CLLMate: A Multimodal Benchmark for Weather and Climate Events Forecasting

Haobo Li[1], **Zhaowei Wang**[1], **Jiachen Wang**[2], **YueYa Wang**[1]
Alexis Kai Hon Lau[1], **Huamin Qu**[1]
[1]Hong Kong University of Science and Technology, [2]Zhejiang University
hliem@connect.ust.hk

Abstract

Forecasting weather and climate events is crucial for making appropriate measures to mitigate environmental hazards and minimize losses. However, existing environmental forecasting research focuses narrowly on predicting numerical meteorological variables (e.g., temperature), neglecting the translation of these variables into actionable textual narratives of events and their consequences. To bridge this gap, we proposed **W**eather and **C**limate **E**vent **F**orecasting (WCEF), a new task that leverages numerical meteorological raster data and textual event data to predict weather and climate events. This task is challenging to accomplish due to difficulties in aligning multimodal data and the lack of supervised datasets. To address these challenges, we present CLLMate, the first multimodal dataset for WCEF, using 26,156 environmental news articles aligned with ERA5 reanalysis data. We systematically benchmark 32 existing models on CLLMate, including closed-source, open-source, and our fine-tuned models. Our experiments reveal the advantages and limitations of existing MLLMs and the value of CLLMate for the training and benchmarking of the WCEF task. The dataset is available at https://github.com/hobolee/CLLMate.

1 Introduction

Weather and climate events, namely discrete episodes of extreme weather or abnormal climate conditions (Hurrell, 2013), pose a significant risk to human society, resulting in potential harm to property, infrastructure, injuries, and even loss of life (Stephenson et al., 2008). Such events (e.g., heatwaves, floods, droughts) have experienced a noticeable surge in frequency, intensity, and duration in recent years due to climate change (Accarino et al., 2023). Significant concerns have been raised among human society. Appropriate measures and optimal strategies are urgently needed to forecast these events and mitigate their negative impacts. Considerable researchers have studied weather forecasting for years. For example, Kang et al. (2020) used factors such as temperature, wind, and pressure data to forecast precipitation. Similarly, Huang et al. (2021) relied on historical radiation data to forecast future radiation.

However, those works only focus on meteorological variables forecasting, missing forecasting textual weather and climate events or mapping the relationship between meteorology and its textual consequence. In the real world, there exists a wide array of weather and climate events, also including their cascading secondary and tertiary consequences. For example, heavy rainfall can lead to waterlogging, infrastructure degradation (e.g., road collapse), traffic disruptions, and human casualties when urban systems lack adaptive resilience to climate change. A critical challenge lies in the disparity between heterogeneous numerical meteorological data and textual descriptions of such events and their consequences. The gap hinders the accurate forecasting of real-world textual events and their downstream impacts based solely on meteorological variables.

To address this critical gap, we propose the **W**eather and **C**limate **E**vent **F**orecasting (WCEF), a novel task in environmental forecasting (Figure 1). Unlike traditional approaches that predict numerical meteorological variables (e.g., precipitation) (Kang et al., 2020; Yang et al., 2024b), WCEF pioneers the generation of natural language descriptions to forecast weather and climate events and their cascading consequences (Figure 1). This represents a fundamental paradigm shift, as it replaces the conventional workflow, where domain experts manually interpret numerical predictions to infer potential events, a process that is labor-intensive, subjective, and lacking geographical adaptability, with an automated, context-aware framework.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 17536–17562
November 4-9, 2025

The complexity of WCEF stems from its dual departure from existing methods: (1) its output requires generating textual event descriptions rather than numerical sequences, and (2) its input demands multimodal integration of numeric, text, and image. Prior efforts, which focus solely on single-modal numerical predictions, fail to address these challenges (Schultz et al., 2021). They neither automate the translation of raw climate data into actionable narratives nor adapt to location-specific vulnerabilities, leaving critical gaps in forecasting precision and practical utility.

The emergence of LLMs and multimodal LLMs (MLLMs) presents significant potential for benchmarking and addressing the WCEF task. To benchmark this task, we collect weather and climate events from environmental news articles, which serve as a valuable complement to meteorological data. These articles provide detailed descriptions of weather and climate events (Roberts, 2023), including temporal and spatial information that can support precise event prediction. However, the sheer volume of unstructured environmental news conceals task-critical information. Additionally, the scarcity of supervised datasets, exacerbated by the novelty of the task and the complexity of curating aligned multimodal pairs (numerical data and event descriptions), requires significant domain expertise. The recent advancements in LLMs (Brown et al., 2020) offer a promising solution. They enable the extraction of structured information from vast amounts of unstructured documents (Dagdelen et al., 2024; Biswas and Talukdar, 2024). Specifically, we utilize GPT-4o-mini (OpenAI, 2025) to process 26,156 news articles, generating structured event representations (e.g., event A, cause, event B, location, and date) from unstructured text. To ensure accuracy, domain experts manually validate the outputs, resulting in the construction of a high-fidelity, spatiotemporally aligned multimodal dataset, CLLMate.

To address the WCEF task, bridging the modality gap between meteorological raster data (high-dimensional spatiotemporal grids) and textual descriptions remains a significant challenge (Liu et al., 2023; Jiang et al., 2024). MLLMs, in particular, offer an opportunity to integrate and align textual narratives with spatiotemporal data for the WCEF task (Xu et al., 2023). To evaluate current MLLMs' capabilities, we conduct extensive experiments on our constructed dataset using 32 MLLMs. The results demonstrate that while these models show potential, they remain constrained in their performance on the WCEF task. The experiment results reveal that current models have the advantage but are limited in the WCEF task. It underscores the need for further work to achieve accurate weather and climate event forecasting.

In summary, the contributions can be outlined:

- We propose the WCEF task to forecast textual events based on numerical meteorological raster data, which differs from traditional meteorological data forecasting.
- We leverage domain knowledge extracted from the news corpus to establish the first multimodal dataset to connect events and meteorology.
- We conduct extensive experiments to evaluate existing MLLMs and our fine-tuned models on the WCEF task, benchmarking their capability in forecasting textual weather and climate events.

2 Related Work

Weather and climate forecasting has long been a research problem in the field of the environment. In this paper, we proposed a more challenging task to predict textual events using MLLMs.

2.1 Weather and Climate Events Forecasting

In the era preceding modern weather prediction, human experience connects diverse natural signs, such as cloud patterns and animal behavior, with weather and their subsequent effects (Risiro et al., 2012). The start of modern weather forecasting was marked with the first modern weather chart (Allaby, 2009; Young and Grahame, 2022). In contemporary studies, a shift towards efficiency has occurred by consolidating various numerical variables into a unified numerical framework. Two primary numerical methodologies are commonly employed: numerical weather prediction, which utilizes numerical simulation methods (Bauer et al., 2015; Lynch, 2008), and AI-based forecasting, which leverages data-driven approaches (Bi et al., 2023; Hewage et al., 2021). Soichiro et al. explored models designed to generate textual comments for four specific weather conditions: sunny, rainy, cloudy, and snowy (Murakami et al., 2021). However, their work is limited by the restricted range of weather conditions and the subjective nature of the comments. In contrast, our study represents the first attempt to address the WCEF task. We construct a valuable dataset containing the numerical meteorology and its corresponding events, aimed at the

Dataset	Meteorological Variables	Textual Events	
		Not Records	Records
SEVIR (Veillette et al., 2020)	Satellite	×	×
Mesogeos (Kondylatos et al., 2024)	ERA5	×	×
Digital Typhoon (Kitamoto et al., 2024)	Satellite	×	×
Vaid et al.'s work (Vaid et al., 2022)	×	×	Social Media
NeuralNERE (Mishra and Mittal, 2021)	×	×	News
ClimateIQA (Chen et al., 2024a)	ERA5	QA generated by LLM	×
WeatherQA (Ma et al., 2024)	ERA5	Expert's pre-analysis	×
CLLMate	ERA5	×	**News, verified by experts**

Table 1: Comparison between CLLMate and existing datasets. CLLMate uniquely integrates ERA5 reanalysis data with expert-verified textual events, addressing the scarcity of aligned multimodal datasets for weather and climate event forecasting. (ERA5: the fifth generation ECMWF atmospheric reanalysis dataset of the global climate.)

relationship between numerical meteorology and its consequential events.

2.2 Multimodal LLM

With the advancement of LLMs, there's been a surge of research on building multimodal LLMs. Their studies (Wu et al., 2023; Zhan et al., 2024) try to incorporate multiple types of data beyond just natural language, such as images, audio, and video. BLIP-2 (Li et al., 2023) has developed a large-scale image captioning dataset, combining a language model with a vision encoder to create a multimodal model. Building on this, LLaVA (Liu et al., 2023) introduces a more cost-effective method for training multimodal models through visual instruction tuning. The following multimodal LLMs, including QwenVL2 (Wang et al., 2024a), CogVLM (Wang et al., 2023a), deepseek-vl (Lu et al., 2024), Intern-vl (Chen et al., 2024b; Dong et al., 2024), etc., all follow a similar architecture of LLaVA. Thanks to the strong ability of multimodal LLMs, an increasing number of works start to seek the assistance of multimodal LLMs (Wang et al., 2025; Sun et al., 2024; Yang et al., 2024a; Chen et al., 2025b). Following them, our work focuses on meteorology.

2.3 Existing Similar Datasets

Prior efforts to model meteorological events fall into three categories (Table 1):

Single-Modality Physical Datasets. Works like SEVIR (storm imagery) (Veillette et al., 2020), Mesogeos (wildfire-related variables) (Kondylatos et al., 2024), and Digital Typhoon (typhoon trajectories) (Kitamoto et al., 2024) focus on meteorological measurements but lack textual event narratives. While valuable for physical modeling, they omit explicit linkages to real-world impacts.

Text-centric datasets. Text-centric datasets such as Vaid et al.'s work (social media events) (Vaid et al., 2022) and NeuralNERE (news-derived knowledge graphs) (Mishra and Mittal, 2021) catalog climate-related events but fail to link them to corresponding meteorological data. This disconnect leaves the causal chain between weather patterns (e.g., extreme rainfall) and societal consequences (e.g., floods) unquantified.

Multimodal Approaches with Limitations. Several other meteorological multimodal dataset construction efforts were conducted around the same time as CLLMate. However, they remain limited by data quality and scope. ClimateIQA (Chen et al., 2024a) pairs ERA5 data with LLM-generated QA pairs but lacks verified event records and impact descriptions. WeatherQA (Ma et al., 2024) aligns ERA5 with expert pre-analysis texts, yet it focuses on meteorological forecasts rather than post-event records and excludes societal impacts. Both datasets prioritize meteorological phenomena over downstream consequences and rely on synthetic or non-verified textual data, limiting their utility for impact-driven research.

CLLMate bridges these gaps by integrating expert-verified post-event records from authoritative news sources with spatiotemporally matched meteorological raster data. Unlike prior works, CLLMate (1) captures both meteorological phenomena (e.g., rainfall) and their downstream consequences (e.g., flooding, infrastructure damage), (2) establishes explicit cross-modal mappings between textual event descriptions (including time and location) and meteorological variables, and (3) ensures reliability through domain expert validation of event reports. CLLMate provides the first high-quality, multimodal corpus for modeling the interplay between weather conditions and real-world impacts, enabling robust analysis of climate-society interactions.

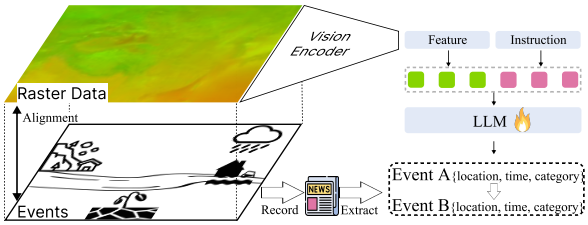


Figure 1: The CLLMate framework is designed to extract textual weather and climate events and align them with meteorological raster data for the WCEF task. The output of the LLLM is the events extracted from news articles, and the input is the corresponding meteorological raster data.

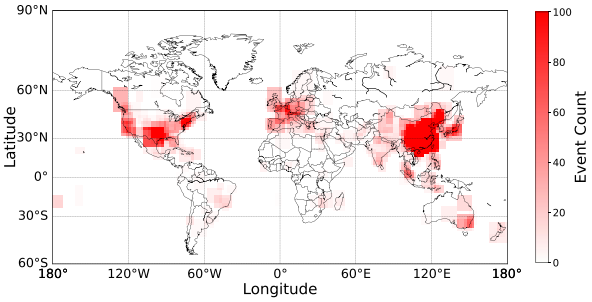


Figure 2: Spatial distribution of extracted events. Each rectangle represents an extracted event. The events span most global regions.

3 Problem Formulation

In this section, we provide an introduction to the data type in our dataset, CLLMate, and outline the formulation of the WCEF task (Figure 1).

3.1 Spatio-Temporal Data

Two types of spatio-temporal data are utilized: meteorological raster data and event data.

3.1.1 Meteorological Raster Data

Meteorological raster data is commonly employed in the environmental domain to depict meteorological variables. For instance, one global meteorological variable can be represented as a three-dimensional tensor $\boldsymbol{R} \in \mathbb{R}^{T\times\Lambda\times\Phi}$, where T denotes the time dimension, and Λ and Φ represent the longitude and latitude dimensions, respectively. Each element $\boldsymbol{r}_{t,\lambda,\phi}$ corresponds to the variable value at a specific time t and location (λ, ϕ).

3.1.2 Event Data

On the other hand, event data $\boldsymbol{E}$ are occurrences of interest at specific points in space and time with the shape of $T \times \Lambda \times \Phi$. For example, a single event can be represented using textual information $\boldsymbol{e}_{t,\lambda,\phi}$, indicating the occurrence at a particular time t and location (λ, ϕ).

Two distinctions exist between raster and event data. First, raster data is numerical, while event data is textual. Next, raster data is dense, whereas event data is sparse, featuring specific instances at discrete locations and times.

3.2 WCEF Task

The existing weather forecasting task involves predicting the future target meteorology variable using either the same or more variables. This task can be formulated as follows:

$$\boldsymbol{R}_{C+1:C+F} = f(\boldsymbol{R}_{C-H+1:C}) \quad (1)$$

where $f(\cdot)$ is the forecasting model, $\boldsymbol{R}_C$ denotes the numerical meteorological variables at the current time, F represents the length of the forecasting data, and H is the length of historical data used.

In contrast, the objective of the proposed WCEF task is to predict textual events using numerical data (Figure 1). These events can be defined based on meteorological characteristics, such as heatwaves and the subsequent consequences they entail. The scope of events encompasses primary occurrences like heatwaves and droughts, as well as secondary and tertiary events (consequences) like landslides and human casualties that arise as a result of the primary events. These examples illustrate the complexity of deriving such events solely from numerical data. It highlights the significance of the historical events and the valuable knowledge they offer. The WCEF task can be formulated:

$$\boldsymbol{E}_{C:C+F-1} = f_\phi(\boldsymbol{R}_{C-H+1:C}, \text{Instruction}) \quad (2)$$

where $f_\phi(\cdot)$ denotes the LLM model, $\boldsymbol{E}_C$ is the textual events happening at the current time, and the instruction is the text prompt. It is worth emphasizing that, unlike existing forecasting tasks utilizing the same source data, we employ C in both the forecasting and historical data. This is because $\boldsymbol{E}_C$ and $\boldsymbol{R}_C$ hold distinct meanings.

4 Dataset Construction Process

In this section, we introduce the meteorological raster dataset and environmental news dataset we used to construct the multimodal instruction dataset for the WCEF task. Then, we explain the pipeline to create the multimodal dataset (appendix A.1).

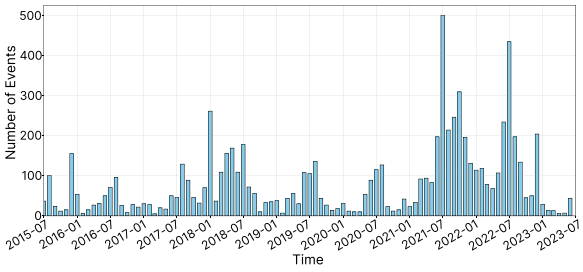


Figure 3: Temporal distribution of extracted events. The events span a long time period, from 2015 to 2023. A notable outlier in the number of events occurred due to the catastrophic flooding in Zhengzhou in July 2021.

4.1 Datasets Used to Construct CLLMate

Meteorological Raster Dataset. We use the open and free ERA5 reanalysis dataset (Hersbach et al., 2020) to obtain the meteorology data (Copernicus License). Its usage in climate research has been widely acknowledged for its quality (Sun et al., 2022). Four variables, namely "2m temperature," "10m u-component of wind," "10m v-component of wind," and "total precipitation," were selected to forecast the following textual weather and climate events. The dataset consists of hourly data spanning from July 2015 to July 2023, with a spatial resolution of $27.75\ km \times 27.75\ km$.

Environmental News Dataset. We acquired environmental news from Wisers (Wisers, 2024) through a procurement process, which consisted of highly environment-related news articles from news publishers. The dataset spans from July 2015 to June 2023. Each news article contains the title, content, character statistics, date, publisher, and media type. The media types encompass both web and publication resources while excluding internet-based media sources primarily reliant on aggregating news reports from official news agencies.

4.2 The Multimodal Dataset

We construct the first multimodal dataset for the WCEF task through three steps.

4.2.1 Event Extraction

Weather and climate events exist within a vast amount of news articles. Research has demonstrated the efficacy of utilizing LLMs for extracting structured information from text (Dagdelen et al., 2024). The initial step of our pipeline is to identify and extract these events and their relationships from each news document. First, we employed GPT-4o-mini to analyze 26,156 environmental news articles, extracting weather and climate event knowledge in structured triples with spatiotemporal metadata (e.g., high temperature causes heatwave, Hong Kong, 2022-08-16). The details of the prompt for extracting events can be found in appendix B.1.

Initial automated labeling identified 6,352 news containing potential event relationships. These articles, along with their extracted triples, locations, and dates, underwent rigorous manual validation by three meteorological domain experts. Events, which are not meteorological events, or lack clear spatial information, or temporal details, were excluded. Following verification, 2,575 news were confirmed to contain accurate event knowledge, yielding a final curated dataset of 7,747 spatiotemporally grounded events. The inter-annotator agreement (IAA) score, measured by the pairwise agreement proportion, was 81.83%, and Fleiss's κ (Fleiss, 1971) was calculated to be 0.63. Please refer to appendix B.2 for more details on the annotation and the agreement metric.

Specifically, as for the spatial aspect, there are 1,049 distinct regions in the dataset. To derive geospatial bounds for these locations, we continue to employ GPT-4o-mini to generate latitude-longitude coordinates, which are subsequently validated by humans to ensure alignment with real-world geographical boundaries by visualizing them on the map. Additionally, we expand the geographical range, with a minimum of 5 degrees, to incorporate surrounding meteorological conditions.

Spatial Distribution. The dataset exhibits a broad geographical representation, covering most global regions (Figure 2). Events are notably concentrated in East Asia, North America, and Europe. The spatial distribution is decided by the database's distribution since we didn't apply filters, avoiding introducing biases.

Temporal Distribution. The temporal coverage spans July 2015 to July 2023, with marked seasonal patterns: event frequency peaks during summer and winter months (Figure 3). The data reveals a temporal trend, with event frequency increasing in more recent years. A significant outlier occurred in July 2021, coinciding with catastrophic flooding since extreme rainfall in Zhengzhou, China, which caused hundreds of fatalities and widespread infrastructure damage. This event garnered exceptional media attention.

4.2.2 Event Categorization

To simplify the evaluation of MLLMs in the WCEF

task, shifting from directly generating meteorological events (experiments can refer to subsection D.1), we collaborated with domain experts to develop the predefined hierarchical categorization of weather and climate events (appendix A.2). This approach aligns with practices in both the NLP domain (Hendrycks et al., 2020; Yue et al., 2024) and meteorology (Ma et al., 2024). By providing this hierarchical categorization for MLLMs to reference, we enhance their ability to forecast extreme weather events more reliably and with greater practical relevance. The categorization combines a bottom-up approach to clustering (using semantic similarity) with a top-down approach (using the taxonomy in the environmental domain) (Gaiteri et al., 2015). Events are classified into two primary categories: meteorological phenomena (3979/7747 events, Figure 4 in Appendix) and their cascading consequences (3768/7747 events, Figure 5 in Appendix). Each category is further divided into fine-grained subclasses, which serve as candidate options for model selection. In the benchmarking, MLLMs are required to select the most probable predefined fine-grained subclass. For more information about the category distribution, please refer to appendix A.3.

4.2.3 Event-Meteorology Alignment

Multimodal dataset construction relies on precise spatiotemporal alignment between events and their meteorological conditions (Figure 1). For each event (appendix A.1), we integrate: (1) textual event, (2) meteorological raster data for the event's specific date and region, and (3) regional statistical context (e.g., max) from the same spatiotemporal origin. Meteorological inputs are preprocessed into RGB-like grids (normalized to [0, 1]):

- Channel 1: 2m temperature.
- Channel 2: 10m wind speed (u- and v-components combined via magnitude norm).
- Channel 3: total precipitation.

This representation enables MLLMs to process spatial climate patterns as visual inputs.

5 Benchmarking Results

5.1 Benchmarked Models

The dataset is partitioned chronologically into training (90%) and testing (10%) sets. We conducted extensive experiments with 32 multimodal models and traditional AI models (after removing multimodal inputs) on the test set, encompassing closed-source, open-source, and our fine-tuned models:

5.1.1 Closed-Source Models

- **GPT Series** (OpenAI, 2025): OpenAI's flagship MLLM with integrated vision capabilities. We evaluated three variants of OpenAI's GPT architecture optimized for multimodal reasoning: GPT-4o, GPT-4o-mini, and GPT-o1.
- **Gemini** (Team et al., 2023): Google's MLLM with enhanced context window and optimized for high-speed processing of text-image inputs. We evaluated Gemini-1.5-pro and Gemini-2.0-flash.
- **Claude-3.5/3.7-Sonnet** (Anthropic, 2025): Anthropic's MLLM with improved analytical capabilities, supporting text and image inputs.

5.1.2 Open-Source Models

- **Deepseek-Janus-Pro-7B** (Chen et al., 2025a): Deepseek's MLLM with both multimodal understanding and text-to-image capabilities.
- **InternVL-8B/38B/78B** (Chen et al., 2024b): A vision-language foundation model with capabilities in handling complex multimodal data.
- **Cambrian-1-8B** (Tong et al., 2024): A fully open multimodal LLM designed with a vision-centric approach.
- **LLaVA-1.5/1.6** (Liu et al., 2023, 2024): Versions of the widely adopted open-source MLLM framework: LLaVA-1.5-7B/13B: MLLMs with standard vision-language alignment. LLaVA-1.6-7B/13B/34B: Enhanced variants MLLMs for fine-grained visual grounding.
- **Qwen2/2.5** (Wang et al., 2024b) utilize dynamic resolution and frame rate training for video understanding. We evaluate Qwen2-VL-7B and Qwen2.5-VL-7/72B.

5.1.3 Fine-tuned Models

To evaluate the performance of conventional visual models, we fine-tuned traditional AI models that rely solely on meteorological images. These models include ResNet50 (He et al., 2016) (IMAGENET1K_V2) and ViT (Dosovitskiy et al., 2020) (vit-base-patch16-224-in21k).

Additionally, we evaluated the performance of CLIP (Radford et al., 2021) as a classifier. For this, we computed image embeddings using meteorological images and text embeddings based on all event descriptions and their contextual information. The event with the highest correspondence to the image was selected as the prediction.

Moreover, we fine-tuned a series of LLaVA models using the training set of the CLLMate to eval-

	Models	Primary Category of Phenomena	Subordinate Category of Phenomena	Primary Category of Consequence	Subordinate Category of Consequence
	Expert (100 cases)	78.00%	54.00%	62.00%	46.00%
	Random guess	33.33%	12.50%	20.00%	5.56%
	Majority guess	40.12%	26.07%	40.11%	4.90%
closed-source	GPT-4o-mini (OpenAI, 2025)	30.28%	13.62%	34.63%	14.13%
closed-source	GPT-4o (OpenAI, 2025)	32.11%	15.65%	39.22%	12.01%
closed-source	GPT-o1 (OpenAI, 2025)	32.93%	10.97%	17.67%	10.60%
closed-source	GPT-o3 (OpenAI, 2025)	30.89%	10.57%	17.67%	2.83%
closed-source	GPT-o4-mini (OpenAI, 2025)	30.28%	10.16%	20.49%	2.47%
closed-source	Gemini-1.5-pro (Team et al., 2023)	20.12%	8.74%	28.98%	10.25%
closed-source	Gemini-2.0-flash (Team et al., 2023)	37.80%	19.72%	30.39%	13.43%
closed-source	Claude-3.5-Sonnet (Anthropic, 2025)	37.60%	19.72%	40.28%	15.90%
closed-source	Claude-3.7-Sonnet (Anthropic, 2025)	40.04%	20.93%	42.05%	15.55%
open-source	Janus-Pro-7B (Chen et al., 2025a)	31.31%	2.64%	21.20%	2.83%
open-source	InternVL2-8B (Chen et al., 2024b)	29.27%	10.37%	21.91%	5.30%
open-source	InternVL3-8B (Chen et al., 2024b)	29.67%	13.62%	20.49%	6.00%
open-source	InternVL3-38B (Chen et al., 2024b)	39.63%	19.51%	33.22%	12.70%
open-source	InternVL3-78B (Chen et al., 2024b)	36.79%	18.09%	42.05%	**17.31%**
open-source	Cambrian-1-8B (Tong et al., 2024)	31.10%	9.56%	31.80%	4.59%
open-source	LLaVA-1.5-7B (Liu et al., 2023)	31.30%	14.02%	20.85%	3.89%
open-source	LLaVA-1.6-vicuna-7B (Liu et al., 2024)	31.30%	13.82%	20.85%	3.89%
open-source	LLaVA-1.6-mistral-7B (Liu et al., 2024)	32.93%	17.28%	39.22%	12.01%
open-source	LLaVA-1.5-13B (Liu et al., 2023)	31.32%	15.04%	28.27%	7.77%
open-source	LLaVA-1.6-vicuna-13B (Liu et al., 2024)	29.47%	14.63%	21.91%	7.42%
open-source	LLaVA-1.6-vicuna-34B (Liu et al., 2024)	42.28%	15.45%	17.67%	3.89%
open-source	Qwen2-VL-7B (Wang et al., 2024b)	32.72%	11.38%	25.80%	5.30%
open-source	Qwen2.5-VL-7B (QwenTeam, 2025)	42.27%	17.48%	21.91%	7.78%
open-source	Qwen2.5-VL-72B (QwenTeam, 2025)	40.85%	20.73%	39.58%	**16.25%**
fine-tuned	Fine-tuned-ResNet50 (He et al., 2016)	38.01%	24.34%	19.08%	6.36%
fine-tuned	Fine-tuned-ViT (Dosovitskiy et al., 2020)	39.43%	23.46%	20.14%	6.48%
fine-tuned	Fine-tuned-Clip (Radford et al., 2021)	42.89%	19.11%	36.40%	6.00%
fine-tuned	Fine-tuned-LLaVA-1.5-7B	**45.93%**	23.37%	43.46%	9.89%
fine-tuned	Fine-tuned-LLaVA-1.6-vicuna-7B	44.91%	**29.67%**	**44.17%**	8.48%
fine-tuned	Fine-tuned-LLaVA-1.6-mistral-7B	45.53%	25.61%	**43.82%**	8.13%
fine-tuned	Fine-tuned-LLaVA-1.5-13B	44.92%	23.78%	43.46%	6.36%
fine-tuned	Fine-tuned-LLaVA-1.6-vicuna-13B	**45.54%**	**27.03%**	42.40%	9.19%

Table 2: The accuracy of benchmarked MLLMs across two granularity levels: primary category classification and subordinate category identification for both meteorological phenomena and their cascading consequence.

uate the value of the dataset and the efficiency of MLLMs in solving the WCEF task. The fine-tuned models include LLaVA-1.5-7B/13B and LLaVA-1.6-vicuna-7B/13B. They are trained on 8 A800 GPUs for one epoch. Please refer to appendix C for more details.

5.2 Prompt Design

We designed specialized prompts for MLLMs to forecast meteorological phenomena and their consequences. The prompts consist of "System Prompt," "Location Information," "Meteorological Parameters," "Meteorological Image," "Options," and "Zero-Shot CoT (Kojima et al., 2022)." As for the Zero-Shot CoT, we used (1) analysis of statistical data, (2) analysis of spatial patterns in the image, and (3) synthesis of findings to select the most probable fine-grained category. Refer to appendix B.3 and appendix B.4 for more details.

5.3 Evaluation Metric

For the benchmarking of multi-choice questions, we employ accuracy as the metric, following the previous research, such as Ge et al.'s work (Ge et al., 2022). Accuracy is a straightforward metric that measures the proportion of correct answers out of the total number of questions.

5.4 Heuristic Baselines

We implement two heuristic baselines:

- Random Guess: Answers are selected uniformly at random from all valid options, reflecting chance-level performance.
- Majority Guess: The most frequent category in the training set is chosen for all test instances.

5.5 Expert Experiment

We conducted the expert experiment to evaluate the task difficulty. We employed two postdocs in the environmental domain to perform the task. They

have access to the meteorology data and the contextual information. However, they were not allowed to conduct any online search. Based on the given information and their domain knowledge, they selected the most likely event category. A total of 50 meteorological phenomena and 50 consequences were randomly selected from the test dataset.

5.6 Main Results

Table 2 summarizes the performance of benchmarked MLLMs across two granularity levels: primary category (coarse-grained) and subordinate category (fine-grained) for meteorological phenomena and their consequences. The results are also shown in Figure 8 and Figure 9.

5.6.1 Meteorological Phenomena Forecasting

Primary Category: Most closed-source and open-source models perform near random guess (30.28–37.80%), with the exception of LLaVA-1.6-vicuna-34B and Qwen2.5 (40.85%-42.28%), which exceed the majority guess baseline (40.12%). Fine-tuned models achieve substantial gains, with Fine-tuned-LLaVA-1.5-7B attaining the highest accuracy (45.93%).

Subordinate Category: While most existing models surpass random guessing, none exceed the majority guess baseline (26.07%). Fine-tuning yields significant improvements (e.g., Fine-tuned-LLaVA-1.6-vicuna-7B: 29.67% vs. 13.82% for its untrained counterpart).

5.6.2 Consequence Forecasting

Primary Category: Claude-3.5-Sonnet leads among untrained models (40.28%), while fine-tuned variants dominate, notably Fine-tuned-LLaVA-1.6-vicuna-7B (44.17%), surpassing the random guess (20.00%) and majority guess baseline (40.11%). This demonstrates the learnability of meteorology-to-impact mappings when models are task-adapted.

Subordinate Category: QWen2.5-VL-72B achieves the highest accuracy (16.25%), marginally exceeding the heuristic baselines (5.56%/4.90%), while fine-tuned models show limited gains.

6 Performance Analysis

We show some findings of the benchmarking. For more analysis, such as case study, ablation study, and scale analysis, please refer to appendix D.

6.1 Expert Performance

From the experiment results, we found that experts perform better than existing LLMs, such as (78.0% vs 45.93% in phenomena forecasting). The relatively low performance of the LLMs is likely due to their limitations. There are a few possible reasons behind this: 1) current MLLMs lack domain-specific knowledge and are not fine-tuned for this task; 2) additionally, their alignment mechanisms may not be well-suited for handling meteorological data. These findings highlight the value of our dataset, which provides events with detailed temporal and spatial information. This extensibility allows for integrating additional meteorological data, exploring alternative alignment methods, and benchmarking against more advanced SOTA MLLMs in the future.

6.2 MLLMs vs. Heuristic Baselines

While MLLMs outperform the random guessing baseline, some of them often fall short of the majority guess baseline, a pattern attributable to an inherent class imbalance in the dataset. The imbalance provides a competitive advantage for naive reliance on prior knowledge of the category distribution. However, fine-tuned models such as Fine-tuned-LLaVA-1.5-7B surpass the majority guess baseline in phenomena prediction (45.93% vs.40.12%) and in consequence forecasting (43.46% vs. 40.11%). Critically, this demonstrates that MLLMs can learn meteorology-to-event mappings, achieving gains through pattern recognition.

6.3 MLLMs vs. Simpler Models

Even after training, traditional models like ResNet perform significantly worse than many zero-shot MLLMs, not to mention fine-tuned MLLMs. This highlights the necessity of MLLMs for effectively bridging the gap between numerical meteorological data and their corresponding events.

6.4 Closed-Source and Open-Source Models

Among closed-source models, Claude-3.5-Sonnet achieves the highest forecasting accuracy. Surprisingly, GPT-o1's performance is lower than that of other models. Open-source models exhibit significant performance fragmentation. While LLaVA-1.6-34B achieves competitive primary phenomena accuracy (42.28%), its consequence forecasting decreases to 17.67%, worse than random guessing (20%). Qwen2.5-VL-72B shows an advantage in consequence forecasting (39.58%/16.25%). The

performance of the best closed-source model is comparable to that of the best open-source model.

6.5 Effectiveness of Task-Specific Fine-tuning

Fine-tuning on CLLMate yields dramatic improvements: Fine-tuned-LLaVA-1.5-7B surpasses Claude-3.5-Sonnet in primary phenomena accuracy (45.93% vs. 37.60%) and consequences (43.46% vs. 40.28%). Subordinate category performance improves by 9–15 percentage points for phenomena (e.g., 23.37% vs. 14.02% in LLaVA-1.5-7B), though subordinate consequence forecasting remains challenging (<10% accuracy). This demonstrates that while MLLMs lack inherent weather reasoning capabilities, task-specific adaptation enables competitive forecasting, demonstrating the dataset's value for climate-aware MLLM adaptation. This underscores the necessity of task-specific alignment for bridging numerical meteorology and textual narratives.

6.6 Limited Performance

While existing or fine-tuned MLLMs outperform the heuristic baselines, they have relatively low accuracy compared to the expert experiment, suggesting current models have the advantage, however, limited, in distinguishing different coarse-grained event types (e.g., Precipitation vs. Wind) and fine-grained event types (e.g., General wind vs. Typhoon). It highlights the need for future work to forecast meteorological events more accurately.

7 Conclusion

We propose the WCEF task, which forecasts weather and climate events using meteorological raster data. To support this task, we present CLLMate, the first supervised multimodal dataset that aligns spatiotemporally grounded event descriptions and their consequences with corresponding meteorological numerical data. We conducted extensive experiments for benchmarking the WCEF task. The results show that while the MLLMs' performance can outperform the heuristic baselines, demonstrating the dataset's value for climate-aware MLLM adaptation.

Limitation

We identify some limitations in this study and future research directions in this emerging field:

First, the inclusion of additional modalities or other domains' data, such as altitude and topography, to offer supplementary insights for forecasting is essential. While this study initially delves into modalities encompassing numerical spatiotemporal data and text, there exist other meteorological data modalities like satellite imagery and time series data (Jin et al., 2023) from global meteorological stations that could contribute to longer-term, more precise forecasting. These data can be easily integrated into the events in CLLMate according to the spatiotemporal information. Further research is warranted to enrich the forecasting capabilities.

Second, since we extracted a large number of triples (event A causes event B), the incorporation of the knowledge graph necessitates additional exploration. Employing more sophisticated techniques to equip the MLLM with prior world knowledge (Do et al., 2024; Wang et al., 2024c) of weather and climate for the events understanding (Wang et al., 2022, 2023b), could potentially enhance the performance.

Last, the events are concentrated in East Asia, North America, and Europe (Figure 2). It reflects socioeconomic disparities in environmental reporting infrastructure and media coverage biases, which is an interesting topic for future research.

Ethics Statement

For the creation of the dataset, we utilized open-source meteorological data and publicly available news reports. All prompts used in this work are climate-related and do not include content that raises ethical concerns. Furthermore, experts were engaged to analyze the meteorological data and reports, and they were fairly compensated at a rate of 20 USD per hour for their contributions. No ethical concerns were identified regarding the treatment of the experts involved in this work.

Acknowledgement

This work was partially supported by the Research Grants Council TRS grant T22-607/24N and the Research Grants Council of Hong Kong (Project Nos. GRF 16302623).

References

Gabriele Accarino, Donatello Elia, Davide Donno, Francesco Immorlano, and Giovanni Aloisio. 2023. A machine learning-powered digital twin for extreme weather events analysis. Technical report, Copernicus Meetings.

Michael Allaby. 2009. *Atmosphere: a scientific history of air, weather, and climate*. Infobase Publishing.

Anthropic. 2025. Anthropic. claude 3.5 sonnet model card addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf. [Online; Accessed 31-January-2025].

Peter Bauer, Alan Thorpe, and Gilbert Brunet. 2015. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538.

Anjanava Biswas and Wrick Talukdar. 2024. Robustness of structured data extraction from in-plane rotated documents using multi-modal large language models (llm). *Journal of Artificial Intelligence Research*.

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

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Zixuan Yuan, Bing Zhu, and Junwei Liang. 2024a. Vision-language models meet meteorology: Developing models for extreme weather events detection with heatmaps. *arXiv preprint arXiv:2406.09838*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025a. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zixin Chen, Sicheng Song, Kashun Shum, Yanna Lin, Rui Sheng, and Huamin Qu. 2025b. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. *arXiv preprint arXiv:2503.18172*.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Quyet V Do, Tianqing Fang, Shizhe Diao, Zhaowei Wang, and Yangqiu Song. 2024. Constraintchecker: A plugin for large language models to reason on commonsense knowledge bases. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 714–731.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Chris Gaiteri, Mingming Chen, Boleslaw Szymanski, Konstantin Kuzmin, Jierui Xie, Changkyu Lee, Timothy Blanche, Elias Chaibub Neto, Su-Chun Huang, Thomas Grabowski, et al. 2015. Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering. *Scientific reports*, 5(1):16361.

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. 2022. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hans Hersbach et al. 2020. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Pradeep Hewage, Marcello Trovati, Ella Pereira, and Ardhendu Behera. 2021. Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, 24(1):343–366.

Liexing Huang, Junfeng Kang, Mengxue Wan, Lei Fang, Chunyan Zhang, and Zhaoliang Zeng. 2021. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Frontiers in Earth Science*, 9:596860.

James W Hurrell. 2013. *Climate Science for Serving Society: research, modeling and prediction priorities*. Springer.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Jinle Kang, Huimin Wang, Feifei Yuan, Zhiqiang Wang, Jing Huang, and Tian Qiu. 2020. Prediction of precipitation based on recurrent neural networks in jingdezhen, jiangxi province, china. *Atmosphere*, 11(3):246.

Asanobu Kitamoto, Jared Hwang, Bastien Vuillod, Lucas Gautier, Yingtao Tian, and Tarin Clanuwat. 2024. Digital typhoon: Long-term satellite image dataset for the spatio-temporal modeling of tropical cyclones. *Advances in Neural Information Processing Systems*, 36.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Spyridon Kondylatos, Ioannis Prapas, Gustau Camps-Valls, and Ioannis Papoutsis. 2024. Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the mediterranean. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Peter Lynch. 2008. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7):3431–3444. Predicting weather, climate and extreme events.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. 2024. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*.

Prakamya Mishra and Rohan Mittal. 2021. Neuralnere: Neural named entity relationship extraction for end-to-end climate change knowledge graph construction. In *Tackling climate change with machine learning workshop at ICML*.

Soichiro Murakami, Sora Tanaka, Masatsugu Hangyo, Hidetaka Kamigaito, Kotaro Funakoshi, Hiroya Takamura, and Manabu Okumura. 2021. Generating weather comments from meteorological simulations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1462–1473, Online. Association for Computational Linguistics.

OpenAI. 2025. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. [Online; Accessed 31-January-2025].

QwenTeam. 2025. Qwen2.5-vl.

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

Joshua Risiro, Dominic Mashoko, T Tshuma, Doreen, and Elias Rurinda. 2012. Weather forecasting and indigenous knowledge systems in chimanimani district of manicaland, zimbabwe. *Journal of Emerging Trends in Educational Research and Policy Studies*, 3(4):561–566.

Michelle Roberts. 2023. What are the heat exhaustion and heatstroke symptoms? https://www.bbc.com/news/health-62120167. [Online; Accessed 08-January-2024].

Martin G Schultz, Clara Betancourt, Bing Gong, Felix Kleinert, Michael Langguth, Lukas Hubert Leufen, Amirpasha Mozaffari, and Scarlet Stadtler. 2021. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A*, 379(2194):20200097.

David B Stephenson, HF Diaz, and RJ Murnane. 2008. Definition, diagnosis, and origin of extreme weather and climate events. *Climate extremes and society*, 340:11–23.

Shuai Sun et al. 2022. Increased moist heat stress risk across China under warming climate. *Scientific Reports*, 12(1):22548.

Yushi Sun, Hao Xin, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2024. Are large language models a good replacement of taxonomies? *Proceedings of the VLDB Endowment*, 17(11):2919–2932.

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

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. 2020. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019.

Jiachen Wang, Zikun Deng, Dazhen Deng, Xingbo Wang, Rui Sheng, Yi Cai, and Huamin Qu. 2025. Empowering multimodal analysis with visualization: A survey. *Computer Science Review*, 57:100748.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023a. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Zhaowei Wang, Quyet V Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2023b. Cola: Contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5253–5271.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024c. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3991–4010.

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2022. Subeventwriter: Iterative sub-event sequence generation with coherence controller. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1604.

Wisers. 2024. Wisers: best media monitoring. https://login.wisers.net/. [Online; Accessed 01-January-2024].

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, et al. 2024a. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490.

Zhiqi Yang, Michael J DeFlorio, Agniv Sengupta, Jiabao Wang, Christopher M Castellano, Alexander Gershunov, Kristen Guirguis, Emily Slinskey, Bin Guan, Luca Delle Monache, et al. 2024b. Seasonality and climate modes influence the temporal clustering of unique atmospheric rivers in the western us. *Communications Earth & Environment*, 5(1):734.

Martin V Young and Nick S Grahame. 2022. The history of uk weather forecasting: the changing role of the central guidance forecaster. part 1: the pre-computer era. *Weather*, 77(10):344–348.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

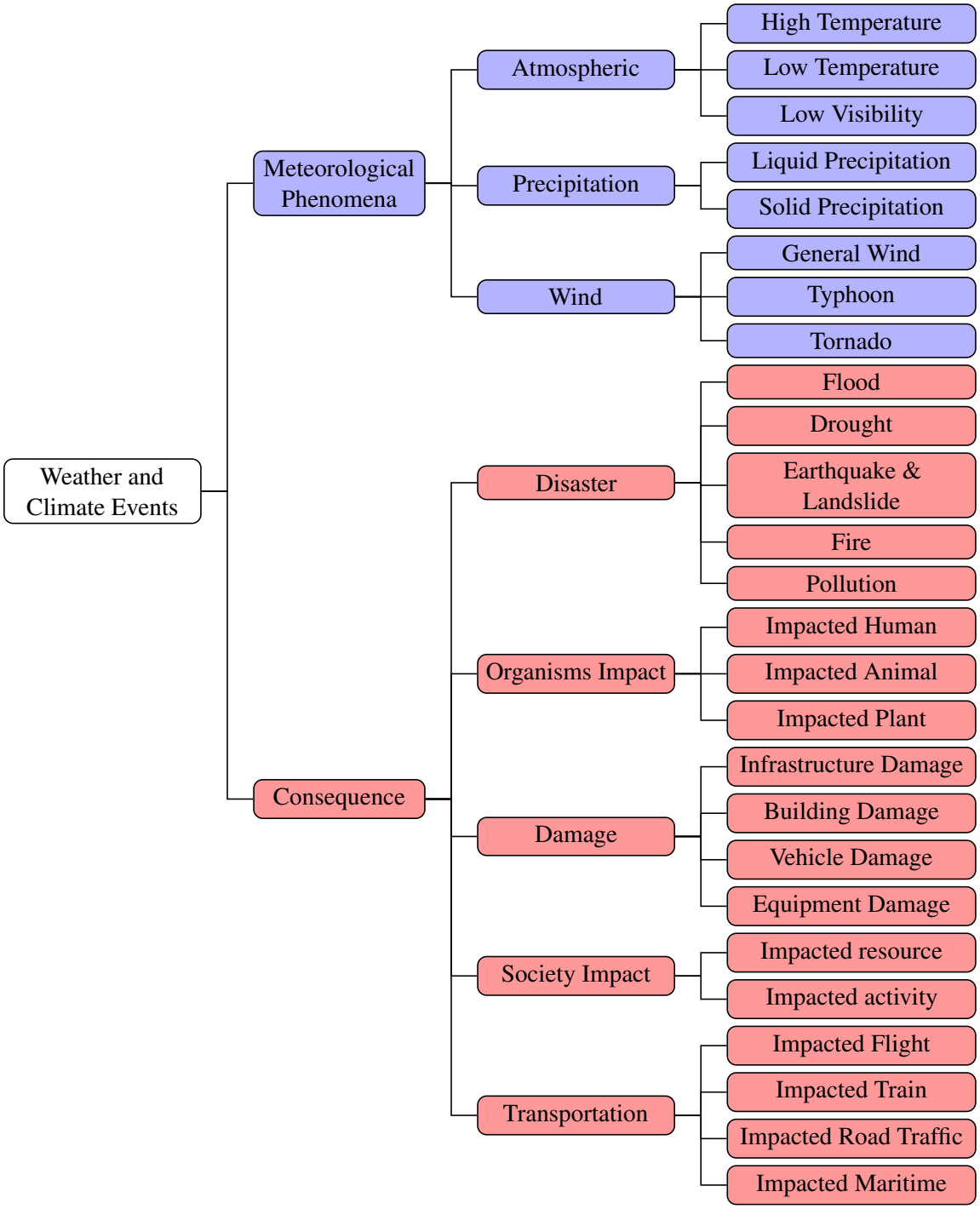
A Dataset Release and License

A.1 CLLMate License and Availability.

The CLLMate dataset is available at https://github.com/hobolee/CLLMate and uses the CC BY 4.0 license. The data format is:


```
{
    ...
    "6973": {
        "caused by": "None",
        "cause": "6970",
        "news_id": "5838681",
        "event": "high temperature",
        "category_name": "phenomena.atmospheric.high temperature",
        "category_index": "A.A.A",
        "time": "2022/07/25",
        "location": "Guangzhou City",
        "coordinate": "[20.69, 110.88, 25.69, 115.88]",
        "image_path": "./image/6973.png",
        "temperature": {
            "max": "302.75", "min": "302.00", "mean": "302.50"},
        "wind speed": {
            "max": "6.56", "min": "4.09", "mean": "5.51"},
        "precipitation": {
            "max": "0.00", "min": "0.00", "mean": "0.00"}
    },
    "6974": {
        "caused by": "None",
        "cause": "None",
        "news_id": "5677320",
        "event": "increased wildfire severity",
        "category_name": "consequence.disaster.fire",
        "category_index": "B.A.D",
        "time": "2022/07/25",
        "location": "California State",
        "coordinate": "[32.5288, -124.4096, 42.0095, -114.1312]",
        "image_path": "./image/6974.png",
        "temperature": {
            "max": "312.50", "min": "295.00", "mean": "306.00"},
        "wind speed": {
            "max": "12.07", "min": "0.00", "mean": "4.01"},
        "precipitation": {
            "max": "0.00", "min": "0.00", "mean": "0.00"}
    },
    ...
}
```

A.2 Category


Weather and Climate Events
Meteorological Phenomena
Atmospheric
High Temperature
Low Temperature
Low Visibility
Precipitation
Liquid Precipitation
Solid Precipitation
Wind
General Wind
Typhoon
Tornado
Consequence
Disaster
Flood
Drought
Earthquake & Landslide
Fire
Pollution
Organisms Impact
Impacted Human
Impacted Animal
Impacted Plant
Damage
Infrastructure Damage
Building Damage
Vehicle Damage
Equipment Damage
Society Impact
Impacted resource
Impacted activity
Transportation
Impacted Flight
Impacted Train
Impacted Road Traffic
Impacted Maritime

A.3 Category Distribution

The dataset is partitioned chronologically into training (90%) and test (10%) sets. The category distributions of the entire dataset are displayed in Figure 5 and Figure 4. There are imbalances in the category distribution, such as for the primary category of meteorological phenomena, there are 43.7% events are atmospheric-related, 40.5% events are precipitation-related, and only 15.8% events are wind-related. That is because all events are extracted from the news and follow the distribution of the recording of events. We also present the category distribution of training (Figure 6) and testing (Figure 7) sets.

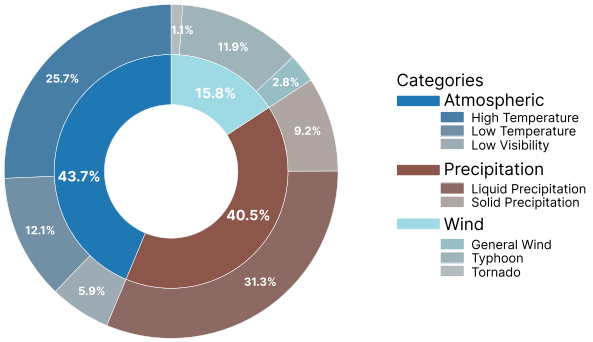


Figure 4: Distribution of categories within the meteorological phenomena category (3,979/7747 events). The distribution is imbalanced, reflecting the nature of event reporting in the news.

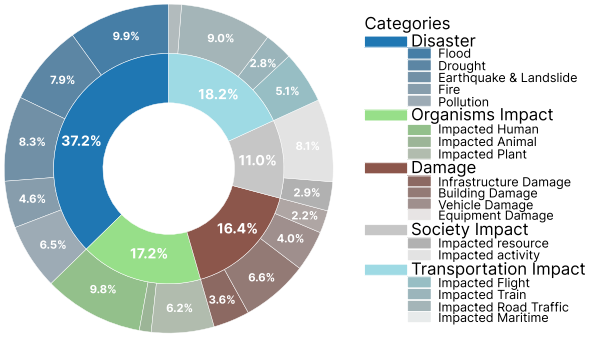


Figure 5: Distribution of categories within the consequences category (3,768/7747 events). The distribution is imbalanced, reflecting the nature of event reporting in the news.

A.4 Generalizability of CLLMate

While the CLLMate benchmark is structured for hierarchical categorization via multiple-choice evaluation, its design inherently supports other discriminative or generative tasks, such as open-ended generation tasks. The inclusion of raw textual event descriptions (subsection A.1) alongside aligned meteorological data enables applications beyond constrained categorization. For example:

- Free-form event narrative generation: Training or evaluating MLLMs to produce human-readable weather/climate event summaries from numerical inputs.
- Causal reasoning exploration: Probing model capabilities to infer and articulate chains of causality between meteorological drivers and societal impacts.
- Automated report drafting: Generating localized risk assessments or mitigation advisories grounded in spatio-temporal climate patterns.
- Open for integrating additional meteorological data (e.g., more meteorological variables) and more modalities (e.g., satellite images). For each event, we provide the location and date, making it easy to align future data.

The generalizability of CLLMate ensures compatibility with both discriminative and generative evaluation paradigms, broadening its utility for research in meteorology-informed language modeling and operational forecasting systems.

A.5 ERA5 Reanalysis Dataset

The link to the ERA5 dataset is https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview. The ERA5 dataset uses a license of Copernicus products. This license permits access to Copernicus Products for any lawful purpose. Authorized uses include, but are not limited to, reproduction, distribution, public dissemination, adaptation, modification, and integration with other data and information.

The meteorological variables of ERA5 used in CLLMate are:

ERA5 Varaibles

- **2m temperature**: This parameter is the temperature of air at 2m above the surface of land, sea or inland waters. 2m temperature is calculated by interpolating between the lowest model level and the Earth's surface, taking account of the atmospheric conditions. This parameter has units of kelvin (K). Temperature measured in kelvin can be converted to degrees Celsius (°C) by subtracting 273.15.

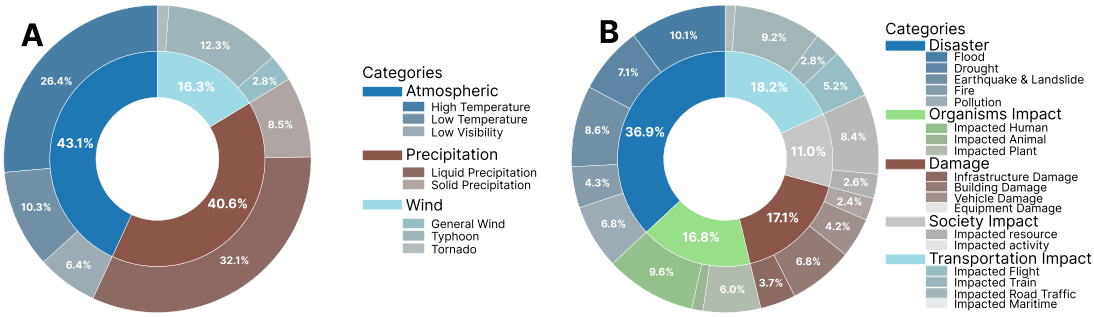


Figure 6: Distribution of categories of the training set. A: meteorological phenomena categories (3487/7747 events). B: consequences categories (3485/7747 events).

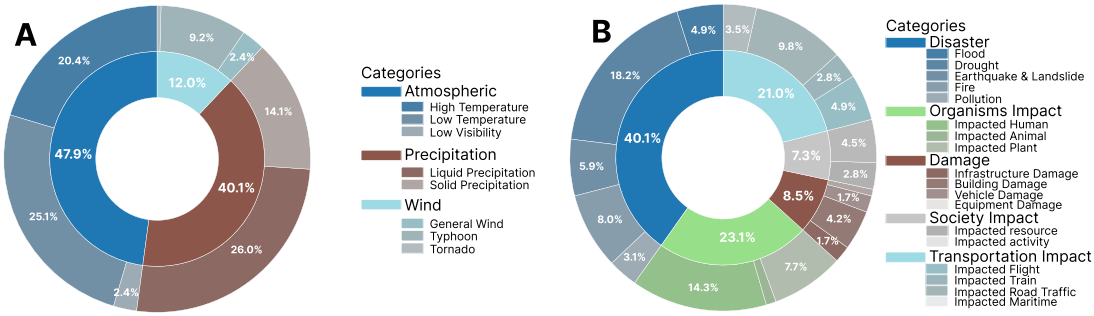


Figure 7: Distribution of categories of the testing set. A: meteorological phenomena categories (492/7747 events). B: consequences categories (283/7747 events).

- **10m u-component of wind**: This parameter is the eastward component of the 10m wind. It is the horizontal speed of air moving towards the east, at a height of ten metres above the surface of the Earth, in metres per second. This parameter can be combined with the V component of 10m wind to give the speed and direction of the horizontal 10m wind. This parameter has units of metre per second (m/s).
- **10m v-component of wind**: This parameter is the northward component of the 10m wind. It is the horizontal speed of air moving towards the north, at a height of ten metres above the surface of the Earth, in metres per second. This parameter can be combined with the U component of 10m wind to give the speed and direction of the horizontal 10m wind. his parameter has units of metre per second (m/s).
- **Total precipitation**: This parameter is the accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface. It is the sum of large-scale precipitation and convective precipitation. Large-scale precipitation is generated by the cloud scheme in the ECMWF Integrated Forecasting System (IFS). This parameter has units of meter (m).

A.6 The environmental news dataset.

The environmental news dataset used to extract meteorological events was acquired through a procurement process from Wisers (Wisers, 2024).

B Prompt and Annotation

B.1 Prompt for Extracting Events

System Prompt

You are an AI assistant with expertise in extracting environmental events. Your task is to analyze the given news article and identify environmental events (and their consequences) and their corresponding locations and time that have actually occurred and are explicitly mentioned in the text. Only extract triples if they pertain to environmental events and actual events. Do not make inferences.

Output Definition

In the extracted triples, the first element should represent the subject, the second element should be the constrained verb 'cause,' and the third element should represent the object. Both the subject and object should relate to environmental events or risks. If the geographic location is not provided, indicate 'no'. If the date is provided, output YYYY/MM/DD. If the date is not provided, indicate 'no'.

Zero-Shot CoT

Initially, evaluate if the news article discusses environmental events. Next, determine if the associated geographic location and time are specified. If both conditions are met, output the list [subject, cause, object, location, time]. The output should be in English.

Few-shot

If the news is <news article containing environmental events>, the output should be <correct events with location and time>. If the news is <news article containing no environmental events>, the output should be "No". Now, the news is: <news to be analyzed>

B.2 Annotation Details

We presented the original news articles along with the extracted events to three domain experts for annotation. The experts are postdocs and PhD candidates in the domain. There are no ethical concerns, as the experts involved do not experience any harm. Furthermore, they were fairly compensated at a rate of 20 USD per hour for their contributions. Each expert assessed the extracted events based on the corresponding news article, assigning a label of 1 (accurate) or 0 (inaccurate). The final annotation for each event was determined by majority vote. Out of 6,352 news articles, the extracted events from 2,575 articles (40.54%) were deemed accurate. The inter-annotator agreement (IAA) score, measured by the pairwise agreement proportion, was 81.83%, and Fleiss's κ (Fleiss, 1971) was calculated to be 0.63.

B.3 Prompt for Phenomena Forecasting

System Prompt

You are an AI assistant with expertise in weather analysis. Your task is to interpret meteorological data and figures, analyze weather variables, and provide accurate insights into weather conditions.

Location Information

The provided figure corresponds to <city_name> city, covering the geographic area defined by:Latitude: [<lat_min>, <lat_max>], Longitude: [<lon_min>, <lon_max>].

Meteorological Parameters

The daily mean meteorological parameters I provided are as follows:
The 2m temperature: the temperature of air at 2m above the surface of land, sea or inland waters. This parameter has units of kelvin (K). Average daily mean temperature of that area: <t_mean> K. Max daily mean temperature of that area: <t_max> K. Min daily mean temperature of that area: <t_min> K.
The 10m wind: the speed of air moving towards the north, at a height of ten metres above the surface of the Earth, in metres per second. Average daily mean wind speed of that area: <w_mean> m/s. Max daily mean wind speed of that area: <w_max> m/s. Min daily mean wind speed of that area: <w_min> m/s.
Precipitation: the accumulated liquid and frozen water, comprising rain and snow, that falls to the Earth's surface. The units of this parameter are depth in metres of water equivalent. Average daily mean precipitation of that area: <p_mean> m. Max daily mean precipitation of that area: <p_max> m. Min daily mean precipitation of that area: <p_min> m.

Meteorological Image

The figure provided is an RGB image derived from spatiotemporal meteorological

data for the region. Each channel represents: channel 1: Temperature. channel 2: Wind Speed. channel 3: Precipitation.

Options (The order is random for different instances.)

The possible answers are structured into main categories with sub-options:

- A: Atmospheric
  - A.1: High Temperature
  - A.2: Low Temperature
  - A.3: Low Visibility
- B: Wind
  - B.1: General Wind
  - B.2: Tornado
  - B.3: Typhoon
- C: Precipitation
  - C.1: Solid Precipitation
  - C.2: Liquid Precipitation

Zero-Shot CoT

To determine the most likely meteorological scenario, follow these steps: 1. Analyze Statistical Data: Evaluate the statistical information (mean, max, and min values) for temperature, wind speed, and precipitation. Identify any conditions that suggest extreme weather or notable patterns. 2. Analyze Spatial Patterns in the Image: Examine the RGB image for regional variations in temperature, wind, and precipitation. Consider the spatial distribution of these variables to identify meteorological phenomena. 3. Synthesize Findings: Combine insights from the statistical data and image analysis. Choose the only most appropriate sub-option from the given categories based on your analysis. You must output the serial number of the option.

B.4 Prompt for Consequences Forecasting

For the prompt for the consequences forecasting, we will change the "System Prompt" and "Options" sections by using the corresponding categories.

System Prompt

You are an AI assistant with expertise in weather analysis. Your task is to interpret meteorological data and figures, analyze weather variables, and provide accurate insights into meteorological consequences.

Options (The order is random for different instances.)

The possible answers are structured into main categories with sub-options:

- A: Disaster
  - A.1: Flood
  - A.2: Drought
  - A.3: Earthquake & Landslide
  - A.4: Fire
  - A.5: Pollution, including Air Pollution, Water Pollution, Solid Pollution
- B: Organisms Impact
  - B.1: Impacted Human, including Trap, Injury, Death
  - B.2: Impacted Animal
  - B.3: Impacted Plant
- C: Damage
  - C.1: Infrastructure Damage
  - C.2: Building Damage
  - C.3: Vehicle Damage
  - C.4: Equipment Damage
- Society Impact
  - D.1: Impacted resource, including Food Shortage, Power Shortage
  - D.2: Impacted activity, including Work Activity, School Activity
- Transportation Impact
  - E.1: Impacted Flight
  - E.2: Impacted Train
  - E.3: Impacted Road Traffic
  - E.4: Impacted Maritime

C Visual Instruction Tuning Models

We fine-tuned the LLaVA models on the training set of CLLMate. The fine-tuned models include LLaVA-1.5-7B/13B and LLaVA-1.6-vicuna-7B/13B.

C.1 Vision Encoder and Adapter

We maintain consistency with LLaVA on the vision encoder and adapter. For the vision encoder, we utilized CLIP ViT-L/14 (Radford et al., 2021) to extract the spatial patterns of meteorological images. For the vision-language adapter, we used the two-layer MLP vision-language adapter provided

by LLaVA. The vision encoder and the adapter are frozen during the fine-tuning.

C.2 Training Configuration

In terms of MLLMs, we fine-tuned LLaVA-7B and 13B models for one epoch using DeepSpeed ZeRO-3 on 8×A800 GPUs, with a learning rate of 2e-5 and batch size of 16. Training required 1 hour (7B) and 2 hours (13B) for one epoch. For the traditional AI models, the training parameters are: epochs: 50, batch Size: 32, optimizer: AdamW, learning Rate: 3e-4, loss Function: Cross Entropy. As for traditional AI models, we trained them using 1x4090 GPUs. The total API cost for closed-source models amounts to approximately $200.

D More Evaluation

D.1 Generation Study

We conducted experiments to directly generate meteorological events without providing predefined categories. To evaluate performance, domain experts manually compared the generated events with the labeled ones. The results (Table 3) indicate that MLLMs currently lack the ability to directly predict extreme weather events. This underscores the necessity of providing comprehensive event categories as a reference for MLLMs.

D.2 Category-wise Performance

To further complement our analysis, we calculated class-wise recall scores, as false negatives are particularly critical in high-stakes prediction tasks. We analyzed the performance of three representative LLMs for the primary category of meteorological phenomena (Table 6) and the primary category of meteorological consequence (Table 7) prediction (one closed-source, one open-source, and one fine-tuned model). The results show that LLMs perform poorly on transportation forecasting (significantly worse than other categories) and perform better on tasks like disaster forecasting. This discrepancy may be caused by the stronger relationship between meteorology and disasters. In contrast, the relationship between meteorology and transportation (e.g., flight delay) is more complex and likely influenced by human decisions based on meteorology and social factors.

D.3 Precision/Recall Results

In addition to the accuracy, we also present the recall (Table 4) and precision (Table 5) scores of the selected models.

D.4 Case Study

To illustrate the model's capabilities on the WCEF task, we analyze the representative examples (event 6973 and event 6974 in subsection A.1) with the original news article and MLLMs' output.

D.4.1 Case 1: Phenomena Forecasting (Event 6973)

We present some typical MLLMs's output in subsection A.1, with correct analyses highlighted in blue and incorrect ones in red, as assessed by a domain expert.

Current MLLMs all provide structured, human-aligned rationales and identify meteorological thresholds (e.g., 29.35°C). All models correctly excluded precipitation (C.1/C.2) and wind extremes (B.2/B.3); however, only Claude-3.5-Sonnet explicitly justified these exclusions using statistical ranges. The fine-tuned model, Fine-Tuned-LLaVA-1.6-vicuna-7B, provided direct answers without offering reasoning.

Regarding multimodal data, these models demonstrated accurate analysis of meteorological images. For instance, QWen2.5-VL-72B uniquely interpreted the meaning of colors, while LLaVA-34B concentrated on analyzing spatial uniformity.

Case 1: Event 6973 (shown in subsection A.1)

The news article:
Guangzhou issued its first high temperature red warning this year. Guangzhou will continue to be sunny and hot in the next three days.
Southern Metropolis Daily
Original Report | July 25, 2022, 16:24

Guangzhou has issued its first red high-temperature warning of the year. As of 4 p.m., seven districts in Guangzhou, excluding Conghua, Zengcheng, Nansha, and Panyu, have issued red warnings for high temperatures. The highest temperatures in these districts are expected to reach around 39°C today and tomorrow. According to the provincial government's defense guidelines, outdoor activities should be minimized during the daytime, and outdoor work in open areas should be suspended, except for spe-

Models	Phenomena	Consequence
Gemini-2.0-flash (Team et al., 2023)	1.42%	1.77%
LLaVA-1.6-mistral-7B (Liu et al., 2024)	1.02%	4.59%

Table 3: The accuracy of MLLMs in directly forecasting meteorological events without providing predefined categories.

Models	Primary Category of Phenomena	Subordinate Category of Phenomena	Primary Category of Consequence	Subordinate Category of Consequence
Claude-3.7-Sonnet (Anthropic, 2025)	40.04%	20.93%	42.05%	15.55%
Qwen2.5-VL-72B (QwenTeam, 2025)	40.04%	20.93%	42.05%	15.55%
Fine-tuned-LLaVA-1.6-vicuna-13B	45.53%	27.03%	42.40%	9.19%

Table 4: The recall of benchmarked MLLMs across two granularity levels: primary category classification and subordinate category identification for both meteorological phenomena and their cascading consequence.

Models	Primary Category of Phenomena	Subordinate Category of Phenomena	Primary Category of Consequence	Subordinate Category of Consequence
Claude-3.7-Sonnet (Anthropic, 2025)	40.40%	21.15%	42.32%	17.71%
Qwen2.5-VL-72B (QwenTeam, 2025)	40.58%	21.13%	39.98%	18.34%
Fine-tuned-LLaVA-1.6-vicuna-13B	45.22%	26.21%	42.99%	9.08%

Table 5: The precision of benchmarked MLLMs across two granularity levels: primary category classification and subordinate category identification for both meteorological phenomena and their cascading consequence.

Models	Atmospheric	Precipitation	Wind
Claude-3.7-Sonnet (Anthropic, 2025)	43.51%	37.18%	39.56%
Qwen2.5-VL-72B (QwenTeam, 2025)	39.61%	32.69%	48.90%
Fine-tuned-LLaVA-1.6-vicuna-13B	41.56%	33.33%	59.34%

Table 6: The category-wise recall scores of benchmarked MLLMs across the primary category of meteorological phenomena.

Models	Disaster	Organisms Impact	Damage	Society Impact	Transportation
Claude-3.7-Sonnet (Anthropic, 2025)	49.15%	41.18%	45.61%	50.00%	26.56%
Qwen2.5-VL-72B (QwenTeam, 2025)	42.37%	41.18%	42.11%	44.23%	29.69%
Fine-tuned-LLaVA-1.6-vicuna-13B	44.07%	39.22%	43.86%	48.08%	37.50%

Table 7: The category-wise recall scores of benchmarked MLLMs across the primary category of meteorological consequence.

cial industries.
Under the influence of a strong subtropical high-pressure system, Guangzhou is expected to experience widespread sunny and hot weather with high temperatures over the next three days.

Ground Truth: A.1: High Temperature (option in the question)
Event: high temperature
Category_name: phenomena.atmospheric.high temperature
Category_index: A.A.A
Time: 2022/07/25
Location: Guangzhou City

GPT-4o's output:

- 1. Analyze Statistical Data:
  - Temperature: Average daily mean temperature is 302.50 K, with a max of 302.75 K and a min of 302.00 K. These values suggest relatively high temperatures.
  - Wind Speed: Average daily mean wind speed is 5.51 m/s, with a max of 6.56 m/s and a min of 4.09 m/s. These figures suggest moderate wind speeds.
  - Precipitation: All values show 0.00 m, indicating no precipitation.
- 2. Analyze Spatial Patterns in the Image: The image predominantly shows a strong red hue, indicating high temperatures. There is also a noticeable green component, indicating the presence of notable wind speeds.
- 3. Synthesize Findings: The data indicates high temperatures without any precipitation and moderate winds.

Conclusion: The most appropriate option based on the analysis is: A.1: High Temperature.

Claude-3.5-Sonnet's output:

Let me analyze the meteorological data and image systematically:

- 1. Statistical Data Analysis:
  - Temperature: 302.5K (29.35°C) with small variation (0.75K range)
  - Wind speed: Average 5.51 m/s with moderate variation (2.47 m/s range)
  - Precipitation: 0.00m (no precipitation)
- 2. Image Analysis: The image shows a predominantly yellow/orange color. High intensity in the red channel (temperature). Moderate intensity in green channel (wind). No visible blue component (confirming zero precipitation).
- 3. Synthesis: The high temperature (302.5K/29.35°C) and uniform yellow/orange coloring indicates hot conditions. Wind speeds are moderate but not extreme. No precipitation is present. The spatial pattern shows consistent high temperature across the region. Based on these observations: The dominant feature is the high temperature. No extreme wind events (ruling out B options). No precipitation (ruling out C options). The uniform high temperature pattern best matches A.1.

Therefore, I select option A.1 (High Temperature) as the most appropriate classification for these conditions. Answer: A.1

LLaVA-1.6-vicuna-34B's output:

Based on the provided meteorological data and the RGB image, the most likely meteorological scenario for the region is: A: Atmospheric. A.1: High Temperature. The average daily mean temperature of 302.50 K (29.50°C) is relatively high, and the maximum daily mean temperature of 302.75 K (29.75°C) is even higher. The image shows a uniform distribution of high temperatures across the region, with no significant variations in temperature. The wind speed and precipitation are relatively low and do not show any extreme patterns. The high temperatures suggest a warm and sunny climate, which is consistent with the image's depiction of high temperatures.

Qwen2.5-VL-72B's output:

A.1. Analysis:

- Step 1: Analyze Statistical Data
  - Temperature: Average daily mean temperature: 302.50 K (approximately 29.35°C). Max daily mean temperature: 302.75 K (approximately 29.60°C). Min daily mean temperature: 302.00 K (approximately 28.85°C). These temperatures indicate relatively high and consistent warmth across the region, suggesting a hot day.
  - Wind Speed: Average daily mean wind speed: 5.51 m/s. Max daily mean wind speed: 6.56 m/s. Min daily mean wind speed: 4.09 m/s. The wind speeds are moderate but

not indicative of extreme weather like tornadoes or typhoons.
  - Precipitation: Average daily mean precipitation: 0.00 m. Max daily mean precipitation: 0.00 m. Min daily mean precipitation: 0.00 m. There is no recorded precipitation, indicating dry conditions.
- Step 2: Analyze Spatial Patterns in the Image. The image is uniformly yellow, which suggests a consistent pattern across the region. Given the color coding: Red (Temperature): The yellow hue indicates a high temperature since red is associated with temperature. Green (Wind Speed): The lack of green suggests moderate wind speeds, consistent with the statistical data. Blue (Precipitation): The absence of blue confirms the lack of precipitation.
- Step 3: Synthesize Findings. Combining the statistical data and the spatial patterns observed in the image, it is clear that the region experienced high temperatures with no significant precipitation and moderate wind speeds.

Therefore, the most appropriate sub-option is A.1: High Temperature.

Fine-Tuned-LLaVA-1.6-vicuna-7B's output:
A.1: High Temperature

D.4.2 Case 2: Consequence Forecasting (Event 6974)

We present some typical MLLMs's output in subsection A.1, with correct analyses highlighted in blue and incorrect ones in red, as assessed by a domain expert.

This case demonstrates critical disparities in multimodal causal reasoning for fire risk prediction. While GPT-4o and Claude-3.5-Sonnet identified fire-prone conditions (high temperatures, moderate winds (should be strong wind), and zero precipitation), both erroneously selected A.4: Drought rather than A.5: Fire. In contrast, Qwen2.5-VL-72B successfully synthesized all variables, especially interpreting wind spikes as fire accelerants, to correctly predict A.5: Fire.

Case 2: Event 6974 (shown in subsection A.1)

The news article:
Wildfire spreads near Yosemite National Park in the United States

On July 25 local time, the spread of a wildfire near Yosemite National Park in California slowed down, the containment rate increased, and firefighting efforts showed signs of progress.
The California Department of Forestry and Fire Protection (Cal Fire) stated on the evening of the 25th that it was a "successful day" for firefighting efforts. From morning to evening, strong winds caused the burned area to expand from 16,791 acres (approximately 68 square kilometers) to 17,241 acres (approximately 70 square kilometers), while the containment rate rose from 10% to 16%. The agency reported that the fire had destroyed 55 structures. As firefighting efforts advanced, evacuation orders in some areas were downgraded to evacuation warnings, and certain roads were reopened for local residents.
On the 25th, a China News Service reporter observed at the fire site that firefighters had cleared a large number of firebreaks through the forested areas. In areas near major roads, the fire was no longer severe, but thick smoke continued to rise from the hillsides and mountain tops. Several aircraft took turns dropping water or fire retardants onto larger hotspots, while fire trucks were dispersed across various locations to respond to potential flare-ups. On the charred land, small groups of firefighters were extinguishing smaller fires.
According to Cal Fire, as of the evening of the 25th, nearly 3,000 firefighters, 24 helicopters, 302 fire engines, and 68 water trucks had been deployed to the fire site. On that day alone, helicopters dropped 300,000 gallons (approximately 1.136 million liters) of water onto the fire.
With climate change causing increasingly hotter and drier conditions in the western United States, wildfires in California have become more severe. This wildfire

near Yosemite National Park started in the Midpines area of Mariposa County and spread rapidly due to strong winds, becoming California's largest wildfire so far this year. U.S. media have reported that this wildfire is likely the beginning of another severe wildfire season in California.

Ground Truth: A.5: Fire (option in the question)
Event: increased wildfire severity
Category_name:consequence.disaster.fire
Category_index: B.A.D
Time: 2022/07/25
Location: California State

GPT-4o's output: To determine the most likely meteorological scenario, let's analyze the provided data and image:

- 1. Analyze Statistical Data:
  - Temperature: The average daily mean temperature is 306.00 K, with a max of 312.50 K. This indicates high temperatures, potentially indicative of heat waves.
  - Wind Speed: The average daily mean wind speed is 4.01 m/s with peaks at 12.07 m/s. This suggests mild to moderate winds, with occasional stronger gusts.
  - Precipitation: There is no precipitation recorded (0.00 m), indicating very dry conditions.
- 2. Analyze Spatial Patterns in the Image: The image is predominantly yellow with reddish tones, indicating high temperatures (red channel is for temperature). This suggests a uniform region of warm temperatures. The lack of blue and green tones indicates negligible precipitation and moderate to minimal wind activity.
- 3. Synthesize Findings: Temperature and Precipitation: The very high temperature coupled with the lack of precipitation suggests a dry heat scenario. Impact Analysis: The conditions are suggestive of potential drought and fire risks due to high temperatures and no rainfall. Considering the provided categories: Drought (A.4): Prolonged high temperatures and no precipitation are indicative of drought conditions, impacting water availability, agriculture, and potentially leading to wildfires.

Given this analysis, the most appropriate sub-option based on the scenario is: A.4: Drought.

Claude-3.5-Sonnet's output:

Let me analyze the meteorological data and image systematically:

- 1. Statistical Data Analysis:
  - Mean: 306.00 K (32.85°C). Max: 312.50 K (39.35°C). Min: 295.00 K (21.85°C). These are very high temperatures, indicating extreme heat conditions.
  - Wind: Mean: 4.01 m/s. Max: 12.07 m/s. Min: 0.00 m/s. Moderate to high wind speeds present.
  - Precipitation: All values at 0.00 m. Indicating completely dry conditions with no rainfall.
- 2. Image Analysis: Dominant yellow-orange coloring indicates high temperatures. Some variations in green suggesting wind speed variations. No blue tinting, confirming zero precipitation.
- 3. Synthesis: The combination of: Extremely high temperatures (approaching 40°C). Complete absence of precipitation. Presence of wind. Geographic location (California). These conditions strongly indicate a severe drought situation with high fire risk. The complete lack of precipitation, combined with high temperatures and wind, creates perfect conditions for drought conditions.

Based on the given options and this analysis, the most appropriate classification would be: A.4 (Drought). The conditions present clear indicators of drought conditions, which can lead to water shortages and increased fire risk. This is supported by both the statistical data showing zero precipitation and very high temperatures, as well as the im-

age showing warm conditions across the region.

LLaVA-1.6-vicuna-34B's output:
Based on the provided meteorological data and the RGB image, I will analyze the situation and provide the most appropriate sub-option from the given categories.

- 1. Analyze Statistical Data: The average daily mean temperature of the area is 306.00 K, which is equivalent to 32.83 degrees Celsius. The max daily mean temperature is 312.50 K (34.38 degrees Celsius), and the min daily mean temperature is 295.00 K (27.27 degrees Celsius). The average daily mean wind speed is 4.01 m/s, with a max daily mean wind speed of 12.07 m/s and a min daily mean wind speed of 0.00 m/s. The average daily mean precipitation is 0.00 m, with a max daily mean precipitation of 0.00 m and a min daily mean precipitation of 0.00 m. From the statistical data, we can observe that the area experiences relatively high temperatures, with a range of 7.56 degrees Celsius between the max and min daily mean temperatures. The wind speeds are relatively low, with a range of 12.07 m/s between the max and min daily mean wind speeds. The precipitation levels are very low, with no precipitation recorded.

- 2. Analyze Spatial Patterns in the Image: The RGB image shows variations in temperature, wind speed, and precipitation across the region. The red channel (temperature) shows a gradient of temperature, with warmer areas in the south and cooler areas in the north. The green channel (wind speed) shows a gradient of wind speed, with higher wind speeds in the south and lower wind speeds in the north. The blue channel (precipitation) shows no precipitation across the region. From the image analysis, we can observe that the region experiences a gradient of temperature and wind speed, with warmer and windier conditions in the south and cooler and calmer conditions in the north. However, there is no precipitation recorded in the image.

- 3. Synthesize Findings: Based on the statistical data and image analysis, the most appropriate sub-option from the given categories is: D: Organisms Impact D.1: Impacted Plant. The reason for this choice is that the region experiences high temperatures and low precipitation, which can lead to drought conditions. Drought can have a significant impact on plant life, as it can cause stress, wilting, and even death. The lack of precipitation in the image further supports this conclusion.

Therefore, the serial number of the most appropriate sub-option is: D.1.

Qwen2.5-VL-72B's output:

Based on the provided meteorological data and the analysis of the RGB image:

- Step 1: Analyze Statistical Data
  - Temperature: Average daily mean temperature: 306.00 K (approximately 32.85°C). Maximum daily mean temperature: 312.50 K (approximately 39.35°C). Minimum daily mean temperature: 295.00 K (approximately 21.85°C). These temperatures indicate relatively warm conditions, with some areas experiencing high temperatures.
  - Wind Speed: Average daily mean wind speed: 4.01 m/s. Maximum daily mean wind speed: 12.07 m/s. Minimum daily mean wind speed: 0.00 m/s. The average wind speed is moderate, but there are areas with higher wind speeds, which could be significant.
  - Precipitation: Average daily mean precipitation: 0.00 m. Maximum daily mean precipitation: 0.00 m. Minimum daily mean precipitation:

0.00 m. There is no recorded precipitation, indicating dry conditions.

- Step 2: Analyze Spatial Patterns in the Image. Temperature (Red Channel): The red color dominates the image, suggesting widespread high temperatures across the region. Wind Speed (Green Channel): The green color is present but less dominant compared to red, indicating variable wind speeds with some areas experiencing higher wind speeds. Precipitation (Blue Channel): The absence of blue suggests no precipitation across the region.
- Step 3: Synthesize Findings. Given the high temperatures, lack of precipitation, and variable wind speeds, the most likely meteorological scenario is one associated with dry and hot conditions. The combination of high temperatures and wind can lead to increased risk of wildfires.

Conclusion: The most appropriate sub-option based on the analysis is: A.5: Fire.

Fine-Tuned-LLaVA-1.6-vicuna-7B's output:

A.2: Earthquake & Landslide

D.5 Ablation Study

To quantify the contribution of multimodal inputs, we conducted an ablation study with Gemini-2.0-flash (closed-source) and LLaVA-1.6-mistral-7B (open-source) for their good performance (e.g., LLaVA-1.6-mistral-7B has the best performance among 7/8B open-source MLLMs). We ablated three components: meteorological images, statistical summaries, and Zero-Shot CoT prompting. The results are displayed in Table 8.

D.5.1 Closed-Source Models (Gemini-2.0-Flash)

Image Removal: Eliminating meteorological raster data catastrophically degrades phenomena forecasting (37.80% → 25.81%, 19.72% → 14.63%), but marginally improves consequence subordinate accuracy (30.39% → 30.74%, 13.43% → 14.84%). This suggests Gemini-2.0-flash relies heavily on spatial patterns for phenomena forecasting.

Statistics Removal: Removing statistics surprisingly does not significantly impact performance, improving overall accuracy from 30.39% to 38.87% of primary consequence forecasting. This indicates a redundancy between statistical summaries and raster data for Gemini-2.0-flash.

Zero-Shot CoT Removal: Disabling Zero-Shot CoT harms event forecasting, especially for consequence forecasting for Gemini-2.0-flash, underscoring its necessity for causal event differentiation.

D.5.2 Open-Source Models (LLaVA-1.6-Mistral-7B)

Image Removal: Degrades performance universally and largely, with catastrophic drops in consequence forecasting (39.22% → 13.78%). Unlike Gemini-2.0-flash, LLaVA-1.6-mistral-7B cannot compensate for missing spatial data via text-based statistics.

Statistics Removal: Improves primary consequence accuracy (39.22% → 44.52%) while harming phenomena prediction (32.93% → 26.22%), suggesting statistical summaries introduce noise for impact forecasting but aid event detection for LLaVA-1.6-mistral-7B.

Zero-Shot CoT Removal: Erases Zero-Shot CoT, reducing the performance of LLaVA-1.6-mistral-7B with the exception of primary phenomena forecasting.

D.5.3 Key Insights

Modality Asymmetry: MLLMs exhibit flexible modality compensation (e.g., statistics → images) in part of tasks but fail to compensate for all tasks. For example, LLaVA-1.6-mistral-7B has a performance drop for primary phenomena forecasting but not for consequence forecasting when ablating statistics.

Reasoning Bottleneck: Zero-Shot CoT prompting provides gains, especially in subordinate accuracy, proving essential for parsing causal meteorology-event relationships.

D.6 Impact of Model Scale

We conducted the scale study on the open-source models, the results are shown in Figure 8 and Figure 9.

D.6.1 Scale

Contrary to expectations, increasing the model scale does not guarantee improved performance:

LLaVA-1.6-34B achieves state-of-the-art primary phenomena accuracy (42.28%) but catastrophically fails at primary consequence forecasting (17.67% vs. 20.85% for its 7B variant), sug-

Models	Primary Category of Phenomena	Subordinate Category of Phenomena	Primary Category of Consequence	Subordinate Category of Consequence
Gemini-2.0-flash (Team et al., 2023)	37.80%	19.72%	30.39%	13.43%
w/o image	25.81%	14.63%	30.74%	14.84%
w/o statistics	37.80%	17.48%	38.87%	13.43%
w/o Zero-Shot CoT	33.94%	11.99%	17.67%	3.89%
LLaVA-1.6-mistral-7B (Liu et al., 2024)	32.93%	17.28%	39.22%	12.01%
w/o image	27.64%	13.41%	13.78%	4.95%
w/o statistics	26.22%	15.45%	44.52%	7.77%
w/o Zero-Shot CoT	33.33%	10.77%	20.49%	3.89%

Table 8: The results of the ablation study show the accuracy of benchmarked MLLMs across two granularity levels: primary category classification (coarse-grained) and subordinate category identification (fine-grained) for both meteorological phenomena and their cascading consequences. w/o image: remove the meteorological image from the prompt. w/o: remove the statistics of meteorological variables from the prompt. w/o: remove the Zero-Shot CoT from the prompt.

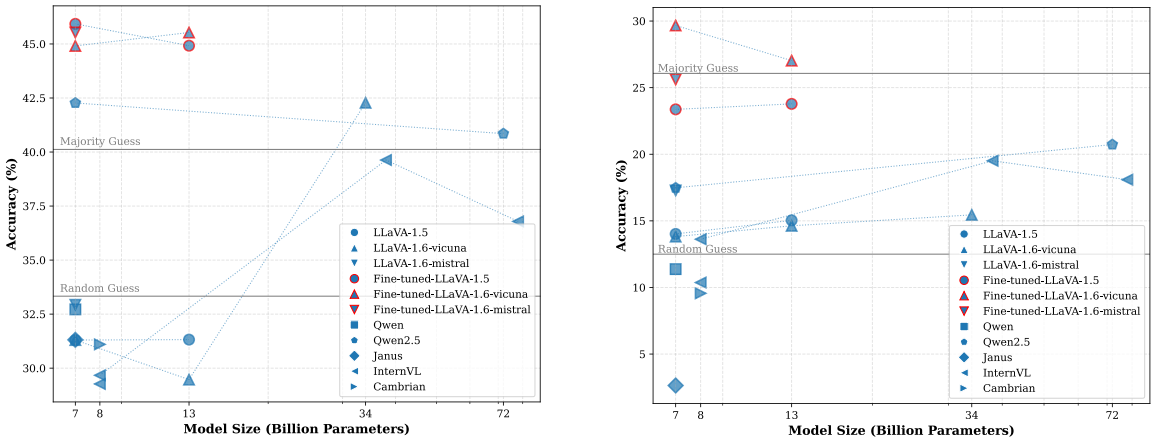


Figure 8: The line chart of the accuracy of open-sourced MLLMs. Left: Primary category of phenomena forecasting. Right: Subordinate category of phenomena forecasting. The x-axis is in the log scale.

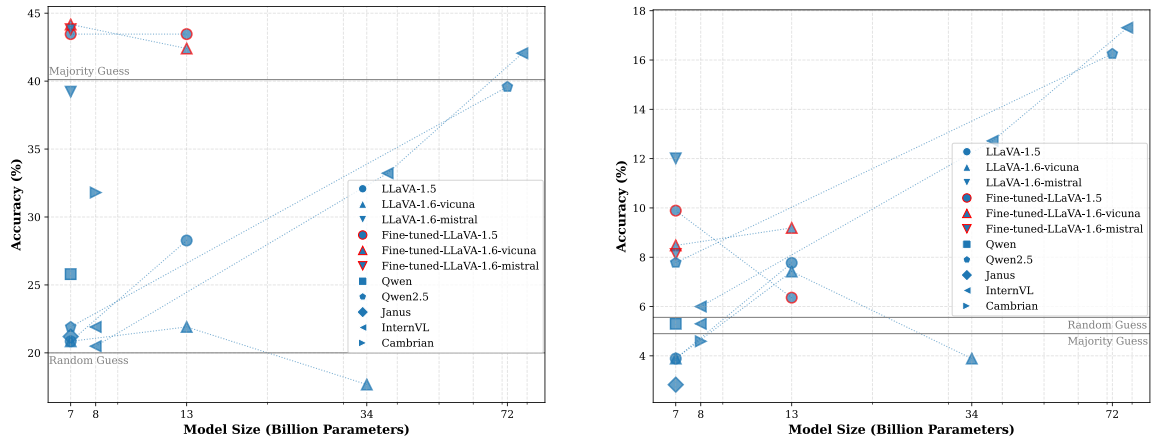


Figure 9: The line chart of the accuracy of open-sourced MLLMs. Left: Primary category of consequence forecasting. Right: Subordinate category of consequence forecasting. The x-axis is in the log scale.

gesting larger models overfit to meteorological patterns while losing causal reasoning capabilities of consequence.

The Qwen2.5-VL and InternVL3 series demon-

strate scale-dependent tradeoffs. For example, Scaling from 7B to 72B, Qwen2.5-VL improves subordinate consequence accuracy by 17.67% in primary consequence forecasting (21.91% $\rightarrow$ 39.58%) and 8.47% in subordinate consequence forecasting (7.78% $\rightarrow$ 16.25%) but reduces primary phenomena forecasting (42.27% $\rightarrow$ 40.85%).

D.6.2 Architectural Refinements

Architectural improvements (LLaVA-1.5 $\rightarrow$ 1.6, Qwen2 $\rightarrow$ Qwen2.5) yield significant gains. The LLaVA-1.6-Mistral-7B variant outperforms LLaVA-1.5-7B in all tasks. The Qwen2.5-VL-7B outperforms the Qwen2-VL-7B except for the primary consequence forecasting.

D.6.3 Fine-Tuning as a Scaling Alternative

Task-specific adaptation eclipses scale advantages: Smaller fine-tuned models consistently outperform larger untrained counterparts (13B/34B), with Fine-tuned-LLaVA-1.6-vicuna-7B surpassing the 34B untrained variant by 26.50% in primary consequence forecasting (44.17% vs. 17.67%). Fine-tuned-LLaVA-1.6-vicuna-7B also achieves 29.67% subordinate phenomena accuracy vs. 15.45% for LLaVA-1.6-vicuna-34B. Fine-tuned 7B models outperform larger untrained counterparts, proving task alignment outweighs raw scale for meteorology-to-text translation.