WeatherBench-R: A Text-Only Benchmark for Evaluating Large Language Models over U.S. Weather Events

Anonymous ACL submission

Abstract

Recent advances in data-driven weather modeling have enabled accurate numerical forecasts, whose outputs are often summarized as natural-language descriptions for interpretation and decision making. While large language models (LLMs) show promise in scientific reasoning, their ability to reason over text-only meteorological summaries, under physical constraints, incomplete evidence, and inherent uncertainty, remains poorly understood. Existing benchmarks primarily rely on multimodal inputs or fact verification, leaving this gap unaddressed. We introduce WeatherBench-R, a large-scale text-only benchmark for meteorological reasoning over U.S. weather events, constructed from ERA5 reanalysis summaries aligned with real-world NOAA storm records. WeatherBench-R decomposes reasoning into three complementary tasks: physical plausibility recognition from aggregate trends, consistency verification under partial and underspecified evidence, and counterfactual evidence reasoning that probes uncertainty awareness and explanation quality. The benchmark comprises **13,116** event-centered summaries spanning diverse event types and trend patterns. A systematic evaluation of LLMs reveals fragmented strengths across tasks, substantial performance degradation under counterfactual perturbations, and distinct failure modes in plausibility calibration and uncertainty handling. The code and data: https://anonymous.4open.science/r/WeatherBench-R-1ED9/.

1 Introduction

In recent years, rapid advances in data-driven weather and climate modeling (Oskarsson et al., 2024; Nguyen et al., 2024; Price et al., 2025) have led to the development of highly effective forecasting systems such as GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2022), and FGN (Alet et al., 2025). Although these models are capable of generating accurate numerical predictions, their outputs are often summarized as natural-language descriptions that remain cognitively demanding to interpret. Large Language Models (LLMs) (Zhao et al., 2023; Naveed et al., 2025) offer a promising direction for addressing this, given their demonstrated capabilities in scientific understanding across domains, such as medical diagnostics (Nazi and Peng, 2024; Yang et al., 2022) and analytical reasoning in economic and legal contexts (Sun, 2023; Guo and Yang, 2024). These advances motivate a question: Can LLMs accurately interpret meteorological textual summaries and provide scientifically grounded reasoning?

To advance the integration of LLMs into climate and weather intelligence, several benchmark datasets have recently been proposed. CLLMate (Li et al., 2025b), MeteorPred (Tang et al., 2025), and ClimateBench-M (Fu et al., 2025) formulate climate understanding as multimodal learning problems, enabling MLLMs (Wu et al., 2023; Yin et al., 2024) to reason over physical meteorological variables for forecasting, event prediction, and downstream applications. ClimateIQA (Chen et al., 2025) further introduces a large-scale heatmap-based VQA benchmark for fine-grained anomaly detection and spatial reasoning. In contrast, benchmarks such as ATMOSSCI-BENCH (Li et al., 2025a) primarily evaluate physical reasoning rather than natural-language weather reasoning, while climate-text datasets including Climate-FEVER (Diggelmann et al., 2020) and Climabench (Spokoyny et al., 2023) focus on fact verification. These limitations leave open the question of how well LLMs can reason over textual meteorological information, motivating the need for a dedicated text-only weather reasoning benchmark.

Despite the clear need for text-only evaluation, constructing a reliable benchmark for meteorological language reasoning is non-trivial. Unlike classical NLP tasks, weather reasoning is governed

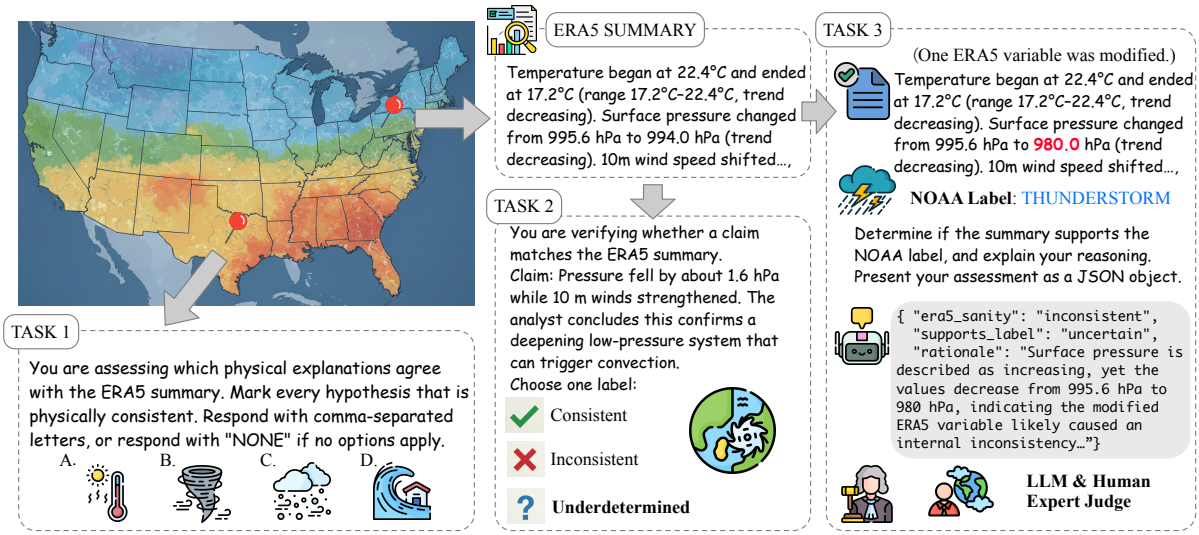

Figure 1: An overview of WeatherBench-R.

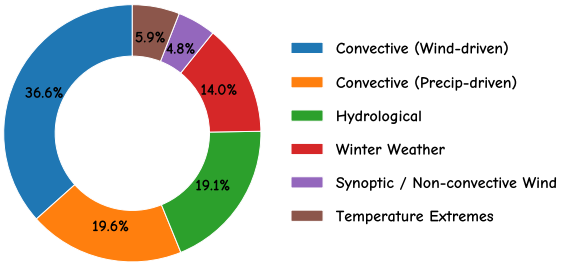

Figure 2: Event Type Distribution in WeatherBench-R.

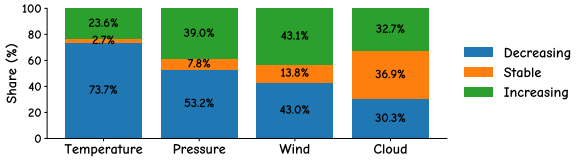

Figure 3: ERA5 Trend Distribution in WeatherBench-R.

by physical constraints while operating under incomplete and ambiguous evidence. Textual summaries derived from reanalysis data compress high-dimensional spatiotemporal fields into aggregate trends, forcing models to assess physical plausibility without access to raw grids. Moreover, real-world descriptions often omit decisive variables, leaving claims genuinely underdetermined rather than strictly supported or contradicted. Existing benchmarks rarely capture this regime, implicitly encouraging overconfident predictions. Finally, multiple atmospheric processes can produce similar aggregate signatures, making uncertainty expression a core component of scientifically valid reasoning. These characteristics suggest that text-only meteorological reasoning cannot be evaluated by deterministic accuracy alone. A suitable benchmark must preserve physically meaningful trends, explicitly model underdetermination, and assess whether models express calibrated uncertainty under incomplete or counterfactual evidence.

To meet these requirements, we introduce **WeatherBench-R**, a text-only benchmark for meteorological reasoning built from ERA5-derived summaries aligned with real-world severe weather events. As depicted in Figure 1, we provides three examples of each tasks, decomposing reasoning into three complementary competencies: recognizing physically plausible scenarios from aggregate trends, verifying claim consistency under partial evidence, and reasoning under counterfactual perturbations that probe uncertainty awareness. We summarize our key contributions as follows:

1. **A large-scale text-only dataset for meteorological reasoning,** constructed from ERA5 reanalysis and NOAA storm events, enabling controlled evaluation under physically grounded yet incomplete evidence.

2. **The first comprehensive benchmark for text-only weather reasoning,** featuring three tasks that jointly assess plausibility recognition, consistency verification, and counterfactual uncertainty reasoning.

3. **A systematic evaluation of LLMs under underspecified conditions,** revealing distinct failure modes in selection bias and uncertainty miscalibration beyond aggregate accuracy.

2 Dataset Construction

WeatherBench-R is constructed from **13,116** event-centered meteorological windows, each corresponding to a real-world NOAA severe weather record (distributed as Figure 2) aligned with a 12-hour ERA5 observation window (distributed as Figure 3). Each window serves as the basis for generating one structured natural-language summary, which is subsequently reused across all three reasoning tasks. This section outlines the procedure used to construct these event-centered windows and their corresponding natural-language summaries.

2.1 Data collection

We obtained meteorological data from the ERA5 hourly reanalysis dataset (Hersbach et al., 2020), produced by the Copernicus Climate Change Service (C3S) and the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 provides hourly estimates of a wide range of atmospheric and oceanic variables on a consistent global grid with approximately $0.25°$ spatial resolution. It covers the period from 1940 to the present, making it a widely adopted dataset in climate and weather research. For this study, we selected six meteorological variables that are commonly used in weather analysis: "2m temperature", "surface pressure", "10m u-component of wind", "10m v-component of wind", "total precipitation", and "total cloud cover". These variables capture key conditions relevant to severe weather characterization.

Our spatial domain was restricted to the contiguous United States, and the temporal range spanned from January 2016 through December 2021 to align with the availability of both ERA5 reanalysis and observational storm records. To map ERA5's latitude–longitude grid to county-level locations within the United States, we leveraged publicly available parquet-formatted datasets that provide county identifiers corresponding to ERA5 grid cells (Fu et al., 2025). Finally, we incorporated records of severe weather events from the NOAA Storm Events Database, which compiles county-level reports of storms and other significant weather phenomena across the United States.

To associate each severe weather record from the NOAA Storm Events Database with corresponding meteorological data, we standardized all event timestamps to Coordinated Universal Time (UTC), the temporal reference system used by ERA5. And we further extracted a 12-hour temporal window of ERA5 observations centered on the event time. Specifically, for an event occurring at time t, we collected ERA5 variables spanning the interval $[t-6, t+6]$ hours. This window captures trend of atmospheric conditions associated with the event, enabling LLMs to reason not only about the state at the event time but also its temporal dynamics.

2.2 Data resampling

To correct for the strong geographical imbalance in the raw storm-event distribution, we apply a three-stage resampling process designed to achieve a more even spatial representation across the United States. First, we assign each U.S. state a target sampling quota proportional to its share of nationwide events, while enforcing lower and upper bounds to prevent either sparsely populated or highly active states from dominating the benchmark. Second, within each state, we distribute the quota across counties according to their relative event frequency, again applying a small upper cap to avoid counties with extremely dense reporting histories from overwhelming the dataset. Finally, for each county, we randomly select the required number of events.

Specifically, let $\mathcal{D}$ denote the full set of storm events with $|\mathcal{D}| = T$, and let each event $i \in \mathcal{D}$ be associated with a state label $s(i)$ and county label $c(i)$. For each state s, we define the total number of events $n_s = |\{i \in \mathcal{D} : s(i) = s\}|$. For each county c within state s, we define $m_{s,c} = |\{i \in \mathcal{D} : s(i) = s, c(i) = c\}|$. In the state-level, we allocate a target quota proportional to its event share as follows: $q_s = \min(\max(\frac{n_s}{\sum_{s'} n_{s'}} T, 300), 600)$. And in the county-level, we allocate events by $s_{s,c} = \min(\lfloor q_s \frac{m_{s,c}}{\sum_{c'} m_{s,c'}} \rfloor, 150, m_{s,c})$. Thus, for each county (s, c), we draw $s_{s,c}$ samples uniformly without replacement, and construct the union of all county selections by

$$\mathcal{D}^* = \bigcup_s \bigcup_c \text{Sample}(\mathcal{D}_{s,c}, s_{s,c}).$$

2.3 Text Summary Generation

For each climate event, we derive a textual description from a 12-hour window of ERA5 observations centered on the event time. Rather than providing the raw hourly values directly, which can be numerically dense and difficult for language models to interpret, we convert each temporal sequence into a concise natural-language summary.

Formally, for each meteorological variable v noted above, let X_v and Y_v denote the values at the

start and end of the temporal window, respectively. We compute the net change $\Delta_v = X_v - Y_v$ and categorize the temporal trend based on predefined thresholds θ_v:

$$T_v = \begin{cases} \text{increasing}, & \Delta_v > \theta_v, \\ \text{decreasing}, & \Delta_v < -\theta_v, \\ \text{stable}, & |\Delta_v| \leq \theta_v. \end{cases}$$

These trend labels, along with the initial and final values (X_v, Y_v), are then filled into a structured template to form a weak-supervised text summary. This approach preserves essential meteorological signal while reducing numerical complexity, enabling LLMs to focus on reasoning over atmospheric behavior rather than parsing raw spatiotemporal data.

3 Task Design

WeatherBench-R introduces a three-task evaluation suite designed to probe complementary aspects of text-only meteorological reasoning. All tasks operate on the same ERA5-derived textual summaries, ensuring a consistent input representation while isolating different reasoning skills. Each task is formulated with explicit output spaces and evaluation metrics to discourage overconfident hallucination and reward evidence-grounded inference.

3.1 Task 1: Physical Plausibility Recognition

Input. Each instance consists of a structured ERA5-derived textual summary describing temperature, pressure, wind, cloud cover, and precipitation trends over a 12-hour window.

Output Format. The model is presented with four candidate weather scenarios (A–D), sampled from a controlled hypothesis set (e.g., convection, cold frontal passage, heat-wave, fog, windstorm). The task is multi-label: models must select all plausible options. If none are supported, the correct output is the special token `[NONE]`.

Definition of Plausibility. A scenario is considered plausible if it does not violate established meteorological relationships. For example, cooling with rising pressure may support cold-frontal activity, whereas strong convection without moisture or wind signals must be rejected.

Evaluation Metrics. Let Y be the ground-truth option set and $\hat{Y}$ the predicted set. We use the following three metrics to evaluate the correctness:

- Exact Match (EM): $\mathbf{1}[\hat{Y} = Y]$
- Set Macro-F1: precision/recall
- Over-selection rate: $\mathbf{1}[\hat{Y} \supset Y]$
- Under-selection rate: $\mathbf{1}[\hat{Y} \subset Y]$

3.2 Task 2: Consistency Verification

Input. Each example provides (i) an ERA5-based textual summary, and (ii) a single declarative claim describing atmospheric state or interpretation.

Output Categories. The model must classify the claim into one of three mutually exclusive labels: *Consistent*: the claim follows directly from the summary, with its numerical descriptors and directional trends aligned. *Inconsistent*: the claim conflicts with the summary in value, direction, or implied meteorological interpretation. *Underdetermined*: the summary does not contain sufficient information to confirm or refute the claim.

Evaluation Metrics. We report *Accuracy*, *Macro-F1*, and *Confusion matrix* to analyze class-wise error patterns.

3.3 Task 3: Counterfactual Evidence

Input. Each instance provides (i) the original NOAA event label and (ii) a counterfactually edited ERA5 text summary in which exactly one meteorological variable has been modified.

Output Format. The model must return a JSON object containing three fields, including `era5_sanity`, `supports_label`, and `rationale`. These outputs correspond to identifying the edited field, evaluating physical consistency, determining whether the altered state still supports the NOAA label, and providing a short explanation.

Evaluation Metrics. Evaluation is decomposed into three components: (1) *Sanity Accuracy.* Correctness of the `era5_sanity` judgment against rule-based physical constraints. (2) *Label Support Decision.* Accuracy and Macro-F1 over the 3-way `supports_label` classification. (3) *Rationale Quality.* Rationales are scored on a 1–5 scale using an LLM-based judge calibrated with human expert annotations. We report: *Mean rationale score* and *pass@k*, the fraction of explanations with score $\geq k$ (default $k = 4$). To improve reliability, rationales are evaluated twice and Cohen's κ is reported on a sampled subset (see Appendix).

LLM Judge Calibration. To align the LLM judge with expert judgment, we provide the judge with a small set of few-shot examples annotated by a human meteorological expert. Each example consists of a model-generated rationale paired with

a human-assigned score in the 1–5 scale. These examples are included in the judge prompt and remain fixed across all evaluations.

4 Experimental Setup

4.1 Baseline Models

We evaluate WeatherBench-R on a diverse set of state-of-the-art LLMs to cover (i) closed-source general-purpose models, (ii) open-source general-purpose models, and (iii) lightweight baselines.

Closed-source LLMs. We include representative models from the GPT family (GPT-5-mini-medium) (OpenAI, 2025) and the Gemini family (Gemini-3-flash-preview) (Team et al., 2023). For brevity, we refer to these models as GPT-5 and Gemini-3, respectively.

Open-source LLMs. We evaluate models from DeepSeek and multiple scales of Qwen3, including Qwen3-14B, Qwen3-32B, and Qwen3-235B (Yang et al., 2025), and DeepSeek-v3 (Liu et al., 2024).

Non-LLM baselines. To assess potential template leakage or shallow statistical correlations, we include Majority and Random baselines for each task.

Model access and versions. For proprietary models, we report the model identifier, release version, and query date. For open-source models, we document checkpoints, decoding parameters, and hardware configurations.

4.2 Prompting and Inference Protocol

For fair comparison, we adopt a prompting protocol and formatting constraints across tasks.

Decoding. We use greedy decoding with temperature set to 0 to minimize stochastic variance.

Context length and truncation. All instances are designed to fit within model context limits. When stricter limits apply, we prioritize retaining (i) the ERA5 summary, (ii) the claim or event label, and (iii) the answer format specification; exemplars are truncated first. We set the maximum token budget to 2048 for Task 1 and Task 2, and 4096 for Task 3. For reasoning-oriented models that exceed token limits and produce incomplete outputs, only fully valid responses are included in the evaluation.

4.3 Task Generation

Task 1 and Task 2 are automatically generated from the extracted event-centered ERA5 summaries using deterministic generation scripts, each yielding 12,362 task instances that cover 94.25% of all extracted events. For Task 1, we report results on all 12,362 generated instances, enabling comprehensive evaluation of multi-label plausibility reasoning across the dataset. For Task 2, although the full set of 12,362 instances is generated, we report results on 6,009 instances to balance evaluation coverage with computational cost, while preserving the overall distribution of claim types. Task 3 is constructed for a subset of events for which valid single-variable counterfactual perturbations can be defined without introducing trivial inconsistencies, resulting in 7,638 task instances (58.23% of extracted events). Due to the higher computational cost associated with structured outputs and explanation evaluation, we sample and report results on 1,222 Task 3 instances in this paper.

4.4 Evaluation Questions

We seek to address the following research questions to evaluate the LLM performance on our benchmark: **RQ1:** How well do LLMs perform on text-only meteorological reasoning tasks? **RQ2:** What types of reasoning errors do LLMs make under partial and underspecified meteorological information? **RQ3:** Can LLMs appropriately express uncertainty and avoid overconfidence when evidence is counterfactually perturbed?

5 Evaluation Results

5.1 Evaluation of LLM Performance

To address **RQ1**, we evaluate the overall performance of LLMs on WeatherBench-R across all three tasks. Table 1 summarizes the main quantitative results, demonstrating that text-only meteorological reasoning remains challenging for current LLMs, with no single model achieving consistently strong performance across all tasks and metrics. On Task 1, performance varies substantially across models. Qwen3-235B achieves the best Exact Match and Macro-F1, suggesting stronger ability to identify the full plausible option set. Notably, several LLMs obtain lower Exact Match than the Majority baseline despite comparable Macro-F1, indicating that they often capture partial plausibility signals but fail to precisely calibrate the complete multi-label set. On Task 2, Qwen3-32B attains the highest Accuracy and Macro-F1, outperforming both larger and proprietary models. The gap between Accuracy and Macro-F1 for GPT-5, Gemini-3, and Qwen3-235B further suggests non-uniform behavior across the three classes, consistent with the challenge of correctly handling *Underdeter-*

Table 1: Main results on WeatherBench-R. Best results are bolded.

Model	Task 1		Task 2		Task 3	
	EM ↑	Macro-F1 ↑	Acc ↑	Macro-F1 ↑	Sanity Acc ↑	Support Acc ↑
GPT-5	31.02	47.59	67.77	57.09	**75.78**	**72.01**
Gemini-3	36.69	53.58	59.39	47.70	74.22	69.72
Qwen3-14B	25.58	42.14	75.49	69.40	50.04	46.52
Qwen3-32B	15.51	54.92	**87.15**	**80.00**	53.24	50.45
Qwen3-235B	**49.49**	**57.10**	63.61	56.88	65.23	56.38
DeepSeek-v3	15.11	42.08	72.19	67.53	36.82	29.13
Majority	40.35	46.89	59.51	24.87	52.50	50.80
Random	6.000	32.23	33.13	28.99	32.80	37.20

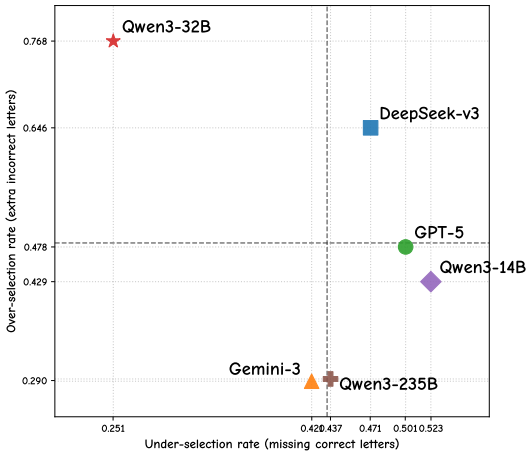

Figure 4: Trade-off between over-selection and under-selection errors across LLMs. The bold dashed line indicates the mean value to divide quadrants.

mined claims. Across models, Task 2 scores are generally higher than Task 1 Exact Match, suggesting that ternary consistency judgments in climate reasoning are easier for LLMs than multi-label plausibility reasoning. On Task 3, performance drops substantially for open-source models, with smaller models showing the most pronounced degradation; DeepSeek-v3 is close to the Random baseline. In contrast, GPT-5 and Gemini-3 achieve the strongest results on both Sanity and Label Support, indicating that counterfactual reasoning under perturbed evidence is the most demanding setting in WeatherBench-R. Within the Qwen3 family, performance generally improves with model scale on Task 3. Overall, these results show that current LLMs exhibit fragmented strengths across different reasoning skills, and that improvements in model scale or general capability do not uniformly translate into robust text-only meteorological reasoning.

5.2 Evaluation of Reasoning Errors

To address **RQ2**, we analyze systematic reasoning errors using selection behavior in Task 1 and confusion patterns in Task 2. Figure 4 illustrates the trade-off between over-selection and under-selection rates across models. A higher over-selection rate indicates a tendency to include unsupported scenarios, while a higher under-selection rate reflects conservative behavior that omits potentially valid options. We observe that Qwen3-32B exhibits a high over-selection rate coupled with a low under-selection rate, suggesting an overconfident strategy that frequently introduces extra incorrect options. DeepSeek-v3 shows high values on both dimensions, indicating unstable selection behavior that combines over-generation with inconsistent omission. These patterns help explain the relatively poor Task 1 Exact Match performance of both models in Table 1. In contrast, Gemini-3 and Qwen3-235B maintain comparatively low rates on both axes, reflecting more balanced plausibility calibration, while GPT-5 and Qwen3-14B occupy intermediate positions. To further examine error sources under partial evidence, we analyze the Task 2 confusion matrix shown in Figure 5. Distinct error profiles emerge between closed-source and open-source models. GPT-5 and Gemini-3 frequently misclassify *Underdetermined* claims as *Inconsistent*, indicating a tendency to interpret missing evidence as contradiction. GPT-5, in particular, shows weaker performance on correctly identifying *Consistent* claims. Conversely, DeepSeek-v3 achieves high accuracy on *Consistent* and *Underdetermined* cases but degrades substantially on *Inconsistent* claims, often predicting them as supported. Taken together, these results reveal two complementary failure modes in text-only meteorological

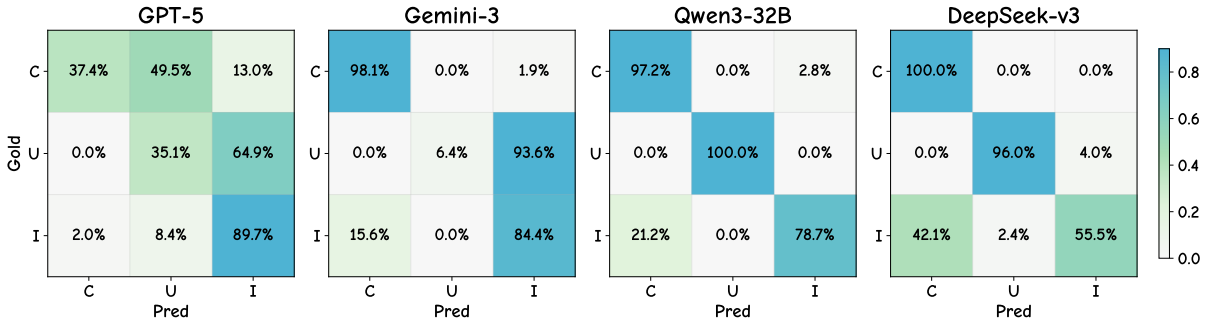

Figure 5: Task 2 confusion matrix. **C**, **U**, **I** stand for *Consistent*, *Underdetermined* and *Inconsistent* respectively.

Table 2: Rationale quality on Task 3 evaluated by an LLM judge calibrated with expert annotations.

Model	Score	Pass@4	Pass@5
GPT-5	**3.68**	51.84	**41.61**
Gemini-3	3.65	**55.82**	35.49
Qwen3-14B	2.57	18.06	8.370
Qwen3-32B	2.89	26.13	13.93
Qwen3-235B	3.09	34.77	21.19
DeepSeek-v3	2.16	4.260	1.640

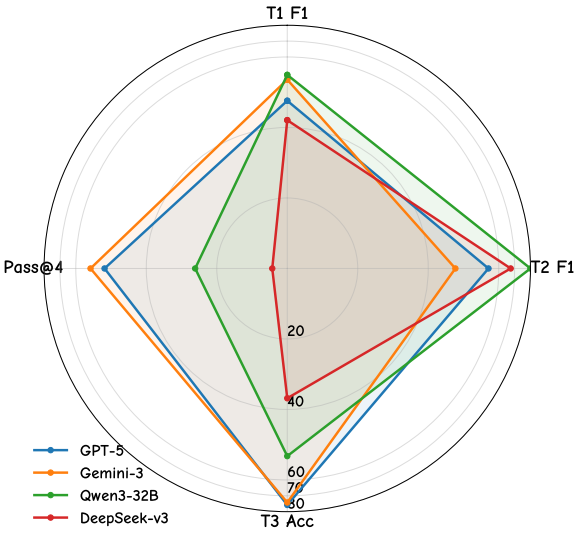

Figure 6: Radar profiles across reasoning dimensions: Task 1 Macro-F1 (T1 F1), Task 2 Macro-F1 (T2 F1), Task 3 Sanity Accuracy (T3 Acc), and rationale Pass@4.

reasoning. In multi-label plausibility recognition (Task 1), some models tend to over-generate plausible scenarios beyond the evidence, while others adopt overly cautious strategies. In ternary consistency verification (Task 2), open-source models primarily err by accepting contradictory claims, whereas closed-source models more often collapse underdetermination into inconsistency.

5.3 Evaluation of Explanation Quality

To address **RQ3**, we examine explanation quality in Task 3, which requires models to reason under counterfactual perturbations and justify their judgments with natural-language rationales. Table 2 reports rationale scores evaluated by an LLM-based judge calibrated with expert annotations. Overall, GPT-5 and Gemini-3 achieve the strongest performance, indicating superior ability to articulate physically grounded explanations in counterfactual settings. GPT-5 attains the highest mean rationale score and Pass@5 rate, suggesting that it more frequently produces detailed and well-justified explanations that fully satisfy expert-aligned criteria. Gemini-3, in contrast, achieves the highest Pass@4 score, reflecting consistent generation of generally sound explanations, albeit with slightly less depth or specificity. Within the Qwen3 family, rationale quality improves steadily with model scale, consistent with trends observed in Task 3 accuracy. However, even the largest Qwen3-235B model remains substantially behind GPT-5 and Gemini-3, indicating that increased scale alone does not close the gap in counterfactual explanation quality. DeepSeek-v3 performs worst across all rationale metrics, aligning with its near-random performance on Task 3 in Table 1 and suggesting limited sensitivity to counterfactual evidence. Figure 6 further illustrates cross-task capability profiles. While GPT-5 and Gemini-3 dominate on rationale quality and Task 3 Sanity Accuracy, both models underperform relative to others on Task 2, highlighting a trade-off between expressive explanation generation and conservative consistency judgment. Conversely, some open-source models achieve stronger consistency verification but fail to produce reliable explanations under perturbation. These findings indicate that explanation fluency and physical correctness are not tightly coupled: models capable of generating

persuasive rationales may still struggle with uncertainty calibration, while models with stronger verification performance may lack explanatory depth.

6 Related Work

Meteorological Benchmarks. Recent work has expanded meteorological machine learning beyond numerical forecasting toward benchmark-driven evaluation across multimodal and textual settings. On the multimodal side, CLLMate (Li et al., 2025b), MeteorPred (Tang et al., 2025), and ClimateBench-M (Fu et al., 2025) integrate gridded atmospheric variables with downstream tasks such as climate event forecasting, severe weather prediction, warning generation, and applied applications including medium-range forecasting and agricultural segmentation. ClimateIQA (Chen et al., 2025) further introduces a heatmap-based visual question answering benchmark that emphasizes anomaly detection and spatial reasoning over high-resolution atmospheric fields. Complementary to these efforts, several benchmarks focus on climate-related text understanding from linguistic and discourse perspectives. Climate-FEVER (Diggelmann et al., 2020) adapts fact-verification frameworks to climate change claims using Wikipedia-derived evidence, while Climabench (Spokoyny et al., 2023) aggregates multiple climate text tasks such as topic classification, stance detection, and document alignment across policy reports, corporate disclosures, and social media. Beyond classification and verification, NeuralNERE (Mishra and Mittal, 2021) targets climate knowledge extraction by constructing knowledge graphs from news articles. While these benchmarks advance multimodal learning and climate text processing, they primarily emphasize factual verification, discourse analysis, or visual–textual alignment, rather than physically grounded reasoning over natural-language meteorological event descriptions.

Scientific Reasoning with LLMs. Recent progress has demonstrated the potential of large language models to perform domain-specific reasoning across increasingly diverse fields. In biomedical research, LLMs have been applied to clinical decision support and mechanistic interpretation of electronic health records, showing strong ability to extract structured insights from medical narratives (Nazi and Peng, 2024; Yang et al., 2022). Similar developments have emerged in economics, finance, and legal analysis, where LLMs assist in document reasoning, causal inference, and regulatory interpretation (Sun, 2023; Guo and Yang, 2024). Within atmospheric science, ATMOSSCI-BENCH (Li et al., 2025a) evaluates whether LLMs can answer domain questions and interpret physical processes grounded in atmospheric theory. In the social sciences, Political-LLM (Li et al., 2024) investigates how language models reason about political ideology, argumentation, and decision logic. Together, these efforts reflect a broader shift toward domain-targeted scientific reasoning benchmarks, yet current work does not address reasoning over natural-language meteorological event descriptions.

7 Conclusion

In this study, we introduce WeatherBench-R, a text-only benchmark for evaluating meteorological reasoning over U.S. weather events. By grounding structured natural-language summaries in ERA5 reanalysis and aligning them with real-world NOAA storm records, WeatherBench-R isolates reasoning under physical constraints, incomplete evidence, and inherent underdetermination. The benchmark decomposes this challenge into three complementary tasks, physical plausibility recognition, consistency verification, and counterfactual evidence reasoning, enabling diagnostic evaluation beyond aggregate accuracy. A systematic assessment of LLMs reveals fragmented strengths across tasks, sensitivity to selection bias, and pronounced degradation under counterfactual perturbations, highlighting gaps in uncertainty handlin. Rationale analysis further exposes a disconnect between explanation fluency and physically grounded reasoning. Together, these findings underscore that text-only meteorological reasoning remains a multi-dimensional challenge not resolved by model scale alone. WeatherBench-R provides a rigorous, reproducible framework for probing these capabilities and offers practical insights for risk-aware deployment of language models in weather-related text understanding, while complementing existing multimodal benchmarks.

8 Limitations

WeatherBench-R is restricted to U.S. weather events, as it is constructed from NOAA Storm Events aligned with ERA5 reanalysis over the contiguous United States. While this design ensures consistent event definitions and reliable annotations, it may limit the direct applicability of the benchmark to regions with different climatological regimes. In addition, the evaluation of explanation quality in Task 3 relies on an LLM-based judge calibrated with a small set of expert-annotated examples. Although this approach enables scalable and consistent assessment of rationales, it may still reflect biases inherent to the judge model and does not fully substitute for large-scale human expert evaluation. Finally, the scope of empirical evaluation is constrained by practical resource and funding considerations, resulting in a representative but limited set of evaluated models and evaluation configurations. Expanding model coverage and human-in-the-loop assessment remains an important direction for future work.

References

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and 1 others. 2025. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2022. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. 2025. Climateiqa: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5322–5333.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Dongqi Fu, Yada Zhu, Zhining Liu, Lecheng Zheng, Xiao Lin, Zihao Li, Liri Fang, Katherine Tieu, Onkar Bhardwaj, Kommy Weldemariam, and 1 others. 2025. Climatebench-m: A multi-modal climate data benchmark with a simple generative method. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6367–6371.

Yue Guo and Yi Yang. 2024. Econnli: evaluating large language models on economics reasoning. *arXiv preprint arXiv:2407.01212*.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, and 1 others. 2020. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, and 1 others. 2023. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421.

Chenyue Li, Wen Deng, Mengqian Lu, and Binhang Yuan. 2025a. Atmossci-bench: Evaluating the recent advance of large language model for atmospheric science. *arXiv preprint arXiv:2502.01159*.

Haobo Li, Zhaowei Wang, Jiachen Wang, YueYa Wang, Alexis Kai Hon Lau, and Huamin Qu. 2025b. Cllmate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17547–17573.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, and 1 others. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Rao Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. 2024. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *Advances*

in Neural Information Processing Systems, 37:68740–68771.

OpenAI. 2025. Introducing gpt-5. Accessed: 2026-01-04.

Joel Oskarsson, Tomas Landelius, Marc Deisenroth, and Fredrik Lindsten. 2024. Probabilistic weather forecasting with hierarchical graph neural networks. *Advances in Neural Information Processing Systems*, 37:41577–41648.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, and 1 others. 2025. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90.

Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards answering climate questionnaires from unstructured climate reports. *arXiv preprint arXiv:2301.04253*.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Shuo Tang, Jian Xu, Jiadong Zhang, Yi Chen, Qizhao Jin, Lingdong Shen, Chenglin Liu, and Shiming Xiang. 2025. Meteorpred: A meteorological multimodal large model and dataset for severe weather event prediction. *arXiv preprint arXiv:2508.06859*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, and 1 others. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

A More Statistics of WeatherBench-R

This section reports additional dataset and task statistics to complement the main paper.

A.1 Scale and Coverage

Spatiotemporal coverage. Event start times span 2016–2022 (UTC) with strong seasonality (summer peak), and geographical coverage across 39 U.S. states. Each event is aligned to a fixed 12-hour ERA5 window (t−6h to t+6h), yielding 12–13 hourly rows per event in the vast majority of cases.

A.2 Temporal Distribution

The temporal distribution of the dataset is illustrated in Figure 7, with a detailed breakdown of event counts per year and month provided in the supplementary analysis. The dataset covers the period from 2016 to 2022, showing a notable concentration of events during the summer months.

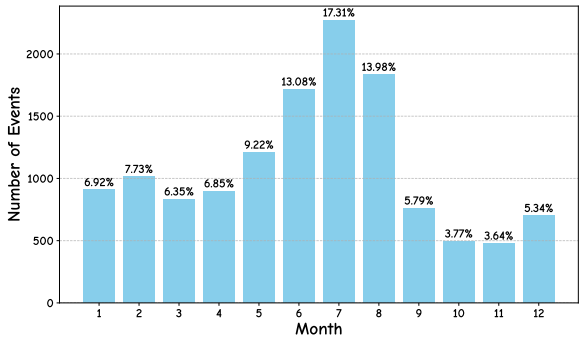

Figure 7: Seasonality of events by month, with each bar representing the number of events occurring in that month. The percentage annotations on top of each bar indicate the share of total events for each month.

A.3 Geographic Distribution

The geographic distribution of events across the United States is summarized in Table 4 and visualized in Figure 8. The dataset spans 39 states and 259 state county regions, reflecting a broad spatial coverage after the resampling process.

A.4 ERA5 Window Properties

Value ranges. Across all per-hour rows, the dataset includes both extreme cold and heat, heavy precipitation, and strong winds (e.g., temperature down to $-34.63\ ^{\circ}C$ and gusts up to 41.99 m/s).

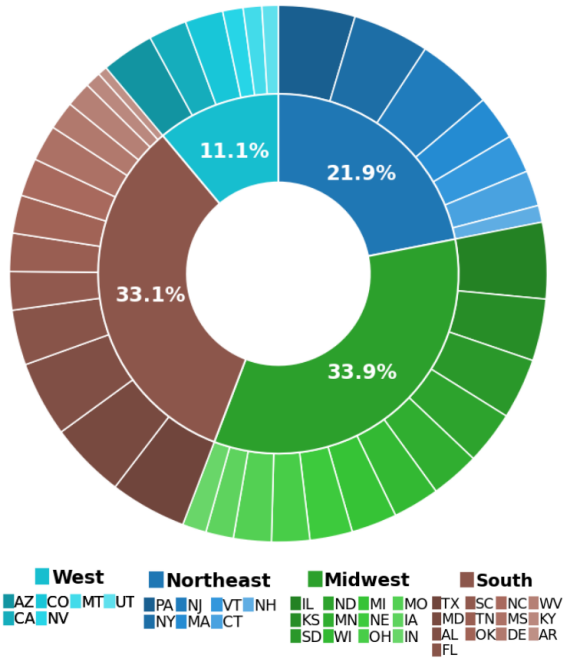

Figure 8: Geographic distribution of events by U.S. region and state. The inner ring shows regional totals (Northeast, Midwest, South, West), while the outer ring displays individual state contributions. Percentage annotations indicate each region's of total events.

A.5 Task3 Counterfactual Edit Distribution

Task3 counterfactuals modify a single ERA5 variable. The edited-field distribution (full Task3) is summarized in Table 7.

A.6 Quality Views and Filtering

Task coverage. Task1 and Task2 cover 12,362 unique events (94.25% of extracted events), while Task3 covers 7,638 events (58.23%) due to additional support filtering.

B Method Details

B.1 Data Alignment and Window Extraction

Sources. We align official NOAA Storm Events (event type + metadata) with ERA5 reanalysis (hourly atmospheric variables).

Time normalization. NOAA timestamps are recorded in local time with a timezone field. We convert to UTC using a fixed offset mapping.

Fixed information budget. For each event with UTC start time t, we extract a 12-hour ERA5 window $[t - 6\text{h},\ t + 6\text{h}]$. This design enforces a consistent evidence budget across models and across event types.

Artifact	#Events	Notes
Raw extracted events	13,116	2016–2022; 39 labels; 39 states; 259 state_county regions
Task1/Task2 (full)	12,362	After low-signal convective filtering
Task2 (sampled)	2,003	6,009 claims (3 per event), 16.20% retention
Task3 (full)	7,638	After rule-based label-support check
Task3 (sampled)	1,223	16.01% retention

Table 3: WeatherBench-R scale and coverage.

State	Events	Share
Pennsylvania	602	4.59%
Texas	601	4.58%
Illinois	600	4.57%
New York	600	4.57%
New Jersey	599	4.57%
Maryland	598	4.56%
Alabama	592	4.51%
Kansas	487	3.71%
South Dakota	477	3.64%
Florida	434	3.31%

Table 4: Top states by event count after resampling.

NOAA EVENT_TYPE	#Events	Share
Thunderstorm Wind	4,248	32.39%
Hail	1,782	13.59%
Flash Flood	1,254	9.56%
Flood	888	6.77%
Winter Weather	885	6.75%
Winter Storm	530	4.04%
Tornado	376	2.87%
High Wind	376	2.87%
Heavy Rain	300	2.29%
Lightning	270	2.06%

Table 5: Top NOAA event labels in the extracted dataset.

B.2 Balanced Sampling for the Released Subset

To obtain a geographically balanced subset, we apply a three-stage quota-and-sampling procedure. Let N_s be the number of events in state s, and N be the total events.

Stage 1: state quota.

$$q_s \leftarrow \mathrm{clip}\left(15000 \cdot \frac{N_s}{N},\, 300,\, 600\right)$$

Stage 2: county quota within state. For a county c in state s with $N_{s,c}$ events:

$$r_{s,c} \leftarrow q_s \cdot \frac{N_{s,c}}{\sum_{c'} N_{s,c'}},$$

$$s_{s,c} \leftarrow \min(\mathrm{round}(r_{s,c}),\, 150,\, N_{s,c})$$

Stage 3: random sampling. Sample $s_{s,c}$ events uniformly within each county.

B.3 Low-Signal Convective Filtering

To avoid degenerate cases where a "convective" NOAA label has no discernible atmospheric signal in the aligned window, we apply a conservative pre-filter for a small set of convective event types. The convective set is: {`Thunderstorm Wind`, `Hail`, `Tornado`, `Flash Flood`, `Heavy Rain`, `Strong Wind`, `Marine Thunderstorm Wind`}. For a convective-labeled event, if none of the following signals are present, the event is skipped before Task1/2 generation, and consequently absent from Task3 as well.

B.4 Task3 Label-Support Filtering

Task3 further filters events using a rule-based "label support" check. Given the summary features, we infer a set of plausible hazard tags and compare them against a coarse mapping from NOAA labels to hazard tags. If the window is internally inconsistent (sanity failure), we force the support decision to `uncertain`.

B.5 NOAA Event Type Vocabulary

We restrict events to a canonical label set, reproduced below for completeness:

B.6 Natural-Language Summaries from ERA5 Windows

We compress each ERA5 window into a compact text summary to avoid exposing raw high-dimensional tables.

Summary schema. Each summary contains: (i) temperature start/end/min/max and trend; (ii) surface pressure start/end and trend; (iii) 10m wind speed start/end and trend (computed from east-/north components); (iv) cloud cover start/end and trend; (v) total precipitation (window sum) and peak hourly precipitation; (vi) dewpoint depression mean, peak relative humidity, and peak gust. Trend labels are computed deterministically: `stable` if $|x_{\mathrm{end}} - x_{\mathrm{start}}| < \tau$; `increasing` if $x_{\mathrm{end}} - x_{\mathrm{start}} \geq \tau$; otherwise `decreasing`, with variable-specific thresholds τ (Table 14).

Trend (start→end)	Increasing	Decreasing	Stable
Temperature	3,097 (23.61%)	9,660 (73.65%)	359 (2.74%)
Pressure	5,118 (39.02%)	6,972 (53.16%)	1,026 (7.82%)
Wind speed	5,660 (43.15%)	5,639 (42.99%)	1,817 (13.85%)
Cloud cover	4,294 (32.74%)	3,978 (30.33%)	4,844 (36.93%)

Table 6: Per-event trend distributions computed from start/end values of the 12-hour window.

Edited field	#Items	Share
total_precipitation	2,607	34.13%
precip_peak	1,551	20.31%
temperature	1,299	17.01%
wind_speed	1,100	14.40%
surface_pressure	999	13.08%
cloud_cover	82	1.07%

Table 7: Distribution of the modified variable in Task3 counterfactuals.

Heuristic support status	Events	Share
supported	7,837	59.75%
not_supported	3,540	26.99%
uncertain	1,739	13.26%

Table 8: Heuristic check of whether ERA5-window signals plausibly support the NOAA label.

Top reason	Events	Share
supported	7,837	59.75%
signals_point_elsewhere	3,540	26.99%
no_mapping	722	5.50%
insufficient_signals	539	4.11%
no_signal	443	3.38%
era5_inconsistent	35	0.27%

Table 9: Reasons produced by the heuristic label-support checker.

EVENT_TYPE (skipped)	Events	Share
Thunderstorm Wind	360	47.75%
Hail	204	27.06%
Flash Flood	145	19.23%
Heavy Rain	29	3.85%
Tornado	15	1.99%
Strong Wind	1	0.13%

Table 10: Labels of 754 events removed by the low-signal convective filter before Task1/2 generation.

Derived quantities. Wind speed is computed from 10m vector components as $\sqrt{u^2 + v^2}$. Total precipitation is converted to mm and summed over the window; the "peak hourly rate" is the max per-step precipitation. Relative humidity is computed via a standard August–Roche–Magnus saturation-vapor-pressure approximation using temperature and dewpoint (details in code).

B.7 Task Construction

B.7.1 Task1: Physical Plausibility (Multi-Select)

Given an ERA5 summary, the model must select *all* hypotheses that are physically consistent with the summary. Options are generated from a small set of scenario templates (e.g., deep convection, cold front, fog, windstorm, insufficient signal) using thresholded feature flags. We enforce at least one correct and one incorrect option among the four choices.

B.7.2 Task2: Event–Summary Consistency Verification (Tri-Class)

Given an ERA5 summary and a claim, the model predicts one label: `Consistent`, `Inconsistent`, or `Underdetermined`. Each event yields three claims: a pressure–wind causal claim, a calm/dry ridge claim, and a deliberately underdetermined claim.

B.7.3 Task3: Counterfactual Evidence–Label Rule Check (Structured JSON)

Task3 edits exactly one summary variable to create a counterfactual and asks the model to: (1) judge whether the edited summary is internally sane; (2) judge whether the (counterfactual) evidence supports the NOAA label; and (3) provide a short rationale grounded in both metrics and rule hints.

Counterfactual construction. We map NOAA labels to coarse hazard tags and sample an editable variable conditioned on the tag. Edits are constrained to plausible ranges and enforce a minimum magnitude to avoid trivial detection.

B.7.4 End-to-End Pipeline (Pseudo-code)

C Experimental Details

C.1 Inference Protocol

We evaluate models via an OpenAI-compatible chat-completions interface. Key settings: temperature $= 0$, `max_output_tokens`=4096, timeout 60s, and up to 6 retries with exponential backoff.

Format guard. To make evaluation reliable, we inject a task-specific system instruction that enforces strict output formats: Task1 returns only a comma-separated set of letters or `NONE`; Task2 re-

event_id	EVENT_TYPE	state_county	precip_total (mm)	precip_peak (mm)
914434	Tornado	texas_cameron	139.48	20.19
664158	Flash Flood	south_carolina_beaufort	128.46	18.68
663379	Flash Flood	north_carolina_wayne	118.89	20.20
920428	Flash Flood	florida_bay	113.33	23.99
721630	Flood	florida_lee	112.11	21.34

Table 11: Top event windows by total precipitation over the 12-hour window.

CZ_TIMEZONE	Offset (hours)
AST-4	-4
EDT-4	-4
EST-5	-5
CDT-5	-5
CST-6	-6
PDT-7	-7
MST-7	-7
PST-8	-8
AKST-9	-9
HST-10	-10
SST-11	-11
GST10	10

Table 12: Timezone offsets used to convert NOAA local timestamps to UTC.

turns only one of three labels; Task3 returns only a JSON object with a fixed schema.

System instructions (verbatim).

C.2 Scoring Rules

Task1 (multi-select). Let gold set Y_i and prediction $\hat{Y}_i$ be subsets of {A,B,C,D}. We compute micro precision/recall/F1 by aggregating TP/FP/FN across all instances: $P = \frac{\sum_i |Y_i \cap \hat{Y}_i|}{\sum_i |\hat{Y}_i|}, R = \frac{\sum_i |Y_i \cap \hat{Y}_i|}{\sum_i |Y_i|}, F1 = \frac{2PR}{P+R}$. We additionally report exact set match rate and over-/under-selection diagnostics.

Task2 (tri-class). We report accuracy, confusion counts, and macro-F1 over the three labels.

Task3 (heuristic correctness). We parse the model JSON and compute: (i) ERA5 sanity accuracy against a rule-based sanity classifier (e.g., flagging `precip_peak` > `precip_total`); (ii) label-support accuracy against a rule-based mapping from NOAA label → coarse hazard tag and inferred plausible tags from the summary. We also measure rationale coverage (mentions variable + at least one indicator + at least one rule keyword).

Parsing and normalization details. Scoring is intentionally conservative to avoid accidental credit from verbose or malformed outputs:

Algorithm 1: WeatherBench-R task generation pipeline for each `state_county` region.

Input: `events.csv` and `era5.csv` for a `state_county` region
Output: Grouped JSONL files for Task 1–Task 3
foreach `state_county` *region* **do**
  $E \leftarrow$ LOADEVENTS(`events.csv`)
  $W \leftarrow$ LOADERA5(`era5.csv`)
  foreach *event* $e \in E$ **do**
    $X \leftarrow$ EXTRACTWINDOW(W, e.id)
     `// fixed 12-hour window; hourly records`
    $s \leftarrow$ SUMMARIZEERA5(X)
     `// structured text summary`
    if LOWSIGNALCONVECTIVE*(e.*type*, X)* **then**
      continue
       `// skip degenerate convective cases`
    EMITTASK1(s)
       `// multi-select plausibility`
    EMITTASK2(s)
       `// three consistency claims`
    if LABELSUPPORTED*(s, e.*type*)* ∨ `allow_unsupported_task3` **then**
      $(s', f) \leftarrow$ COUNTERFACTUALEDIT(s, e.type)
         `// edit one variable; record field` f
      EMITTASK3(e.type, s')
         `// counterfactual evidence check (JSON)`
EXPORTGROUPEDJSONLS()
  `// write per-task grouped outputs`

- **Task1**: letter parsing prefers a clean answer-only line (e.g., `A,C` or `NONE`) and falls back to an explicit `Answer: ...` prefix; it avoids extracting stray "A/B/C/D" from prose.

- **Task2**: label normalization accepts exact labels, single-letter shorthands (`C/I/U`), and weak lexical cue.

- **Task3**: JSON is parsed either directly or by extracting the outermost {...} span; `modified_variable` is canonicalized with an alias map.

Signal	Condition	Source features
Sustained wind	max wind ≥ 8.0 m/s	10m east/north components
Wind gust	gust max ≥ 12.0 m/s	10m gust (maximum)
Total precip	window sum ≥ 0.5	total precipitation
Peak precip rate	max rate ≥ 0.5	maximum precipitation rate
Pressure drop	drop ≥ 2.0 hPa	surface pressure start/end
Moisture (near sat.)	$\min(T - T_d) \leq 3.0$ degC	temperature + dewpoint
High cloud increase	Δ high cloud ≥ 0.3	high cloud cover start/end

Table 13: Signals used by the low-signal convective filter. The filter skips an event only when *all* signals fail.

Quantity	Trend threshold τ
Temperature (degC)	0.2
Surface pressure (hPa)	0.3
10m wind speed (m/s)	0.3
Cloud cover (0–1)	0.05

Table 14: Trend thresholds used in the ERA5 summarizer.

Heuristic sanity checks (Task3). The sanity classifier flags internal contradictions in the summary (e.g., peak precipitation exceeding total, heavy precipitation with near-clear skies), returning `sane` vs. `inconsistent` with reasons.

C.3 Reproducibility Commands

Dependencies. The dataset builder and analysis scripts require common Python scientific packages (e.g., `pandas`, `numpy`, `pyarrow`); model inference uses an OpenAI-compatible HTTP API (optionally the `openai` Python SDK).

C.4 LLM-as-Judge for Task3 Rationale Quality

For Task3, we optionally score rationale quality via an LLM judge using a 1–5 rubric. We run two independent judge passes and invoke an arbitration pass when they disagree, producing a final judge JSON per item.

D Data Consent

In this work, we utilize meteorological and severe weather event data from two publicly accessible scientific data sources:

NOAA Storm Events Database: The severe weather event records used in this study are obtained from the Storm Events Database maintained by the National Centers for Environmental Information (NCEI) under the U.S. National Oceanic and Atmospheric Administration (NOAA). This database provides chronological listings of various storm and weather phenomena across the United States, including tornadoes, thunderstorms, floods, hurricanes, and other significant events, and is freely accessible to the research community and public for scientific use.

ERA5 Reanalysis Data: The atmospheric reanalysis variables used for text-summary generation are derived from the ERA5 dataset produced by the Copernicus Climate Change Service (C3S) and distributed via the Climate Data Store (CDS). Access to ERA5 data requires acceptance of the Licence to Use Copernicus Products, which permits free and non-exclusive use of the data for research, publication, and redistribution, subject to the obligation to provide clear attribution to the Copernicus Climate Change Service. Any publication or public dissemination of derived data should include appropriate source acknowledgment, and explicitly note that neither the European Commission nor the European Centre for Medium-Range Weather Forecasts (ECMWF) is responsible for how the data are used.

By using these datasets in WeatherBench-R, we adhere to their respective terms and licensing frameworks and ensure proper citation and acknowledgment in all related outputs.

E Use Of Ai Assistants

During the preparation of this paper, we used LLM-based tools for grammar checking and language polishing throughout the paper.

F Instructions Given To Participants

We report the full text of instructions given to participants for writing rationales in Figure 9.

G Data Examples

The released JSONL tasks follow the exact prompt templates in Appendix B.7. Below we include template-faithful examples to illustrate required outputs. The released JSONLs store prompts as UTF-8 strings (including degree symbols in the ERA5 summary); if compiling with pdfLaTeX, ensure UTF-8 support (or use XeLaTeX).

Option tag	Condition for being marked correct
deep_convection	`wet` $\wedge$ (`pressure_fall` $\vee$ `windy`)
cold_front	`cooling` $\wedge$ (`pressure_rise` $\vee$ `windy`)
heat_wave	`very_warm`
fog	`calm` $\wedge$ `cloudy` $\wedge$ `humid_fog` $\wedge\neg$ `very_warm`
windstorm	`windy` $\vee$ `gusty_storm`
insufficient	`ambiguous`
spurious_heat	always false (adversarial distractor)

Table 15: Task1 option correctness conditions.

Claim family	Gold label rule
pressure–wind causal claim	`Consistent` iff pressure_drop $>$ 1.5 hPa *and* wind_change $>$ 0; else `Inconsistent`
calm/dry ridge claim	`Consistent` iff `calm` and not `wet` and $\lvert$temp_change$\rvert$ $<$ 2.0 degC and $\lvert$pressure_change$\rvert$ $<$ 1.0 hPa; else `Inconsistent`
underdetermination claim	always `Underdetermined`

Table 16: Task2 claim construction and gold labeling rules.

Hazard tag	NOAA labels mapped to this tag
deep_convection	Thunderstorm Wind; Marine Thunderstorm Wind; Tornado; Hail; Tropical Storm; Tropical Depression; Hurricane (Typhoon)
heavy_precip	Flash Flood; Flood; Heavy Rain
windstorm	Strong Wind; High Wind; Marine Strong Wind; Marine High Wind; Wind
heat_wave	Heat; Excessive Heat
drought	Drought
cold_front	Cold/Wind Chill; Extreme Cold/Wind Chill
fog	Dense Fog; Fog
freezing_fog	Freezing Fog
winter_precip	Winter Storm; Heavy Snow; Blizzard; Ice Storm; Sleet; Lake-Effect Snow; Winter Weather

Table 17: NOAA label $\rightarrow$ hazard tag mapping used in Task3.

Hazard tag	Editable variables (candidates)
deep_convection	total_precipitation; precip_peak; wind_speed; surface_pressure
heavy_precip	total_precipitation; precip_peak
windstorm	wind_speed
heat_wave	temperature; total_precipitation
cold_front	temperature
fog	wind_speed; cloud_cover
winter_precip	temperature; total_precipitation
drought	total_precipitation; cloud_cover; temperature
freezing_fog	temperature; wind_speed; cloud_cover

Table 18: Editable-variable candidates per hazard tag.

Variable	Edit operator (range)
temperature	add a constant Δ to {start,end,min,max}; $\lvert\Delta\rvert \geq 4$ degC; clamp temps into $[-35, 45]$ degC
surface_pressure	set end pressure to start$+\delta$; $\delta \in [1.5, 4.0]$ (rise) or $\delta \in [-4.0, -1.5]$ (fall) or $\delta \in [-0.2, 0.2]$ (stable); clamp to [950,1050] hPa
wind_speed	resample start/end: calm (start 0.3–2.5, end within $\pm$0.5, clamp to 0–5) or windy (start 11–16, end shift $[-2, +4]$, clamp to 5–25)
cloud_cover	resample start/end: clear (0.0–0.2) or cloudy (0.8–1.0) or moderate (0.3–0.7)
total_precipitation	set precip_total: dry (0.0–0.2), wet (10.0–25.0), or moderate (0.5–4.0) mm
precip_peak	set precip_peak: dry (0.0–0.05), wet (8.0–15.0), or moderate (0.1–1.0) mm

Table 19: Task3 counterfactual edit operators. Direction choices are tag-conditioned.

Flag	Definition (from summary features)
`windy`	max wind ≥ 10.0 m/s
`calm`	max wind ≤ 1.0 m/s
`wet`	precip_total ≥ 1.0 mm *or* precip_peak ≥ 0.25 mm
`heavy_precip`	precip_total ≥ 10.0 mm *or* precip_peak ≥ 7.6 mm
`dry`	precip_total ≤ 0.2 mm
`very_warm`	mean temperature ≥ 25.0 degC (computed as $(T_{\text{start}} + T_{\text{end}})/2$)
`below_freezing`	temp_min < 0.0 degC
`near_freezing`	temp_min ≤ 2.0 degC
`cloudy`	max(cloud_start, cloud_end) ≥ 0.8
`foggy_clouds`	max(cloud_start, cloud_end) ≥ 0.6
`clearing`	cloud_end$-$cloud_start ≤ -0.2
`humid_fog`	dewpoint depression mean < 2.0 degC *and* RH max $\geq 95\%$
`gusty_storm`	gust max ≥ 15.0 m/s
`ambiguous`	|temp_change| < 6.0 degC and |pressure_change| < 2.0 hPa and max_wind< 7.0 m/s and precip_total< 1.5 mm

Table 20: Core feature flags used by Task1/2/3 generation.

Instructions for Participants:

Participants are tasked with evaluating weather event summaries based on the given meteorological data. For each event, please read the summary and rationale and assign a score based on the justification provided. Below, you will find the instructions and example rationale cases that you need to use when generating your own rationale examples.

1. **Read the weather event summary** carefully. Pay attention to the key meteorological data, including temperature, surface pressure, wind speed, cloud cover, and precipitation.
2. **Evaluate the consistency of the event summary** with the event label. This will involve checking whether the conditions outlined in the summary meet the necessary thresholds for the given event type (e.g., 'heavy_precip' for flash flood).
3. **Generate a rationale** explaining why the event summary supports or does not support the given event label. Your rationale should consider:
   - Whether the meteorological conditions meet the threshold criteria for the event (e.g., sufficient precipitation for a flash flood).
   - Any counterfactual changes in the summary (e.g., how changes in surface pressure affect the event label).
4. **Assign a score from 1 to 5,** based on the quality of the rationale:
   - **5 points:** The rationale fully supports or refutes the event label with a clear, well-supported explanation. All conditions are thoroughly explained, and no additional clarification is needed.
   - **4 points:** The rationale is mostly correct and provides a valid explanation, but some minor detail or reasoning may be lacking or slightly unclear.
   - **3 points:** The rationale is partially correct, but some key conditions or important reasoning are missing or only partially explained. More explanation or evidente is needed to fully support the event label.
   - **2 points:** The rationale is weak and may have some major errors in reasoning or fails to address key aspects of the event summary. The explanation is incomplete or contradictory.
   - **1 point:** The rationale is either incorrect or unsupported, failing to justify the event label in a meaningful way.

Figure 9: Instructions for writing rationales.

[NOAA event labels (alphabetical)]

- Astronomical Low Tide
- Avalanche
- Blizzard
- Coastal Flood
- Cold/Wind Chill
- Debris Flow
- Dense Fog
- Dense Smoke
- Drought
- Dust Devil
- Dust Storm
- Excessive Heat
- Extreme Cold/Wind Chill
- Flash Flood
- Flood
- Freezing Fog
- Frost/Freeze
- Funnel Cloud
- Hail
- Heat
- Heavy Rain
- Heavy Snow
- High Surf
- High Wind
- Hurricane (Typhoon)
- Ice Storm
- Lake-Effect Snow
- Lakeshore Flood
- Lightning
- Marine Hail
- Marine High Wind
- Marine Strong Wind
- Marine Thunderstorm Wind
- Rip Current
- Seiche
- Sleet
- Storm Surge/Tide
- Strong Wind
- Thunderstorm Wind
- Tornado
- Tropical Depression
- Tropical Storm
- Tsunami
- Volcanic Ash
- Waterspout
- Wildfire
- Winter Storm
- Winter Weather

Task 1 Prompt Template

[System]
You are assessing which physical explanations agree with the ERA5 summary. The response must be true and accurate, and no additional content should be output.

[Task Description]
Your task is to select all hypotheses that are physically consistent with the provided ERA5 summary.

[Restriction]
Mark every hypothesis that is physically consistent. Respond with comma-separated letters (e.g., `"A,C"`), or respond with `"NONE"` if no options apply. Only include options that can be determined confidently from the ERA5 summary.

- If the ERA5 summary does not provide enough evidence to judge an option, do not select it.
- Do not add explanations, reasoning, or any text beyond the required output format.

[Response Examples]
`A,C`
`NONE`

[Context Input]
`<ERA5 summary>`

A. `<hypothesis>`
B. `<hypothesis>`
C. `<hypothesis>`
D. `<hypothesis>`

Task 2 Prompt Template

[System]
You are verifying whether a claim matches the ERA5 summary. The response must be true and accurate, and no additional content should be output.

[Task Description]
Given an ERA5 summary and a claim, determine whether the claim is supported by the summary.

[Restriction]
Choose exactly one label from: `Consistent`, `Inconsistent`, `Underdetermined`.
Output only the label (no explanation, no extra text).

- `Consistent`: The ERA5 summary provides clear evidence that the claim holds.
- `Inconsistent`: The ERA5 summary provides clear evidence that contradicts the claim.
- `Underdetermined`: The ERA5 summary does not contain sufficient information to decide confidently.

[Response Examples]
`Consistent`
`Inconsistent`
`Underdetermined`

[Context Input]
`<ERA5 summary>`

Claim: `<statement>`

Task 3 Prompt Template

[System]
You are evaluating a counterfactual ERA5 summary for a NOAA-labeled weather event. The response must be true and accurate, and no additional content should be output.

[Context Input]
NOAA event label: `<event_label>`
ERA5 summary (counterfactual):
`<edited_summary>`

[Task Description]
Exactly one ERA5 variable was modified to create this counterfactual summary. Use *only* the summary above to complete all four items below.

[Tasks]
Answer *all four* items:

1. **Identify**: `modified_variable` (variable name only; do not infer original values).
2. **ERA5 Sanity**: `era5_sanity` ∈ {`sane`, `inconsistent`, `uncertain`}.
3. **Support NOAA label**: `supports_label` ∈ {`supported`, `not_supported`, `uncertain`}.
   If `era5_sanity` is `inconsistent`, set `supports_label` to `uncertain`.
4. **Reasoning**: `rationale` (must mention the modified variable, cite at least one ERA5 metric, and cite at least one rule hint, if provided).

[Output Format]
Respond as a *single* JSON object with *exactly* the following keys:
`modified_variable`, `era5_sanity`, `supports_label`, `rationale`.
Do not include any additional text.

[JSON Schema Example]
```
{"modified_variable": "...", "era5_sanity": "...", "supports_label": "...", "rationale": "..."}
```

[Rule Hints] (optional)
`<optional_rule_hints>`

Example: Task 1 (Multi-select)

ERA5 summary:
Temperature began at 27.0 °C and ended at 24.5 °C (range 22.8–28.1 °C; trend decreasing).
Surface pressure changed from 1007.2 hPa to 1001.8 hPa (trend decreasing).
10 m wind speed shifted from 6.2 m/s to 11.4 m/s (trend increasing).
Cloud cover moved from 0.35 to 0.88 (trend increasing).
Total precipitation over the period was 18.40 mm, with a peak hourly rate of 8.20 mm.
Dewpoint depression averaged 1.4 °C. Relative humidity peaked at 98%. Peak wind gusts reached 18.3 m/s.

Question (Task1):
Mark every hypothesis that is physically consistent. Respond with comma-separated letters (e.g., `"A,C"`), or respond with `"NONE"` if no options apply.

Hypotheses:
A. Deep convection is plausible: pressure dropped, winds strengthened, and heavy precipitation occurred.
B. A stagnant dry ridge is confirmed because winds stayed calm and precipitation was minimal.
C. The signals are weak and conflicting, so no single hazard can be confirmed.
D. Gradient wind damage is expected because gusts were very strong despite no precipitation.

Gold answer: `A`

Example: Task 2 (Tri-class)

ERA5 summary:
Temperature began at 27.0 °C and ended at 24.5 °C (range 22.8–28.1 °C; trend decreasing).
Surface pressure changed from 1007.2 hPa to 1001.8 hPa (trend decreasing).
10 m wind speed shifted from 6.2 m/s to 11.4 m/s (trend increasing).
Cloud cover moved from 0.35 to 0.88 (trend increasing).
Total precipitation over the period was 18.40 mm, with a peak hourly rate of 8.20 mm.
Dewpoint depression averaged 1.4 °C. Relative humidity peaked at 98%. Peak wind gusts reached 18.3 m/s.

Claim:
Pressure fell by about 5.4 hPa while 10 m winds strengthened. The analyst concludes this confirms a deepening low-pressure system that can trigger convection.

Gold label: `Consistent`

Example: Task 3 (Counterfactual JSON)

NOAA event label: `Flash Flood`

ERA5 summary (counterfactual):
Temperature began at 27.0 °C and ended at 24.5 °C (range 22.8–28.1 °C; trend decreasing).
Surface pressure changed from 1007.2 hPa to 1001.8 hPa (trend decreasing).
10 m wind speed shifted from 6.2 m/s to 11.4 m/s (trend increasing).
Cloud cover moved from 0.35 to 0.88 (trend increasing).
Total precipitation over the period was 0.15 mm, with a peak hourly rate of 8.20 mm.
Dewpoint depression averaged 1.4 °C. Relative humidity peaked at 98%. Peak wind gusts reached 18.3 m/s.

Rule hint excerpt:
```
- heavy_precip: precip_total >= 10.0 mm or precip_peak >= 7.6 mm
```

Gold JSON:
```
{
   "modified_variable": "total_precipitation",
   "era5_sanity": "inconsistent",
   "supports_label": "uncertain",
   "rationale": "..."
}
```

System instructions (verbatim)

[System Constraints]
The model must return *only* the final answer in the required format. No explanations, reasoning steps, or additional text are allowed.

[Task1: Physical Plausibility (Multi-Select)]
Return **ONLY** a comma-separated list of letters (`A-D`), or `NONE` if no options apply.

[Task2: Consistency Verification (Tri-Class)]
Return **ONLY** one label from: `Consistent`, `Inconsistent`, `Underdetermined`.

[Task3: Counterfactual Evidence Check (JSON)]
Return **ONLY** a valid JSON object with *exactly* these keys: `modified_variable`, `era5_sanity`, `supports_label`, `rationale`.
No extra keys, no surrounding text, and no trailing commas.

- `modified_variable` ∈ {`temperature`, `surface_pressure`, `wind_speed`, `cloud_cover`, `total_precipitation`, `precip_peak`}.
- `era5_sanity` ∈ {`sane`, `inconsistent`, `uncertain`}.
- `supports_label` ∈ {`supported`, `not_supported`, `uncertain`}.
- `rationale`: keep concise; do not include any other text beyond the JSON value string.

Judge prompt (verbatim)

[System]
You are an expert meteorological judge. You must grade both correctness and rationale quality. Use the provided rule hints as guidance, but allow `uncertain` when evidence is insufficient or thresholds are ambiguous. Return **ONLY** valid JSON (no extra text).

[User Input Template]
Task: `counterfactual_evidence_check`
NOAA label: `<event_label>`
Counterfactual ERA5 summary (exactly one variable was modified):
`<counterfactual_summary>`
Optional rule hints (if provided):
`<rule_hints>`
Model output (expected JSON, may be malformed):
`<model_output_raw>`

[Reference (Ground Truth)]

```
{
   "modified_variable": "<gt_modified_variable>",
   "era5_sanity": "<gt_era5_sanity>",
   "supports_label": "<gt_supports_label>"
}
```

[Rationale Quality Rubric (1–5)]

- **5**: very professional; explains key evidence and why other (unchanged) variables do not conflict; no hallucinations.
- **3**: roughly correct but shallow; mostly repeats variable changes; limited causal chain.
- **1**: wrong direction; ignores key factor(s); may sound plausible to non-experts.
- **2,4**: intermediate quality.

Key requirements for a score of 5:

- Must explicitly identify the modified variable (or clearly anchor the evidence chain to it).
- Must use rule hints / thresholds / a causal chain to justify `supports_label`.
- Must explain why other (unchanged) variables do *not* create conflicts.
- Must not introduce variables or phenomena not present in the summary or rule hints (hallucination).

[Output JSON Schema]
Return a single JSON object with exactly these keys:

- `json_parseable` (boolean): whether `<model_output_raw>` can be parsed as JSON.
- `schema_valid` (boolean): whether parsed JSON contains exactly the required keys and valid value types.
- `model_fields` (object): extracted fields `modified_variable`, `era5_sanity`, `supports_label`, `rationale`.
- `field_errors` (array of strings): schema/type/enum violations (empty if none).
- `modified_variable_correct` (boolean).
- `era5_sanity_correct` (boolean).
- `supports_label_correct` (boolean).
- `rationale_score_1to5` (integer in [1,5]).
- `rationale_issues` (array of strings): e.g., hallucination, missing metric, ignores rule hints, etc.
- `notes` (string): brief judge notes (keep concise).

Return JSON (template):

```
{
   "json_parseable": true,
   "schema_valid": true,
   "model_fields": {"modified_variable": "...", "era5_sanity": "...", "supports_label": "...", "rationale": "..."},
   "field_errors": [],
   "modified_variable_correct": true,
   "era5_sanity_correct": true,
   "supports_label_correct": true,
   "rationale_score_1to5": 5,
   "rationale_issues": [],
   "notes": "..."
}
```