Aquilon: Towards Building Multimodal Weather LLMs

Sumanth Varambally [1] **Veeramakali Vignesh Manivannan** [2] **Yasaman Jafari** [2] **Luyu Han** [2] **Zachary Novack** [2] **Zhirui Xia** [2 3] **Salva Rühling Cachay** [2] **Srikar Eranky** [2] **Ruijia Niu** [2] **Taylor Berg-Kirkpatrick** [2] **Duncan Watson-Parris** [1 4] **Yian Ma** [1 2] **Rose Yu** [2]

Abstract

Recent advancements in weather foundation models—pre-trained on vast amounts of structured numerical data—have set new standards in weather forecasting accuracy. However, their lack of language-based reasoning capabilities leaves a critical opportunity untapped for human-in-the-loop analysis systems. In contrast, large language models (LLMs) excel at understanding and generating text, but they struggle with high-dimensional weather inputs like meteorological datasets. In this work, we take a significant step towards bridging this gap by enabling multimodal LLMs to reason over complex weather data. We address two fundamental challenges: (1) the absence of large-scale, multitask, multimodal datasets for weather reasoning, and (2) the lack of methods for embedding multi-channel weather data into LLM-compatible representations. To tackle these, we introduce a scalable data generation pipeline that constructs diverse question-answer pairs across a wide spectrum of weather-related tasks, from basic lookups to advanced forecasting and extreme event detection. We also leverage pre-trained weather foundation models to extract low-dimensional embeddings of weather fields, enabling their integration with LLMs. Our experiments reveal that multimodal weather reasoning is a challenging problem that current models only partially address—highlighting the need for more effective weather representations and richer training data to fully unlock the potential of LLMs in meteorological applications.

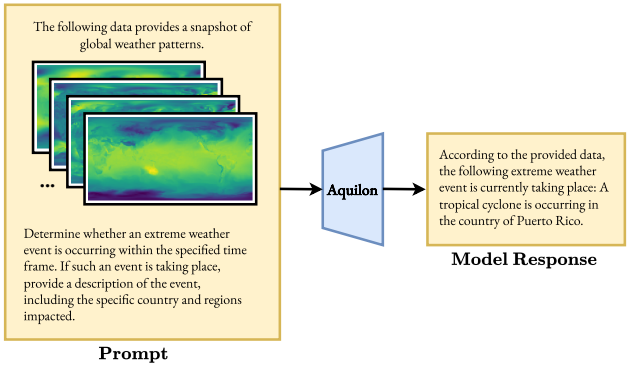

Figure 1: **Aquilon pipeline for multimodal weather reasoning.** High-dimensional weather data is encoded alongside natural language questions, enabling a natural language interface with meteorological data.

1. Introduction

Accurate weather modeling is an important scientific problem with wide ranging implications, ranging from agriculture and disaster preparedness to transportation and energy management (Alley et al., 2019). Traditional meteorological systems rely on physics-based numerical simulations (Molteni et al., 1996; Bauer et al., 2015), which are computationally expensive and typically require expert interpretation to produce actionable forecasts, identify extreme weather events, and assess climate risks. In recent years, machine learning has emerged as a powerful tool in augmenting or even replacing traditional weather modeling methods. Foundation models for weather forecasting (Nguyen et al., 2023; Kurth et al., 2023; Lam et al., 2023; Bi et al., 2023; Nguyen et al., 2024) have achieved state-of-the-art performance in medium-range forecasting. However, these models are trained solely on structured numerical weather data (e.g., reanalysis data) and are not designed to incorporate important information sources from other data modalities like textual weather bulletins, observational metadata, or weather reports from field stations. Crucially, they lack the ability to support interactive, language-based querying or reasoning over weather data, limiting their utility in human-in-the-loop or decision-support systems.

Meanwhile, large language models (LLMs) have shown

*Equal contribution [1]Halıcıoğlu Data Science Institute, UC San Diego [2]Computer Science and Engineering, UC San Diego [3]Department of Mathematics, UC San Diego [4]Scripps Institution of Oceanography, UC San Diego. Correspondence to: Sumanth Varambally <svarambally@ucsd.edu>.

ICML 2025 Workshop on Assessing World Models.

great promise across a variety of scientific domains (Birhane et al., 2023), including drug discovery (Zheng et al., 2024; Wu et al., 2024b), materials science (Lei et al., 2024; Jablonka et al., 2023), and network biology (Theodoris et al., 2023). LLMs excel at processing textual sources such as research literature, source code (Jiang et al., 2024), and even tabular or structured data (Zhang et al., 2024). However, their ability to reason about numerical weather data remains limited, particularly in domains like meteorology where the data are not naturally represented as text.

There are several challenges that impede LLMs from effectively using weather data. Meteorological data are high dimensional, dynamic, and complex, far exceeding the context length of any existing LLMs. Numerical data is continuous-valued as opposed to discrete text tokens. Textual data and numerical weather data also describe phenomena at vastly different spatial and temporal scales. These challenges raise a critical question - How can we design architectures that allow LLMs to work with multimodal weather data? While recent advances in multimodal LLMs have shown impressive capabilities in integrating text with images (Wang et al., 2022; Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023c; Li et al., 2023; Liu et al., 2023a; 2024), video (Zhao et al., 2022; Maaz et al., 2023; Zhang et al., 2023; Cheng et al., 2024; Lin et al., 2024; Zhang et al., 2025), and audio (Chu et al., 2023; 2024; Défossez et al., 2024; Wu et al., 2024a; 2025; Doh et al., 2025; Ghosh et al., 2025), these modalities are not well suited for processing global, multi-channel weather data. Recent works (Chen et al., 2024a; Li et al., 2024; Ma et al., 2024) have applied vision-language models to specific meteorological tasks. However, they focus on specific applications, such as extreme weather event prediction, and utilize only a limited subset of atmospheric variables. As a result, these models are not designed for general-purpose, multitask reasoning across diverse meteorological queries, limiting their broader applicability.

In this work, we take the first step towards enabling multimodal LLMs for weather applications by addressing two fundamental challenges: (1) the absence of a suitable multitask, multimodal dataset for training and evaluating such models, and (2) the lack of effective methods for embedding multi-channel weather data into representations compatible with LLMs. To tackle the first challenge, we propose a scalable data generation pipeline that constructs diverse question-answer pairs across weather-related tasks of varying complexity. These tasks are designed to progressively demand more advanced reasoning capabilities, ranging from straightforward lookups and basic forecasting to extreme event prediction and counterfactual analysis. We construct our dataset by pairing ERA5 reanalysis data (Hersbach et al., 2020) — sourced from WeatherBench 2 (Rasp et al., 2024) at 6-hourly intervals — with question-answer pairs spanning 7 distinct task types. This ensures a sufficiently large and diverse dataset to support both model training and rigorous evaluation across a broad range of meteorological applications.

For the second challenge, we demonstrate how pretrained weather foundation models, such as Stormer (Nguyen et al., 2024), can be leveraged to extract low-dimensional representations of weather data, enabling effective integration with LLMs. We introduce the Aquilon pipeline (Figure 1), which incorporates these embeddings into a multimodal LLM framework, allowing the model to process both textual questions and structured weather inputs. Through comprehensive experiments, we evaluate Aquilon alongside a range of baselines, including a fine-tuned text-only LLM and the training-free Program-Aided Language (PAL) model. We find that while our approach shows promise on some tasks, challenges remain in more complex meteorological problems such as extreme weather event detection. These findings underscore the complexity of multimodal weather reasoning and highlight several directions for future research. Our contributions include:

1. We curate a comprehensive multitask, multimodal benchmark dataset spanning 7 distinct weather tasks, consisting of 439,900 training samples, 1,000 validation samples, and 3,344 test samples.
2. We design Aquilon, a novel architecture that integrates low-dimensional representations from pretrained weather models with large language models.
3. Our experiments reveal the limitations of current models on complex meteorological reasoning tasks and highlight clear directions for future work.

2. Related Work

Weather Foundation Models. Large-scale pretrained weather models (Lam et al., 2023; Price et al., 2025; Bi et al., 2023; Pathak et al., 2022; Nguyen et al., 2023; Bodnar et al., 2024; Nguyen et al., 2024) are reshaping numerical weather forecasting by outperforming traditional PDE-based systems (Molteni et al., 1996) with significantly greater efficiency. However, these models are typically limited to single-task forecasting and lack capabilities for natural language interaction or general-purpose multitask reasoning.

General-Purpose Vision-Language Models. Vision-language models (VLMs) (Li et al., 2021; Alayrac et al., 2022; Li et al., 2022; 2023; Liu et al., 2023c;b;a; 2024) have shown impressive performance on a wide range of multimodal benchmarks. Yet, their applicability to scientific domains remains limited. Most VLMs operate on 3-channel RGB inputs and struggle with tasks requiring precise numerical reasoning. In contrast, weather data comprises structured, multivariate fields that demand specialized architectures for effective integration with LLMs.

Multimodal Weather Datasets. Recent efforts have introduced multimodal datasets for weather and earth observation. Terra (Chen et al., 2024b) pairs geo-tagged imagery with text, but places limited emphasis on meteorological data. ClimateIQA (Chen et al., 2024a) targets localized extreme event detection using wind gust data, while WeatherQA (Ma et al., 2024) focuses on severe weather reasoning using satellite imagery and expert discussions. CLLMate (Li et al., 2024) aligns news articles with ERA5 data for event categorization. Despite these contributions, existing datasets are narrow in scope, typically targeting a single task or using only a small subset of ERA5 variables. However, weather is inherently complex, involving interactions across many atmospheric variables and spatial scales. To address this gap, we introduce a scalable, multitask, multimodal dataset and benchmark, built from 69-channel ERA5 data to support diverse weather reasoning tasks.

3. Data Curation

At the moment, there does not exist a large-scale dataset that combines structured weather data with textual or task-specific annotations. To address this, we design a scalable data curation pipeline.

We leverage the ERA5 reanalysis dataset (Hersbach et al., 2020), specifically from WeatherBench 2 (Rasp et al., 2024) which provides global atmospheric data at a spatial resolution of 1.5°. We include 4 surface variables and 5 atmospheric variables with 13 levels each, for a total of 69 channels. To create task-specific examples, we define natural language templates for each task type, with placeholders such as location, variable, and time window. These placeholders are filled by randomly sampling inputs, and the corresponding ground truth is computed deterministically using human-written code applied to the raw ERA5 data. To add more diversity to the training dataset, we use a small LLM (specifically, Phi-4 (Abdin et al., 2024)) to reword questions generated by our data curation pipeline. Figure 2 shows an example template, and a sample generated from it.

We structure the dataset around two tiers or levels of tasks. Level 1 consists of simple retrieval and aggregation tasks, while Level 2 includes forecasting and extreme weather event detection tasks. Table 1 provides an overview of the task types we implement.

To enable location-based queries, we introduce the **Geolocator**, a wrapper around the Natural Earth dataset (Natural Earth, 2024) that maps ERA5 grid points to natural language location names such as countries, states, and water bodies. For extreme weather event tasks (Tasks 6 and 7), we use records from the EM-DAT international disaster database (Delforge et al., 2025), matching event entries by date and location to the ERA5 data.

Id	Level	Task Description
1-a	1	Determine which location has the highest/lowest value for a variable
1-b	1	Determine maximum, minimum or average value of a variable at a location
1-c	1	Determine which location within a geofeature has the highest/lowest value for a variable
2-a	2	Predict future value of a variable at a location
2-b	2	Predict when a variable at a location reaches its maximum or minimum value
2-c	2	Detect if an extreme weather event is currently happening
2-d	2	Predict upcoming extreme weather events based on current weather state

Table 1: **Summary of curated multimodal weather tasks.** Level 1 consists of retrieval and aggregation tasks, while Level 2 includes forecasting and extreme weather event detection tasks.

Using our framework, we construct a dataset comprising 439,900 training samples based on weather snapshots from 1979 to 2020, 1,000 validation samples from 2021, and 3,344 test samples from 2022. For a detailed breakdown of dataset statistics, please refer to Appendix A.1.

4. Evaluation Metrics

Since all our tasks are designed around weather tasks with objectively correct answers, we design an evaluation pipeline to evaluate the scientific correctness of the answers produced by the models. The model answers fall into three primary categories: **numeric**, **temporal**, and **spatial (location-based)**. Given that model outputs are in natural language, we evaluate them through a multi-stage process:

1. **Verification:** Determine whether the model's response contains a relevant and valid answer, using an LLM to assess correctness. (In this case, `gpt-4.1-mini`).
2. **Extraction:** Extract the specific answer from the model response using another LLM prompt.
3. **Scoring:** Apply scoring methods specific to the type of question, which are detailed below.

Numeric Answers. For numeric responses, we compute the Standarized Mean Squared Error (Std. MSE) between the predicted and reference values. To normalize across variables with different scales or units, we divide each MSE by the variance of the corresponding variable in the dataset. The final metric is the mean of these standardized MSEs across all tasks.

Time-based Answers. We evaluate tasks with time values as responses using MSE. We omit the standarization step, since all the answers are in the same units (that is, hours).

Location-based Answers. For questions whose answers are geographic locations, we first match the extracted location name to one of the expected entries from the NaturalEarth dataset (e.g., mapping "USA" to "United States of America"). For countries, we use the `country_converter` library (Stadler, 2017). For other geographic entities such as continents and water bodies, we apply fuzzy string matching (Bachmann et al., 2023), accepting matches above a predefined similarity threshold.

To quantitatively assess the geographic deviation between predicted and reference locations, we employ the Earth Mover's Distance (EMD) (Monge, 1781) as a primary evaluation metric. We begin by generating surface area-weighted masks over a latitude–longitude grid for both the predicted and reference locations. These masks are normalized to form probability distributions. To account for the curvature of the Earth, we compute pairwise distances between grid points using geodesic distance. The EMD is then calculated using the `POT` library (Flamary et al., 2021). As a complementary metric, we also report Location Accuracy, which simply measures whether the predicted and reference location strings are an exact match.

Extreme-Weather Tasks. In order to evaluate the extreme-weather tasks, we report two metrics: (1) F1 score, which only assesses whether the model correctly predicts the *occurrence* of an extreme event anywhere in the world, without considering event type or exact location. (2) Earthmover's Distance, which measures the agreement between the reference and predicted list of countries.

5. Methods

We evaluate the capability of language models to perform weather reasoning by benchmarking a diverse set of architectures, ranging from training-free baselines to custom multimodal large language models (LLMs). Below, we describe each model and training setup in detail.

Pre-trained Frontier LLM. As a zero-shot baseline, we evaluate a pre-trained frontier language model (Open AI `gpt-4.1`) on its ability to answer weather reasoning questions without direct access to any structured weather data. The model receives only the natural language metadata and the target question, without seeing the underlying numerical inputs from the dataset.

PAL. The Program-Aided Language (PAL) model is a training-free baseline designed to solve weather-related questions by generating and executing Python code. To answer a given question, PAL constructs an input prompt that includes documentation for the Geolocator object (which provides geographic utility functions), descriptions of variables and keys from the WeatherBench2 dataset, and the question itself. This prompt is then passed to a pretrained LLM (OpenAI `o4-mini` (OpenAI, 2025)), which generates a Python program intended to solve the question. The generated code is executed in a sandboxed environment, using the relevant data from the question and the Geolocator object as inputs. If the code fails—due to errors or timeouts—PAL employs a retry mechanism. It prompts the LLM again with additional context, such as error tracebacks or efficiency suggestions, to regenerate and rerun the code. This process is repeated up to a maximum of three times (`max_attempts`=3).

Fine-tuned text-only LLM. To assess how well language models can learn statistical patterns from the textual portion of our dataset, we fine-tune a text-only LLaMA 3.1-8B model (Grattafiori et al., 2024). The model takes as input the natural language metadata along with the target question. Fine-tuning is performed using LoRA with $r = 32$ and $\alpha = 32$. As the model is trained on our training dataset, it can learn recurring associations and trends present in the text. This setup serves as a strong text-only baseline.

Aquilon: Multimodal LLMs with Weather Encoders. We propose to integrate high-dimensional weather data using pretrained encoders that convert 69-channel weather inputs into sequences of weather embeddings. These embeddings replace the metadata of the weather data in the input prompt. To delineate the time series segment, we wrap the weather embeddings with special tokens with learnable embeddings: `<W_START>` at the beginning and `<W_END>` at the end. The resulting sequence—comprising both the weather embeddings and the user's question—is passed to the LLM as a single input stream (see Figure 3).

1. U-Net Encoder. We use a U-Net model (Dhariwal & Nichol, 2021) pretrained on the ERA5 dataset to forecast 6 hours ahead. We use the encoder's intermediate activations (of shape $(1024, 30, 16)$) by flattening and projecting them through a two-layer MLP, resulting in 1024 tokens of dimension 4096.

2. Stormer Encoder. Stormer (Nguyen et al., 2024) is a ViT-style model tailored to weather forecasting. We remove its unpatchify layers and retain the output of the final self-attention layer (shape $(2048, 1024)$). These are projected via a two-layer MLP to 2048 tokens with embedding dimension 4096.

In both models, encoder backbones remain trainable. The LLM component is fine-tuned with LoRA using the same configuration as the text-only baseline. This setup mirrors approaches like LLaVA (Liu et al., 2023c) for images, treating weather data as a separate modality integrated through

Model	% Valid Outputs (↑)	MSE (Time Tasks, in hr^2) (↓)	Std. MSE (Numerical) (↓)
GPT 4.1	88.60	573.57	328.89
PAL	95.3	759.60	**0.2393**
FT-LLaMA 3.1 8B	100	544.00	0.2612
Aquilon U-Net	100	**507.13**	0.2540
Aquilon Stormer	99.97	529.46	0.2992

Table 2: Output validity and regression metrics for numerical answers. Std. MSE stands for standarized MSE. FT-LLaMA 3.1 8B refers to a LLaMA 3.1 8B model finetuned on our training dataset.

Model	Location Accuracy (%)(↑)	EMD (km) (↓)	Extreme Weather F1 (↑)
GPT 4.1	6.8	5644.38	0.025
PAL	**74.3**	**1684.50**	0.347
FT-LLaMA 3.1 8B	29.8	4131.86	**0.450**
Aquilon U-Net	30.4	4121.15	0.004
Aquilon Stormer	28.0	4189.00	0

Table 3: Location metrics for location answer-based questions. EMD stands for Earthmovers Distance. FT-LLaMA 3.1 8B refers to a LLaMA 3.1 8B model finetuned on our training dataset.

specialized encoders. All training methods are optimized using cross-entropy loss computed over the predicted answers, conditioned on the input question and its corresponding weather data context.

6. Experimental Results

We evaluate model performance across all task types introduced in Section 3. All models, except for PAL (which is training-free), are trained for three epochs. Tables 2 and 3 report results on the held-out test set for all models described in Section 5.

The PAL model demonstrates strong performance across a variety of tasks. It achieves the lowest mean squared error (MSE) on numerical questions, the lowest Earthmover's Distance (EMD) for location questions, and the highest scores in both location accuracy and Extreme Weather F1. Its advantage on location-based tasks is likely due to its use of the Geolocator, enabling explicit reasoning over spatial inputs. Aquilon models outperform all others on time prediction tasks. However, the improvement over the text-only LLaMA model is marginal, indicating that the current form of weather embeddings contributes only modestly to performance gains. It is interesting to note that all models—including Aquilon and PAL—struggle with extreme weather detection. Notably, the text-only model achieves the highest F1, and no model is able to correctly localize an extreme event in most cases (see Tables 11 and 12), underscoring the inherent difficulty of reasoning about rare and spatially complex meteorological phenomena. For a detailed breakdown of performance by task and level, please refer to Appendix A.3.

7. Conclusion

We tackled the challenge of enabling LLMs to reason over high-dimensional, complex weather data by developing a scalable data generation pipeline and curating a large, diverse dataset for multimodal weather reasoning. We evaluated multiple baselines, including PAL—a training-free program synthesis model—and a fine-tuned text-only LLM. We also introduced the Aquilon architecture, integrating pretrained weather foundation models with LLMs. While results are promising on certain tasks, significant challenges remain—especially in spatially grounded tasks like extreme weather detection. Future work could improve on both the data and modeling fronts. On the dataset side, expanding the range of tasks and incorporating textual weather reports or observational metadata could better support real-world applications. On the modeling side, developing more effective, text-aware weather embeddings may offer improved performance. Finally, extending the setting to include temporal context in the input—i.e., modeling full spatiotemporal sequences—presents a promising direction for capturing dynamic weather phenomena.

Acknowledgements

Sumanth would like to thank Chinmay Talegaonkar for several useful discussions about this work. This work was supported in part by the U.S. Army Research Office under Army-ECASE award W911NF-07-R-0003-03, the

U.S. Department Of Energy, Office of Science, IARPA HAYSTAC Program, and NSF Grants #2205093, #2146343, #2134274, CDC-RFA-FT-23-0069, DARPA AIE FoundSci and DARPA YFA.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Alley, R. B., Emanuel, K. A., and Zhang, F. Advances in weather prediction. *Science*, 363(6425):342–344, 2019.

Bachmann, M., layday, Kokkinou, G., Fihl-Pearson, J., dj, Schreiner, H., Sherman, M., Górny, M., pekkarr, Delfini, Hess, D., Rosin, G., Moine, H. L., Tang, K., Renkamp, N., H, T., glynn, and odidev. rapidfuzz/rapidfuzz: Release 3.6.1, December 2023. URL https://doi.org/10.5281/zenodo.10440201.

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

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Birhane, A., Kasirzadeh, A., Leslie, D., and Wachter, S. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, 2023.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024.

Chen, J., Zhou, P., Hua, Y., Chong, D., Cao, M., Li, Y., Yuan, Z., Zhu, B., and Liang, J. Vision-language models meet meteorology: Developing models for extreme weather events detection with heatmaps. *arXiv preprint arXiv:2406.09838*, 2024a.

Chen, W., Hao, X., Wu, Y., and Liang, Y. Terra: A multimodal spatio-temporal dataset spanning the earth. *Advances in Neural Information Processing Systems*, 37: 66329–66356, 2024b.

Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Delforge, D., Wathelet, V., Below, R., Lanfredi Sofia, C., Tonnelier, M., van Loenhout, J. A. F., and Speybroeck, N. EM-DAT: The Emergency Events Database. *International Journal of Disaster Risk Reduction*, pp. 105509, 2025. doi: 10.1016/j.ijdrr.2025.105509. URL https://doi.org/10.1016/j.ijdrr.2025.105509.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Doh, S., Choi, K., and Nam, J. Talkplay: Multimodal music recommendation with large language models. *arXiv preprint arXiv:2502.13713*, 2025.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Ghosh, S., Kong, Z., Kumar, S., Sakshi, S., Kim, J., Ping, W., Valle, R., Manocha, D., and Catanzaro, B. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730): 1999–2049, 2020.

Jablonka, K. M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J. D., Bran, A. M., Bringuier, S., Brinson, L. C., Choudhary, K., Circi, D., et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital discovery*, 2(5):1233–1250, 2023.

Jiang, J., Wang, F., Shen, J., Kim, S., and Kim, S. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., and Anandkumar, A. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Lei, G., Docherty, R., and Cooper, S. J. Materials science in the era of large language models: a perspective. *Digital Discovery*, 3(7):1257–1272, 2024.

Li, H., Wang, Z., Wang, J., Lau, A. K. H., and Qu, H. Cllmate: A multimodal llm for weather and climate events forecasting. *arXiv preprint arXiv:2409.19058*, 2024.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023c.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ma, C., Hua, Z., Anderson-Frey, A., Iyer, V., Liu, X., and Qin, L. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, 122(529):73–119, 1996.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Natural Earth. Natural earth data. https://www.naturalearthdata.com/, 2024. Accessed: 2024-11-15.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Maulik, R., Kotamarthi, R., Foster, I., Madireddy, S., and Grover, A. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *Advances in Neural Information Processing Systems*, 37:68740–68771, 2024.

OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025. Accessed: 2025-05-21.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6): e2023MS004019, 2024.

Stadler, K. The country converter coco - a python package for converting country names between different classification schemes. *Journal of Open Source Software*, 2(16):332, 2017. doi: 10.21105/joss.00332. URL https://doi.org/10.21105/joss.00332.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022.

Wu, J., Novack, Z., Namburi, A., Dai, J., Dong, H.-W., Xie, Z., Chen, C., and McAuley, J. Futga: Towards fine-grained music understanding through temporally-enhanced generative augmentation. *arXiv preprint arXiv:2407.20445*, 2024a.

Wu, J., Novack, Z., Namburi, A., Dai, J., Dong, H.-W., Xie, Z., Chen, C., and McAuley, J. Futga-mir: Enhancing fine-grained and temporally-aware music understanding with music information retrieval. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.

Wu, K., Xia, Y., Deng, P., Liu, R., Zhang, Y., Guo, H., Cui, Y., Pei, Q., Wu, L., Xie, S., et al. Tamgen: drug design with target-aware molecule generation through a chemical language model. *Nature Communications*, 15 (1):9360, 2024b.

Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Zhang, H., Li, X., and Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023.

Zhang, X., Zhang, J., Ma, Z., Li, Y., Zhang, B., Li, G., Yao, Z., Xu, K., Zhou, J., Zhang-Li, D., et al. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*, 2024.

Zhao, Y., Misra, I., Krähenbühl, P., and Girdhar, R. Learning video representations from large language models. In *arXiv preprint arXiv:2212.04501*, 2022.

Zheng, Y., Koh, H. Y., Yang, M., Li, L., May, L. T., Webb, G. I., Pan, S., and Church, G. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *arXiv preprint arXiv:2409.04481*, 2024.

A. Appendix

A.1. Dataset Statistics

Table 4 reports the exact number of samples generated for each of the seven tasks across training, validation, and testing sets. Tasks 1-a to 1-c belong to Level 1, simple retrieval and aggregation, while Tasks 2-a to 2-d comprise Level 2, forecasting and extreme event detection. Each task has been allocated roughly 60,000-67,000 training instances, with smaller but proportionally balanced validation and test sets. This split strategy ensures each task level is well represented during model development and evaluation.

Id	Level	Training	Validation	Test
1-a	1	66530	140	500
1-b	1	66894	157	500
1-c	1	66576	127	500
2-a	2	59866	126	500
2-b	2	60117	171	500
2-c	2	59947	135	370
2-d	2	59970	144	474
Total Level 1		200000	424	1500
Total Level 2		239900	576	1844
Overall Total		439900	1000	3344

Table 4: Dataset statistics by task, level, and data split.

For the test set, we first randomly generate approximately 13500 test samples to ensure sufficient diversity in locations, geographical features, and weather variables, and to maintain a balanced number of samples per task. Then, to preserve representativeness while improving evaluation efficiency, we remove duplicate samples and discard redundant samples that target the same location or variable. We also selectively eliminate no-event cases to increase the proportion of samples with extreme weather events, thereby better evaluating the model's ability to identify such events across regions and time spans. Specifically, in Task 2-c, the proportion of extreme event samples increased from 0.071 to 0.16, with 61 extreme event cases. In Task 2-d, it rose from 0.30 to 0.86, with 410 extreme event cases. Finally, we balance the sample counts across all tasks, resulting in a total of 3,344 test samples. Table 4 shows the task-wise distribution of the samples across training, validation and test sets.

A.2. Example from the dataset

The following data shows a snapshot of the global weather fields.

{data}

Based on the above data, answer the following question:

Which {geofeature} experienced the {extremum_direction} average {variable}?","Based on the provided data, {answer} experienced the {extremum_direction} average {variable} over the specified time-period, with an average {variable} of {answer_numeric}."

Example Template

The following data shows a snapshot of the global weather fields.

{'type': 'wb2', 'variables': ['mean_sea_level_pressure', '10m_u_component_of_wind', '10m_v_component_of_wind', '2m_temperature', 'geopotential', 'specific_humidity', 'temperature', 'u_component_of_wind', 'v_component_of_wind'], 'time_indices': '54746:54747:1'}

Based on the above data, answer the following question: Which continent experienced the highest average Surface temperature?

Based on the provided data, Africa experienced the highest average Surface temperature over the specified time-period, with an average Surface temperature of 303.5 K.

Generated Sample

Figure 2: (left) Example Template from which samples are generated (right) A sample generated using the template.

A.3. Addititional Results

Table 5 summarizes model performance on the test set, broken down by task level, while Tables 6, 7, 8, 9, 10, 11 and 12 present detailed results for each individual task. Several notable patterns emerge:

PAL shows a stark contrast in performance between Level 1 and Level 2 tasks. It performs exceptionally well on Level 1 tasks but struggles considerably more on Level 2 tasks. This performance drop is unique to PAL—other models do not show such a pronounced disparity between levels. Notably, PAL also generates a disproportionately high number of invalid outputs on Task ID 3. Upon closer inspection, this issue appears to stem from PAL's inability to correctly interact with the Geolocator API in these instances.

More broadly, all models face significant challenges on Level 2 tasks. Task 2-c, in particular, proves especially difficult, with no model achieving meaningful performance. Interestingly, in Task 2-d, the best-performing baseline is the one that does not even use weather data, highlighting the complexity of this task and suggesting that the current methods for incorporating weather information remain insufficient.

Model	Std. MSE on numerical tasks (Level 1) (↓)	Std. MSE on numerical tasks (Level 2) (↓)	EMD (Level 1)(↓)	EMD (Level 2) (↓)
GPT 4.1	222.87	401.15	5638.42	6495.06
PAL	**0.0379**	0.4439	**883.24**	8968.75
LLaMA3.1 8B-LoRA	0.2254	0.2969	4099.71	**4368.29**
Aquilon U-Net	0.2123	**0.2957**	4116.88	8373.66
Aquilon Stormer	0.2864	0.3121	4119.00	-

Table 5: Level-wise comparison of model performance. Std. MSE stands for standarized MSE.

Model	Valid Outputs (%)	Location Accuracy (%)	Earth Mover's Distance Score
GPT 4.1	85.00	1.40	8310.64
PAL	99.40	**86.40**	**863.44**
LLaMA 3.1 8B-LoRA	100.00	15.20	5959.11
Aquilon U-Net	100.00	13.60	6230.75
Aquilon Stormer	100.00	12.20	6293.78

Table 6: Results for task ID 1-a: Determine which location has the highest/lowest value for a variable.

Model	Valid Outputs (%)	Std. Numerical MSE
GPT 4.1	63.40%	222.87
PAL	99.60%	**0.0379**
LLaMA 3.1 8B-LoRA	100.00%	0.2254
Aquilon U-Net	100.00%	0.2123
Aquilon Stormer	100.00%	0.2864

Table 7: Results for task ID 1-b: Determine maximum/minimum/average value of a variable at a location

Model	Valid Outputs (%)	Location Accuracy (%)	Earth Mover's Distance Score
GPT 4.1	86.80	12.20	2966.21
PAL	75.20	**62.20**	**903.03**
LLaMA 3.1 8B-LoRA	100.00	44.40	2240.32
Aquilon U-Net	100.00	47.20	2003.02
Aquilon Stormer	99.80	43.80	2084.21

Table 8: Results for task ID 1-c: Determine which location within a geofeature has the highest/lowest value for a variable

Model	Valid Outputs (%)	Numerical MSE
GPT 4.1	86.20	401.1512
PAL	98.20%	0.4439
LLaMA 3.1 8B - LoRA	100.00%	0.2969
Aquilon U-Net	100.00%	**0.2957**
Aquilon Stormer	100.00%	0.3121

Table 9: Results for task ID 2-a: Predict future value of a variable at a location

Model	Valid Outputs (%)	Time MSE
GPT 4.1	99.60%	573.57
PAL	95.40%	759.61
LLaMA 3.1 8B-LoRA	100.00%	544.10
Aquilon U-Net	100.00%	**507.14**
Aquilon Stormer	100.00%	529.47

Table 10: Results for task ID 2-b: Predict when a variable at a location reaches its maximum/minimum value

Model	Extreme Weather F1	**EMD** (km)
GPT 4.1	**0.14**	6495.0567
PAL	0.101	12395.1000
LLaMA 3.1 8B - LoRA	0.00	-
Aquilon U-Net	0.00	-
Aquilon Stormer	0.00	-

Table 11: Results for task ID 2-c: Detect if an extreme weather event is currently happening

Model	Extreme Weather F1	**EMD** (km)
GPT 4.1	0.00	-
PAL	0.404	8771.07
LLaMA 3.1 8B - LoRA	**0.50**	**4368.29**
Aquilon U-Net	0.005	8373.66
Aquilon Stormer	0.00	-

Table 12: Results for task ID 2-d: Predict upcoming extreme weather events based on current weather state

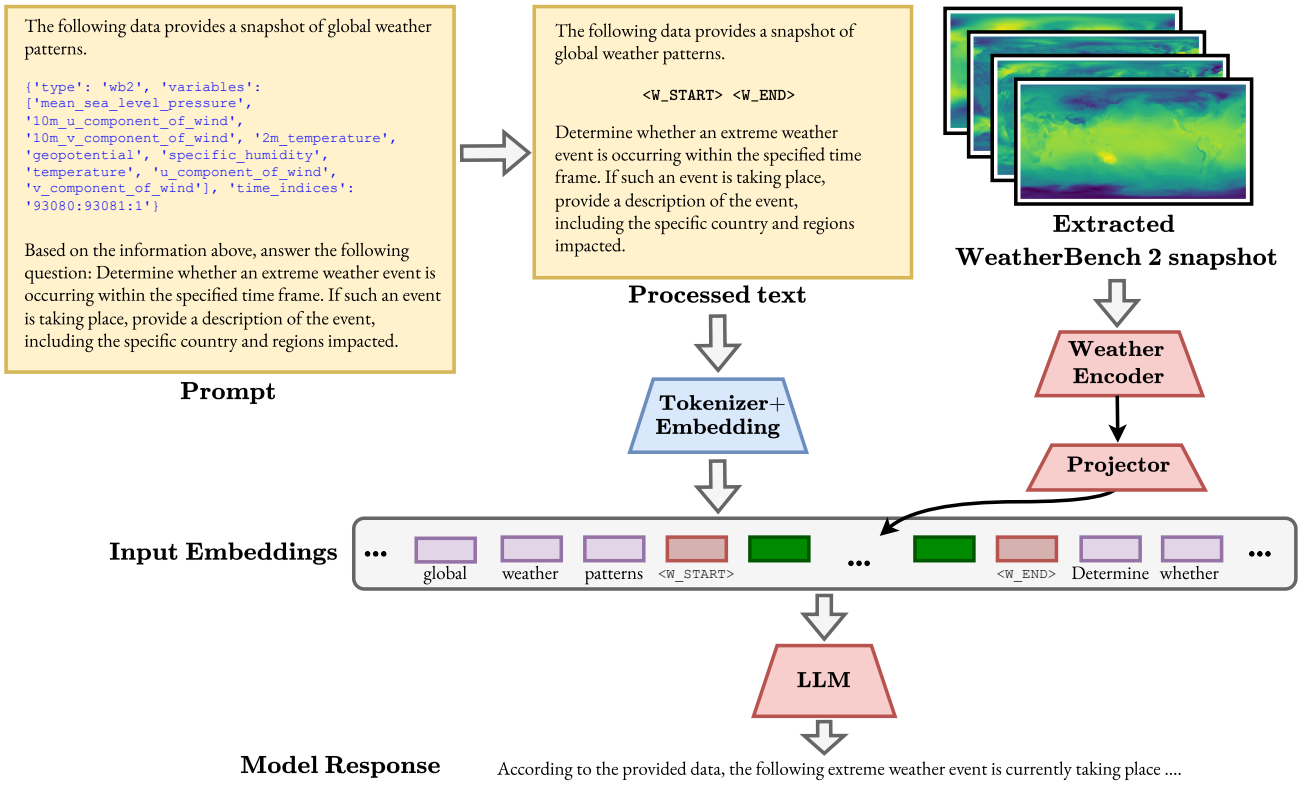

Figure 3: Overview of the Aquilon architecture. A natural language prompt containing a weather-related question and metadata is processed into token embeddings. The corresponding WeatherBench 2 snapshot is extracted and passed through a weather encoder and projector to produce dense weather embeddings. These embeddings are inserted into the token stream between special <W_START> and <W_END> tokens. The combined sequence is then fed into an LLM to enable language-based reasoning over structured weather data.