CLIMATEVIZ: A Benchmark for Statistical Reasoning and Fact Verification on Scientific Charts

Ruiran Su[1], **Jiasheng Si**[2], **Zhijiang Guo**[3,4], **Janet B. Pierrehumbert**[1]
[1]University of Oxford
[2]Qilu University of Technology (Shandong Academy of Sciences)
[3]Hong Kong University of Science and Technology
[4]Hong Kong University of Science and Technology (Guangzhou)
ruiran.su@trinity.ox.ac.uk, jiashengsi@qlu.edu.cn,
zhijiangguo@hkust-gz.edu.cn, janet.pierrehumbert@oerc.ox.ac.uk

Abstract

Scientific fact-checking has largely focused on textual and tabular sources, neglecting scientific charts—a primary medium for conveying quantitative evidence and supporting statistical reasoning in research communication. We introduce CLIMATEVIZ, the first large-scale benchmark for scientific fact-checking grounded in real-world, expert-curated scientific charts. CLIMATEVIZ comprises 49,862 claims paired with 2,896 visualizations, each labeled as support, refute, or not enough information. To enable interpretable verification, each instance includes structured knowledge graph explanations that capture statistical patterns, temporal trends, spatial comparisons, and causal relations. We conduct a comprehensive evaluation of state-of-the-art multimodal large language models, including proprietary and open-source systems, under zero-shot and few-shot settings. Our results show that current models struggle to perform fact-checking when statistical reasoning over charts is required: even the best-performing systems, such as Gemini 2.5 and InternVL 2.5, achieve only 76.2–77.8% accuracy in label-only output settings, which is far below human performance (89.3% and 92.7%). While few-shot prompting yields limited improvements, explanation-augmented outputs significantly enhance performance in some closed-source models, notably o3 and Gemini 2.5. We released our dataset and code alongside the paper.[1]

1 Introduction

Scientific fact-checking—the task of assessing the validity of scientific claims through cross-referencing with established literature, empirical observations, or experimental data (Wadden et al., 2020; Vladika and Matthes, 2023)—is essential for maintaining the integrity of research findings, combating misinformation, and preserving public confidence in scientific discourse (Wadden et al., 2022). This challenge is particularly acute in the visual domain, where data visualizations have become a battleground for controversial scientific understandings. During the COVID-19 pandemic, for instance, coronavirus skeptics actively created and circulated their own visualizations, often using the same official datasets as health authorities, to argue that the crisis was exaggerated and public health measures were unnecessary (Lee et al., 2021). These actors frequently employ what are described as 'counter-visualizations': charts that use orthodox, scientifically-sound methods to promote unorthodox arguments and misinformation. Worryingly, research shows that the majority of charts used to support misleading arguments online do not contain visual tricks like truncated axes but are, in fact, faithfully plotted visualizations taken from reputable sources and reframed with a misleading narrative (Lisnic et al., 2023).

However, the rapid accumulation of scholarly findings and the increasing demand for domain-specific expertise often exceed the capacity of manual verification, making scientific fact-checking a critical focus in the NLP community.

Significant progress has been made with the development of benchmarks such as SciFact (Wadden et al., 2020), SciFact-Open (Wadden et al., 2022), and SciTab (Lu et al., 2023). Despite these advances, existing resources exhibit critical limitations in scope. Specifically, prior benchmarks predominantly focus on verifying scientific claims against *textual* (Wadden et al., 2020; Diggelmann et al., 2021; Sarrouti et al., 2021; Saakyan et al., 2021) or *tabular* (Mohr et al., 2022; Lu et al., 2023) evidence, particularly from literature abstracts or tables. These claims are typically validated using *semantic* or *structural* logical reasoning between claims and corresponding evidence. In contrast, real-world scientific findings often involve claims that are intrinsically tied to quantitative data. In such contexts, charts serve as both visual and sta-

[1]https://github.com/Albasu120491/ClimateViz

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23436–23458
November 4-9, 2025

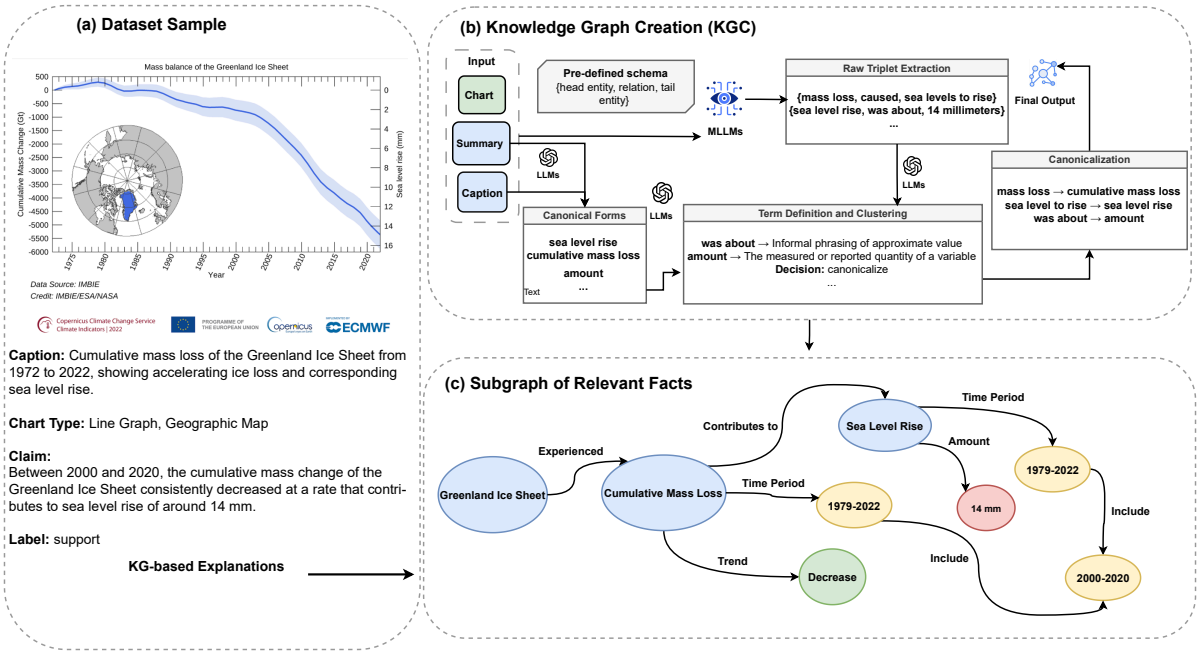

Figure 1: (a) A sample from CLIMATEVIZ showing a scientific chart, caption, claim, and its label. (b) Knowledge graph creation pipeline: raw triplets are extracted by a multimodal LLM and canonicalized. (c) Subgraph of relevant facts representing structured reasoning from the chart to the claim.

tistical representations, summarizing complex numerical information, revealing trends, supporting quantitative reasoning, and effectively communicating scientific insights (Huang et al., 2024). However, chart-based verification is largely absent from prior fact-checking benchmarks—despite requiring explicit *chart understanding* and *statistical reasoning* over visualized data, beyond what semantic or structural inference alone can handle (Akhtar et al., 2023b, 2024).

More specifically, scientific claims often involve not only raw data observations but also interpretations of patterns, anomalies, correlations, and statistical aggregation. For example, the claim in Figure 1 implies a statistical relationship between *ice sheet mass loss* and *sea-level rise*. Verifying such information is inherently challenging and requires: (i) *visual understanding* of charts with non-uniform temporal granularity (e.g., dual-axis time series), spatial dimensions (e.g., geographic insets), and uncertainty bands; (ii) *nuanced statistical reasoning* to interpret temporal trends, quantify cumulative change, and relate ice mass loss to sea-level rise; and (iii) *cross-modal reasoning* to establish logical coherence between the linguistic embedding of scientific claims and the perceptual features of the chart. This example illustrates the depth of reasoning needed to understand complex scientific findings, yet the absence of benchmarks that require such statistical reasoning severely limits the evaluation of models designed for general scientific understanding. To address these challenges and systematically evaluate model capabilities, our work is guided by the following core research questions:

(**RQ1**) How accurately can state-of-the-art multimodal large language models (MLLMs) perform fact-checking that requires complex statistical reasoning over real-world scientific charts?

(**RQ2**) To what extent do intermediate representations, such as structured knowledge graph explanations or extracted data tables, improve model performance and interpretability?

In this paper, we introduce CLIMATEVIZ, a novel dataset sourced from reputable climate institutions (e.g., the National Oceanic and Atmospheric Administration[2] and the UK Met Office[3]), designed to advance scientific fact-checking with a focus on statistical reasoning over charts. CLIMATEVIZ comprises 49,862 *claim–chart–knowledge graph (KG)* triplets, each accompanied by relevant *metadata* (e.g., chart caption, chart type). Claims are systematically constructed based on

[2] https://www.noaa.gov/
[3] https://www.metoffice.gov.uk/

2,896 expert-curated scientific charts and annotated with one of three labels: *support*, *refute*, or *not enough information (NEI)*. A key innovation of the dataset is the inclusion of chart-specific knowledge graphs that provide structured, interpretable explanations for fact verification. These KGs capture core scientific information aligned with the claims, such as quantities, trends, spatial and temporal contexts, and causal relationships—enabling explicit multi-hop reasoning. To construct CLIMATEVIZ, we launched a large-scale project on the citizen science platform Zooniverse[4], which ensured scientifically literate annotations. Each chart was independently annotated by six contributors, and the resulting claims were reviewed and verified by two domain experts to ensure correctness and quality.

We utilize CLIMATEVIZ as a diagnostic benchmark to evaluate the zero-shot and in-context learning capability across a varied range of state-of-the-art models, including open- and closed-source language models, and chart-based vision-language models. Comprehensive experiments reveal that all models struggle with verifying claims over scientific charts when statistical reasoning is necessary. Furthermore, the integration of a chart-specific knowledge graph proves beneficial when models are provided with both scientific charts and supplementary KG data. In addition, while models generate semantically plausible explanatory triplets, they typically fail to produce properly canonicalized outputs. These findings underscore the unique challenges posed by CLIMATEVIZ and highlight the need for further advances in models capable of statistical reasoning, structured explanation generation, and deep scientific understanding from visual evidence.

2 Related Work

Scientific Fact-checking Benchmarks. Several benchmarks have been proposed to advance automated scientific fact-checking, primarily focusing on textual evidence (see Table 1). SciFact (Wadden et al., 2020) introduced claim verification against biomedical research abstracts, while Climate-FEVER (Diggelmann et al., 2021) extended claim verification to the climate domain using Wikipedia articles. Other datasets, such as HealthVer (Sarrouti et al., 2021) and COVID-Fact (Saakyan et al., 2021), collected claims from health news and pandemic-related sources, respectively. More recently, CoVERT (Mohr et al., 2022) and SciTab (Lu et al., 2023) shifted toward structured evidence using tables from social media and scientific papers. However, these benchmarks largely target shallow reasoning tasks, often allowing claims to be verified through direct evidence matching rather than deeper inferential processes. Moreover, they rely exclusively on textual or tabular evidence and overlook scientific charts, which are central to communicating empirical findings in scientific domains. In contrast, CLIMATEVIZ introduces a large-scale, expert-verified benchmark grounded in high-quality scientific charts, requiring statistical reasoning for claim verification and aiming to more closely reflect real-world scientific fact-checking scenarios.

Fact-checking over Structured and Visual Data. Beyond textual evidence, fact-checking over structured formats such as tables and visualizations has attracted growing attention. TabFact (Chen et al., 2020) introduced a benchmark for fact verification against Wikipedia tables, while FEVEROUS (Aly et al., 2021) extended claim verification to semi-structured tables. Models such as TAPAS (Herzig et al., 2020) and DePlot (Liu et al., 2023a) enable direct reasoning over tabular data by treating tables as inputs to pretrained language models. In parallel, chart understanding has emerged as a distinct challenge, with datasets like PlotQA (Methani et al., 2020a), ChartQA (Masry et al., 2022a), and ChartBench (Xu et al., 2024) focusing on data extraction or question answering over charts. However, these efforts typically rely on synthetic charts and frame the task narrowly, limiting their relevance for real-world fact-checking (Guo et al., 2022). ChartCheck (Akhtar et al., 2024) is a more recent dataset that targets fact-checking over Wikimedia charts, but its reliance on non-curated, relatively simple visualizations—mostly line and bar graphs—and its focus on shallow observational claims restricts its depth and utility. In contrast, CLIMATEVIZ introduces high-quality, expert-curated scientific charts exhibiting greater structural and semantic complexity, and frames fact-checking as a task requiring statistical reasoning over visualized data, offering a significantly more realistic and challenging benchmark for scientific verification.

Statistical Reasoning in NLP. Statistical reasoning refers to the process of interpreting, analyzing, and drawing inferences from quantitative data—often involving trends, comparisons, vari-

[4] https://www.zooniverse.org/

Dataset	Modality	Domain	#Claims	Source
SciFact (Wadden et al., 2020)	Text	Biomedical	1.4K	Medical literature
Climate-FEVER (Diggelmann et al., 2021)	Text	Climate	1.5K	Wikipedia
HealthVer (Sarrouti et al., 2021)	Text	Health	14K	News
COVID-Fact (Saakyan et al., 2021)	Text	COVID-19	4.1K	News
CoVERT (Mohr et al., 2022)	Table	COVID-19	10K	Social media
SciTab (Lu et al., 2023)	Table	CS	1.2K	Scientific papers
CLIMATEVIZ	**Chart**	**Climate**	**49.8K**	**Expert-curated scientific charts**

Table 1: Comparison of scientific fact-checking datasets by modality, domain, claim volume, and source. **CLIMATEVIZ** is the first chart-based benchmark at this scale, grounded in real expert-curated scientific charts.

ability, and uncertainty—to reach logically sound conclusions (Fertig, 1958). Unlike general reasoning (Chen et al., 2024, 2025a; Peng et al., 2025), which may rely on commonsense or world knowledge, statistical reasoning demands precise, data-grounded inference directly from observed evidence. This capability is particularly critical in scientific fact-checking, where verifying claims derived from charts requires understanding and interpreting complex quantitative patterns. Challenges in visual reasoning are well-established; early work on compositional question answering over real-world images, such as the GQA dataset, demonstrated that models struggle significantly with multi-step relational and spatial reasoning, even when the required logic is non-quantitative (Hudson and Manning, 2019). While reasoning tasks have been extensively studied in NLP, existing benchmarks rarely require statistical reasoning over charts; most focus on discrete, categorical reasoning (Pan et al., 2023; Akhtar et al., 2023a; Glockner et al., 2024) rather than interpreting continuous data distributions. Techniques such as few-shot prompting (Brown et al., 2020) have shown promise in improving performance on symbolic and arithmetic reasoning tasks, but our experiments demonstrate that few-shot prompting yields minimal gains on CLIMATEVIZ —underscoring the unique challenges posed by statistical reasoning over scientific charts.

3 CLIMATEVIZ: Dataset Construction

3.1 Annotation

We manually selected 2,896 diverse scientific charts from six respected open-domain climate sources, each accompanied by metadata[5]. These charts—spanning topics such as temperature anomalies, CO_2 concentrations, precipitation trends, and sea level rise—were used to design a three-task annotation project on Zooniverse, a well-established and influential citizen science platform (Fortson et al., 2011; Simpson et al., 2014). We provided annotators with a comprehensive field guide, golden samples labeled by the authors for pre-annotation training, and a live discussion board to assist with challenging cases during the annotation process (see Appendix A). Each chart was independently annotated by six contributors.

3.1.1 Chart Type Annotation

In the first task, annotators were asked to identify the chart type by selecting from a set of predefined categories: line graph, pie chart, scatter plot, geographic map, or other (see Figure 4). Given the complexity of many scientific charts, such as those with overlapping modalities or multiple subplots (see Figure 1), annotators were permitted to select multiple chart types for a single instance. This task aimed to categorize chart forms to support downstream analysis and model conditioning.

3.1.2 Caption Annotation

In the second task, annotators were instructed to write or revise the caption associated with each chart, ensuring clarity, accuracy, and conciseness in describing its content. In addition, annotators were asked to compose at least one true claim per chart that required statistical reasoning and was directly verifiable from the visualized data. We applied an automated preprocessing step to filter out incomplete or overly short claims (fewer than 10 words). The remaining claims were manually validated by two domain experts according to two

[5] https://www.noaa.gov/, https://www.metoffice.gov.uk/, https://www.copernicus.eu/, https://earthobservatory.nasa.gov/, https://www.climate.gov/, https://climatereanalyzer.org/

criteria: (i) factual correctness independent of external context, and (ii) direct verifiability using only the information presented in the chart. Claims that met both criteria were labeled as "keep"; those that did not were discarded.

To generate refuted claims, we employed GPT-4o (OpenAI, 2024), prompting it to apply common data fallacy strategies including trend modification, exaggeration, and metric swaps (Akhtar et al., 2024; Xu et al., 2024). To ensure each generated claim was semantically contradictory to the original, we filtered 20,148 candidates from 23,190 generated refuted claims using DeBERTa-Large-MNLI (Laurer, 2022). Outputs passing this stage were then reviewed by domain experts to verify grammaticality and falsifiability with respect to the associated chart.

For NEI (Not Enough Information) claims, we employed conceptual generalization (Drchal et al., 2024), transforming specific factual details into broader or unverifiable language (e.g., "Florida" → "a coastal region"). We combined 200 manually authored NEI examples with GPT-4o-generated variants, additionally prompting entity replacements (e.g., "average" → "maximum" anomaly) to increase linguistic and semantic diversity. All NEI claims were independently verified by two domain experts to ensure that they were plausible yet unverifiable based on the chart. See Table 7 for examples of refuted and NEI claims.

3.1.3 Knowledge Graph-Based Explanation

We propose a method for generating structured explanations in chart-based fact-checking by constructing chart-specific knowledge graphs (KGs) composed of canonicalized (h, r, t) triplets. In contrast to prior work that applies LLMs to general-purpose knowledge graph construction (Bi et al., 2024; Li et al., 2023), we use a multimodal LLM (GPT-4o (OpenAI, 2024)) to extract factual triplets.

The pipeline begins by parsing each chart and its caption, followed by aggregating all supported claims to construct a unified chart summary. GPT-4o is then prompted with this context—chart image, metadata, and summary—under a loosely defined schema to extract factual triplets that reflect the chart's content. To reduce ambiguity and improve consistency, we apply a self-canonicalization stage inspired by the Extract, Define, Canonicalize (EDC) framework (Zhang and Soh, 2024), which standardizes the representation of entities and relations across triplets.

Statistic	Value
Supported claims	15,100
NEI claims	15,258
Refuted claims	19,504
Total claims	**49,862**
charts	2,896
Avg. tokens per claim	19.0
Avg. claims per chart	17.2

Table 2: Dataset statistics for CLIMATEVIZ. NEI stands for *Not Enough Information*.

Annotation Task	Randolph's Kappa
Chart Type Annotation	82.9
Caption Annotation	68.3
Claim Generation	76.5

Table 3: Randolph's Kappa values for IAA across tasks.

These chart-derived triplets serve as structured and interpretable explanations that support fact-checking decisions. For schema details and representative examples, see Appendix D.

3.2 Dataset Analysis

3.2.1 Dataset Statistics

CLIMATEVIZ comprises a total of 49,862 claims labeled as *support*, *refute*, or *not enough information (NEI)* against 2,896 expert-curated charts from the Climate field. The statistics of our CLIMATEVIZ are shown in Table 2.

We computed inter-annotator agreement scores using Randolph's Kappa (Randolph, 2005) across the three annotation tasks. For the first and second tasks, agreement was measured among six annotators per chart, while for the third task, agreement was calculated between two domain experts responsible for validating the final set of claims. The resulting scores (see Table 3) indicate substantial agreement across all tasks (Landis and Koch, 1977).

3.2.2 Statistical Reasoning in CLIMATEVIZ

We randomly sampled a balanced subset of 300 claims from the CLIMATEVIZ dataset, covering a diverse range of chart types. Each claim was manually annotated by an author with the types of statistical reasoning required for its verification, following the taxonomy defined in Table 4. We observe that temporal comparison, value extraction, and anomaly detection are the most prevalent reasoning types in our dataset.

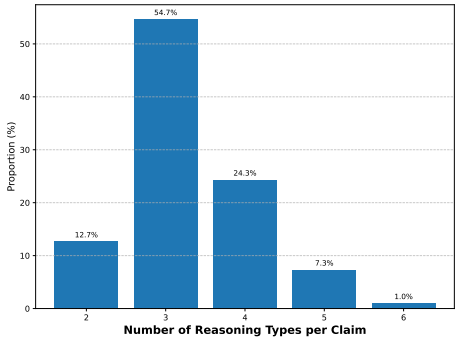

Figure 2: Distribution of statistical reasoning complexity in CLIMATEVIZ claims.

We further analyze the complexity of claims by examining the number of statistical reasoning types required per instance (Figure 2). Our analysis reveals that multi-hop reasoning is prevalent: a majority of claims (79.0%) require three or four distinct types of statistical reasoning. This highlights the inherently compositional nature of scientific fact verification in CLIMATEVIZ.

This deeper analysis demonstrates that CLIMATEVIZ is not merely a collection of simple lookup tasks. Rather, it challenges models to perform compositional statistical inference—often across varying temporal scales, spatial contexts, and measurement units. As such, CLIMATEVIZ serves as a rigorous benchmark for evaluating a model's ability to perform complex, multi-faceted statistical reasoning over scientific visual evidence.

4 Experimental Setup

4.1 Task Settings

We define two input settings for the chart-based fact-checking task:

- **Chart + Text (CT):** The model $\mathcal{M}$ receives a chart $\mathcal{C}_{\text{chart}}$, an associated caption $\mathcal{T}_{\text{caption}}$, and a claim $\mathcal{T}_{\text{claim}}$, and predicts a fact-checking label $\mathcal{Y} \in \{\text{support, refute, NEI}\}$.
- **Chart + Table + Text (CTT):** We apply a chart-to-table conversion model, DePlot (Liu et al., 2023b), to extract a structured table $\mathcal{T}_{\text{table}}$ from $\mathcal{C}_{\text{chart}}$. The model $\mathcal{M}$ then receives $(\mathcal{C}_{\text{chart}}, \mathcal{T}_{\text{table}}, \mathcal{T}_{\text{caption}}, \mathcal{T}_{\text{claim}})$ as input and predicts $\mathcal{Y}$.

We further consider two output settings:

- **Label-Only Output:** The model outputs only the fact-checking label $\mathcal{Y}$:

$$\mathcal{F}(\text{inputs}) \rightarrow \mathcal{Y}.$$

- **Explanation-Augmented Output:** The model outputs both a set of structured explanatory triplets $\mathcal{E}$ and the final label $\mathcal{Y}$:

$$\mathcal{F}(\text{inputs}) \rightarrow (\mathcal{E}, \mathcal{Y}).$$

We evaluate label classification using accuracy and macro F_1. We further evaluate generated triplets using BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020).

To validate the use of DePlot-generated tables, we conducted a manual evaluation of 50 chart-to-table conversions, sampled across five chart types. Each output was assessed according to three criteria: fidelity to the original chart, omission of relevant information, and presence of hallucinated content. Our analysis (see Appendix E) shows that DePlot produces high-fidelity tables for line and bar charts, and reasonably accurate tables for pie charts, scatter plots, and maps. These results support the reliability of $\mathcal{T}_{\text{table}}$ as a structured input in our experimental settings.

4.2 Baselines

We evaluate a suite of state-of-the-art multimodal models across multiple configurations.

Open-source models. We include three publicly available vision-language models: LLaMA-4-Maverick-17B (MetaAI, 2025), InternVL-2.5-78B (Chen et al., 2025b), and Qwen2.5-VL-72B (Bai et al., 2025), evaluated under both zero-shot and few-shot settings.

Closed-source models. We evaluate three proprietary multimodal large language models (MLLMs): o3 (OpenAI, 2025), GPT-4o (OpenAI, 2024), and Gemini 2.5 (DeepMind, 2025), under both zero-shot and few-shot settings.

Chart-specific vision-language models. We include several variants of Matcha (Liu et al., 2023c), an open-source model designed specifically for chart understanding. In particular, we evaluate two off-the-shelf variants: Matcha-ChartQA, pretrained on the ChartQA benchmark (Masry et al., 2022b), and Matcha-PlotQA, trained on the PlotQA dataset (Methani et al., 2020b), targeting chart question answering and plot comprehension, respectively. Additionally, we fine-tune the base Matcha model on the CLIMATEVIZ training and development sets for the fact-checking task.

Statistical Reasoning	Prop. (%)	Definition	Example
Temporal Comparison	75.7%	Compare values across time points.	temperature in 2020 higher than 2010
Value Extraction	63.0%	Read exact values from charts.	CO_2 level was 412 ppm in 2020
Anomaly Detection	52.3%	Spot unexpected patterns or values.	a sudden spike in temperature
Temporal Aggregation	49.3%	Summarize data over periods.	average rainfall over a decade
Spatial Comparison	35.7%	Compare across different regions.	England warmer than Scotland in July
Trend Detection	26.3%	Identify rising or falling trends.	CO_2 emissions rising over time
Unit Interpretation	14.0%	Understand and convert units.	mm of rain converted to inches
Uncertainty	13.0%	Interpret variability or error bars.	temperature estimate: $20^\circ \pm 1^\circ C$

Table 4: Statistical reasoning types required in CLIMATEVIZ claims. "Prop. (%)" denotes the proportion of sampled claims.

Human performance. To establish an upper bound for model performance, we include a human evaluation baseline. We randomly sample 150 examples for each setting from the CLIMATEVIZ benchmark. Each example is annotated by a human with expertise in both climate science and natural language processing, using the same input modalities as the corresponding model configuration.

Evaluation protocol. The CLIMATEVIZ dataset is split into training (70%), development (10%), and test (20%) subsets. All models are evaluated on the same held-out test set to ensure fair and consistent comparison across model types and input settings.

5 Results

We present the main findings from our experiments in Tables 5 and 6, which report results for both label classification and explanation generation.

Scientific chart-based fact-checking remains challenging for current models. Despite recent advances in multimodal reasoning (Wang et al., 2024, 2025) and chart understanding (Akhtar et al., 2024; Xu et al., 2024), a substantial performance gap remains between models and human annotators. Human evaluators achieve 89.3% accuracy in the **Chart + Text (CT)** setting and 92.7% in the **Chart + Table + Text (CTT)** setting, outperforming all model variants across both input conditions. These results highlight the continued difficulty of scientific chart-based fact verification and the limitations of current models in capturing nuanced statistical reasoning.

Explanation-augmented output improves closed–source model performance. Closed-source models such as o3 and Gemini 2.5 show notable gains when generating structured explanations alongside label predictions. For example, o3 achieves the highest explanation-augmented performance in the CT setting, with 84.6% accuracy and a macro F1 score of 83.1—outperforming all other models. These results suggest that incorporating intermediate reasoning steps enables closed-source models to better ground their predictions, particularly when interpreting complex scientific visualizations.

CTT setting significantly boosts the performance of open-source models under few-shot prompting. All open-source models—including LLaMA-4, InternVL 2.5, and Qwen 2.5—achieve their highest label accuracy and F1 scores in the CTT setting when using few-shot prompting. For example, Qwen 2.5 and InternVL both reach 77.8% accuracy in the CTT few-shot condition, outperforming their CT counterparts. These results highlight the value of structured tabular inputs and prompt-based adaptation for improving factual reasoning in resource-constrained models.

Few-shot prompting offers limited benefit for scientific fact-checking over charts in the CT setting. While in-context learning is widely adopted to improve model performance, its impact on scientific fact-checking over charts is inconsistent. Notably, several closed-source models (e.g., GPT-4o and Gemini 2.5) exhibit degraded performance under few-shot prompting in the CT and label-only setting, particularly in explanation and triplet generation tasks. Open-source models also show only marginal gains except for LLaMA-4-Maverick, indicating that a few-shot prompting alone is insufficient to support complex reasoning over scientific visual data.

Fine-tuned Matcha-CLIMATEVIZ performs best among chart-specific models but still lags behind multimodal LLMs. Among chart-specific baselines, the fine-tuned Matcha-CLIMATEVIZ model achieves the highest accuracy

Category	Model	Setting	CT				CTT			
			Acc-L	F1-L	Acc-E	F1-E	Acc-L	F1-L	Acc-E	F1-E
Closed-source	o3	Zero-shot	59.3	58.9	**84.6**	**83.1**	64.0	63.3	68.9	67.8
	o3	Few-shot	61.3 ↑	61.0 ↑	**67.5** ↓	**67.0** ↓	65.5 ↑	64.9 ↑	65.4 ↓	64.5 ↓
	GPT-4o	Zero-shot	67.8	67.5	**68.1**	**68.2**	64.3	64.0	60.2	59.2
	GPT-4o	Few-shot	63.3 ↓	59.5 ↓	64.3 ↓	64.9 ↓	**68.3** ↑	**67.9** ↑	62.8 ↑	61.8 ↑
	Gemini 2.5	Zero-shot	**76.2**	**75.9**	73.2	71.2	57.6	57.0	**85.7**	57.3
	Gemini 2.5	Few-shot	57.4 ↓	53.8 ↓	**73.3** ↑	**73.9** ↑	56.6 ↓	56.2 ↓	70.4 ↓	70.3 ↑
Open-source	LLaMA-4-Maverick-17B	Zero-shot	39.4	29.7	47.4	43.6	47.2	45.3	52.5	49.4
	LLaMA-4-Maverick-17B	Few-shot	54.5 ↑	51.3 ↑	37.8 ↓	29.7 ↓	**79.4** ↑	**76.9** ↑	57.9 ↑	53.0 ↑
	InternVL 2.5-78B	Zero-shot	65.8	65.7	54.6	50.4	61.3	59.8	63.3	60.9
	InternVL 2.5-78B	Few-shot	61.3 ↓	61.4 ↓	63.8 ↑	62.5 ↑	**77.8** ↑	**75.5** ↑	76.4 ↑	73.2 ↑
	Qwen 2.5-VL-72B	Zero-shot	**68.3**	**68.3**	54.3	53.8	60.8	57.9	54.3	47.7
	Qwen 2.5-VL-72B	Few-shot	67.3 ↓	68.0 ≈	65.8 ↑	64.3 ↑	**77.8** ↑	**75.3** ↑	72.1 ↑	70.8 ↑
Chart-specific	Matcha-ChartQA	Zero-shot	34.6	33.2	–	–	31.3	30.2	–	–
	Matcha-PlotQA-V1	Zero-shot	21.3	21.7	–	–	24.5	22.4	–	–
	Matcha-PlotQA-V2	Zero-shot	32.4	30.6	–	–	33.4	32.9	–	–
	Matcha-CLIMATEVIZ	Fine-tuned	**51.2**	**50.7**	–	–	**50.4**	**48.6**	–	–
Human Performance			**89.3**	**89.3**	–	–	**92.7**	**88.6**	–	–

Table 5: Accuracy and Macro-F1 scores (%) on the **CLIMATEVIZ** fact-checking benchmark across two input settings and two output formats. **CT** (Chart+Text): chart image + caption + claim; **CTT** (Chart+Table+Text): chart image + extracted table + caption + claim. **Acc-L/F1-L**: label-only output; **Acc-E/F1-E**: explanation-augmented output. **Bold** indicates the best score per column. ↑ / ↓ / ≈ indicate intra-model differences.

Model	CT				CTT			
	BLEU	METEOR	ROUGE-L	BERTScore	BLEU	METEOR	ROUGE-L	BERTScore
o3 (ZS)	20.2	66.0	57.3	92.6	21.8	66.2	56.3	92.6
o3 (FS)	10.3↓	53.8↓	43.2↓	90.3↓	11.3↓	52.9↓	42.5↓	91.6↓
GPT-4o (ZS)	**48.4**	**77.2**	**73.6**	92.7	**46.2**	**73.4**	**67.4**	93.2
GPT-4o (FS)	13.8↓	51.3↓	38.0↓	87.0↓	14.7↓	55.4↓	43.8↓	89.4↓
Gemini 2.5 (ZS)	37.8	68.6	61.0	90.9	34.7	72.1	65.6	**92.7**
Gemini 2.5 (FS)	15.2↓	57.6↓	50.2↓	89.7↓	15.6↓	58.9↓	53.8↓	89.9↓
LLaMA-4-Maverick-17B (ZS)	35.3	68.3	60.3	92.3	34.8	60.2	60.2	91.3
LLaMA-4-Maverick-17B (FS)	13.0↓	52.2↓	41.5↓	90.3↓	13.1↓	48.8↓	38.9↓	89.4↓
InternVL 2.5-78B (ZS)	30.8	65.2	56.6	91.4	27.6	68.1	61.2	93.1
InternVL 2.5-78B (FS)	20.7↓	65.2≈	53.6↓	90.6↓	23.9↓	67.9↓	54.1↓	91.2↓
Qwen 2.5-VL-72B (ZS)	36.8	66.2	57.5	91.9	35.1	70.5	61.8	**93.5**
Qwen 2.5-VL-72B (FS)	25.7↓	57.3↓	45.7↓	89.3↓	9.6↓	39.8↓	29.9↓	88.3↓

Table 6: Explanatory triplet generation results on CLIMATEVIZ. Models are evaluated in both zero-shot (ZS) and few-shot (FS) settings across CT and CTT inputs.↑ / ↓ / ≈ indicate intra-model change. Bold indicates the best in each column.

(51.2% in CT, 50.4% in CTT), outperforming zero-shot variants like Matcha-ChartQA and Matcha-PlotQA. However, its performance remains substantially below that of general-purpose multimodal LLMs. This performance gap suggests that while task-specific fine-tuning improves chart understanding, chart-specialized models still lack the general reasoning capabilities and scalability of large multimodal LLMs.

Explanatory triplet generation. Across both **CT** and **CTT** settings, few-shot prompting consistently degrades triplet quality for all models. GPT-4o remains the strongest performer overall, achieving the highest scores in BLEU, METEOR, and ROUGE-L. While all models attain consistently high BERTScore values—indicating semantic plausibility—their BLEU scores remain low, suggesting that models often generate logically correct triplets but fail to produce outputs in a standard, canonicalized format.

Isolating the Contribution of the Visual Modality. To quantify the unique contribution of the visual input, we conducted an ablation study where the chart image was removed, leaving only the DePlot-generated table and text as inputs (TT setting). The results, detailed in Table 9 , reveal a substantial drop in performance across all models, underscoring the indispensability of the visual modality. This gap demonstrates that while structured tables can convey raw data, they fail to capture the relational, spatial, and contextual patterns that are visually encoded in the chart and are essential for robust scientific fact-checking.

6 Broader Implications

NLP for High-Stakes Domains. Despite recent advances, current models fall short of human performance in verifying claims from scientific charts, highlighting the need for NLP systems that are both trustworthy and verifiable in high-stakes domains like science communication and policy.

Multimodal and Spatio-Temporal Reasoning. CLIMATEVIZ goes beyond text and tables, requiring reasoning over visual, spatial, and temporal patterns. Current models struggle with this complexity, especially in statistical interpretation, motivating new architectures that unify multimodal reasoning.

Model Explainability. CLIMATEVIZ supports joint evaluation of predictions and reasoning via explanatory triplets. Explanation-augmented outputs improve accuracy in closed-source models, while the gap between BERTScore and BLEU reveals a need for better canonicalization of semantically correct outputs.

Countering Sophisticated Visual Misinformation. The COVID-19 pandemic demonstrated that malicious actors do not necessarily lie with 'visual tricks' but rather by reframing well-designed, faithfully plotted charts from authoritative sources to support false narratives (Lee et al., 2021; Lisnic et al., 2023). They achieve this by making flawed causal inferences, cherry-picking data from interactive dashboards, and failing to account for statistical nuance, thereby creating plausible but misleading arguments. Developing AI systems capable of robust statistical reasoning, as benchmarked by CLIMATEVIZ, is a critical step towards automatically identifying and flagging such misleading claims.

7 Conclusion

We introduce CLIMATEVIZ, the first large-scale benchmark for scientific fact-checking grounded in real-world expert-curated charts. By evaluating a diverse range of state-of-the-art models, we reveal limitations in multimodal factual reasoning, especially when statistical interpretation is required. Our findings demonstrate that current models still lag behind human performance, and that in-context learning alone offers limited gains. However, explanation-augmented outputs show promise in improving model reliability and interpretability. CLIMATEVIZ establishes a new foundation for building multimodal systems that reason faithfully, communicate transparently, and support scientific decision-making in high-stakes domains.

Limitations

While CLIMATEVIZ introduces a comprehensive benchmark for scientific fact-checking over real-world charts and supports structured explanation through knowledge graphs, our study has several limitations.

First, our experiments focus primarily on in-context learning under zero-shot and few-shot settings. We do not explore more advanced prompting strategies such as chain-of-thought (CoT) prompting (Wei et al., 2022), tree-of-thought (ToT) reasoning (Yao et al., 2023), or program-guided verification (Pan et al., 2023), which may further improve performance on compositional and multi-hop reasoning tasks. This restricts our ability to fully characterize model capabilities in structured reasoning scenarios.

Second, we evaluate factuality and explanation quality using predefined structured output formats (triplets), but our automatic metrics (e.g., BLEU, METEOR) may not fully capture factual soundness or semantic coherence of the generated explanations (Schlichtkrull et al., 2023). Future work could incorporate human evaluations or more targeted reasoning metrics.

Lastly, while the dataset spans a wide range of climate topics and chart types, it is domain-specific. Generalization to other scientific disciplines with different conventions, terminologies, or visual formats remains untested.

Ethics Statement

We recognize the importance of ethical considerations in our work.

All charts included in the ClimateViz benchmark are sourced from publicly accessible, reputable scientific institutions, and no proprietary or confidential data was used. The associated claims were annotated through a large-scale citizen science campaign on Zooniverse, with additional quality control by domain experts. Annotators were fully informed about the research purpose and provided their consent voluntarily. No personally identifiable or sensitive information was collected.

While our goal is to foster trustworthy AI, we acknowledge potential risks. A primary concern is **automation bias**: a model trained on CLIMAT-

EVIZ, even at SOTA performance, may still make significant errors. Over-reliance on such a system for automated fact-checking without human oversight could lead to the unintentional amplification of misinformation or the discrediting of valid scientific claims. Furthermore, there is a risk of adversarial vulnerability, where malicious actors could devise novel manipulation strategies not covered in our benchmark to evade detection. Besides, we recognize a potential for **dual-use**, where the very models designed to verify claims could be repurposed to generate plausible-sounding, but ultimately false claims to accompany scientific charts, thereby accelerating targeted disinformation campaigns.

To mitigate these risks, we emphasize that CLIMATEVIZ is intended for research purposes to advance model capabilities, not for deployment in unchecked, real-world applications. Outputs from models trained or evaluated on ClimateViz should not be used in isolation for critical decision-making.

To support reproducibility and responsible use, we have released the CLIMATEVIZ dataset under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, and all associated code under the MIT License on GitHub.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive comments. Our work would not have been possible without the Zooniverse platform and the dedicated efforts of the many citizen science volunteers who contributed to the annotation of the CLIMATEVIZ dataset. This research was supported by the National Natural Science Foundation of China (Grant No. 62402258), the Taishan Scholars Program of Shandong Province (Grant No. tsqn202507242), and the Natural Science Foundation of Shandong Province (Grant No. ZR2024QF099).

References

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023a. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Mubashara Akhtar, Michael Sejr Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023b. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5430–5448. Association for Computational Linguistics.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. ChartCheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937, Bangkok, Thailand. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2024. Codekgc: Code language model for generative knowledge graph construction. *Preprint*, arXiv:2304.09048.

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. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards

reasoning era: A survey of long chain-of-thought for reasoning large language models. *Preprint*, arXiv:2503.09567.

Qiguang Chen, Libo Qin, Jiaqi Wang, Jingxuan Zhou, and Wanxiang Che. 2024. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In *Advances in Neural Information Processing Systems*.

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. Conference paper.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2024-05-01.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*.

Jan Drchal, Herbert Ullrich, Tomáš Mlynář, and Václav Moravec. 2024. Pipeline and dataset generation for automated fact-checking in almost any language. *Neural Comput. Appl.*, 36(30):19023–19054.

John W. Fertig. 1958. Introduction to statistical reasoning. *American Journal of Public Health*, 48:533–533.

Lucy Fortson, Karen Masters, Robert Nichol, Kirk Borne, Edd Edmondson, Chris Lintott, Jordan Raddick, Kevin Schawinski, and John Wallin. 2011. Galaxy zoo: Morphological classification and citizen science. *Preprint*, arXiv:1104.5513.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kung-Hsiang Huang, Hou Pong Chan, May Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Trans. on Knowl. and Data Eng.*, 37(5):2550–2568.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Moritz Laurer. 2022. Deberta-large-mnli model. https://huggingface.co/MoritzLaurer/DeBERTa-Large-MNLI. Accessed: 2025-05-12.

Crystal Lee, Tanya Yang, Gabrielle D Inchoco, Graham M. Jones, and Arvind Satyanarayan. 2021. Viral visualizations: How coronavirus skeptics use orthodox data practices to promote unorthodox science online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *Preprint*, arXiv:2304.11633.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Maxim Lisnic, Cole Polychronis, Alexander Lex, and Marina Kogan. 2023. Misleading beyond visual tricks: How people actually lie with charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023b. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023c. 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.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

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

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

MetaAI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/meta-llama-4-and-code-llama-3. Accessed: 2025-05-01.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020a. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1527–1536.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020b. Plotqa: Reasoning over scientific plots. In *Proceedings of the 28th ACM International Conference on Multimedia*.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

OpenAI. 2024. Gpt-4o: A new frontier in openai's multimodal models. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-05-01.

OpenAI. 2025. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-05-01.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

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

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. LMM-R1: empowering 3b lmms with strong reasoning abilities through two-stage rule-based RL. *Preprint*, arXiv:2503.07536.

Justus J. Randolph. 2005. Free-marginal multirater kappa: An alternative to fleiss' fixed-marginal multirater kappa. https://eric.ed.gov/?id=ED490661. ERIC Document ED490661.

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.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Robert Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd*

International Conference on World Wide Web, WWW '14 Companion, page 1049–1054, New York, NY, USA. Association for Computing Machinery.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. 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.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *Preprint*, arXiv:2503.12605.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *Preprint*, arXiv:2401.06805.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2024. Chartbench: A benchmark for complex visual reasoning in charts. *Preprint*, arXiv:2312.15915.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. *Preprint*, arXiv:2404.03868.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Annotation for CLIMATEVIZ

A.1 Before Annotation: Preparation Phase

Before starting the annotation process, we conducted extensive preparation to ensure that annotators had the necessary guidance, tools, and understanding of the scientific charts. We began with an internal review involving climate science experts and NLP practitioners. This was crucial to refine the scope of the tasks, establish clear goals, and identify potential challenges in the annotation of complex visual scientific data.

Then, a beta test was conducted with a small group of experienced annotators who provided early feedback on the clarity and difficulty of the tasks. This helped identify areas where instructions or task complexity needed adjustment. Following the beta test, we gathered feedback through detailed forms, allowing us to iteratively improve the task definitions and annotation interface.

The finalized workflows and task requirements were then implemented on the Zooniverse platform's dedicated webpage, which served as the main point of interaction for annotators.

A.2 During Annotation: Annotation Phase

The annotation phase was designed to facilitate a smooth and productive experience for annotators, equipping them with the resources necessary to accurately interpret and label the charts. The tools used during this phase include:

A.2.1 Field Guide

A comprehensive field guide was provided to the annotators, covering the different types of data representations commonly found in the charts. This guide includes:

Types of Visuals: Examples of bar charts, line graphs, pie charts, scatter plots, geographic maps, and others, helping annotators become familiar with each format.

Key Definitions: Explanations of essential concepts, such as "anomalies" or "trends," that might be important when describing climate-related visuals.

A.2.2 Instructions

Each task was accompanied by explicit, step-by-step instructions. This was especially important for the third task, which involved summarizing factual information from the charts. Annotators were instructed to focus on objective descriptions, providing factual statements that require statistical reasoning regarding the chart without interpretation or bias.

The following are the intructions shown to the annotators for our three tasks.

Task 1: Write a Clear and Informative Caption for the Scientific Chart

Welcome! Your task is to write a straightforward, clear caption that accurately describes the main content of the scientific chart. This caption should help a reader quickly understand what the chart shows, without needing to read all the details.

Guidelines for Writing the Caption:

- **Summarize the Main Information:** Focus on the key message or trend shown in the chart. What is the chart primarily about?
- **Use Straightforward English:** Write in plain, clear language without jargon. Your caption should be understandable even to readers outside the field.
- **Ignore Sources and Logos:** Do not include any references to logos, footnotes, or sources. We assume the charts are from reliable resources.
- **If the Chart is Unclear:** If you cannot determine what the chart shows, type "NA" as the caption.
- **If Multiple Messages Appear:** If the chart covers multiple topics, focus on the most important or prominent trend or finding.

Where to Look for Clues:

- **Chart Title:** Use the chart title if available, but rewrite it slightly to form a complete, descriptive sentence if necessary.
- **Axes Labels:** Look at the x-axis and y-axis labels to understand what is being measured over what range.
- **Legend and Annotations:** If the chart includes a legend or text annotations, use them to guide your description.

Task 2: Identify the Data Representation Used in the Chart

Your next task is to specify how the data in the chart is represented. Some charts use only one form of representation, while others may use several types together.
Click on all types of data representation that you observe.
Refer to the Field Guide on the right side for detailed descriptions of various data representations used.
Available Options:

- Bar Chart
- Line Graph
- Pie Chart
- Scatter Plot
- Geographic Map
- Other

Task 3: Write Claims Using Statistical Reasoning Based on the Scientific Chart

Your final task is to carefully study the graphic and write one or more factual claims that use **statistical reasoning**. Each claim should be based directly on what the chart shows.
Imagine you are **explaining the information to someone who cannot see the graphic**. Your claims should summarize important quantitative patterns, relationships, or trends, using **straightforward English** and **specific details** (such as location, time, measurements, and units).
Guidelines for Writing Claims:

- **Base Claims on the Graphic Only:** Use only the information visible in the graphic. Do not rely on outside knowledge.
- **State Quantitative Information Clearly:** Use numbers, percentages, or comparisons whenever possible.
- **Focus on Statistical Trends and Relationships:**
  - Changes over time
  - Comparisons between groups
  - Visible correlations
- **Be Specific and Detailed:** Include location, time period, and units.
- **One Sentence per Claim:** Write each claim clearly and concisely.
- **Avoid Vague Statements:** Prefer specific, measurable facts.
- **If the Chart is Ambiguous:** Write "NA" if you cannot state any confident claims.

Examples of Good Claims:

- This line graph shows that the average annual temperature in Paris increased from approximately 12°C in 1970 to 15°C in 2020.
- The pie chart indicates that over 60% of global renewable energy production in 2022 came from solar and wind sources combined.

A.2.3 Tutorials

We created interactive tutorials that walked annotators through example charts and tasks. These tutorials emphasized how to identify and describe elements like key data points, trends, or anomalies.

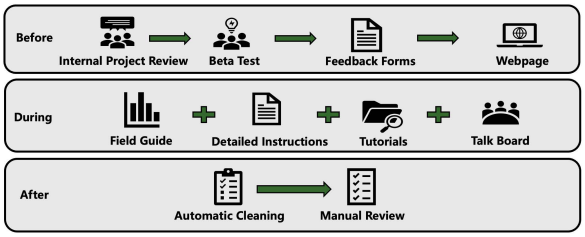

Figure 3: Quality Control Process: before, during, and after annotation

A.2.4 Talk Board

The Zooniverse platform also included an active "Talk Board" during annotation, where annotators could discuss uncertainties, ask questions, and receive support from both project moderators and their peers. The authors of this paper play an active role in explaining our tasks and discussing how to annotate some particularly complex charts. This collaborative environment was instrumental in resolving ambiguous cases and ensuring consistency across annotations.

A.3 Post Annotation: Quality Assurance Phase

Once the annotations were completed, an extensive quality assurance phase was implemented to verify the accuracy and reliability of the collected data.

A.3.1 Automatic Cleaning

Initially, automated data cleaning scripts were run to detect potential issues such as outlier annotations, incomplete tasks, or incorrect data types. Also, we removed annotations less than 10 words for the "fact" task, with the assumption that they are not informative enough.

A.3.2 Manual Review

Following the automated cleaning, the data underwent a manual review by domain experts with a PhD degree in climate science and NLP. During this review, we scrutinized the flagged annotations for

correctness and consistency. We also went through each claim to make sure it contained the necessary context, which makes it a claim by itself. This dual-step process was critical in catching errors that may have been overlooked by automated methods and ensuring that the dataset retained a high level of reliability.

B Statistics for the Scientific Charts

The scientific charts used in CLIMATEVIZ were manually selected from reputable public sources to serve as high-quality, trustworthy visual evidence for fact-checking. All charts were curated to ensure interpretability, sufficient information density, and alignment with key indicators of climate science.

Figure 4 presents the distribution of chart sources and chart types. The majority of charts were obtained from two primary sources: climateanalyzer(52.2%) and the UK Met Office (40.5%). A smaller proportion of charts come from organizations such as Copernicus, NASA's Earth Observatory, Skeptical Science, and Climate.gov, each contributing less than 4%.

In terms of visual representation, line graphs dominate the dataset, comprising 68.7% of all charts. Bar charts are the second most common (24.2%), while scatter plots, maps, pie charts, and other types collectively account for the remaining 7.1%. This reflects the prevalent use of time-series and trend-based data visualization in scientific charts.

By incorporating a wide variety of scientifically valid visualizations from trusted institutions, CLIMATEVIZ ensures that models are evaluated on realistic and diverse chart-based evidence, closely mirroring the data presentation formats encountered in real-world scientific communication and policymaking.

C Details for Refute and NEI claims

C.1 Refuted Claim Generation

Refuted claims are created by systematically modifying supported claims to introduce factual inaccuracies while maintaining grammatical plausibility. We apply three complementary strategies to generate diverse and realistic refutations:

Trend Modification: Directional trends in the original claim are reversed to contradict the data. This includes altering keywords such as "increased" to "decreased," or "rising" to "falling." These changes invert the implied statistical direction while preserving the overall structure of the sentence.

Exaggeration: Numerical values and descriptive language are amplified to misrepresent the magnitude of a change. Quantities such as temperature or precipitation are scaled by random multipliers, and qualitative modifiers (e.g., "slight," "moderate") are replaced with more extreme terms (e.g., "severe," "dramatic").

Metric Swap: The core metric or variable in the claim is replaced with a similar but distinct one, preserving the sentence form while altering the underlying meaning. For example, "mean maximum temperature" might be swapped with "mean minimum temperature," or "sunshine duration" replaced by "cloud cover."

Following generation, we use the DeBERTa-Large-MNLI model (Laurer, 2022) to verify that the modified claim contradicts the original. Each claim pair is scored by the model, which classifies their relationship as entailment, neutral, or contradiction. Only claims labeled as contradiction with high confidence (score > 0.8) are retained.

All accepted refuted claims are then manually reviewed by two domain experts to ensure: (i) the claim is grammatically and semantically well-formed, and (ii) the statement is clearly refuted by the corresponding chart evidence.

C.2 NEI Claim Generation

NEI (Not Enough Information) claims are constructed to appear plausible while being unverifiable based on the chart alone. We adopt a multi-step generation strategy combining conceptual generalization, entity replacement, and LLM-based generation:

Conceptual Generalization: Specific, verifiable references in factual claims are replaced with broader or vaguer terms to obscure direct traceability to the chart. For instance, geographic entities like "Florida" are generalized to "a coastal region," and temporal references such as "July 2020" are broadened to "a recent summer month."

Entity Replacement: Key variables or metrics are substituted with related but unverifiable alternatives. For example, "average temperature anomaly" may be swapped with "maximum temperature anomaly," or "total precipitation" replaced with "cloud cover." This ensures the claim remains

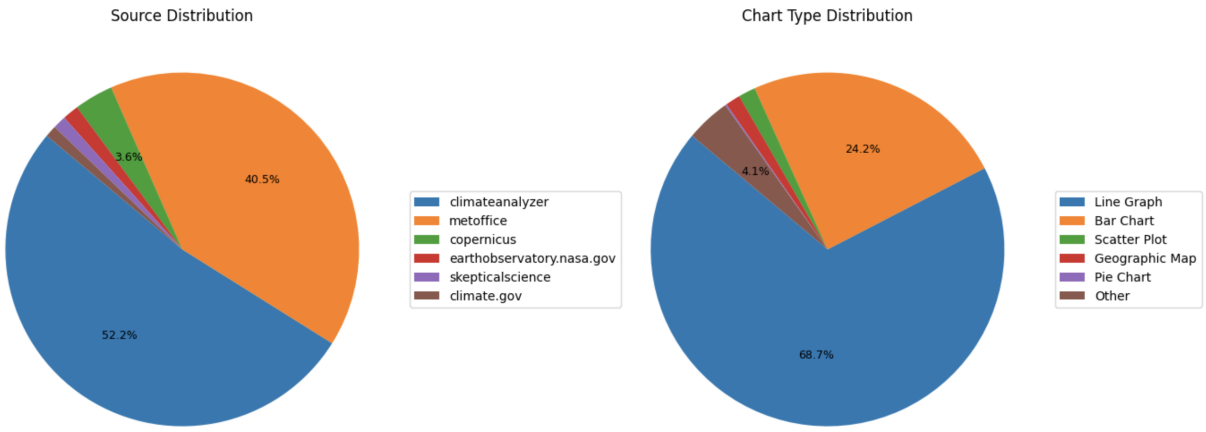

Figure 4: Distributions of source and type of the charts

topically relevant but cannot be definitively supported or refuted by the chart.

LLM-Based Generation: We manually curated 200 NEI claims and then used GPT-4o (OpenAI, 2024) to generate additional NEI examples by prompting the model with existing NEI instances and instructing it to maintain plausibility while avoiding chart-verifiable details. We included prompts to encourage diversity in language and structure while preserving the overall scientific tone.

All generated NEI claims were filtered to remove overly vague or clearly irrelevant instances. Two domain experts manually validated each claim using the following criteria: (i) the claim must be semantically and grammatically correct, and (ii) it must not be directly classifiable as support or refute based on the chart. Only claims meeting both criteria were included in the final NEI set.

Here are some real examples in the dataset, see Table 7.

D Knowledge Graph-Based Explanation

To support interpretable and structured scientific fact verification, we construct a knowledge graph (KG) for each chart in CLIMATEVIZ. These graphs consist of canonicalized triplets of the form (h, r, t)—representing factual assertions extracted from chart content. This appendix details our triplet-centric schema, the construction pipeline, and the canonicalization process.

D.1 Triplet-Aligned Schema

Each KG is structured as a set of atomic triplets (h, r, t), where:

- **head** (h)**:** a scientific entity (e.g., "Greenland ice sheet"),
- **relation** (r)**:** a semantic predicate (e.g., "contributes to", "amount", "experienced"), and
- **tail** (t)**:** a value, indicator, or another entity (e.g., "sea level rise", "3900 Gt").

To preserve key scientific details, each triplet is accompanied by a `metadata` object that captures contextual qualifiers such as time period, unit, trend, and uncertainty. This separation enables clear logical reasoning while preserving fidelity to the original chart.

Full triplets based on Figure 1 are shown below:


```
{
  "triplets": [
    {
      "head": "Greenland Ice Sheet",
      "relation": "experienced",
      "tail": "cumulative mass loss",
      "metadata": {
        "head_type": "Region",
        "tail_type": "Indicator",
        "time_range": "1979--2022",
        "temporal_granularity": "yearly"
      }
    },
    {
      "head": "cumulative mass loss",
      "relation": "trend",
      "tail": "decreasing",
      "metadata": {
        "head_type": "Indicator",
        "tail_type": "Trend",
        "time_range": "2000--2020"
      }
    },
    {
      "head": "cumulative mass loss",
      "relation": "contributes to",
      "tail": "sea level rise",
      "metadata": {
        "head_type": "Indicator",
        "tail_type": "Indicator",
```

Original Claim	Method	Refuted or NEI Claim
Refuted Claims		
The mean winter temperature in Wales has shown an upward trend from 1890 to 2020.	Trend Modification	The mean winter temperature in Wales has shown a downward trend from 1890 to 2020.
The general trend line for sunshine duration in Northern Ireland during spring suggests a slight upward shift over time since 1890.	Exaggeration	The general trend line for sunshine duration in Northern Ireland during spring suggests a significant upward shift over time since 1890.
Average sunshine duration in August 2021 for England was approximately 186.6 hours.	Metric Swap	Average maximum temperature in August 2021 for England was approximately 186.6 hours.
NEI (Not Enough Information) Claims		
The average temperature anomaly in April 2015 in Florida was around +3°F.	Conceptual Generalization	The average temperature anomaly in April 2015 in a coastal region was around +3°F.
The average temperature anomaly in April 2015 in Florida was around +3°F.	Conceptual Generalization	The average temperature anomaly in 2015 in Florida was around +3°F.
The average temperature anomaly in April 2015 in Florida was around +3°F.	Entity Replacement	The maximum temperature anomaly in April 2015 in Florida was around +3°F.

Table 7: Examples of generating **refuted** and **NEI (not enough information)** claims from original climate statements using different perturbation strategies. Color highlights the modified elements (red for refuted, blue for NEI).


```
        "time_range": "2000--2020"
      }
    },
    {
      "head": "sea level rise",
      "relation": "amount",
      "tail": "14 mm",
      "metadata": {
        "head_type": "Indicator",
        "tail_type": "Physical Measurement",
        "unit": "mm",
        "time_range": "2020",
        "temporal_granularity": "yearly",
        "uncertainty": "1 mm"
      }
    }
  ]
}
```


D.2 KG Construction Pipeline

We construct triplets automatically using GPT-4o (OpenAI, 2024), using the chart, caption, and the set of supported claims as the chart summary as inputs. We formulate prompts using a lightly constrained schema, instructing the model to extract semantically grounded (h, r, t) triplets with associated metadata fields.

D.3 Self-Canonicalization with LLMs

Following extraction, we canonicalize the surface forms of both entities and relations. Inspired by the Extract, Define, Canonicalize (EDC) framework (Zhang and Soh, 2024), we prompt the model to define and normalize semantically equivalent terms. The canonicalized form is extracted from chart captions and summaries. For instance:

- "was about" → `amount`
- "led to" → `contributes to`
- "Greenland" → `Greenland Ice Sheet`

This normalization enables consistency across charts and supports downstream evaluation using structured matching.

D.4 Coverage and Format

Triplets are generated only for supported claims to ensure factual consistency with the chart evidence. On average, each chart yields 6–8 canonicalized triplets. The final knowledge graphs are stored in structured JSON files, with each entry linked to its corresponding chart ID. These structured explanations serve dual purposes: (i) enhancing model interpretability and (ii) supporting multi-hop reasoning during fact verification.

These triplets capture causal and quantitative relationships essential for verifying scientific claims and provide a structured representation of the underlying chart semantics.

D.5 Limitations and Future Work

While the pipeline produces semantically coherent triplets, errors may arise from ambiguous captions or overloaded visual encodings. In future work, we aim to improve schema alignment with external

scientific ontologies, introduce confidence scoring per triplet, and extend the pipeline to cover refuted and NEI claims for contrastive reasoning.

E Manual Evaluation for Chart-to-Table Conversion

This design is motivated by recent work showing that supplementing visual inputs with structured tabular representations improves multimodal reasoning abilities. We aim to study whether this trend also holds for the task of chart-based fact-checking.

To ensure the reliability of our chart-to-table conversions using DePlot, we conduct a manual evaluation of a representative subset of charts in the CLIMATEVIZ dataset. We randomly sample 50 charts, stratified by chart type: 10 each from *line graph*, *bar chart*, *scatter plot*, *map*, and *pie chart*.

Each generated table is evaluated against three criteria:

Fidelity Does the table faithfully represent all relevant data values from the chart (e.g., axes, legends, numerical values)?

Omission/Misread Does the table omit or misinterpret any visual content (e.g., missing labels or incorrect numeric entries)?

Hallucination Does the table introduce spurious values or labels not present in the original chart?

Based on these criteria, we assign each chart-table pair to one of three categories: *Fully Accurate*, *Minor Issues*, or *Major Issues*.

We find that DePlot performs reliably on line and bar charts, where the data structure is linear and labeling is clear. It struggles more with pie charts, maps, and scatter plots, often due to overlapping text, spatial encoding, or small font sizes. This evaluation provides a level of trust in DePlot outputs while acknowledging limitations, especially for spatial or complex chart types.

F Experiments

The total computational time for all evaluations was approximately 300 GPU hours.

F.1 Prompt Templates

This appendix provides the full prompt templates used in our experiments across different settings. Each template reflects the exact structure used to prompt models in zero-shot and few-shot configurations. For few-shot settings, we include two demonstrations per class label (support, refute, and not enough information) to ensure balance.

F.1.1 Zero-Shot Prompt (CT, Label-Only Output)

Instruction: You are a scientific fact-checking assistant. Based on the chart caption and the claim, determine whether the claim is supported by the chart, refuted by the chart, or if there is not enough information. Respond with one of: `support`, `refute`, or `not enough information`.

Caption: Between 2000 and 2020, the Greenland Ice Sheet experienced accelerating mass loss, contributing to sea level rise.
Claim: The Greenland Ice Sheet saw stable mass over the period 2000–2020.
Answer:

F.1.2 Few-Shot Prompt (CT, Label-Only Output)

Includes two examples per label. The final query appears after all six examples.

Example 1
Caption: Average CO_2 levels rose from 370 ppm in 2000 to 412 ppm in 2020.
Claim: CO_2 levels have increased between 2000 and 2020.
Answer: `support`

Example 2
Caption: In 2022, the UK experienced the highest annual mean temperature on record.
Claim: The UK recorded its coldest year in 2022.
Answer: `refute`

Example 3
Caption: The Arctic sea ice extent varied significantly between 1979 and 2020, with notable seasonal fluctuations.
Claim: Arctic sea ice was 5 million sq km in 1999.
Answer: `not enough information`

Example 4
Caption: Annual rainfall in Southern England fluctuated with no clear trend over the last 50 years.
Claim: Annual rainfall in Southern England decreased significantly since 1970.
Answer: `refute`

Chart Type	# Samples	Fully Accurate	Minor Issues	Major Issues
Line Graph	10	8	1	1
Bar Chart	10	7	2	1
Scatter Plot	10	5	3	2
Map	10	4	4	2
Pie Chart	10	5	3	2
Total	**50**	**29**	**13**	**8**

Table 8: Manual evaluation of DePlot's chart-to-table outputs across five chart types. "Fully Accurate" indicates complete table fidelity; "Minor Issues" include small omissions or rounding mismatches; "Major Issues" involve missing core information or hallucinated values.

Example 5
Caption: Spring temperature anomalies in Scotland increased slightly between 1960 and 2020.
Claim: Scotland saw the largest anomaly in 1978.
Answer: `not enough information`

Example 6
Caption: Average summer temperature in Wales increased by 1.2°C from 1980 to 2020.
Claim: Summer temperature in Wales has warmed in the past 40 years.
Answer: `support`

Final Query
Caption: [Tcaption]
Claim: [Tclaim]
Answer:

F.1.3 Few-Shot Prompt (CTT, Explanation-Augmented Output)

Includes two examples per label with structured triplet explanations. The model must generate both reasoning triplets and the final label.

Example 1
Caption: Annual mean surface temperature in England from 2015 to 2020.
Table:

Year	Temperature (°C)
2015	9.5
2016	9.7
2017	9.6
2018	9.9
2019	10.1
2020	10.2

Claim: England's mean surface temperature rose steadily from 2015 to 2020.
Triplets:

- (England, Experienced, Surface Temperature Increase)
- (Surface Temperature, Trend, Increasing)
- (Time Period, Range, 2015–2020)
- (Temperature, Start Year Value, 9.5°C)
- (Temperature, End Year Value, 10.2°C)
- (Increase Amount, Computed Difference, 0.7°C)

Label: `support`

Example 2
Caption: Total rainfall in Wales from 2010 to 2015.
Table:

Year	Rainfall (mm)
2010	1400
2011	1380
2012	1450
2013	1390
2014	1420
2015	1410

Claim: Wales received significantly less rainfall in 2015 compared to earlier years.
Triplets:

- (Wales, Experienced, Rainfall)
- (Time Period, Range, 2010–2015)
- (Rainfall, Value in 2015, 1410 mm)
- (Rainfall, Mean Value 2010–2014, 1408 mm)
- (Rainfall in 2015, Comparative Trend, No Significant Decrease)

- (Rainfall Comparison, Difference from Average, +2 mm)

Label: `refute`

Example 3
Caption: Annual sunshine duration in Scotland between 1995 and 2000.
Table:

Year	Sunshine Hours
1995	1100
1996	1080
1997	1095
1998	1120
1999	1090
2000	1085

Claim: Scotland had the highest annual sunshine duration on record in 2001.
Triplets:

- (Scotland, Recorded, Sunshine Duration)
- (Time Period, Table Coverage, 1995–2000)
- (Sunshine Duration in 2001, Availability, Missing)
- (Max Sunshine in Table, Year, 1998)
- (Assertion Year 2001, Not Covered in Table, True)

Label: `not enough information`

Example 4
Caption: Annual CO_2 concentrations globally from 2010 to 2015.
Table:

Year	CO_2 (ppm)
2010	390.1
2011	392.6
2012	395.4
2013	397.9
2014	399.8
2015	402.3

Claim: CO_2 levels increased each year from 2010 to 2015.
Triplets:

- (Global Atmosphere, Measured, CO_2)
- (Time Period, Range, 2010–2015)
- (CO_2, Trend, Increasing)
- (CO_2 in 2010, Value, 390.1 ppm)
- (CO_2 in 2015, Value, 402.3 ppm)

Label: `support`

Example 5
Caption: Average spring temperatures in Northern Ireland from 2000 to 2005.
Table:

Year	Temperature (°C)
2000	8.2
2001	8.3
2002	8.5
2003	8.7
2004	8.9
2005	9.0

Claim: Spring temperatures in Northern Ireland gradually increased from 2000 to 2005.
Triplets:

- (Northern Ireland, Experienced, Spring Temperature Increase)
- (Time Period, Range, 2000–2005)
- (Spring Temperature, Trend, Increasing)
- (Spring Temperature in 2000, Value, 8.2°C)
- (Spring Temperature in 2005, Value, 9.0°C)
- (Increase Amount, Computed, 0.8°C)

Label: `support`

Example 6
Caption: Monthly average rainfall in Scotland in 2022.
Table:

Month	Rainfall (mm)
Jan	120
Feb	115
Mar	90
Apr	85
May	75
Jun	65
Jul	70
Aug	80
Sep	95
Oct	110
Nov	125
Dec	130

Claim: Rainfall in Scotland was highest in winter months during 2022.
Triplets:

- (Scotland, Observed, Monthly Rainfall)
- (Winter Months, Include, Dec–Feb)
- (Rainfall, Highest Values, Dec:130mm, Jan:120mm, Nov:125mm)
- (Winter Rainfall, Compared to, Higher than Summer)
- (Time Period, Year, 2022)
- (Rainfall, Seasonal Trend, Peak in Winter)

Label: `support`

Final Query
Caption: [Tcaption]
Table: [Ttable]
Claim: [Tclaim]
Triplets:
Label:

F.2 Table + Text Only Ablation.

To evaluate the importance of visual input in scientific chart-based fact-checking, we conduct an ablation study by removing the chart image and providing only the structured table (generated via DePlot) along with the chart caption and the claim as model input. This setting, denoted as **Table + Text**, isolates the contribution of the tabular and textual modalities, allowing us to assess whether models can accurately verify claims without access to the original chart. While this setup preserves key quantitative patterns through table representations, it lacks access to spatial, visual, and stylistic cues embedded in the chart. Our results indicate that performance drops noticeably compared to the full **Chart + Table + Text** (CTT) setting, highlighting the complementary role of visual features in supporting accurate and interpretable fact verification.

Table 9 presents the ablation results under the Table + Text (TT) setting, where the chart image is omitted. Across both open- and closed-source models, we observe that performance declines moderately compared to the full CTT setting, indicating that visual input contributes complementary information beyond structured tabular data. InternVL 2.5 and Qwen 2.5 achieve strong performance, with InternVL reaching the highest accuracy (55.6%) in the few-shot condition. Interestingly, Gemini 2.5 yields the best zero-shot results (53.5% accuracy, 53.4% F1), but performance degrades slightly with few-shot prompting—mirroring the instability seen in other few-shot settings. Among open-source models, all benefit consistently from few-shot prompting, whereas closed-source models exhibit marginal gains or regressions. These results confirm that while structured table representations alone are effective, incorporating chart visuals remains essential for optimal fact verification performance.

Model	Setting	Acc-L (TT)	F1-L (TT)
Closed-source			
o3	Zero-shot	47.6	43.7
	Few-shot	49.2 (↑)	45.1 (↑)
GPT-4o	Zero-shot	51.5	48.0
	Few-shot	52.9 (↑)	49.2 (↑)
Gemini 2.5	Zero-shot	**53.5**	**53.4**
	Few-shot	52.2 (↓)	52.5 (↓)
Open-source			
LLaMA-4-Maverick-17B	Zero-shot	47.2	45.3
	Few-shot	52.5 (↑)	49.7 (↑)
InternVL 2.5-78B	Zero-shot	53.3	49.1
	Few-shot	**55.6** (↑)	**51.0** (↑)
Qwen 2.5-VL-72B	Zero-shot	52.8	48.8
	Few-shot	54.4 (↑)	51.2 (↑)

Table 9: Ablation study results for the **Table + Text (TT)** setting, where models receive only the DePlot-generated table, chart caption, and claim, omitting the chart image. Acc-L and F1-L denote label-only accuracy and macro F1, respectively. Arrows indicate intra-model performance change from zero-shot to few-shot. Bolded values are the best per column.