiable from the given two pieces of evidence. All the sampled data will require joint interpretation from both modalities and none of the generated evidence in the sampled data has conflicts with the original table evidence. This demonstrates that the proposed pipeline can generate high-quality multi-modal multi-hop fact verification dataset over structured table evidence.

Split	Compare	Coref	Math	Total
train	2924	2015	2579	7518
val	172	118	151	441
Test	344	238	305	897

Table 2: M^2-TabFact Statistics on the distribution of different compositional claims and corresponding train, validation, and test split.

3.6 Dataset Statistics

Table 2 gives an overview of our M^2-TabFact dataset. For each unique table, we generate one claim, visual evidence piece, and textual evidence piece, resulting in a total number of 8,856 unique data samples. There are 4,397 supported and 4,449 refuted claims, which are relatively balanced. The table summarizes the distribution of the three different multi-hop compositional claims defined in section 3.2. We separate the data into train, val and test split using a 85/5/10 ratio.

4 Experiments and Results

4.1 Task Definition

We define our task of verifying factual claims against multi-modal structured table evidence as follows. Each instance i is represented by the tuple (c_i,v_i, t_i, y_i) consisting of a natural language claim c_i, a piece of visual evidence representing a structured table v_i, textual evidence representing data from a structured table t_i, and a claim label $y_i \in \{0, 1\}$ which represents whether c_i is supported ($y_i = 1$) or refuted ($y_i = 0$) by the two pieces of evidence (v_i, t_i). Each claim is also associated with two questions (q_i^v, q_i^t) and their corresponding answers (a_i^v, a_i^t), where each question asks about the key information to verify the claim from a piece of uni-modal evidence. These uni-modal question-answer pairs form a subtask which is useful in verifying whether the result of the final claim is supported by the correct intermediate reasoning.

4.2 Baselines

We evaluate several strong vision and language methods on M^2-TabFact, which can be grouped into two categories: (1) domain-specific chart-based vision language models (C-VLMs) that are tailored towards chart understanding tasks; or (2) large foundational vision language models (LVLMs) that are universally powerful generalists for a diverse set of multi-modal tasks.

The C-VLMs include Pix2struct (Lee et al., 2023), MATCHA (Liu et al., 2023), and, UniChart (Masry et al., 2023). All three models use a similar generative encoder-decoder architecture with different pre-training tasks. Pix2struct is pre-trained on HTML code generation from web screenshots and achieves strong performance across multiple document understanding tasks. MATCHA is a version

Model	Setting	Compare	Coref	Math	Overall
Pix2Struct	SFT	48.8	48.7	52.8	50.2
MATCHA	SFT	45.1	75.2	52.1	55.6
UniChart	SFT	52.3	76.5	53.4	59.2
LLAVA	Prompt	49.7	48.3	51.6	50.0
Gemini	Prompt	55.7	48.1	48.4	51.1
GPT-4o	Prompt	62.5	48.3	51.8	55.0
Human	-	90.4	89.1	86.5	88.2

Table 3: Fact Verification Results on M^2-TabFact with different VLM Methods and Human Evaluator.

of Pix2Struct which is pre-trained with two additional tasks, chart-to-table translation and mathematical reasoning, and further fine-tuned for chart reasoning. UniChart is pre-trained on a set of diverse Chart understanding tasks which achieves strong performance on multiple chart understanding datasets, including ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020), and Chart-to-Text (Kantharaj et al., 2022).

For LVLMs, we evaluate GPT-4o (Achiam et al., 2023), Gemini 1.5 Pro (Gemini Team et al., 2024), and LLAVA 1.6 (Liu et al., 2024)[1], which represents a group of frontier generalist vision and language models on various benchmarks.

4.3 Experimental Set-up

All baseline models are evaluated on the primary Multi-Modal Fact Verification task which evaluates c_i against the two pieces of multi-modal evidence (v_i, c_i). The distribution of labels for supprted and refuted claims are fairly balanced (49.6% vs 50.4%), we measure task performance in terms of accuracy.

The C-VLMs are fine-tuned with cross-entropy loss (SFT) on the training split of M^2-TabFact and then evaluated on the test split. For all models, we fine-tune for 10,000 steps, using a batch size of 8, on 4 NVIDIA Titan RTX GPUs. We use AdaFactor with a learning rate of $1e-5$, and use cosine scheduling with 1000 warm-up steps.

The LVLMs are directly evaluated with zero-shot inference on the test-split. The LVLM is asked to provide the final verification result after generating an intermediate reasoning chain, similar to work on unimodal reasoning (e.g. Wei et al. (2022)). The prompt is available in Appendix B. We also evaluate human-level performance on this task by asking two human evaluators to each verify 50 claims sampled from the test split. We ensure each type of compositional claim is equally represented in this test batch and the distribution of supported and reputed claims is also balanced. We find that human-level performance on this task is 88.2 %. We also observe agreement on over 90 % of their decisions on claim verification. The failed cases are mainly due to the difficulty of identifying the correct data values from visual evidence and sometimes the gap between the incorrect value in the refuted claim is too close to the correct value.

[1]We use the LLAVA 1.6 13b model for our experiment

4.4 Results and Discussion

Table 3 provides an overview of all of our benchmarking results on our multi-modal, multi-hop claim classification task. We analyze both the overall performance and the performance on the individual claim types.

Overall, we find that this task is quite challenging for existing VLMs. Even highly specific Chart-VLMs with further fine-tuning on our downstream task can only reach 59.2% accuracy, and frontier LVLMs are only able to reach 55.0% accuracy. In contrast, humans can achieve close to 90% on this task, indicating that the task is solvable with robust reasoning.

We also find that claims that require different compositions of multi-modal evidence demonstrate different degrees of challenges to the baselines. For frontier LVLMs like Gemini and GPT-4o, the claims that require comparison of data value across the multi-modal evidence seem to be the easiest type of claim to handle, whereas mathematical reasoning appears to be more challenging . This also aligns with the pattern of human performance across the three different types of claims. We think this is because comparison only requires a coarse-level estimation of data while arithmetic operations will need accurate computation for every data value. However, we observe a different pattern on the performance on the fine-tuned CVLMs. We observe that fine-tuning C-VLMs like MATCHA and UniChart is specifically helpful to the performance of co-reference evaluation where they both outperform the best performing LVLM by at least 26.9%. LVLM struggles to perform well on co-reference type claims due to a bias in its reasoning process to collect key information of a queried subject from one evidence. While supervised finetuning are useful for improving the multi-modal reasoning capabilities on these tasks and even surpassing frontier LVLMs with smaller specialized models, the current gap from human-level performance indicates that the standard tuning approaches are still insufficient for learning multi-modal multi-hop reasoning.

Model	V-QA	T-QA	Accuracy
MATCHA*	18.9	18.7	54.5
UniChart*	19.8	18.5	57.3
Gemini	29.5	18.1	51.1
GPT-4o	41.6	40.3	55.0

Table 4: Evaluation results on uni-modal evidence question answering for M^2-TabFact. We present the relaxed accuracy for uni-modal evidence question answering and the overall accuracy of final claim verification. The fine-tuning process of Matcha* and Unichart* is different from that in table 3, where we finetune the model on both uni-modal question answering tasks and final claim verification tasks.

4.4.1 Understanding of Uni-modal Evidence

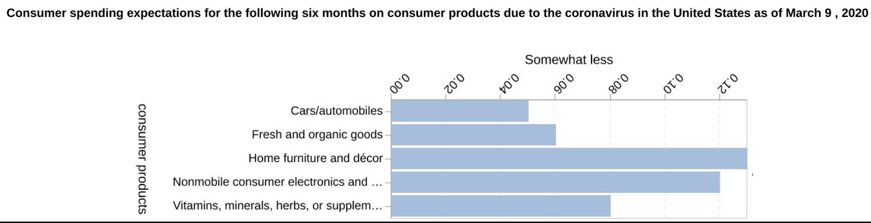
The ability to verify each claim against multi-modal evidence requires the model to first accurately verify the key sub-components of the claim against its corresponding uni-modal evidence (i.e. either the chart or the text paragraph). Thus, we also evaluate the model's capability to predict the key sub-information from the final claim that corresponds to uni-modal evidence via the question-answering task. An example is included in Figure 1. To verify the truthfulness of the final claim, the model needs to be able to answer the question *what is the number of motor engines produced in January?* via checking the chart. Additionally, the model should also be able to tell the number of engines produced in September from the text evidence.

Table 4 summarizes our findings on the uni-modal reasoning capability of the best-performing models on our task. We measure the model's ability to accurately answer the question against the chart or text evidence. For questions with golden answer of numerical value , we allow an error tolerance of 5% of the generated answer to be considered correct following (Masry et al., 2022). For all other answer types, we require an exact match in order for a candidate prediction to be considered correct.

We find that GPT-4o displays the strongest uni-modal evidence understanding capability, Although it is worse than Uni-Chart when it comes to producing the final prediction, we find that the prediction result of GPT-4o is more consistent with its intermediate uni-modal evidence understanding and the Uni-Chart final verification performance is less reliable. We also notice that textual understanding is comparatively more difficult than the visual understanding in our task. Gemini's performance on textual evidence question answering is 9.4% lower than its performance on chart evidence answering, and both GPT-4o and UniChart display a 1.3% performance degradation on the textual evidence question-answering task. The difficulty of the textual evidence may be due to the multi-hop reasoning required to identify the key information which was introduced in the data generation procedure in Section 3.3.

4.4.2 Qualitative Error Analysis

To further identify the limitations of existing VLMs, we have gone through 60 examples where the best-performing VLM: GPT-4o fails to make the correct verification. From these failure predictions, we observe that there are two major limitations of the existing VLMs on this task. (1) **Failure to extract correct data value from charts**: we find that most of the time when the model fails to verify a claim that requires accurate numerical operation, it is due to the model's inaccurate data extraction from chart evidence. For example, in Fig 4, GPT-4o mistakenly interpret the data value for home furniture and decor as 0.12 which should be 0.13 instead. This is especially true when more than 10 data points are contained in the chart evidence. (2) **Limited capability to compare data points across multiple pieces of evidence**: these examples indicate that GPT-4o may only account for data points that originate from the same source, and mistakenly think that the data value of the subject from another evidence source is not provided. As shown in Fig 5, in the claim "The difference between Count von Count and Oscar the Grouch is 0.05". The model



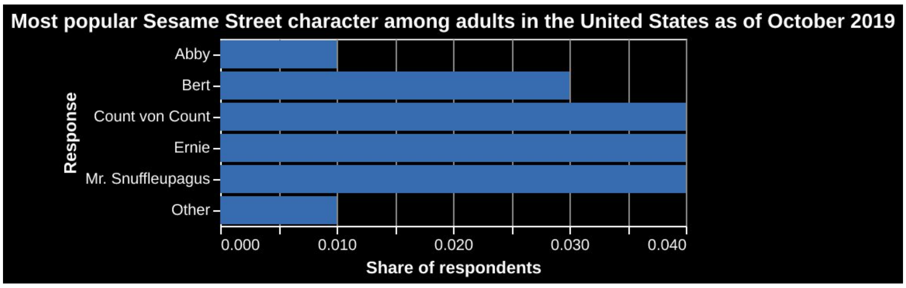
Text Evidence: As the United States grapples with the impacts of the coronavirus pandemic, consumer spending expectations for various products have shown notable fluctuations. Across all consumer categories, these expectations have varied within a range of 0.06, with a minimum value of 0.12 observed. Notably, the highest consumer spending expectation was recorded for Outerwear/jackets/coats, indicating a potential uptick in demand for these items. In contrast, spending expectations for Luxury/fashion items fell short of the peak, being less than the maximum by 0.04999999999999999. This divergence highlights shifting consumer priorities as the nation navigates the ongoing public health crisis.

Claim: The difference between home furniture and décor and luxury/fashion is 0.0.

GPT-4o Prediction: False

GPT-4o Reasoning: The chart shows that the consumer spending expectation for home furniture and décor is approximately 0.12. From the text evidence we can estimate the spending forr luxury fashion is 0.13, which is 0.0499999 lower than maximum which is 0.18. The difference should be 0.12-0.13 = 0.01.

Figure 4: An example of GPT-4o failed claim verification due to its in accurate prediction of the data value from the chart. The error in the reassoning chain is highlighted in red



Text Evidence: In October 2019, a survey revealed intriguing insights into the popularity of Sesame Street characters among American adults. Notably, Cookie Monster emerged as the most beloved character, reaching a peak popularity that highlighted his enduring appeal. In contrast, Oscar the Grouch lagged behind, with his popularity falling short of Cookie Monster's by 0.07. Overall, the data showed that the popularity of these characters fluctuated within a range of 0.11, with the lowest recorded value being 0.05. This variance underscores the nostalgic connection many adults still have with these iconic figures from their childhood.

Claim: The difference between Count von Count and Oscar the Grouch is 0.05.

GPT-4o Prediction: False

GPT-4o Reasoning: The chart does not provide any information about Oscar the Grouch's popularity. Therefore, it is not possible to verify the claim with the given evidence.

Figure 5: An example of GPT-4o failed claim verification due to limited capability to interpret cross-modal information. The error in the reasoning chain is highlighted in red

thinks that the claim is unverifiable as the information about Oscar the Grouch's popularity is not available on the chart, although the statistic can be found in the evidence in the text evidence. We hope these findings can inspire future work to focus on enhance VLM's capability on accurate data extraction from charts and joint interpretation of multiple evidences presented in different modalities.

4.4.3 Ablation Studies

Additional Transfer Learning A common strategy for C-VLMs is to transfer knowledge from a set of large-scale and diverse chart-understanding tasks, according to the intuition that different chart-understanding tasks can mutually benefit each other. Such strategies are commonly used even for already pre-trained models (i.e., prefinetuning; Aghajanyan et al. (2021)) for multiple modalities (Chen and Yu, 2023)). Existing work finds that the largest performance improvements are yielded only from the most highly related tasks (Chen and Yu, 2023; Padmakumar et al., 2022). We thus explore whether pre-finetuning on other chart understanding tasks can benefit their performance on the multi-modal chart fact verification task. We pre-finetune MATCHA and UniChart on three highly related tasks — Chart Question Answering (Masry et al., 2022), Chart Summarization(Kantharaj et al., 2022), and Chart Fact Checking(Akhtar et al., 2023b) — and then finally finetune them on our multi-modal fact verification task.

Model	Base	ChartQA	Chart2Text	ChartCheck
Matcha	55.6	59.2	54.6	53.0
UniChart	59.2	51.9	56.4	57.5

Table 5: Evaluation results on M^2-TabFact with different chart understanding tasks to pre-finetune

Table 4 summarizes the result of pre-finetuning on each related task.

We observe that pre-finetuning on other individual chart understanding tasks frequently leads to degraded downstream task performance. Pre-finetuning on ChartQA improves downstream verification performance for MATCHA but leads to a significant performance drop on UniChart. One possible explanation for this degradation is the multi-hop nature of our multi-modal verification task. While the majority of the existing chart understanding tasks focus strictly on information extraction from the chart, M^2-TabFact requires the model to additionally compare and reason over information from two equally weighted modalities.

Challenges from Multiple Modalities and Multiple Documents We further study the challenges arising from multi-modality and multi-document reasoning. In Table 6, we compare the performance of GPT-4o and Unichart when it is provided on three different evidence formats: (1) Single-modality and single-document, where the evidence is provided as the original table or a chart plot of the table. (2) Single-modality and multi-document, where the original tables are split into two sub-tables and then the two sub-tables are directly provided as the evidence or converted into two chart plots to serve as visual evidence. (3) Multi-modality and multi-documents, where the model is provided a pair consisting of a sub-table and the chart plot of the other sub-table.

If all the evidence is contained in a single document with one modality, we observe that GPT-4o is more capable of modeling tables than charts. Even GPT-4o is pre-trained on a more comprehensive set of datasets and tasks, it is clear that it is still bottlenecked by their ability to handle visual context compared to textual context. However, for Chart specialized VLM like Unichart, the table appears to be the harder modality to handle.

When the uni-modal evidence is split into two pieces , we observe consistent performance degradation. Splitting one chart to two causes around 3$\%$ performance degradation for GPT-4o. This shows that LVLM may encounter challenges combining the information extracted from multiple documents even if there is no additional information compared to the single document evidence. Finally, when the evidence is split into two pieces with different modalities, we see that for LVLM the performance of the model degrades close to the performance on the single document setting for the more challenging modality. For CVLM like UniChart, they are more vulnerable to multi-modal evidences. This shows that current VLMs additionally lack cross-model understanding, and may be bottlenecked by both their imbalanced capabilities across modalities and their ability to aggregate information from multiple documents.

Evidence Format	UniChart	GPT-4o
Chart	68.3	59.6
Table	57.7	67.1
Table-Table	58.0	65.3
Chart-Chart	58.3	56.7
Chart-Table	51.2	59.8

Table 6: Comparing model's performance when structured data are presented in different settings of sources and modalities

5 Conclusion

We introduce M^2-TabFact, the first multi-document multi-modal fact-checking dataset over structured data to simulate the complex real-world fact-checking scenario on multi-source table evidence. We evaluate SOTA models including chart-focused VLMs and powerful foundational VLMs in a fine-tuned setting and a zero-shot setting. Our best baseline achieves 59.2 $\%$ accuracy, which is still lagging far behind human's performance. We identify the major bottom neck of existing VLMs'low performance on this dataset is the unbalanced capability to handle structured data in different modalities. We hope our research can inspire the development of robust fact-checking system against various structured data representations.

Limitation

There are several limitations exist in this research work. First, as we generate our text evidences and claims with predefined template paraphrased by a LLM, there is certain language diversity limitation and bias from LLM that leads to a gap compared to human-written evidence and claims. In the future, we plan to further augment the quality of the dataset by having human annotators to re-edit the existing claims and evidences. Second, although M^2-TabFact includes tables that cover a wide range of topics and various chart types, certain table topics (e.g., the biomedical domain) and chart types (e.g., heat maps) are not covered. Addressing a broader range of table themes and chart types is an important future research direction.

Ethical Consideration

Our dataset is created from the tables of public dataset that is free to be reused for research purpose based on their license: GPL-3.0. Our proposed dataset is intended for research purposes, not as a tool to evaluate any real-world applications. We do not intend to have anyone to train models for making decision on the truthfulness of a claim against real-world context. We informed the human evaluator about all data being collected and its purpose. We hire students from our lab to conduct the evaluation. We pay the human evaluators above the minimum wage and decide the payment based on their working hours on accomplishing the evaluation task. To support the future research, we plan to release the dataset as well as the code script to evaluate different VLMs.

References

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

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023a. Reading and reasoning over chart images for evidence-based automated fact-checking. pages 399–414.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2023b. Chartcheck: An evidence-based fact-checking dataset over real-world chart images. *CoRR*, abs/2311.07453.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *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 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, , and Yiannis Kompatsiaris. 2015. Verifying multimedia use at mediaeval 2015.

Maximillian Chen and Zhou Yu. 2023. Pre-finetuning for few-shot emotional speech recognition. In *INTERSPEECH*. International Speech Communication Association.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact : A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *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 1683–1698, Online. Association for Computational Linguistics.

Jie Gao, Hella-Franziska Hoffmann, Stylianos Oikonomou, David Kiskovski, and Anil Bandhakavi. 2021. Logically at factify 2022: Multimodal fact verification. *arXiv preprint arXiv:2112.09253*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srini-

vasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mah-

moud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jin-

hyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa, Tanmoy, and Chakraborty. 2020. Newsbag: A benchmark multimodal dataset for fake news detection. In *SafeAI@AAAI*.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18893–18912. PMLR.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

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

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

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

Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:651–662.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, A. Sheth, and Asif Ekbal. 2022. Factify: A multi-modal fact verification dataset. In *DE-FACTIFY@AAAI*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

Vishakh Padmakumar, Leonard Lausen, Miguel Ballesteros, Sheng Zha, He He, and George Karypis. 2022. Exploring the role of task transferability in large-scale multi-task learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2542–2550.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2019. A corpus of debunked and verified user-generated videos. *Online Inf. Rev.*, 43:72–88.

Chahat Raj and Priyanka Meel. 2021. Arcnn framework for multimodal infodemic detection. *Neural Networks*, 146.

Frédéric Rayar, Mathieu Delalandre, and Van-Hao Le. 2022. A large-scale tv video and metadata database for french political content analysis and fact-checking. In *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, CBMI '22, page 181–185, New York, NY, USA. Association for Computing Machinery.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *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 2116–2129, Online. Association for Computational Linguistics.

Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350.

Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

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

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

A Dataset Creation

In this section, we introduce details not covered in our main sections for the automatic data generation process including: (1) Table Title Re-Writing (2) Text Evidence Generation (3) Prompt for LLM paraphrasing.

A.1 Table Title Re-writing

As introduced in section 3.1, we need to rewrite the original tile in certain cases such as the titles covering time-range or the titles that cover categorical values mentioned in the row header or column header. The title for the sub-table needs to be adjusted based on the time range or the categorical values of the sub-table data. To check whether the table covers a time range, we check if there is a match of key word like "year", "month" in row header or column header. For titles that cover categorial values across row headers or column headers, we also check every category value in these headers against the original table title. Once one of these two case is detected, we prompt a large language model, GPT-4o to rewrite the title using the following prompt:

```
You are the table title editor. Your job is to edit a given title to adapt it to a provided table.
Given the original table:
<Original Table>
and the original title:
<Original Title>
please edit the original title accordingly to the new table:
<New Table>
and generate the title in the following format: 'The new title is: ....'
```

A.2 Visual Themes

The full list of 24 visual themes we use are: 1.bmh; 2.classic; 3.dark background; 4.fast; 5.fivethirtyeight; 6.ggplot; 7.grayscale, 8.seaborn v08; 9.seaborn v08 brigh; 10.seaborn v08 colorblind; 11.seaborn v08 dark; 12.seaborn v08 dark palette; 13.seaborn v08 darkgrid; 14.seaborn v08 deep; 15.seaborn v08 muted; 16.seaborn v08 notebook; 17.seaborn v08 paper; 18.seaborn v08 pastel; 19.seaborn v08 poster; 20.seaborn v08 talk; 21.seaborn v08 ticks; 22.seaborn v08 white; 23.seaborn v08 whitegrid; 24.tableau colorblind10.

A.3 Text Evidence Generation Procedure

An overview of text evidence generation process is displayed in figure 6. Given the source table, the text evidence is mainly composed of two types of information: (1) the general facts about the source table such as the variation trend, maximum value, and average value. (2) the key data information from the source table that is required for the final claim assessment.

The types of general fact we extract from the table is listed as following where the definition of each type is followed by an example template:

Range the data value range of table.
The production of motor engines experienced fluctuations within a range of <RANGE VALUE> from June to September.

Min Value the minimum data value of the table.
The minimum number of motor engines is produced in June as <MIN VALUE>

Max Value the maximum data value of the table.
The maximum number of motor engines is produced in September as <MAX VALUE>

Average Value the averrage data value of the table.
The average number of motor engines is produced from June to September is <Average VALUE>

Variation Trending the overall trend of data change across time.
The number of motor engines produced is increased from June to September We sample one or two types of general fact out of all the categories every time to create the textual evidence.

After summarizing the general fact of the table, we will include the key data information from the source table that is used for final claim verification. Instead of directly summarizing the key data into natural language sentence, we identify the numerical connection between the key data with the extracted general fact and present the key information indirectly. For example, in figure 6, we summarize the key data information as *September produced <NUM DIFF> fewer motor engines when compared to the month with the maximum production.*

Finally we prompt GPT-4o to convert the summarized general facts and key data into a natural text paragraph with the following instruction:

```
You are writing a news report in four to five sentences to draw conclusions on the
```

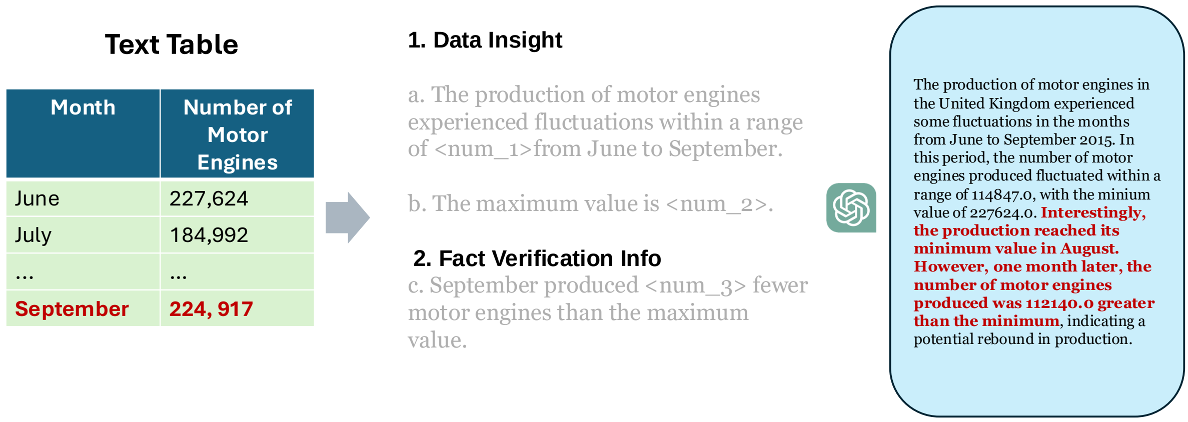


Figure 6: The text evidence generation process includes three steps: (1)summarize general fact of the table with predefined template. (2) summarize the numerical connection between the genral table fact with the key data information for final claim verification. (3) paraphrase the extracted summary with a LLM.

data provided about: <Table Title>.
Your news report should focus on presenting the following fact:
<EXTRACTED FACT>
1. You can add any relevant context to the topic, and add any transition sentences.
2.Don't simply copy the given facts into the final paragraph.

A.4 Prompt for paraphrasing

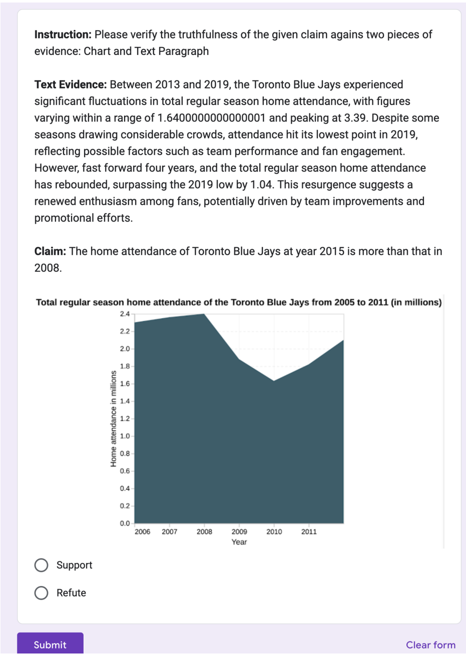
Prompt: Without changing the meaning or sentence structure, rewrite the provided sentence into a more natural one with correct grammar and spelling. Examples:
Original: The Year that Services has 37.2% Share in gross domestic product (GDP), Agriculture has 14.09% Share in gross domestic product (GDP) .
Rewritten: The year that Service sector had a 37.2% share in gross domestic product (GDP), the Agriculture sector had 14.09% share in GDP.
Original: The Year that Imports has 18 in Million dozen, Exports has 333 in Million dozen.
Rewritten: The year that Imports had 18 million dozen, Exports had 333 million dozen.
Original: The year that Photo had 3% in distribution of worldwide mobile app revenues in the Apple App Store from 2018 to 2024, in U.S. dollars, Music had 5% in distribution of worldwide mobile app revenues in the Apple App Store from 2018 to 2024, in U.S. dollars.
Rewritten: The year that Photo had 3% in distribution of worldwide mobile app revenues in the Apple App Store, Music had 5% in distribution of worldwide mobile app revenues.
2 Let's Start!
Original:

B Fact Checking Instruction for LVLMs

Here we provided the instruction template we use to prompt the LVLMs: GPT-4o, LLAVA, and Gemini to solve M^2-TabFact .
You are given a text claim and two pieces of evidence: a chart and a text article. The verification of the claim will require jointly interpreting both two evidences. Your task is to verify the claim against the two evidences and determine whether the claim is factually consistent with the given two evidences.
<CLAIM>
<Text Evidence>

You must respond in a structured JSON format that can be directly parsed with json.loads. Your response should contain two fields and two fields only:
"verification": Answer "Yes" if the two pieces of evidence factually support the claim. Answer "No", if you think the claim is not factually supported.
"explanation": an explanation of your verification result

C Interface for Human Evaluation


Instruction: Please verify the truthfulness of the given claim agains two pieces of evidence: Chart and Text Paragraph

Text Evidence: Between 2013 and 2019, the Toronto Blue Jays experienced significant fluctuations in total regular season home attendance, with figures varying within a range of 1.6400000000000001 and peaking at 3.39. Despite some seasons drawing considerable crowds, attendance hit its lowest point in 2019, reflecting possible factors such as team performance and fan engagement. However, fast forward four years, and the total regular season home attendance has rebounded, surpassing the 2019 low by 1.04. This resurgence suggests a renewed enthusiasm among fans, potentially driven by team improvements and promotional efforts.

Claim: The home attendance of Toronto Blue Jays at year 2015 is more than that in 2008.

Support

Refute

Submit

Clear form

Figure 7: Screenshot of Human Evaluation Task for M^2-TabFact