ClimateViz: A Dataset for Evaluating the Fact-checking and Reasoning Abilities of LLMs using Facts Extracted from Scientific Graphs

Anonymous ACL submission

Abstract

This paper introduces ClimateViz, the largest dataset to date for evaluating the fact-checking and reasoning capabilities of large language models (LLMs) in the climate science domain. ClimateViz comprises claims extracted by humans from high-quality scientific graphics, and checked for acccuracy and domain relevance. To advance the SOTA in NLP for fact-checking, we develop a robust pipeline that systematically generates claims that are highly similar, but false. Additionally, we introduce ReasonClim, a complementary benchmark built using graph-based methods to evaluate spatial, temporal, and spatio-temporal reasoning tasks. To assess LLM's performance on these tasks, we conduct a comprehensive evaluation of the state-of-the-art models. Our findings demonstrate that LLMs struggle with detecting certain types of false claims, especially those generated through exaggeration. The results also highlight significant challenges in fact verification and reasoning over climate data, particularly in temporal reasoning tasks. By providing a benchmark for evaluating LLMs on real-world climate data, ClimateViz and ReasonClim support the development of more reliable AI systems for climate science applications.

1 Introduction

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities across a broad range of natural language processing (NLP) tasks (Gemini et al., 2024). Despite these successes, LLMs continue to struggle with fact verification and complex reasoning, in domains such as climate science (Manivannan et al., 2024). Accurately assessing LLMs in this domain is critical, as climate misinformation can significantly impact public perception and policy decisions (Diggelmann et al., 2021). However, current benchmarks remain limited, focusing primarily on text-based fact-checking while overlooking multi-modal evidence, spatio-temporal reasoning, and structured knowledge representations.

Dataset	Category	Size	Text Fact-Checking	Multimodal Evidence	Multi-step Reasoning
FEVER	General	185k	✓	✗	✗
PlotQA	General	28.9M	✗	✓	✗
ClimateFEVER	Domain-Specific	1.5k	✓	✗	✗
GeoQA	Domain-Specific	5k	✗	✗	✓
ChartQA	General	23k	✗	✓	✗
Factify	General	50k	✓	✓	✗
MMMU	General	11.5k	✓	✓	✗
TRAM	General	526k	✗	✗	✓
MMFakeBench	General	11k	✓	✓	✗
CV+RC†	Domain-Specific	20k	✓	✓	✓

Table 1: Comparison of existing fact-checking and reasoning datasets. † **CV+RC** refers to ClimateViz + ReasonClim. Multi-step Reasoning includes both Spatio-Temporal and Graph-Based Reasoning.

Climate science relies heavily on scientific graphics—such as bar charts, line graphs, and maps—to communicate trends, anomalies, and forecasts (Xu et al., 2024). Yet, existing datasets rarely evaluate LLMs' ability to verify claims from such graphics. Additionally, climate-related reasoning often involves spatial dependencies (e.g., temperature variations across regions) and temporal trends (e.g., long-term CO_2 emissions), which current fact-checking datasets fail to systematically capture (Cheng et al., 2024; Chu et al., 2024). Furthermore, multi-step reasoning, which is essential for rigorous inference in scientific domains, remains largely unexplored in existing benchmarks.

To bridge these gaps, we introduce ClimateViz, the first dataset designed to evaluate fact-checking based on information in scientific publications including the information in scientific graphics, as shown in Table 1. ClimateViz also includes a pipeline for generating realistic false claims, improving robustness in evaluating LLMs' misinformation detection capabilities.

Complementing ClimateViz, we also present ReasonClim. ReasonClim contains a knowledge representation graph that integrates facts from Cli-

mateViz, plus a test suite of questions and answers that leverages this graph to evaluate spatial, temporal, and spatio-temporal reasoning. ReasonClim integrates spatial and temporal knowledge from climate indicators, allowing for a more comprehensive assessment of LLM's ability to infer, predict, and verify claims grounded in climate data.

In addition to dataset constrution, we conduct a comprehensive evaluation of 7 state-of-the-art LLMs to assess their ability to verify claims from ClimateViz and perform spatial and temporal reasoning over ReasonClim. While some models exhibit strong fact-checking abilities, they struggle with specific reasoning tasks, particularly temporal reasoning and detecting exaggerated claims.

By providing a new benchmark for evaluating LLMs on real-world climate data, ClimateViz and ReasonClim offer a testbed for developing agents capable of robust fact verification, domain-specific reasoning, and misinformation detection in climate science.

2 Relation to Previous Work

2.1 Fact-Checking and Misinformation in NLP

The field of automated fact-checking has evolved significantly. Early benchmarks like LIAR (Wang, 2017) and FEVER (Thorne et al., 2018) focus on text-based claim verification. While these foundational datasets have advanced the field, the reliance on textual data limited their applicability to multimodal contexts. To address this, recent datasets such as MMFakeBench (Anonymous, 2025) and Factify (Suryavardan et al., 2023) integrate text and visual data, enabling multi-modal misinformation detection. Other domain-specific efforts, such as MM-COVID (Li et al., 2020) and Fauxtography (Zlatkova et al., 2019), highlight the growing importance of multi-modal fact-checking in medical and general domains.

ClimateFEVER (Diggelmann et al., 2021) introduced the first climate-specific fact-checking dataset. Its claims are based on reputable sources such as Wikipedia. However, it is limited to textual information; and it does not address the challenges of processing claims from scientific graphics.

ClimateViz builds on these advances by offering the largest gold standard dataset in the climate domain to date. It contains over 15,000 true claims and related 5,000 false claims. The true claims are derived from highly validated sources. The false claims are constructed over the same vocabulary and are guaranteed to be false. We built the dataset to incorporate scientific graphics as a primary modality, bridging the gap between textual and visual fact-checking.

2.2 Multi-Modal and Visual Information Extraction

The integration of visual data in NLP tasks has led to the development of datasets like ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020), which focus on extracting and reasoning over data from charts and plots. While these resources address general-purpose visual reasoning, they do not account for domain-specific challenges, such as those found in climate science.

ClimateViz builds upon these efforts by emphasizing climate-specific data representations, such as temperature anomalies and carbon emissions trends, and by incorporating human annotations to ensure domain relevance.

2.3 Spatial and Temporal Reasoning in NLP

Spatial and temporal reasoning are crucial for analyzing real-world phenomena (Smith et al., 2023), particularly in domains like climate science, where relationships between time and space drive critical insights. In NLP, temporal reasoning tasks often focus on understanding chronological data and extracting time-sensitive information (Qin et al., 2021), as exemplified by work such as TRAM (Wang and Zhao, 2024), Similarly, spatial reasoning benchmarks like GeoQA (Chen et al., 2022) assess models' ability to comprehend geographic relationships in textual data.

Because these benchmarks primarily operate within text-based domains, they have limited capacity to evaluate reasoning over combined spatial and temporal dimensions, particulary in domain-specific contexts where such information is typically presented in graphics that complement the text. For instance, while temporal reasoning in TRAM explores event-based temporal relations, it does not extend to analyzing trends or anomalies over time and space in scientific data.

ReasonClim addresses these limitations by introducing spatio-temporal reasoning tasks specifically tailored to climate science. Using a knowledge graph constructed from climate-related claims, ReasonClim generates structured reasoning tasks that evaluate LLMs' ability to navigate and analyze relationships across spatial (e.g., geographic re-

gions), temporal (e.g., specific years or seasons), and spatio-temporal dimensions (e.g., anomalies occurring in specific regions over time). By integrating information derived from graphics, ReasonClim enables nuanced assessment of LLMs in complex reasoning scenarios relevant to climate science.

2.4 Benchmarks and Evaluation of LLMs in Climate Science

The application of NLP to climate science is an emerging area of research (Leippold et al., 2024). Early efforts, such as ClimateFEVER (Diggelmann et al., 2021), introduced text-based fact-checking datasets tailored to the climate domain. Topic modeling for climate-related text corpora, as demonstrated in (Gokcimen and Das, 2024), and trend analysis using news dataset (Dorfleitner and Zhang, 2024), showcased the potential of NLP for climate communication. However, these approaches remain limited to textual analysis, excluding the critical visual and spatio-temporal dimensions often found in the graphics that complement the text. Multi-modal approaches are promising but have yet to be fully realized in climate science. For example, benchmarks like MMMU (Yue et al., 2024) and MMBench (Liu et al., 2024) demonstrate multi-modal capabilities, but they lack the domain-specific focus needed to address complex relationships inherent in climate science.

ClimateViz and ReasonClim aim to bridge these gaps by providing the first comprehensive benchmark tailored to fact-checking and spatio-temporal reasoning in climate science. ClimateViz incorporates scientific graphics as a primary modality, advancing the application of NLP in addressing critical challenges in climate science.

3 Methodology

3.1 Dataset Creation and Annotation

We scraped the scientific graphics from 6 respected open-domain sources with metadata. [1] We used them to design a three-tasked project on Zooniverse[2], a well-established and influential citizen-science project, c.f (Fortson et al., 2011; Simpson et al., 2014). The first two tasks are to define titles of the graphic and to choose the types of data representation of the graphic. The third task is to summarize true facts based on the graphic alone (and not on the annotators' background knowledge). The annotators are prompted to objectively describe significant elements observed in the graphic, such as key data points, trends, comparisons, and any notable anomalies or patterns. The answers to the first two tasks provide context for the third task, which is the source of the text in the ClimateViz dataset.

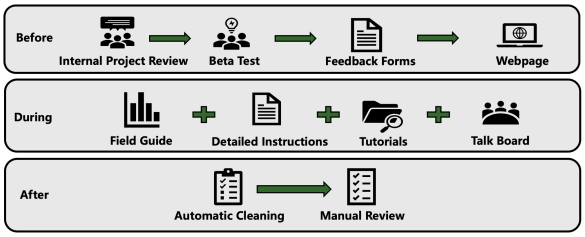

Figure 1: Quality Control Process: before, during, and after annotation

Our carefully designed quality-control process involved three distinct phases: before annotation, during annotation, and after annotation (see Figure 1 and Appendix A). Post-annotation, we scrutinized the annotations for correctness and consistency. We also went through each claim to make sure it contained sufficiently precise context to sustain multi-step reasoning. We ended up with 15,100 true claims in ClimateViz.

3.2 False Statement Generation via Augmentation Pipeline

To evaluate the capabilities of LLMs in discerning factual information from misinformation, we developed an augmentation pipeline to systematically generate false statements based on the true climate-related claims as shown in Figure 2. The generated false facts are designed to resemble real-world misinformation scenarios that could lead to misinterpretations.

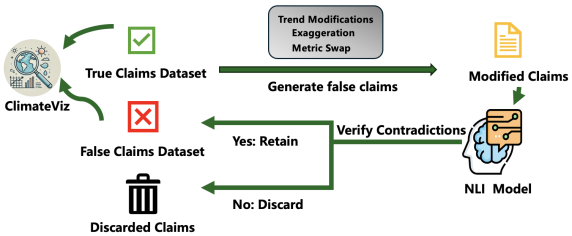

Figure 2: False Statement Generation Pipeline

Once false facts were generated, we employed DeBERTa-Large-MNLI to ensure that the false

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

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

claims generated had a high likelihood of being semantically contradictory with original ones. We set a confidence threshold of 0.8 so that only clear contradictions were retained. Taking together the false claims and the 15,100 true claims, ClimateViz contains 20,119 claims in total (See Table 2).

Statistic	Value
True claims	15,100
False claims	5,109
False Claims by Method	
Trend modification	328
Exaggeration	3,496
Metric swap	1,285
Total claims	**20,119**

Table 2: Statistics of true and false claims in ClimateViz, including a breakdown of false claims generated by different methods.

3.3 Graph-Based Reasoning Tasks and Groundtruth Generation

ReasonClim is a complimentary dataset derived from ClimateViz using a graph-based approach to capture complex relationships inherent in climate data, including spatial, temporal, and spatio-temporal interconnections. We employed GPT-4o (Islam and Moushi, 2024) to break complex claims into single claims, making sure that each single claim retains enough contextual information (see Figure 3).

We constructed a directed knowledge graph $G = (N, E)$, where N represents the set of nodes and E represents the set of edges. The graph encodes relationships between climate-related entities to facilitate spatio-temporal reasoning and fact-checking tasks.

The nodes in N include: **Region nodes** n_{region}, **ClimateIndicator nodes** n_{climate}, **Record nodes** n_{record}, and **TimePeriod nodes** n_{time}, each uniquely identified and characterized by properties such as values, units, or descriptions.

The edges in E are:

$$(n_{\text{region}}, n_{\text{climate}}, \text{EXPERIENCED}) \quad (n_{\text{climate}}, n_{\text{record}}, \text{HAS_RECORD})$$
$$(n_{\text{record}}, n_{\text{region}}, \text{RECORDED_AT}) \quad (n_{\text{record}}, n_{\text{time}}, \text{OCCURRED_IN})$$

These define the relationships between nodes shown in Figure 3 . This graph structure enables complex reasoning and querying across climate data by traversing the relationships.

We employed spaCy model to extract geographic locations from each fact and pattern matching to identify time periods (e.g., years, seasons, months), climate indicators and records. Edges are created based on relevant source and target.

This structured approach enabled us to create a rich, interlinked graph where each climate-related fact was represented in a way that is easy for graph queries. In the graph we build, there are 12147 edges and 1508 nodes.

Reasoning questions were then generated through graph traversal techniques that utilized various types of nodes and edges, each representing a different aspect of climate data relationships. Specifically, we aimed to generate three categories of reasoning tasks: spatial, temporal, and spatio-temporal, each task was paired with a ground truth answer and a detailed explanation to ensure both interpretability and reliability of the generated data. Temporal reasoning tasks focused on understanding changes across different years or seasons, while spatial tasks were geographically oriented, and spatio-temporal tasks involved multidimensional reasoning that integrated both elements. Our goal was to provide a challenging benchmark for evaluating LLMs' abilities to reason about climate data beyond basic factual recall.

Temporal Questions: For each climate indicator n_{climate}, we query edges:

$$E_{\text{has_record}} = \{(n_{\text{climate}}, n_{\text{record}}, \text{HAS_RECORD})\},$$

retrieve records n_{record}, and query edges of type OCCURRED_IN:

$$E_{\text{occurred}} = \{(n_{\text{record}}, n_{\text{time}}, \text{OCCURRED_IN})\}.$$

We filter by time period n_{time} of type "Year" and ensure the match in the record description. The generated question is:

$$q_{\text{temporal}} : \text{In which year did } n_{\text{climate}} \text{ occur in } n_{\text{region}}?$$

with ground truth $g_{\text{temporal}} =$ year of occurrence.

Spatial Questions: For spatial questions, we focus on whether a climate anomaly occurred in a region during a given time period. We query edges:

$$E_{\text{experienced}} = \{(n_{\text{region}}, n_{\text{climate}}, \text{EXPERIENCED})\},$$

retrieve records and check edges of OCCURRED_IN:

$$E_{\text{occurred}} = \{(n_{\text{record}}, n_{\text{region}}, \text{RECORDED_AT})\}.$$

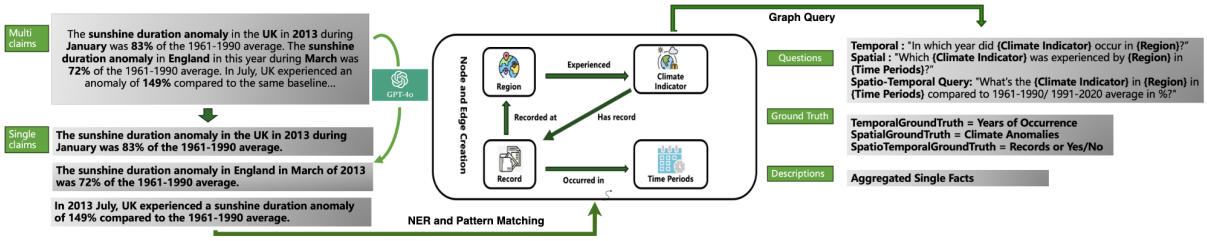

Figure 3: Graph Query for ReasonClim Generation

If the record contains region and time period matches, the generated question is:

q_{spatial} : Which climate anomaly was experienced by n_{region} in n_{time}?

with ground truth $g_{\text{spatial}} = n_{\text{climate}}$.

Spatio-Temporal Questions: For each climate indicator n_{climate} and region n_{region}, we query edges:

$$E_{\text{experienced}} = \{(n_{\text{region}}, n_{\text{climate}}, \text{EXPERIENCED})\},$$

retrieve records n_{record}, and check edges of OCCURRED_IN:

$$E_{\text{occurred}} = \{(n_{\text{record}}, n_{\text{time}}, \text{OCCURRED_IN})\}.$$

If the record contains anomaly values, there are two types of question being generated:

$q_{\text{spatio_temporal}}$: "What's the n_{climate} in n_{region} in n_{time} compared to the average?"

or "Did n_{region} experience n_{climate} in n_{time}?"

Ground truth can either be a percentage value or "Yes"/"No" depending on the context.

Ground Truth and Explanation Generation To enhance transparency, each reasoning task was paired with an explanation that described the graph traversal used to reach the answer. These explanations included references to the specific nodes and relationships in the knowledge graph. This approach was designed to foster not only the accuracy of the answer, but also the interpretability, which is crucial to ensuring trust in AI-generated conclusions, especially in high-stakes domains such as climate science.

In ReasonClim, we compiled 294 spatial questions, 294 temporal questions, and 528 spatio-temporal questions with ground truth and explanations.

4 Task Design and Baselines

To evaluate the performance of LLM in the climate science domain, we designed tasks targeting fact-checking and reasoning capabilities.

4.1 Fact-Checking Task

Fact-checking is essential in climate science due to the prevalence of misinformation (Leippold et al., 2024), which can undermine public understanding and policy development. This task evaluates the models' ability to classify climate-related claims as true or false, a critical competency for verifying insights from scientific graphics and textual data. We sampled 1,000 claims from the ClimateViz, maintaining a 7:3 proportion of true to false claims to simulate real-world scenarios. Additionally, false claims were randomly selected from each generation method to ensure balanced evaluation of model sensitivity to different types of misinformation.

4.2 Reasoning Tasks

The reasoning tasks are designed to evaluate models' capacity to perform spatial, temporal, and spatio-temporal reasoning. These dimensions are critical in climate science, where understanding relationships across geographic regions, timeframes, and combined spatio-temporal patterns is fundamental to interpreting trends and anomalies. By addressing these reasoning tasks, we benchmark models' ability to not only extract information but also analyze and synthesize it in a structured and meaningful way.

4.3 Selection of Models

We evaluated 7 models: GPT-4o (Islam and Moushi, 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), O1 (Zhong et al., 2024), GPT-4 (OpenAI et al., 2024), GPT-3.5 (Ye et al., 2023), Phi-3 (Abdin et al., 2024), and Llama-3-8B (Grattafiori et al., 2024) These models were selected based on their state-of-the-art performance in language understanding and reasoning.

Our tasks are very challenging, but we note three ways that the models could succeed. Deepseek-R1, which is the only model evaluated that have the ability to search the web on-the-fly to answer questions could access the exact high-profile sources that we took our graphics from, and process either the graphics themselves or the accompanying .csv files. The training sets for the others would have in principle had access to a substantial fraction of these on-line materials, and implicitly represented world knowledge that could be derived from them. Alternatively the correct answers to some of the questions could be derived as entailments from other sources.

5 Experimental Setup

5.1 Model Configuration

The experiments are running on a NVIDIA A100 Tensor Core GPU and approximately 24 GB of available RAM.

5.2 Evaluation Metrics

To comprehensively evaluate model performance, we employ the following metrics:

5.2.1 Fact-Checking Metrics.

Abstention Rate (%): The percentage of questions where the model abstains from answering.

Balanced Accuracy (%): Accounts for class imbalance by averaging recall across true and false claims:

$$\text{Balanced Accuracy} = \frac{\text{Recall}_{\text{true}} + \text{Recall}_{\text{false}}}{2}$$

Recall (%): Measures the proportion of correctly identified true claims.

F1-Score (%): Balances precision and recall, providing a single measure of classification performance.

5.2.2 Reasoning Task Metrics.

Reasoning tasks are evaluated using the following metrics:

Abstention Rate (%): The percentage of questions where the model abstains from answering.

Mean Evaluation (%): The average evaluation score for non-abstained questions, calculated using task-specific criteria:

- For *Spatial and Temporal Reasoning*: Weighted Partial Match is used:

$$\text{Score} = \frac{|P \cap G|}{|P \cup G|}$$

- For *Spatio-Temporal Reasoning*:
  - Numerical questions: Predictions within 10% of the ground truth are scored as correct.
  - Binary classification questions: Exact matches are scored as correct.

Evaluation Standard Deviation (%): Captures the variability in model performance across answered questions.

Task-Specific Aggregation. For overall metrics across reasoning tasks, weighted averages are computed:

$$\text{Overall Metric} = \frac{\sum_{t=1}^{T} N_t \cdot M_t}{\sum_{t=1}^{T} N_t}$$

where M_t is the metric value for task t and N_t is the number of questions in task t.

These metrics ensure a holistic evaluation, capturing both accuracy and consistency across diverse task types.

6 Results

6.1 Model Evaluation for Fact-checking Task

As shown in Table 3, GPT-4o has the highest balanced accuracy in the fact-checking task but also has a high abstention rate, which implies the model may only provide answers when it is confident.

Llama-3 has the highest recall and a strong F1-score, indicating it is particularly good at identifying true positives while maintaining a good balance between precision and recall. That implies, Llama-3 is highly sensitive to identifying correct facts, however, it might struggle with classifying negative instances.

6.2 Model Evaluation for Reasoning Tasks

To evaluate the performance of language models across reasoning tasks, we calculate overall metrics by aggregating results across spatial, temporal, and spatio-temporal reasoning tasks (See Table 4).

Overall, temporal reasoning is the hardest task. All models exhibited their lowest scores in the temporal reasoning task. There is a trade-off between

Metric	o1	DeepSeek-R1	GPT-4o	GPT-4	GPT-3.5	Phi-3	Llama-3
Abstention Rate(%)	0.9	1.60	9.3	1.4	**0.00**	7.2	**0.00**
Balanced Accuracy(%)	64.69	61.92	**66.81**	56.43	65.71	58.57	63.48
Recall(%)	43.52	39.39	48.89	16.28	73.43	35.67	**84.29**
F1-Score(%)	58.19	53.93	62.87	27.66	76.72	49.57	**80.71**

Table 3: Evaluation metrics for all models on the Fact-checking task.

Metric	o1	DeepSeek-R1	GPT-4o	GPT-4	GPT-3.5	Phi-3	Llama-3
Spatial Reasoning							
Abstention Rate (%)	2.17	**0.00**	3.26	**0.00**	**0.00**	**0.00**	**0.00**
Mean Evaluation (%)	41.48	4.98	37.83	**45.65**	20.29	43.82	31.16
Evaluation Std Dev (%)	31.92	**17.47**	35.69	24.19	24.53	37.60	31.57
Temporal Reasoning							
Abstention Rate (%)	14.29	14.29	**0.00**	7.14	**0.00**	**0.00**	**0.00**
Mean Evaluation (%)	0.56	8.68	2.42	3.89	5.72	2.93	**11.92**
Evaluation Std Dev (%)	**1.92**	6.26	3.41	5.58	5.29	3.54	10.44
Spatio-Temporal Reasoning							
Abstention Rate (%)	20.08	**0.00**	4.92	5.49	**0.00**	4.92	**0.00**
Mean Evaluation (%)	29.15	31.25	**37.85**	24.85	37.12	14.34	30.11
Evaluation Std Dev (%)	45.50	46.4	48.55	43.26	48.36	**35.09**	45.92
Overall Metrics							
Abstention Rate (%)	17.35	0.32	4.57	4.73	**0.00**	4.10	**0.00**
Mean Evaluation (%)	30.31	26.94	**37.06**	27.41	33.98	18.37	31.06
Evaluation Std Dev (%)	43.31	41.32	46.40	40.60	45.17	**35.11**	45.92

Table 4: Metrics of different models across spatial, temporal, and spatio-temporal reasoning tasks, along with overall metrics.

accuracy and consistency. Models such as GPT-4o achieve high accuracy but exhibit higher variability (e.g., high standard deviation in spatio-temporal task), while models like Phi-3 demonstrate more consistent performance but lower overall accuracy.

GPT-4o shows the best performance across all the reasoning tasks. While almost all models struggle with temporal reasoning tasks (See Figure 5), Llama-3 performs better on temporal reasoning, showing a noticeable drop in error rate compared to the other models. DeepSeek-R1 is has the most consistent performance across all tasks, but it has surprisingly low performance on spatial reasoning task, which means the model may lack access to granular historical climate data, especially in European countries. Incorporate temporal embeddings and causal graphs to model time-dependent relationships (e.g., lagged effects of emissions) may mitigate this problem.

False positives are of particular interest because false but plausible claims are the stuff of disinformation campaigns and conspiracy theories. So, we conduct further error analysis by calculating the error rates by false claim categories for the fact-checking task and by reasoning categories for the reasoning task. As shown in Figure 4, Llama3 and GPT-3.5 have particularly high errors rates for false claim. Overall, error rates for the exaggeration and metric-swap claims are higher than for the trend

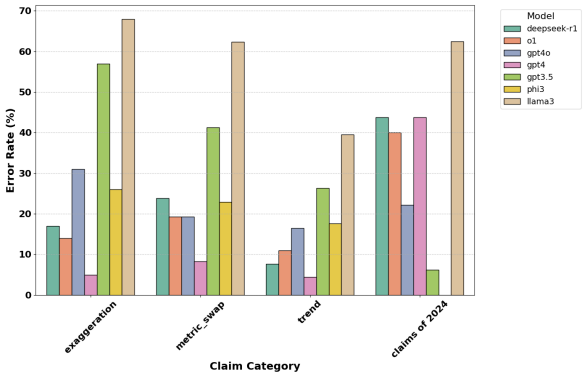

Figure 4: Error Rates for Fact-checking Task Across Models

false claims.

All the models' training data cutoff dates (except for DeepSeek-R1, of which the training data isn't explicitly mentioned) are earlier than our dataset's August 2024 cutoff for scientific graphics. Consequently, models must rely on extrapolation, retrieval-augmented mechanisms, or inherent reasoning abilities to verify 2024 claims.So we also tested the models' fact-checking performance on all claims of the year 2024. As shown in Figure 4, Phi-3's exceptional performance may suggest a training approach that enables robust extrapolation from past data. Despite its older architecture, GPT-3.5 has a significantly lower error rate than

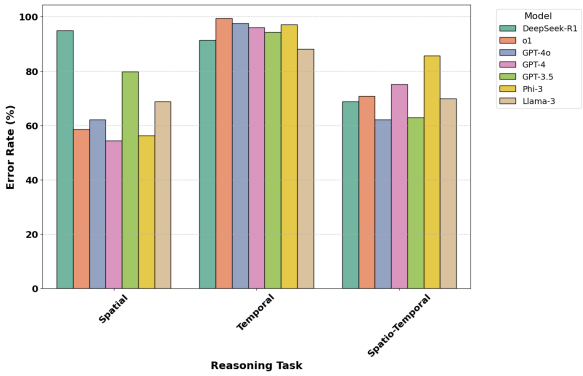

Figure 5: Error Rates by Reasoning Task Across Models

many more advanced models, suggesting a lower tendency to hallucinate or overgeneralize in areas where it lacks knowledge.

6.3 Broader Implications

Advancing NLP in High-Stakes Domains. Climate science is a domain where the stakes of misinformation are particularly high. The datasets and benchmarks introduced in this work support the development of more accurate and robust LLMs capable of processing complex climate-related data. This advancement has the potential to support scientific decision-making, enhance public understanding, and combat climate misinformation effectively.

Multi-Modal and Spatio-Temporal Reasoning. The rather low performance levels displayed in Table 4 indicates that LLMs are unable to extract information from scientific graphics that human readers readily extract. This highlights the importance of moving beyond text-based benchmarks for scientific domains. This work points towards the need for multi-modal NLP systems capable of synthesizing insights across visual and textual data, a requirement for addressing challenges in other domains, such as healthcare, economics, and environmental policy.

Enhancing Model Explainability. The detailed ground truth and explanations provided in ReasonClim emphasize interpretability, fostering trust in AI systems deployed in high-stakes applications.

6.4 Limitations

The datasets are curated from specific gold standard sources. Applying the approach to a broader or different domain would require a fresh curation effort. The variety of scientific graphics in ClimateViz is limited, types of visual images, such as satellite imagery or video data, are not included, presenting opportunities for future expansion. Plus, the temporal reasoning tasks focus on explicit time units (e.g., years, seasons) but do not account for less explicit temporal references that might be comprehensible to humans but still present challenges for LLMs. Additionally, we only evaluate LLMs with text as input in this work and do not test Large Multimodal Models (LMMs) that integrate both text and images. To address these limitations, future work could explore: Expanding the dataset to include more diverse modalities, such as geospatial and sensory data. Incorporating data that require disambiguation of named entities or timepoints. Using graphics as input to evaluate LMMs ability for multimodal fact-checking.

7 Conclusion

In this work, we introduced ClimateViz, the largest dataset to date for evaluating fact-checking capabilities of large language models in climate science, and ReasonClim, a complementary benchmark for spatio-temporal reasoning tasks. Together, these datasets address critical gaps in the NLP landscape by integrating scientific graphics, human-labeled claims, and graph-based reasoning tasks to support robust evaluation across multi-modal and domain-specific challenges.

Our findings demonstrate that current state-of-the-art LLMs struggle with fact-checking and spatio-temporal reasoning tasks in climate science, highlighting the need for more targeted research in these areas. Additionally, the emphasis on interpretability through detailed ground truth and explanations underscores the importance of explainable AI.

We envision ClimateViz and ReasonClim as catalysts for advancing research on multi-modal fact-checking, domain-specific reasoning, and trustworthy AI. Our work aims to bridge the gap between AI capabilities and real-world needs, ultimately contributing to informed decision-making and combating misinformation in the fight against climate change.

Acknowledgments

References

Marah Abdin, Jyoti Aneja, and Hany Awadalla et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Anonymous. 2025. MMFakebench: A mixed-source multimodal misinformation detection benchmark for LVLMs. In *The Thirteenth International Conference on Learning Representations*.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2022. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *Preprint*, arXiv:2105.14517.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision language models. *Preprint*, arXiv:2406.01584.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, and Dejian Yang et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Gregor Dorfleitner and Rongxin Zhang. 2024. Esg news sentiment and stock price reactions: A comprehensive investigation via bert. *Schmalenbach Journal of Business Research*, 76.

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.

Gemini, Rohan Anil, and Sebastian Borgeaud et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Tunahan Gokcimen and Bihter Das. 2024. Exploring climate change discourse on social media and blogs using a topic modeling analysis. *Heliyon*, 10(11):e32464.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2024. Automated fact-checking of climate change claims with large language models. *Preprint*, arXiv:2401.12566.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *Preprint*, arXiv:2011.04088.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. 2024. Climaqa: An automated evaluation framework for climate foundation models. *Preprint*, arXiv:2410.16701.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *Preprint*, arXiv:2203.10244.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. *Preprint*, arXiv:1909.00997.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and Janko Altenschmidt et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. *Preprint*, arXiv:2106.04571.

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.

Michael J Smith, Luke Fleming, and James Geach. 2023. Earthpt: a foundation model for earth observation. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*.

S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A multimodal fake news and satire news dataset. *Preprint*, arXiv:2304.03897.

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.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yuqing Wang and Yun Zhao. 2024. Tram: Benchmarking temporal reasoning for large language models. *Preprint*, arXiv:2310.00835.

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.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Tianyang Zhong, Zhengliang Liu, and Yi Pan et al. 2024. Evaluation of openai o1: Opportunities and challenges of agi. *Preprint*, arXiv:2409.18486.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. *Preprint*, arXiv:1908.11722.

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 climate graphics. 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 visual climate 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 graphics. The tools used during this phase include:

Field Guide: A comprehensive field guide was provided to the annotators, covering the different types of data representations commonly found in the graphics. This guide included:

Types of Visuals: Examples of bar charts, line graphs, pie charts, scatter plots, geographic maps, and box plots, 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.

Detailed 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 graphics. Annotators were instructed to focus on objective descriptions, providing factual statements regarding the graphic without interpretation or bias.

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

Talk Board: The Zooniverse platform also included a dedicated "Talk Board," where annotators could discuss uncertainties, ask questions, and receive support from both project moderators and their peers. 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.

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.

Manual Review: Following the automated cleaning, the data underwent a manual review by domain experts. 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 False Staement Generation

B.1 False Fact Generation

False facts are generated by manipulating true facts through various strategies. Three distinct methods are employed for this purpose:

Trend Modification: This method involves altering the directional trend of a fact, such as changing "increased" to "decreased" or "rising" to "falling." Such modifications reverse the implied trends in the facts.

Exaggeration: This method amplifies numerical values or descriptive terms to exaggerate the magnitude of the claim. Numerical values (e.g., temperature, precipitation) are adjusted using a random multiplier, while adjectives such as "moderate" or "slight" are replaced with more extreme alternatives (e.g., "severe" or "considerable").

Metric Swap: This approach involves replacing specific metrics or variables with similar but distinct ones. For example, replacing "mean maximum temperature" with "mean minimum temperature" or swapping "sunshine duration" with "cloud cover."

Here are some real examples in the dataset, see Table 5.

Each true fact undergoes one of these transformations at random, generating a modified version that is considered a potential false fact.

B.2 Contradiction Detection via NLI Model

Once the false facts are generated, we employed the microsoft/deberta-large-mnli model for verifying contradiction between generated and original claims. The NLI model classifies the relationship between two text statements as entailment, neutral, or contradiction. The false facts are selected only if they are classified as contradictions with a high confidence score (greater than 0.8).

C Climate Knowledge Graph

The Climate Knowledge Graph (CKG) we built in this paper is a directed graph $G = (N, E)$, where N represents the set of **nodes** and E represents the set of **edges**. The graph encodes relationships between climate-related entities (regions, climate indicators, time periods, and numerical records) to facilitate spatio-temporal reasoning and fact-checking tasks.

C.1 Nodes

The nodes in the graph represent distinct entities in the climate domain. Each node $n \in N$ is categorized by its type t.

- **Region nodes** n_{region}: Represent geographical locations (e.g., "London," "Africa").
- **ClimateIndicator nodes** n_{climate}: Represent climate indicators (e.g., "Rainfall Anomaly," "Temperature Anomaly").
- **Record nodes** n_{record}: Represent numerical data associated with climate indicators (e.g., "400 ppm" of CO_2).
- **TimePeriod nodes** n_{time}: Represent temporal entities such as years, months, or seasons (e.g., "2020," "Winter").

Each node n is uniquely identified by a node ID and may contain properties such as numerical values, units, or descriptions.

C.2 Edges

Edges represent the **relationships** between nodes. Each edge $e \in E$ connects two nodes and is characterized by a relationship type r.

Original Claim	Method	False Claim
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.
Sunshine duration in August 2021 for England was 115% compared to the 1961-1990 average.	Metric Swap	Cloud cover in August 2021 for England was 115% compared to the 1961-1990 average.

Table 5: Examples of False Claim Generation Methods

- **EXPERIENCED** edges: Represent the relationship where a Region n_{region} experiences a ClimateIndicator n_{climate}.

$$(n_{\text{region}}, n_{\text{climate}}, \text{EXPERIENCED})$$

- **HAS_RECORD** edges: Represent the relationship where a ClimateIndicator n_{climate} is associated with a Record n_{record}.

$$(n_{\text{climate}}, n_{\text{record}}, \text{HAS_RECORD})$$

- **RECORDED_AT** edges: Represent the relationship where a Record n_{record} is associated with a Region n_{region}, indicating the location of the measurement.

$$(n_{\text{record}}, n_{\text{region}}, \text{RECORDED_AT})$$

- **OCCURRED_IN** edges: Represent the relationship where a Record n_{record} is associated with a TimePeriod n_{time}, indicating when the measurement occurred.

$$(n_{\text{record}}, n_{\text{time}}, \text{OCCURRED_IN})$$

C.3 Graph Construction

The Climate Knowledge Graph is built by iterating over a dataset of climate-related facts. For each fact:

- Entities (regions, climate indicators, records, time periods) are extracted and represented as nodes.
- Relationships between entities (e.g., a region experiencing a climate indicator, a record associated with a climate indicator) are captured as edges.

The graph structure allows for complex queries and reasoning tasks, such as:

- Identifying the occurrence of specific climate indicators in different regions over time.
- Verifying the consistency of climate facts by examining the relationships between indicators, records, regions, and time periods.

This knowledge graph provides a structured representation of climate data, enabling automated analysis and decision-making.

D Pipeline for Question Generation

Temporal Reasoning Temporal questions were constructed to assess a model's ability to reason about climate-related phenomena across specific time periods. This task type was generated by selecting climate indicators, such as "Temperature Anomaly" or "Rainfall Anomaly," and linking them with relevant records indicating the extent or impact of these anomalies. For example, from the knowledge graph, edges of type HAS_RECORD were followed to identify specific climate records, and then edges of type OCCURRED_IN were used to associate these records with their respective temporal periods. The resulting question might be: "In which year did the temperature anomaly occur in Scotland?" The ground truth was directly extracted from the time period node, and an accompanying explanation was included to clarify the climate indicator's historical context and temporal occurrence.

Spatial Reasoning Spatial questions targeted the ability of LLMs to understand geographic relationships in climate data. These questions were generated by navigating from region nodes to associated climate indicators through the EXPERIENCED

edges. For each region, such as "Wales" or "Scotland," we identified climate indicators it experienced and generated questions like, "Which climate anomaly was experienced by Wales in 2003?" The ground truth was derived from the linked climate indicator node, while the explanations contextualized the geographic specifics of the indicator's manifestation, highlighting the regional variance in climate impacts.

Spatio-Temporal Reasoning Spatio-temporal questions were the most complex, requiring models to reason about both spatial and temporal aspects simultaneously. To generate these tasks, we traversed the knowledge graph to identify relationships between regions, climate indicators, and time periods. This involved edges of types such as EXPERIENCED (connecting regions to climate indicators) and OCCURRED_IN (linking climate records to temporal nodes). A typical question might be, "Did England experience a rainfall anomaly in Spring 2019?" Ground truth for such questions was determined by the presence of relevant edges linking the entities. The explanations provided detailed reasoning about both the temporal context (e.g., Spring 2019) and the specific regional climate anomaly, aiming to enhance the interpretability of the answer.

E Experimental Setup

We accessed OpenAI Models (GPT-3.5-turbo, GPT-4-turbo, GPT-4o, o1) with configurations: max_tokens: 100-300 (For o1 model, max_reasoning_tokens =10,000).

We evaluate the following state-of-the-art LLMs:

GPT-4o and GPT-4: These models represent the latest advancements in OpenAI's GPT series, known for their exceptional reasoning capabilities and performance across NLP benchmarks.

GPT-3.5: Included as a comparative baseline to highlight advancements in reasoning and accuracy from prior iterations.

Llama-3-8B: This model is noted for its strong performance in fine-tuned, domain-specific tasks, making it a relevant choice for evaluating climate science data.

Phi-3: Selected for its lightweight architecture and efficiency, allowing an exploration of trade-offs between performance and resource utilization.

O1: As a representative of emerging lightweight models, O1 was included to assess performance scalability in resource-constrained environments.

DeepSeek-R1 An emerging competitive reasoning model.

We also include a table of the training cutoff date of each model (See Table 6).

The inclusion of these diverse baselines ensures a robust evaluation of how different architectures and training paradigms perform on the ClimateViz and ReasonClim datasets. Additionally, their performance across spatial, temporal, and spatio-temporal reasoning tasks provides valuable insights into the challenges posed by each reasoning dimension and helps identify avenues for model improvement. The experiments were conducted for both fact-checking and reasoning tasks, with all computations executed on an NVIDIA A100 Tensor Core GPU (40 GB memory) and approximately 24 GB of available RAM. Below, we describe the data splits, hyperparameters, and evaluation processes for each task.

E.1 Fact-Checking Task

For the fact-checking task, we sampled 1,000 claims from the ClimateViz dataset, maintaining a 7:3 ratio of true to false claims to simulate real-world scenarios. The dataset was fed to the model using structured prompts:

Is the following statement true or false? Reply with 1 for true and 0 for false. Statement: <fact>

Predicted labels were collected for each fact, and results were saved to a CSV file. To ensure robustness against API rate limits, a one-second delay was added between requests. Metrics such as abstention rate, balanced accuracy, recall, and F1-score were calculated to evaluate the model's performance comprehensively.

E.2 Reasoning Tasks

Spatial Questions: Questions required identifying climate anomalies (e.g., "Sunshine Anomaly"). Prompts emphasized direct responses:

Answer only with the climate anomalies experienced, choosing only from "Sunshine Anomaly," "Rainfall Anomaly," or "Temperature Anomaly." Separate multiple anomalies using commas. Do not abstain from answering. Question: <spatial question>

Temporal Questions: Questions focused on identifying specific years. Prompts ensured clarity

Model	Training Cutoff Date	Notes
o1	October 2023	RL-focused, no SFT phase
DeepSeek-R1	Not explicitly stated	Base model (V3) trained in late 2024
GPT-4o	May 2023	Trained from scratch
GPT-4	December 2023	Most recent OpenAI cutoff
GPT-3.5	September 2021	Discrepancies in model responses
Phi-3	October 2023	Microsoft's lightweight model
Llama-3	March/December 2023	Varies by model size

Table 6: Training Cutoff Dates for Evaluated Models

and precision: *Answer only with numbers representing years after 1990, like 1991, 2001, etc. Separate multiple years using commas. Do not abstain from answering. Question: <temporal question>*

Spatio-Temporal Questions: *Prompts varied depending on the format of the question: Numerical questions (e.g., "What's..."): Answer only with a number followed by %, like "138%". No additional text. Do not abstain from answering. Question: <spatio-temporal question>*

Binary questions (e.g., "Did..."): Answer only with "Yes" or "No". No additional text. Do not abstain from answering. Question: <spatio-temporal question>

E.3 Data Splits

Fact-Checking Task: The ClimateViz dataset was used for zero-shot evaluation, with models assessed on a curated set of 1,000 claims (7:3 ratio of true to false claims). No training/testing split was applied, as the task is designed purely for benchmarking.

Reasoning Tasks: The ReasonClim dataset, covering spatial, temporal, and spatio-temporal reasoning, was also used exclusively for evaluation. Since these tasks serve as benchmarks, no separate training or fine-tuning was performed.

E.4 Evaluation Criteria

E.4.1 Metrics for fact-checking task

Abstention Rate: Proportion of claims where the model abstained from providing an answer.

Balanced Accuracy: Average recall across true and false claims to mitigate class imbalance.

Recall and F1-Score: To measure the precision and robustness of predictions.

E.4.2 Metrics for reasoning tasks

Coverage: Percentage of non-abstained responses.

Mean Evaluation: Average score for non-abstained questions, calculated using task-specific criteria (e.g., weighted partial matches for spatial and temporal reasoning).

Evaluation Standard Deviation: To assess variability in performance across tasks.

E.5 Reproducibility

The provided scripts are publicly available and support replicating all fact-checking and reasoning experiments, ensuring that results are reproducible across different environments.

F Metrics Calculation for Each Language Model in Fact-checking Task

To evaluate the performance of the fact-checking model, we compute the following metrics: **Abstention Rate**, **Balanced Accuracy**, **Recall**, and **F1-Score**. These metrics provide a comprehensive understanding of the model's confidence, accuracy, and ability to handle imbalanced datasets. The calculations are detailed below.

F.1 Abstention Rate

The **Abstention Rate** quantifies the proportion of samples for which the model abstains from making a prediction. A prediction is considered abstained if the predicted label is neither 0 (false) nor 1 (true). It is defined as:

$$\text{Abstention Rate} = \frac{N_{\text{abstain}}}{N_{\text{total}}} \times 100 \qquad (1)$$

where N_{abstain} is the number of abstained samples, and N_{total} is the total number of samples. This metric is expressed as a percentage, with higher values indicating greater abstention.

F.2 Balanced Accuracy

The **Balanced Accuracy** accounts for class imbalance by averaging the recall values for both true (1) and false (0) claims:

$$\text{Balanced Accuracy} = \frac{\text{Recall}_{\text{True}} + \text{Recall}_{\text{False}}}{2} \tag{2}$$

with:

$$\text{Recall}_{\text{True}} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Recall}_{\text{False}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

Balanced Accuracy ensures both classes are equally considered, even when one class is underrepresented. It is expressed as a percentage, with 100% indicating perfect performance.

F.3 Recall

The **Recall** metric measures the model's ability to correctly identify true claims (1). It is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

where TP and FN are the true positives and false negatives, respectively. Recall is expressed as a percentage, with 100% indicating that all true claims are correctly identified.

F.4 F1-Score

The **F1-Score** balances Precision and Recall, providing a single measure of the model's classification performance. It is calculated as:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

Here, FP refers to false positives. The F1-Score is expressed as a percentage, with 100% indicating perfect precision and recall.

F.5 Table

Here we list all the metrics for the models we evaluated.

Metric	GPT-3.5 Value (%)
Abstention Rate	0.00
Balanced Accuracy	65.71
Recall	73.43
F1-Score	76.72

Table 7: Evaluation metrics for the GPT-3.5 model on the fact-checking task.

Metric	GPT-4 Value (%)
Abstention Rate	1.4
Balanced Accuracy	56.43
Recall	16.28
F1-Score	27.66

Table 8: Evaluation metrics for the GPT-4 model on the fact-checking task.

G Metrics Calculation for Each Language Model in Reasoning Tasks

To evaluate each language model's performance on reasoning tasks, we use the following metrics: **Abstention Rate**, **Coverage**, **Mean Evaluation**, and **Evaluation Standard Deviation**.

These metrics are detailed below.

G.1 Abstention Rate

The **Abstention Rate** measures the proportion of questions where the model abstains from answering. Let N represent the total number of questions, and N_{abstain} represent the number of abstained questions. The Abstention Rate (A_{rate}) is calculated as:

$$A_{\text{rate}} = \frac{N_{\text{abstain}}}{N} \times 100 \tag{7}$$

This metric is expressed as a percentage.

G.2 Coverage

The **Coverage** metric is complementary to the Abstention Rate and represents the proportion of questions where the model provides an answer. Let $N_{\text{non-abstain}}$ represent the number of non-abstained questions. Coverage (C) is calculated as:

$$C = \frac{N_{\text{non-abstain}}}{N} \times 100 \tag{8}$$

Alternatively, it can be derived directly from the Abstention Rate:

$$C = 100 - A_{\text{rate}} \tag{9}$$

G.3 Mean Evaluation

The **Mean Evaluation** represents the average score of the model on all non-abstained questions. Let E_i represent the evaluation score for the i-th non-abstained question, and $N_{\text{non-abstain}}$ represent the number of non-abstained questions. The Mean Evaluation (μ_E) is calculated as:

$$\mu_E = \frac{1}{N_{\text{non-abstain}}} \sum_{i=1}^{N_{\text{non-abstain}}} E_i \tag{10}$$

The evaluation score E_i depends on the specific reasoning task:

Metric	GPT-4o Value (%)
Abstention Rate	9.3
Balanced Accuracy	66.81
Recall	48.89
F1-Score	62.87

Table 9: Evaluation metrics for the GPT-4o model on the fact-checking task.

Metric	o1 Value (%)
Abstention Rate	0.9
Balanced Accuracy	64.69
Recall	43.52
F1-Score	58.19

Table 10: Evaluation metrics for the o1 model on the fact-checking task.

Metric	llama3 Value (%)
Abstention Rate	0.00
Balanced Accuracy	63.48
Recall	84.29
F1-Score	80.71

Table 11: Evaluation metrics for the llama3 model on the fact-checking task.

Metric	phi3 Value (%)
Abstention Rate	7.2
Balanced Accuracy	58.57
Recall	35.67
F1-Score	49.57

Table 12: Evaluation metrics for the phi3 model on the fact-checking task.

- **Temporal and Spatial Reasoning**: E_i corresponds to the Weighted Partial Match score.
- **Spatio-Temporal (Numerical Questions/What's Questions)**: E_i is 1 if the predicted value is within tolerance ($\pm 10\%$ of the ground truth), otherwise 0.
- **Spatio-Temporal (Binary Classification Questions/Did Questions)**: E_i is 1 for correct binary answers (Yes/No), and 0 otherwise.

G.4 Evaluation Standard Deviation

The **Evaluation Standard Deviation** (σ_E) quantifies the variability in the model's performance on non-abstained questions. It is defined as:

$$\sigma_E = \sqrt{\frac{1}{N_{\text{non-abstain}}} \sum_{i=1}^{N_{\text{non-abstain}}} (E_i - \mu_E)^2} \quad (11)$$

Here, E_i is the evaluation score for the i-th non-abstained question, and μ_E is the Mean Evaluation score.

G.5 Task-Specific Evaluation Scores

Temporal and Spatial Reasoning: Weighted Partial Match. The Weighted Partial Match score is calculated as:

$$E_i = \frac{|P \cap G|}{|P \cup G|} \quad (12)$$

where P is the set of predicted values, and G is the set of ground truth values.

Spatio-Temporal (Numerical Questions/What's Questions): Tolerance-Based Accuracy. A Tolerance-Based Accuracy score is assigned as:

$$E_i = \begin{cases} 1, & \text{if } G \cdot (1-\epsilon) \leq P \leq G \cdot (1+\epsilon) \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

Here, G is the ground truth value, P is the predicted value, and $\epsilon = 0.1$ (10% tolerance).

Spatio-Temporal (Binary Classification Questions/Did Questions): Direct Match Accuracy. A Direct Match Accuracy score is assigned as:

$$E_i = \begin{cases} 1, & \text{if } P = G \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

G.6 Implementation Notes

To compute the metrics, we first filter out abstained questions and operate only on non-abstained rows. The evaluation scores (E_i) are then used to calculate the Mean Evaluation and Standard Deviation. Abstention Rate and Coverage are computed across all questions. These metrics provide a comprehensive assessment of model performance, capturing accuracy, abstention behavior, and variability across different reasoning tasks.

G.7 Table

Here we list all the metrics for the models we evaluated.

H Error Analysis

H.1 Fact-checking Task

The error rate for each generation method in the dataset is computed as the proportion of incorrect predictions made on false claims. The calculation process involves the following steps:

Reasoning Type	Total Qns	Aggregated Qns
Spatial Reasoning	294	92
Temporal Reasoning	294	14
Spatio-Temp. Reasoning	528	528

Table 13: Distribution of questions across reasoning types, showing total and aggregated questions.

Metric	Spatial	Temporal	Spatio-Temporal
Abstention Rate (%)	2.17	14.29	20.08
Coverage (%)	97.83	85.71	79.92
Mean Evaluation (%)	41.48	0.56	29.15
Evaluation Std Dev (%)	31.92	1.92	45.50

Table 14: Metrics for the o1 model across Spatial, Temporal, and Spatio-Temporal reasoning types.

H.2 Filtering False Claims

The dataset is filtered to isolate entries with `ground_truth` $= 0$, representing factually incorrect claims. Let D denote the dataset, and F the subset of false claims:

$$F = \{x \in D \mid \texttt{ground_truth}(x) = 0\}.$$

This subset F is used for further error rate analysis.

H.3 Grouping by Generation Method

The filtered dataset F is grouped by the `generation_method` attribute, which specifies the model or algorithm responsible for generating the predictions. Let G represent the set of unique generation methods, and for each $g \in G$, let F_g denote the subset of F associated with generation method g.

H.4 Error Rate Computation

For each generation method g, the error rate is the mean proportion of incorrect predictions. The error rate for g is computed as:

$$\text{Error Rate}_g = \frac{1}{|F_g|} \sum_{x \in F_g} \mathbb{I}\big(\texttt{label}(x) \neq \texttt{truth}(x)\big), \tag{15}$$

where $|F_g|$ is the total number of predictions for g, and $\mathbb{I}(\cdot)$ is the indicator function (1 if true, 0 otherwise).

The error rate as a percentage is:

$$\text{Error Rate } (\%)_g = \text{Error Rate}_g \times 100.$$

Metric	Spatial	Temporal	Spatio-Temporal
Abstention Rate (%)	0.00	14.29	0.00
Coverage (%)	100.00	85.71	100.00
Mean Evaluation (%)	4.98	8.68	31.25
Evaluation Std Dev (%)	17.47	6.26	46.40

Table 15: Metrics for the DeepSeek-R1 model across Spatial, Temporal, and Spatio-Temporal reasoning types.

Metric	Spatial	Temporal	Spatio-Temporal
Abstention Rate (%)	3.26	0.00	4.92
Coverage (%)	96.74	100.00	95.08
Mean Evaluation (%)	37.83	2.42	37.85
Evaluation Std Dev (%)	35.69	3.41	48.55

Table 16: Metrics for the GPT-4o model across Spatial, Temporal, and Spatio-Temporal reasoning types.

Metric	Spatial	Temporal	Spatio-Temporal
Abstention Rate (%)	0.00	7.14	5.49
Coverage (%)	100.00	92.86	94.51
Mean Evaluation (%)	45.65	3.89	24.85
Evaluation Std Dev (%)	24.19	5.58	43.26

Table 17: Metrics for the GPT-4 model across Spatial, Temporal, and Spatio-Temporal reasoning types.

Metric	Spatial	Temporal	Spatio-Temporal
Abstention Rate (%)	0.00	0.00	0.00
Coverage (%)	100.00	100.00	100.00
Mean Evaluation (%)	20.29	5.72	37.12
Evaluation Std Dev (%)	24.53	5.29	48.36

Table 18: Metrics for the GPT-3.5 model across Spatial, Temporal, and Spatio-Temporal reasoning types.

Metric	Spatial	Temporal	Spatio-Temporal
Abstention Rate (%)	0.00	0.00	4.92
Coverage (%)	100.00	100.00	95.08
Mean Evaluation (%)	43.82	2.93	14.34
Evaluation Std Dev (%)	37.60	3.54	35.09

Table 19: Metrics for the Phi-3 model across Spatial, Temporal, and Spatio-Temporal reasoning types.

Metric	Spatial	Temporal	Spatio-Temporal
Abstention Rate (%)	0.00	0.00	0.00
Coverage (%)	100.00	100.00	100.00
Mean Evaluation (%)	31.16	11.92	30.11
Evaluation Std Dev (%)	31.57	10.44	45.92

Table 20: Metrics for the Meta-Llama-3 model across Spatial, Temporal, and Spatio-Temporal reasoning types.