EXCONAN: A GLOBAL BENCHMARK AND EVIDENCE-GROUNDED MULTI-AGENT FRAMEWORK FOR EXTREME WEATHER REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

While recent large language models have demonstrated remarkable general reasoning capabilities, their potential in Earth system sciences remains bottlenecked by the complexity of meteorological tasks. Existing evaluations primarily target numerical forecasting, leaving a critical gap in assessing LLMs for evidence-grounded extreme weather reasoning that requires multi-source evidence interpretation and domain-specific knowledge. To address this, we introduce **EXConan-Bench**, a comprehensive global benchmark comprising 8,498 expert-curated instances. By integrating multi-source event records with ERA5 meteorological reanalysis data, the benchmark covers eight extreme weather hazards and normal conditions, challenging models to identify events using structured atmospheric evidence and interpretable rationales. Furthermore, we propose **EXConan**, an evidence-grounded multi-agent reasoning framework designed to overcome the limitations of direct prompting. EXConan introduces an agentic paradigm that meticulously coordinates specialized meteorological experts, evidence-based candidate evaluation, and counterfactual verification to ensure reliable decision-making. Extensive experiments across representative closed-source and open-source LLMs reveal the significant limitations of current baselines in meteorological reasoning. Remarkably, our EXConan framework effectively mitigates these shortcomings; it improves Qwen3-30B from 49.33% to 54.63%, with consistent gains in Macro Precision, Weighted-F1, and Kappa. Ultimately, EXConan establishes a scientifically grounded paradigm for reliable weather intelligence.

1 INTRODUCTION

Deep learning (Liang et al., 2025; LeCun et al., 2015) has advanced weather forecasting (Li et al., 2026) and enabled the adaptation of weather foundation models to diverse downstream tasks (Bi et al., 2023; Cao et al., 2026). However, assessing extreme weather events from atmospheric field data, interpreting how meteorological conditions interact and evolve, and identifying the hazards they may trigger still require human expertise. In this context, large language models (LLMs) (Minaee et al., 2024) offer an alternative approach by integrating meteorological evidence with domain knowledge to support event identification and interpretation (Ma et al., 2024; Li et al., 2025).

However, existing benchmarks offer a fragmented basis for evaluating this capability. Numerical forecasting benchmarks primarily assess atmospheric prediction accuracy rather than event recognition by LLMs (Rasp et al., 2024; Ran et al., 2025). Event-oriented benchmarks provide more direct assessments, but their geographic and task coverage remains uneven: WeatherQA focuses on severe-weather reasoning, while MP-Bench covers severe-weather events in China (Ma et al., 2024; Tang et al., 2026a). Other resources address event narratives, weather-report generation, or synoptic forecast discussions through distinct input representations and prediction targets (Li et al., 2025; Zheng et al., 2026; Higgins et al., 2026). Although valuable individually, these different evaluation settings do not provide a common basis for comparing event-recognition capabilities across regions and temporal contexts. This motivates a standardized benchmark combining ❶ *globally distributed events*, ❷ *an extended historical record*, and ❸ *multiple hazards*, as shown in Figure 1.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

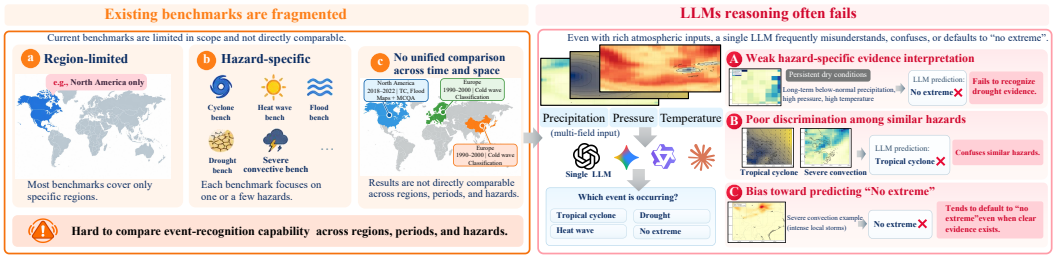


Figure 1: **Motivation for EXConan-Bench and EXConan.** Left: differences in benchmark coverage and evaluation protocols. Right: illustrative LLM failures in interpreting hazard evidence, distinguishing competing events, and rejecting premature *No Extreme* predictions.

To address these gaps, we introduce **EXConan-Bench**, a global benchmark comprising 8,498 instances spanning 2000–2023 and covering eight extreme-weather categories alongside a no-extreme control. We consolidate multi-source event records through metadata harmonization, spatiotemporal deduplication, and cross-source consistency checks, then pair each event with a category-neutral report derived from ERA5 (Hersbach et al., 2020). These reports combine recent atmospheric evolution with seven- and thirty-day antecedent context, with input windows ending 24 hours before the target time. Structural validation, temporal and label-leakage checks, and expert review produce a four-choice task for evaluating event recognition across geographic and temporal settings.

Using EXConan-Bench, we evaluate representative open- and closed-source LLMs. A detailed error analysis of LLMs reveals frequent misclassification of drought and severe convection as No Extreme. Selected cases show that the model interprets persistent dry conditions as an absence of hazards because heavy-rainfall thresholds are not met, or selects a weather category without establishing its distinguishing meteorological conditions. These observations highlight three reasoning challenges, as illustrated in Figure 1: ① *evidence interpretation*, where drought is missed despite persistent dry conditions; ②*hazard discrimination*, where severe convection is mistaken for a tropical cyclone; and ③ *decision verification*, where intense local storms are prematurely classified as **No Extreme**.

Motivated by these diagnostic findings, we propose **EXConan**, an evidence-grounded multi-agent framework. For ① *evidence interpretation*, specialists in synoptic patterns, hazard severity, and temporal evolution independently assess atmospheric organization, intensity, and persistence based on explicit meteorological evidence. For ② *hazard discrimination*, evidence competition and arbitration evaluate and compare the evidence supporting and contradicting each candidate. For ③ *decision verification*, stress testing identifies weakly supported judgments and insufficiently justified *No Extreme* predictions for counterfactual review. This review examines alternative explanations before retaining or revising the initial decision. When all three specialists recommend *No Extreme*, Aurora can also provide the reviewer with supplementary forecast evidence (Bodnar et al., 2025).

Experiments across 15 direct-inference configurations show that substantial errors remain, while stratified analyses reveal performance differences across hazard categories, geographic regions, and time periods. On the full benchmark, EXConan improves the reported aggregate metrics for all four evaluated backbones, increasing Qwen3-30B accuracy from 49.33% to 54.63% (Table 2). Controlled comparisons and ablation studies on Bench-950 further support the contributions of specialist analysis, candidate comparison, and decision review. Our main contributions are:

- ♣ We introduce EXConan-Bench, a global benchmark for extreme-weather reasoning spanning 24 years. It contains 8,498 events from 2000 to 2023, covering eight hazard categories and a no-extreme control category, with a unified decision task and a consistent pre-target.
- ♥ We propose EXConan to address fragmented evidence, weak hypothesis discrimination, and overly conservative predictions. The framework combines complementary meteorological experts with evidence competition, counterfactual review, and decision stress testing.
- ♠ Experiments on 15 representative open- and closed-source LLMs reveal category-specific weaknesses in extreme-weather recognition and demonstrate that EXConan can better unlock the capability of LLMs to predict extreme-weather events.

Table 1: **Weather-benchmark comparison.** Columns summarize evidence, tasks, and coverage. ✓: supported; ▲: partially supported; —: absent or unreported. Multi-window denotes recent and longer-term antecedent inputs; 24-h cutoff denotes a shared valid-time boundary at least 24 hours.

Benchmark	Input evidence	Evaluation task	Global	Long history	Multi-hazard	Multi-window	Event choice	Normal control	24-h cutoff
ExtremeWeather (Racah et al., 2017)	Simulated fields	Detection	▲	✓	✓	—	▲	—	—
SEVIR (Veillette et al., 2020)	Sensor imagery	Nowcasting	—	—	✓	—	—	▲	—
HR-Extreme (Ran et al., 2025)	Weather fields	Forecast scoring	—	—	✓	—	—	—	—
WeatherQA (Ma et al., 2024)	Weather maps	Weather QA	—	—	✓	—	▲	▲	—
CLLMate (Li et al., 2025)	Maps + statistics	Event / impact QA	✓	—	✓	—	✓	—	—
MP-Bench (Tang et al., 2026a)	4D weather fields	Event QA	▲	—	✓	—	✓	✓	—
WSInstruct (Zheng et al., 2026)	Weather maps	Report generation	—	—	▲	—	—	—	—
SynopticBench (Higgins et al., 2026)	Forecast maps	Discussion generation	—	—	▲	—	—	—	—
EXConan-Bench	**Spatiotemporal summaries**	**Event choice**	✓	✓	✓	✓	✓	✓	✓

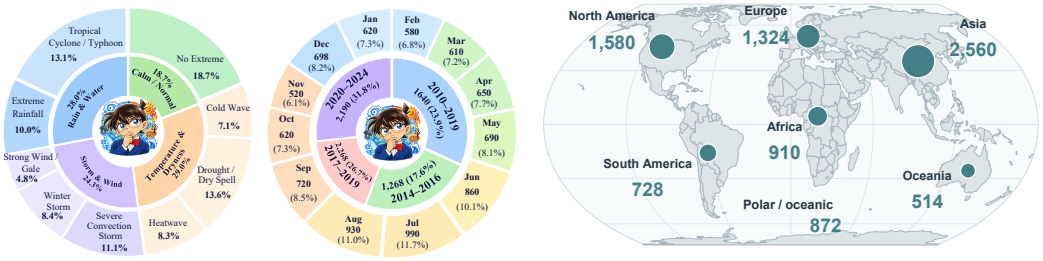
2 Related Work

2.1 Weather Understanding and Evaluation Benchmarks.

Weather benchmarks assess numerical forecasting, event understanding, and report generation. WeatherBench 2 standardizes global forecast evaluation, while HR-Extreme and Extreme Weather Bench emphasize high-impact weather (Rasp et al., 2024; Ran et al., 2025; McGovern et al., 2026). For language and multimodal models, WeatherQA evaluates affected-area prediction and severe-weather concerns using U.S. meteorological maps (Ma et al., 2024). CLLMate aligns globally distributed news-reported events with ERA5 to evaluate weather phenomena and their consequences (Li et al., 2025), whereas MP-Bench supports severe-weather question answering using Chinese warning records and meteorological fields, with an additional U.S. evaluation (Tang et al., 2026a). Other resources examine complementary capabilities: K-MetBench assesses professional meteorological knowledge and visual reasoning, while WSInstruct and SynopticBench evaluate forecast-report and discussion generation (Kim et al., 2026; Zheng et al., 2026; Higgins et al., 2026). These benchmarks differ in geographic coverage, historical periods, input windows, and prediction targets. EXConan-Bench complements them by combining global, multi-decadal, multi-hazard coverage within a unified event-selection task and a common pre-target evidence cutoff.

2.2 Agentic Reasoning for Meteorological Analysis.

Tool interaction, feedback, and collaboration provide complementary approaches to organizing language-model reasoning, as illustrated by ReAct, Reflexion, and multiagent debate (Yao et al., 2022; Shinn et al., 2023; Du et al., 2024). Meteorological agents adapt these approaches to domain-specific data and workflows. Zephyrus provides a code-based environment for weather-data analysis, forecasting, and climate simulation (Varambally et al., 2026). AgentCaster studies interactive interpretation of forecast maps and soundings for probabilistic tornado-risk prediction (Chen, 2025). HVR-Met integrates expert guidelines, diagnostic computation, and plotting within a hypothesis–verification–replanning loop for extreme-weather diagnosis (Tang et al., 2026b). Building on this direction, EXConan combines complementary meteorological specialists, evidence competition, counterfactual review, and decision stress testing to adjudicate competing event categories. Further background on deep learning for weather forecasting appears in Appendix A.



(a) Hazard composition (b) Temporal coverage (c) Geographic coverage

Figure 2: **EXConan-Bench composition.** (a) Category shares, with broader groups in the inner ring. (b) Target-date distributions by calendar month (outer ring) and historical period (inner ring). (c) Geographic coverage with regional instance counts.

3 EXCONAN-BENCH CONSTRUCTION

EXConan-Bench evaluates extreme-weather recognition from antecedent meteorological evidence through a unified four-choice task. It comprises 8,498 region–target-time instances spanning 2000–2023, covering eight hazard categories and a *No Extreme* control across globally distributed regions. Each instance pairs a target region and time with a structured weather report, four candidate categories, and a reference label. Figure 3 summarizes the construction pipeline, from event grounding and evidence-package assembly to benchmark-item authoring and quality assurance. Figure 2 summarizes the hazard composition, temporal coverage, and geographic distribution of the released instances.

3.1 EVENT GROUNDING AND SELECTION

We assemble candidates from an initial benchmark collection, EM-DAT disaster records (Delforge et al., 2025), SPEI-based drought mining (Vicente-Serrano et al., 2010; Beguería et al., 2010), ERA5-based sampling, and model-assisted expansion, combining documented disasters with meteorologically selected cases. EM-DAT records are mapped to geographic regions and the shared hazard taxonomy; drought and heatwave candidates are selected using index thresholds and persistent temperature extremes, respectively. Non-extreme sampling covers quiet conditions and negatives with near-threshold meteorological signals. Reference labels follow source-specific rules in Appendix B.

Candidates are normalized to a common schema and deduplicated by spatial overlap within each date and hazard category, reducing 10,165 candidates to 8,503 valid metadata records before evidence extraction. The units are region–target-time instances; this filtering does not establish independence between physical disasters. Source-level counts and the overlap criterion are provided in Appendix B.

3.2 EVIDENCE AND QUESTION CONSTRUCTION

We extract ERA5 evidence for each target region (Hersbach et al., 2020) under a shared temporal boundary. For target time T_i, every meteorological input has valid time t satisfying

$$t \le C_i, \qquad C_i = T_i - 24\,\mathrm{h}. \tag{1}$$

A 48-hour window captures recent atmospheric evolution, while seven- and thirty-day windows characterize accumulated background conditions; all end at C_i. Labels are assigned retrospectively and excluded from model inputs. This pre-target valid-time constraint neither requires the target to mark event onset nor implies operational availability of reanalysis data at the cutoff.

A common, class-independent template converts surface and pressure-level fields into structured reports. Regional means, spatial quantiles, extrema, and sector summaries describe atmospheric intensity and spatial organization, while six-hour sequences and longer-window summaries describe temporal evolution. Reports cover precipitation, temperature, pressure, wind, humidity, and related diagnostics, with unavailable fields explicitly marked. This shared representation makes different hazards comparable without changing the input format by reference category. Extraction conventions, variables, units, and reporting thresholds are provided in Appendix C.

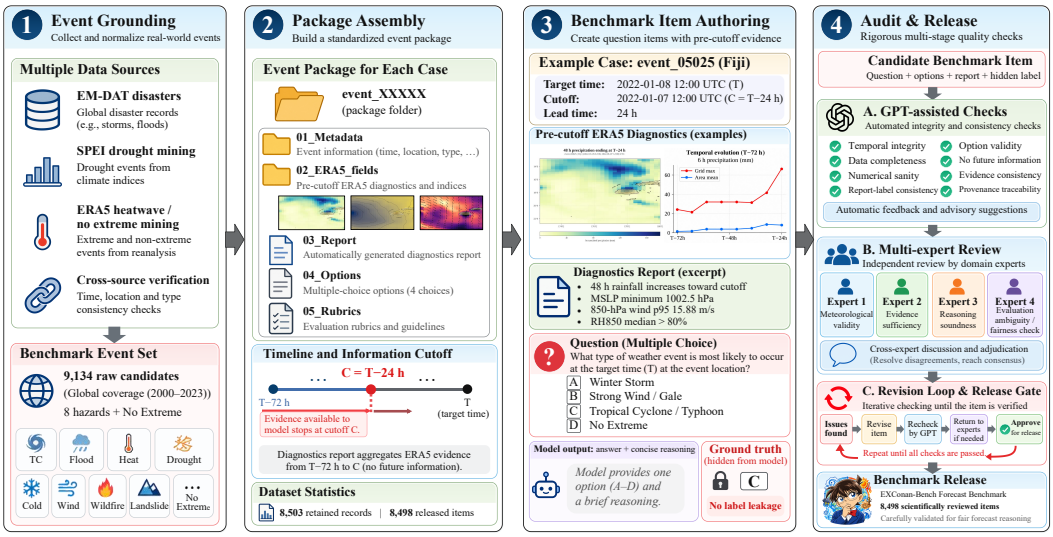


Figure 3: **EXConan-Bench construction.** Event grounding and deduplication are followed by ERA5 report generation, four-option question construction, and quality assurance. Reports combine recent evolution and antecedent conditions up to $C = T - 24\,\mathrm{h}$; reference answers remain hidden. Automated checks and human review form the audit process. The release contains 8,498 instances.

Each question contains four distinct options, including the reference category and *No Extreme*; the remaining options are sampled from the other hazards and their order is reproducibly shuffled. Models receive the region, target time, report, and options, and return an option letter with an explanation. Scoring uses the selected category. Reference answers and sampling metadata are excluded from model-visible messages. Appendix D specifies option construction and the answer protocol.

3.3 Quality Assurance and Release

Automated checks verify metadata validity, temporal and numerical completeness, compliance with the input cutoff, and consistency between options and reference labels. Five records fail report generation, yielding the final 8,498-instance release. GPT-assisted checks complement a staged human review process: a team of three students performs the initial review, and cases they cannot confidently resolve are referred to domain experts for assessment and adjudication.

The release includes eight hazards—extreme rainfall, tropical cyclone, dry spell, severe convective storm, heatwave, winter storm, cold wave, and strong wind/gale—alongside 1,585 non-extreme controls. Appendix D.3 reports class counts and validation details. A class-balanced development subset contains 950 instances drawn from the release; full-benchmark evaluations therefore include these development instances and are distinguished from independent held-out evaluation.

4 EXConan: Evidence-Grounded Meteorological Reasoning

EXConan organizes event recognition around three complementary questions: which weather system explains the observations, how severe its impacts are, and how it evolves toward the target time. As outlined in Figure 4, independent specialists assess a shared evidence record, an evidence-competition layer exposes support and contradictions across candidates, and an arbiter produces an initial decision. Decision stress testing and conditional counterfactual review then examine whether that decision withstands alternative explanations, with particular attention to premature *No Extreme* judgments.

4.1 Evidence Preparation and Complementary Specialists

Each benchmark instance is converted into an inference record containing the target region, target time, pre-cutoff meteorological report, and four semantic candidates. The release identifier and randomized

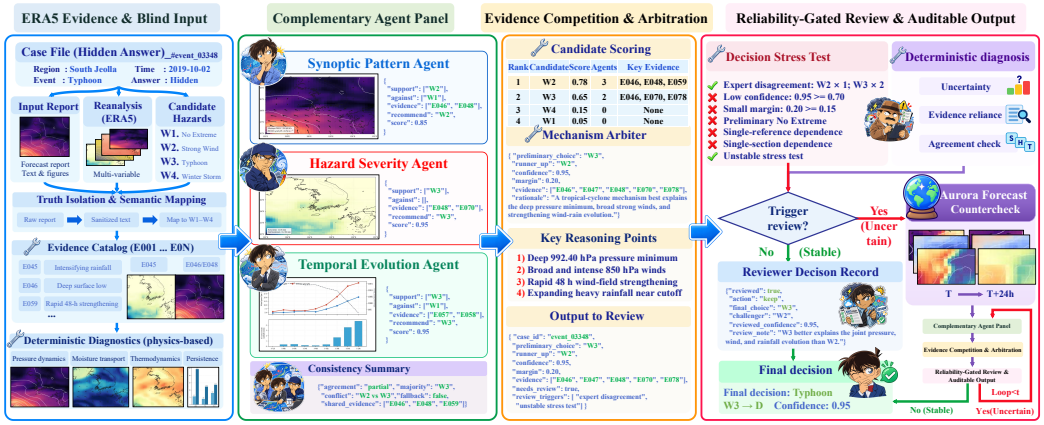


Figure 4: **EXConan workflow.** Three specialists assess synoptic structure, hazard severity, and temporal evolution using numbered evidence, diagnostics, and candidates W1–W4. Evidence competition and arbitration produce an initial decision for stress testing and conditional review. Valid unanimous *No Extreme* recommendations can trigger Aurora, whose available forecasts support only the reviewer. Reference answers remain hidden; the final candidate maps to its answer letter.

A–D mapping are removed; candidates are sorted by label and assigned internal identifiers W1–W4. Non-empty evidence lines receive identifiers E001, E002, and so forth, allowing subsequent assessments to cite specific observations. Reference answers remain outside the inference record. Three deterministic tools extract pressure and circulation diagnostics, precipitation and convection-related diagnostics, and persistence indicators from the report. Their outputs organize the supplied evidence without introducing additional observations.

Three role-conditioned specialists review the same numbered report and diagnostic outputs in parallel, without access to one another's responses. The *synoptic-pattern specialist* assesses mechanisms based on pressure, circulation, moisture, and thermal structure. The *hazard-severity specialist* evaluates intensity and spatial extent, distinguishing weak impacts from evidence against the underlying system. The *temporal-evolution specialist* examines 6-, 24-, and 48-hour changes alongside seven- and thirty-day antecedent conditions. Each assigns every candidate a support score in $[0, 1]$, cites supporting and contradicting evidence, and identifies missing information. All must treat missing diagnostics as uncertainty and explicitly consider hazard alternatives before recommending *No Extreme*.

4.2 Evidence Competition and Arbitration

To make disagreements inspectable, EXConan aggregates assessments at the candidate level. Let $\mathcal{V}$ denote the set of specialists with valid outputs, $s_k(c)$ the support assigned by specialist k to candidate c, and $\mathcal{E}_k^-(c)$ its cited contradicting evidence. The competition score is

$$S(c) = \frac{1}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} s_k(c) - \frac{\lambda}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} \mathbf{1}\left[\mathcal{E}_k^-(c) \neq \varnothing\right], \tag{2}$$

with $\lambda = 0.15$ in the primary configuration. The second term penalizes the fraction of specialists citing a contradiction. Supporting and contradicting roles remain explicit, while missing evidence is retained for qualitative review. Invalid specialist outputs are excluded from aggregation.

An LLM arbiter receives the original evidence, specialist assessments, and competition table. It selects an initial candidate and runner-up, accompanied by decisive evidence references, self-reported confidence and decision margin, and remaining uncertainties. The arbiter is instructed to compare explanatory mechanisms and their associated impacts; the competition ranking informs this assessment without mechanically determining the answer. Thus, candidate disagreement remains available for examination beyond a vote count or aggregate score.

4.3 Decision Stress Testing and Counterfactual Review

Decision stress testing checks whether the initial judgment rests on sufficiently diverse evidence and explicit comparison of alternatives. This deterministic diagnosis flags specialist disagreement, sparse or single-section evidence, low confidence, a small reported margin, and specialist failures. For a *No Extreme* decision, it additionally checks whether every competing hazard has been explicitly contradicted by at least one valid specialist. These checks target unsupported dismissal of extreme events as well as fragile positive predictions. They inspect the existing decision record without perturbing meteorological fields or rerunning the specialists. Exact criteria are provided in Appendix E.8.

Flagged cases undergo a single counterfactual review using the original evidence, compressed specialist assessments, initial decision, competition results, and stress diagnostics. The reviewer seeks the strongest alternative explanation and checks for shared threshold bias before retaining or replacing the initial candidate. In the primary configuration, Aurora is invoked when all three specialists return valid *No Extreme* recommendations. Initialized with fields at $C_i - 6\,\mathrm{h}$ and C_i, it advances four six-hour steps to $T_i = C_i + 24\,\mathrm{h}$. Regional forecast summaries are supplied only to the reviewer as supplementary evidence and do not independently determine the event label.

A valid review can retain the initial choice or replace it with the identified challenger. If review fails, the initial decision is preserved; unavailable forecasts likewise do not imply non-extreme conditions. The final semantic candidate is mapped back to its original answer letter only after inference. Output contracts, forecast configuration, and fallback rules are specified in Appendix E.

5 Experiments

We evaluate three questions: how accurately current LLMs recognize extreme weather from antecedent evidence, whether EXConan improves on direct inference and simpler multi-call strategies, and which components account for its performance. We then examine whether aggregate gains extend across hazards, geographic regions, and time. Figure 5 presents the controlled experiments on Bench-950 and the full-benchmark profiles across hazards, time, and geographic regions.

5.1 Experimental Setup

Benchmark and protocol. We evaluate on EXConan-Bench, comprising 8,498 instances spanning 2000–2023 across eight extreme-weather categories and a no-extreme control. Each instance presents a four-option decision task, with meteorological evidence restricted to valid times at least 24 hours before the target time. We report the results of the full benchmark tests and the ablation studies on Bench-950 (a development subset comprising 950 instances) separately.

Evaluated models. We evaluate 15 representative LLM configurations. The open-weight baselines comprise Qwen3 configurations (Yang et al., 2025), Qwen3.8-27B (Qwen Team, 2026), Phi-4 (Abdin et al., 2024), and Mistral-Small-24B-Instruct (Mistral AI, 2025), covering different model sizes and instruction-following or reasoning configurations. The proprietary baselines include Gemini-3.6-Flash (Google, 2026), GPT-5.4 and GPT-5.5 (OpenAI, 2026a;b), and Claude-Sonnet-5 and Claude-Opus-4.8 (Anthropic, 2026b;a). We instantiate EXConan using various large models in a comprehensive benchmark, comparing each agent against its corresponding direct-answer backbone model on the same instances. The main configuration conditionally accesses Aurora for supplementary forecast evidence (Bodnar et al., 2025). Detailed settings and metrics are provided in Appendix E and F.

5.2 Main Results

Direct inference leaves substantial room for improvement. Table 2 shows variation across models. Qwen3-30B-A3B-Instruct is the strongest open-weight direct baseline, with 49.33% accuracy, followed by Mistral-Small-24B-Instruct at 48.53%. Gemini 3.6 Flash leads the proprietary configurations at 56.53%. Performance does not increase monotonically with model size or a reasoning-oriented configuration: Qwen3-14B scores 36.41%, compared with 37.63% for Qwen3-8B, and Qwen3-30B-A3B-Thinking trails its instruct counterpart by 8.86 pp. These results indicate that general language-model capability alone does not ensure accurate interpretation of meteorological evidence.

Table 2: **Direct and EXConan on all 8,498 instances.** Accuracy, macro precision, and weighted F1 are percentages; κ is unscaled. Parentheses show gains over the corresponding direct-inference backbone in percentage points. Boldface marks EXConan and the best closed-source baseline scores.

Model	Accuracy	Macro Precision	Weighted-F1	κ
Closed-source Models				
Gemini 3.6 Flash	**56.53**	**71.13**	**59.88**	**0.5110**
GPT-5.5	44.32	60.89	47.07	0.3737
Claude Sonnet 5	42.00	66.01	46.28	0.3477
Claude Opus 4.8	41.68	67.41	46.00	0.3442
GPT-5.4	39.68	66.44	43.57	0.3217
Open-source Models				
Qwen3-4B	36.46	47.76	34.74	0.2715
Qwen3-8B	37.63	51.51	34.14	0.2679
Qwen3-14B	36.41	55.53	34.31	0.2529
Qwen3.8-27B (Non-thinking)	44.41	61.75	46.11	0.3427
Qwen3.8-27B (Thinking)	43.53	62.53	45.35	0.3298
Qwen3-30B-A3B-Instruct-2507	49.33	55.25	50.27	0.4214
Qwen3-30B-A3B-Thinking-2507	40.47	56.48	40.41	0.2937
Phi-4 (14B)	27.73	55.93	26.81	0.1869
Phi-4-mini-instruct	27.47	34.52	22.32	0.1846
Mistral-Small-24B-Instruct-2501	48.53	57.38	50.18	0.4043
EXconan				
Qwen3-4B	**38.61** (↑2.15)	**49.49** (↑1.73)	**36.09** (↑1.35)	**0.2910** (↑0.0195)
Qwen3-30B-A3B-Instruct-2507	**54.63**(↑5.30)	**58.29**(↑3.02)	**53.43**(↑3.16)	**0.4915**(↑0.0701)
Phi-4 (14B)	**38.44** (↑10.71)	**59.42** (↑3.49)	**37.30** (↑10.49)	**0.3079** (↑0.1209)
Phi-4-mini-instruct	**30.00** (↑2.53)	**37.54** (↑3.02)	**25.09** (↑2.77)	**0.2125** (↑0.0279)

EXConan improves all three evaluated backbones. With Qwen3-30B, EXConan raises accuracy from 49.33% to 54.63%, weighted F1 from 50.27% to 53.43%, and κ from 0.4214 to 0.4915. Qwen3-4B and Phi-4-14B gain 2.15 and 10.71 pp in accuracy, respectively. All three backbones improve on every reported aggregate metric. The varying gains suggest that EXConan's benefits depend on backbone architecture and model size. These comparisons involve additional inference calls and conditional use of supplementary Aurora evidence, motivating controlled experiments in Section 5.3.

5.3 Ablation and Controlled Comparisons

We conduct controlled comparisons and ablations using Qwen3-4B on Bench-950 to examine the contributions of additional calls, core components, and supplementary forecast evidence. Each intervention is evaluated against the complete configuration within its experiment.

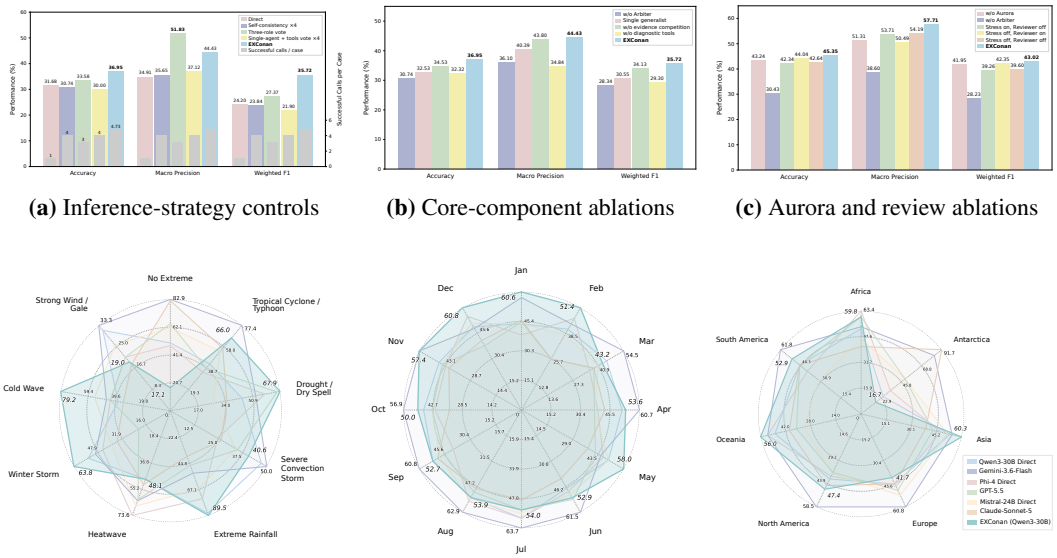
Multi-call controls. EXConan achieves 36.95% accuracy, exceeding direct inference (31.68%), four-sample self-consistency (30.74%), and three-role voting (33.58%) (Figure 5a). Additional calls do not guarantee improvement. Three-role voting outperforms EXConan in macro precision. EXConan averages 4.73 successful calls per case, versus three for role voting and four for repeated sampling. These results suggest a performance advantage under similar computational budgets.

Core-component ablations. Relative to the full configuration (36.95% accuracy), removing the arbiter, removing diagnostic tools, replacing specialized experts with a single generalist, and removing evidence competition reduce accuracy by 6.21, 4.63, 4.42, and 2.42 percentage points, respectively (Figure 5b). All four ablations lower macro precision and weighted F1, supporting these components' contributions to end-to-end performance. Arbiter removal causes the largest accuracy decline.

Aurora and review mechanisms. Figure 5c evaluates a separate full configuration with 45.35% accuracy. Disabling Aurora lowers accuracy to 43.24%, indicating a benefit from supplementary forecast evidence. Disabling the reviewer, stress testing, or both yields 42.34%, 44.04%, and 42.64% accuracy, respectively. Both stages together achieve the highest accuracy, while stress testing alone does not outperform disabling both, suggesting it is most useful when followed by review.

5.4 Performance across Hazards, Regions, and time

We evaluate representative open-weight models, closed-source models, and EXConan across three dimensions: hazard categories, geographic regions, and time. This analysis uses all 8,498 cases in EXConan-Bench to characterize model across weather events and spatiotemporal contexts.



(a) Inference-strategy controls **(b)** Core-component ablations **(c)** Aurora and review ablations

(d) Recall by hazard category **(e)** Accuracy by calendar month **(f)** Accuracy by geographic region

Figure 5: **Controlled experiments and benchmark profiles.** (a–c) Qwen3-4B inference controls, component ablations, and Aurora/review ablations (Bench-950). Colors show accuracy, macro precision, and weighted F1 (%); gray show mean successful calls per case (a, right axis). Full-configuration accuracies: 36.95% (a,b) and 45.35% (c); compare within configurations. (d–f) Category recall, monthly, and regional accuracy for open-weight , closed-source models, and EXConan on 8,498 instances. Thick teal: EXConan (Qwen3-30B). Radar axes are independently scaled.

Hazard categories. Figure 5d highlights distinct strengths across models. EXConan achieves high recall for cold waves (79.2%) and extreme rainfall (89.5%), while Gemini performs better on tropical cyclones, strong winds, and non-extreme conditions. These contrasting profiles reflect both hazard-specific recognition capabilities and the ability to distinguish extreme from normal weather.

Geographic regions. Model rankings vary across regions (Figure 5f): EXConan leads in Asia and Oceania, Phi-4 performs best in Africa, and Gemini leads in Europe and the Americas. This regional breakdown reveals geographic strengths and uneven performance that aggregate scores may obscure.

Temporal analysis. Monthly profiles reveal performance variation across the annual cycle (Figure 5e). EXConan leads in January and December, while Gemini performs best in March and July. Pooling cases from 2000–2023 by month enables the benchmark to assess consistency throughout the year.

Together, these analyses assess meteorological event identification across hazard types, regions, and times of year, revealing baseline strengths and limitations that aggregate scores may obscure.

6 Conclusion

We introduce **EXConan-bench**, a benchmark for extreme-weather event reasoning from spatially structured meteorological reports under a fixed pre-target valid-time cutoff. It comprises 8,498 instances covering eight hazards and non-extreme conditions across diverse regions and time periods. Evaluations of 14 direct-inference configurations from open-weight and closed-source model families reveal distinct strengths and persistent errors across these dimensions, exposing capability differences obscured by aggregate scores. We further introduce **EXConan**, an evidence-driven framework integrating meteorological diagnostics, specialist analysis, arbitration, and decision review. Through various model architectures and parameters, EXConan improves all reported aggregate metrics over direct inference, with ablations supporting the contributions of its core mechanisms and supplementary forecast evidence. Together, EXConan-bench and EXConan provide a comprehensive evaluation resource and an effective baseline for advancing evidence-based extreme-weather reasoning.

AI USE STATEMENT

We used Qwen2.5-72B to assist with category and distractor generation for the ERA5-conditioned expansion component of EXConan-Bench. Dataset construction also involved model-assisted normalization of ambiguous disaster descriptions and GPT-assisted quality checks. The associated validation and human-review procedures are described in Appendices B and D. Codex assisted with bibliographic checking, manuscript consistency checks, metric explanations, and LaTeX drafting, including these statements. The LLMs evaluated as baselines and used within EXConan are documented in the experimental setup and Appendix E. The authors take responsibility for the final manuscript, dataset, analyses, and conclusions, including all AI-assisted content.

ETHICS STATEMENT

EXConan-Bench and EXConan are intended for research on evidence-grounded extreme-weather reasoning. The benchmark combines retrospective disaster records and meteorological reanalysis; performance on this task does not establish readiness for operational forecasting or emergency decision-making. Geographic and class imbalance, heterogeneous labeling rules, and model-assisted annotation may affect benchmark comparisons and generalization. Generated explanations may contain meteorological errors and require expert assessment before operational use. Use and redistribution of upstream datasets and pretrained models remain subject to their respective licenses and access conditions.

REPRODUCIBILITY STATEMENT

Appendices B–D describe the data sources, label construction, deduplication, temporal cutoff, evidence rendering, option generation, and quality-assurance procedures. Appendix E details the agent roles, diagnostic tools, structured output contracts, fallback rules, generation settings, and conditional forecast integration. Appendix F specifies the evaluation cohorts, metric definitions, treatment of invalid outputs, and statistical conventions. Controlled comparisons and generation-seed experiments are documented in Appendix H. These descriptions distinguish full-benchmark evaluation from development-subset experiments and make the evaluation procedures explicit.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, et al. Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*, 2024. URL https://arxiv.org/abs/2412.08905.

Anthropic. Introducing Claude Opus 4.8, 2026a. URL https://www.anthropic.com/news/claude-opus-4-8. Official release announcement. Accessed September 24, 2026.

Anthropic. Introducing Claude Sonnet 5, 2026b. URL https://www.anthropic.com/news/claude-sonnet-5. Official release announcement. Accessed September 24, 2026.

Santiago Beguería, Sergio M. Vicente-Serrano, and Marta Angulo-Martínez. A multiscalar global drought dataset: The SPEIbase: A new gridded product for the analysis of drought variability and impacts. *Bulletin of the American Meteorological Society*, 91(10):1351–1356, 2010. doi: 10.1175/2010BAMS2988.1.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023. doi: 10.1038/s41586-023-06185-3. URL https://doi.org/10.1038/s41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

Shilei Cao, Hehai Lin, Jiashun Cheng, Yang Liu, Guowen Li, Xuehe Wang, Juepeng Zheng, Haoyuan Liang, Meng Jin, Chengwei Qin, Hong Cheng, and Haohuan Fu. Task-adaptive parameter-efficient fine-tuning for weather foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eFExhM3tKr.

Michael Chen. AgentCaster: Reasoning-Guided Tornado Forecasting, 2025. URL https://arxiv.org/abs/2510.03349.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Damien Delforge, Valentin Wathelet, Regina Below, Cinzia Lanfredi Sofia, Margo Tonnelier, Joris A. F. van Loenhout, and Niko Speybroeck. EM-DAT: The emergency events database. *International Journal of Disaster Risk Reduction*, 124:105509, 2025. doi: 10.1016/j.ijdrr.2025.105509.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

GADM. GADM database of global administrative areas, version 4.1. https://gadm.org/data.html, 2022. Accessed: 2026-09-26.

Google. Introducing Gemini 3.6 Flash, 3.5 Flash-Lite, and 3.5 Flash Cyber, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/. Official release announcement. Accessed September 24, 2026.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803. URL https://doi.org/10.1002/qj.3803.

Timothy B. Higgins, Antonios Mamalakis, and Chirag Agarwal. SynopticBench: Evaluating Vision-Language Models on Generating Weather Forecast Discussions of the Future, 2026. URL https://arxiv.org/abs/2604.16451v1.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Soyeon Kim, Cheongwoong Kang, Myeongjin Lee, Eun-Chul Chang, Lee Jaedeok, and Jaesik Choi. K-MetBench: A multi-dimensional benchmark for fine-grained evaluation of expert reasoning, locality, and multimodality in meteorology. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5574–5612, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.275. URL https://aclanthology.org/2026.findings-acl.275/.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation

models for weather and climate. *Nature*, 632(8027):1060–1066, 2024. doi: 10.1038/s41586-024-07744-y. URL https://doi.org/10.1038/s41586-024-07744-y.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://doi.org/10.1126/science.adi2336.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Guowen Li, Xintong Liu, Yang Liu, Mengxuan Chen, Shilei Cao, Xuehe Wang, Juepeng Zheng, Jinxiao Zhang, Haoyuan Liang, Lixian Zhang, Jiuke Wang, Meng Jin, Hong Cheng, and Haohuan Fu. TianQuan-S2S: A subseasonal-to-seasonal global weather model via incorporate climatology state. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7Dvmq7MhwU.

Haobo Li, Zhaowei Wang, Jiachen Wang, Yueya Wang, Alexis Kai Hon Lau, and Huamin Qu. CLLMate: A multimodal benchmark for weather and climate events forecasting. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17536–17562, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.886. URL https://aclanthology.org/2025.emnlp-main.886/.

Haoyuan Liang, Xinyu Zhang, Shilei Cao, Guowen Li, and Juepeng Zheng. Tta-feddg: Leveraging test-time adaptation to address federated domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 18658–18666, 2025.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. WeatherQA: Can Multimodal Language Models Reason about Severe Weather?, 2024. URL https://arxiv.org/abs/2406.11217.

Amy McGovern, Taylor Mandelbaum, Daniel Rothenberg, Nicholas Loveday, Corey Potvin, Montgomery Flora, Linus Magnusson, Eric Gilleland, and John Allen. Extreme Weather Bench: A framework and benchmark for evaluation of high-impact weather, 2026. URL https://arxiv.org/abs/2605.01126v1.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Mistral AI. Mistral-Small-24B-Instruct-2501: Model Card. Hugging Face, 2025. URL https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501. Accessed September 24, 2026.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/nguyen23a.html.

OpenAI. Introducing GPT-5.4, 2026a. URL https://openai.com/index/introducing-gpt-5-4/. Official release announcement. Accessed September 24, 2026.

OpenAI. Introducing GPT-5.5, 2026b. URL https://openai.com/index/introducing-gpt-5-5/. Official release announcement. Accessed September 24, 2026.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, 2022. URL https://arxiv.org/abs/2202.11214.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025. doi: 10.1038/s41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.8-27B: Model Card. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed September 24, 2026.

Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr Prabhat, and Chris Pal. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Advances in neural information processing systems*, 30, 2017.

Nian Ran, Peng Xiao, Yue Wang, Wesley Shi, Jianxin Lin, Qi Meng, and Richard Allmendinger. HR-Extreme: A High-Resolution Dataset for Extreme Weather Forecasting. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3236f661fc6f00e8b16ff42db99acaaf-Abstract-Conference.html.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *Journal of Advances in Modeling Earth Systems*, 16(6), 2024. doi: 10.1029/2023MS004019. URL https://doi.org/10.1029/2023MS004019.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021. doi: 10.1038/s41586-021-03854-z. URL https://doi.org/10.1038/s41586-021-03854-z.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Shuo Tang, Jian Xu, Jiadong Zhang, Yi Chen, Qizhao Jin, Lingdong Shen, Chenglin Liu, and Shiming Xiang. Meteorpred: A meteorological multimodal large model and dataset for severe weather event prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22910–22919, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tang_MeteorPred_A_Meteorological_Multimodal_Large_Model_and_Dataset_for_Severe_CVPR_2026_paper.html.

Shuo Tang, Jiadong Zhang, Gengxian Zhou, Qizhao Jin, Qinxuan Wang, Yi Hu, Ning Hu, Hongchang Ren, Lingli He, Shiming Xiang, Jingtao Ding, Jian Xu, Jiaolan Fu, and Cheng-Lin Liu. HVR-Met: A Hypothesis-Verification-Replanning Agentic System for Extreme Weather Diagnosis, 2026b. URL https://arxiv.org/abs/2603.01121v2.

Sumanth Varambally, Marshall Fisher, Jas Thakker, Yiwei Chen, Zhirui Xia, Yasaman Jafari, Ruijia Niu, Manas Jain, Veeramakali Vignesh Manivannan, Zachary Novack, Luyu Han, Srikar Eranky, Salva Rühling Cachay, Taylor Berg-Kirkpatrick, Duncan Watson-Parris, Yi-An Ma, and Rose Yu. Zephyrus: An Agentic Framework for Weather Science. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.04017.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in neural information processing systems*, 33:22009–22019, 2020.

Sergio M. Vicente-Serrano, Santiago Beguería, and Juan I. López-Moreno. A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. *Journal of Climate*, 23(7):1696–1718, 2010. doi: 10.1175/2009JCLI2909.1. URL https://doi.org/10.1175/2009JCLI2909.1.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, 2022. URL https://arxiv.org/abs/2210.03629v3.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I. Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, 619 (7970):526–532, 2023. doi: 10.1038/s41586-023-06184-4. URL https://doi.org/10.1038/s41586-023-06184-4.

Zinan Zheng, Yang Liu, Nuo Chen, Juepeng Zheng, Hong Cheng, and Jia Li. WeatherSyn: An Instruction Tuning MLLM For Weather Forecasting Report Generation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63486.

Appendix

Contents

A More Related Work **18**

A.1 Deep Learning for Weather Forecasting 18

B Data Sources, Label Construction, and Deduplication **19**

B.1 Task unit and hazard taxonomy 19

B.2 Source integration and accounting 19

B.3 Spatial-overlap deduplication 20

B.4 Target-time precision and region extent 21

C Meteorological Evidence and Report Construction **21**

C.1 One cutoff shared by every evidence window 21

C.2 Variables, units, and transformations 21

C.3 Spatial aggregation and threshold coverage 22

C.4 Temporal evolution and persistence 22

C.5 Report validation 22

D Question Generation, Release Composition, and Quality Assurance **22**

D.1 Four-option construction 22

D.2 Visible and offline fields 23

D.3 Final composition 23

D.4 Development membership 23

D.5 Automated and human quality assurance 24

E EXConan Implementation and Structured Output Contracts **24**

E.1 Versioned implementations 24

E.2 Inference record and evidence identifiers 24

E.3 Deterministic diagnostic tools 24

E.4 Independent specialist roles 24

E.5 Shared prompt principles and output validation 25

E.6 Candidate evidence competition 25

E.7 Arbitration and fallback 25

E.8 Decision stress testing 26

E.9 Conditional counterfactual review 26

E.10 Conditional Aurora countercheck 26

E.11 Generation and transport configuration 26

F Evaluation Cohorts, Metrics, and Statistical Conventions **27**

F.1 Distinct evaluation cohorts . . . 27
F.2 Strict classification metrics . . . 27
F.3 Binary extreme-event measures . . . 28
F.4 Paired comparisons and denominators . . . 28
F.5 Intervals, seeds, and exploratory comparisons . . . 28

G Full-Benchmark and Matched-Backbone Results 29
G.1 Benchmark difficulty across direct models . . . 29
G.2 Primary same-backbone improvement . . . 29
G.3 Where the correction capacity is strongest . . . 30
G.4 Stronger-backbone corroboration on a matched cohort . . . 31
G.5 Distinguishing final results from intermediate snapshots . . . 31

H Frozen Exconan Experiments and Controlled Comparisons 31
H.1 A fixed interface for mechanism analysis . . . 31
H.2 Stability over three generation seeds . . . 31
H.3 Inference strategy and budget controls . . . 31
H.4 Specialists, tools, arbitration, and competition . . . 33
H.5 Input evidence and option-order controls . . . 33

I Regional, Temporal, and Source-Stratified Analyses 34
I.1 Analysis scope . . . 34
I.2 Geographic groups . . . 34
I.3 Six historical periods . . . 34
I.4 Source components . . . 35
I.5 Confusion structure . . . 35

J Six Verified Corrections of Direct Inference 36
J.1 Selection and interpretation . . . 36
J.2 Case 1: Recognizing a tropical-cyclone system beyond its wind impact . . . 36
J.3 Case 2: Recovering severe convection from an inconsistent No Extreme judgment . . . 37
J.4 Case 3: Distinguishing sustained cold from an upper-air wind signal . . . 37
J.5 Case 4: Identifying a winter-storm system before the target time . . . 38
J.6 Case 5: Correctly reading an expanding heavy-rainfall footprint . . . 38
J.7 Case 6: Rejecting a false gale alarm with consistent forecast counterevidence . . . 39

K Reviewer Transitions, Forecast Availability, and Cost 39
K.1 Primary EXConan execution record . . . 39
K.2 Reviewer behavior with the reliable EXConan interface . . . 40
K.3 Binary recall and false-alarm tradeoff . . . 40

K.4 Aurora coverage and a separate cache diagnostic 40
K.5 Computation and conditional paths 41

L Limitations **41**

A More Related Work

A.1 Deep Learning for Weather Forecasting

Global atmospheric-field prediction. Data-driven weather forecasting learns the evolution of atmospheric states from large meteorological archives. FourCastNet uses adaptive Fourier neural operators for high-resolution global forecasting, demonstrating that learned models can generate forecasts and ensembles at low inference cost (Pathak et al., 2022). Pangu-Weather incorporates three-dimensional atmospheric structure through an Earth-specific Transformer and reduces accumulated errors through hierarchical temporal aggregation (Bi et al., 2023). GraphCast uses graph neural networks to model atmospheric interactions across spatial scales and produces medium-range forecasts of surface and upper-air variables (Lam et al., 2023). These approaches establish different architectural routes to efficient global prediction. Their outputs describe the evolving atmospheric state, providing numerical evidence from which downstream applications can identify hazardous conditions, track weather systems, or derive risk indicators.

Extreme weather, uncertainty, and physical structure. Deep learning has also advanced forecasting tasks that directly concern hazardous weather. For precipitation nowcasting, DGMR generates plausible future radar sequences probabilistically, while NowcastNet combines physical evolution with conditional generative learning to forecast extreme precipitation (Ravuri et al., 2021; Zhang et al., 2023). At global scales, GenCast models a distribution of future atmospheric trajectories through ensemble generation, supporting uncertainty-aware evaluation of weather extremes (Price et al., 2025). NeuralGCM takes a complementary approach by coupling a differentiable atmospheric dynamical core with learned parameterizations of unresolved processes (Kochkov et al., 2024). Together, these studies show that weather prediction extends beyond deterministic regression toward the representation of uncertainty and physical structure. They also establish that learned forecasting systems already support extreme-weather applications. EXConan-Bench examines a subsequent decision problem: whether a language model can use the available meteorological evidence to distinguish among competing hazard categories and a no-extreme alternative.

Foundation models and task adaptation. Weather foundation models seek reusable representations across variables, resolutions, and downstream objectives. ClimaX pretrains on heterogeneous climate simulations and supports adaptation to weather forecasting and climate projection tasks (Nguyen et al., 2023). Aurora extends this direction through large-scale pretraining on diverse atmospheric datasets and adaptation to applications including high-resolution weather prediction, air quality, and ocean waves (Bodnar et al., 2025). Such breadth makes adaptation efficiency an important consideration. WeatherPEFT addresses differences in variables, spatial resolution, and spatiotemporal coverage through task-adaptive dynamic prompting and stochastic Fisher-guided parameter selection (Cao et al., 2026). These methods concern how numerical forecasting capabilities are learned and transferred. EXConan addresses how meteorological evidence is interpreted and compared at inference time. The two directions can operate together: its conditional Aurora component supplies additional forecast evidence, while specialist analysis and candidate comparison determine the final event category.

Longer forecast horizons and temporal context. Subseasonal-to-seasonal prediction introduces a different temporal challenge from short- and medium-range forecasting. TianQuan-S2S combines initial weather states with climatological means and an uncertainty-augmented Transformer to improve prediction at these longer lead times (Li et al., 2026). This line of research concerns how far into the future atmospheric conditions can be predicted. EXConan-Bench distinguishes forecast lead time from two other temporal dimensions: the historical period from which instances are sampled and the antecedent windows supplied for each decision. Its 2000–2023 coverage spans 24 calendar years, while each instance combines recent atmospheric evolution with 7- and 30-day background conditions ending 24 hours before the target. This design supports evaluation across a broad historical event record and multiple antecedent durations within a common event-selection task. It complements numerical forecast evaluation by measuring the accuracy of the resulting event judgments.

B Data Sources, Label Construction, and Deduplication

B.1 Task unit and hazard taxonomy

The fundamental unit in EXConan-Bench is a *region–target-time instance*: a geographic bounding box, a target time, a report constructed from antecedent meteorological fields, four candidate categories, and an offline reference answer. A release identifier uniquely indexes this unit. It does not identify a verified independent physical disaster. A single weather system may affect several regions or dates, and the archive does not supply a source-independent physical-event identifier.

The label space contains eight hazards—Cold Wave, Drought / Dry Spell, Extreme Rainfall, Heatwave, Severe Convection Storm, Strong Wind / Gale, Tropical Cyclone / Typhoon, and Winter Storm—and a No Extreme control. Labels are semantic categories rather than a globally uniform set of operational warning thresholds. In particular, a tropical cyclone describes a weather system, whereas strong wind and extreme rainfall can describe its impacts. EXConan explicitly compares these explanations within the four candidates supplied for each instance.

B.2 Source integration and accounting

Seven source components contribute candidates. Their complementary roles are historical event grounding, targeted sampling of persistent hazards, and construction of both quiet and difficult negative conditions. Table S1 accounts for all candidates, validation removals, duplicate removals, and report failures. This accounting distinguishes metadata acceptance from successful evidence rendering.

Table S1: Complete source accounting from candidates to rendered instances.

Source	Initial	Invalid	Dup.	Kept	Fail	Final
Legacy benchmark	3542	1	1030	2511	2	2509
SPEI drought	700	0	21	679	0	679
EM-DAT	4043	0	588	3455	2	3453
ERA5 heatwave	210	0	0	210	0	210
Difficult negatives	570	0	22	548	0	548
High-latitude negatives	800	0	0	800	1	799
Model-assisted expansion	300	0	0	300	0	300
Total	10165	1	1661	8503	5	8498

The initial pool contains 10,165 candidate records. One candidate fails structural validation and 1,661 are removed by the spatial-overlap rule, leaving 8,503 metadata records. Five records subsequently fail evidence rendering, yielding 8,498 released instances. The excluded release identifiers are `event_00879`, `event_00880`, `event_03396`, `event_06203`, and `event_07805`. The archived errors concern opening required ERA5 files with the available xarray I/O backends. They are preprocessing failures and are distinct from model-output failures during evaluation.

Legacy benchmark component. The legacy collection contributes 2,509 rendered instances after centralized filtering. The source answer is inherited from its semantic option mapping. The available archive does not preserve a complete instance-level external provenance or human-adjudication chain for this component. We retain its identity in the release manifest so that its performance and potential source effects remain separately inspectable.

Historical disaster records. The EM-DAT pipeline uses the disaster archive described by Delforge et al. (2025) and contributes 3,453 rendered instances. It excludes nonmeteorological disasters and flood records, resolves affected places using GADM 4.1 (GADM, 2022), and maps documented disaster subtypes to the shared taxonomy. Records with one to three resolved locations can yield separate regional instances; records with four or more locations are combined into a larger region. Thus, a source disaster record and a benchmark instance need not have a one-to-one relationship.

Explicit source subtypes are mapped deterministically: tropical cyclones to Tropical Cyclone / Typhoon; severe weather, lightning/thunderstorms, tornado, hail, and derecho to Severe Convection Storm; blizzard/winter storm and extra-tropical storm to Winter Storm; and cold wave, heat wave, and drought to their corresponding categories. Ambiguous subtypes, including general storms and severe winter conditions, undergo model-assisted classification constrained to the canonical categories or rejection. This is a source-normalization operation, not evidence that every resulting label has been independently verified by experts.

Drought mining. The SPEI component uses the Standardized Precipitation Evapotranspiration Index (Vicente-Serrano et al., 2010) and its gridded global dataset, SPEIbase (Beguería et al., 2010). Our mining pipeline uses monthly SPEI at a one-month accumulation scale on a 0.5° grid for 2005–2018. Candidate extraction begins at $\mathrm{SPEI} \leq -1.5$, retains the driest grid point in each coarse 5° cell per month, limits absolute latitude to 55°, and requires an eligible second-level administrative region within 100 km. The final intensity-ranked selection of 700 candidates all satisfies $\mathrm{SPEI} \leq -2.0$. Each selected point defines a roughly $2^\circ \times 2^\circ$ region. Central deduplication leaves 679 instances. These labels reflect an accumulated moisture-deficit index; the model-visible report does not directly expose the reference SPEI label.

Heatwave mining. The dedicated heatwave pipeline uses ERA5 two-metre temperature and a 1981–2000 reference period, separate from its 2005–2018 selection period. For each calendar day it estimates a local 95th percentile using a ± 15-day neighborhood. A candidate must exceed this local percentile and an absolute $35^\circ\mathrm{C}$ threshold for at least three consecutive days. Land, latitude, proximity-to-administrative-region, per-location, and soft per-country constraints reduce concentration in a few places. The pipeline contributes 210 accepted and rendered instances. The climatology used to *mine labels* must not be confused with climatological features supplied to the reasoning model; the released reports mark local-seasonal diagnostics as unavailable.

Negative controls. The first ERA5 negative-sampling pipeline contributes 548 difficult controls after deduplication. Its sampling targets include weak low pressure without strong impacts, cloudy conditions without substantial rain, near-heatwave conditions, near-gale conditions, and moist cold conditions insufficient for the source's winter-storm criteria. The second pipeline samples high-latitude $5^\circ \times 5^\circ$ regions across 2000–2023 and filters on maximum precipitation no greater than 0.002 m, maximum ten-metre wind no greater than 10 m/s, and two-metre temperature between 243.15 and 308.15 K; 799 of its 800 metadata records render successfully. The remaining No Extreme instances come from the legacy and model-assisted components. These sources do not implement one universal definition of normal weather.

Model-assisted expansion. The final component combines ERA5-conditioned sampling with Qwen2.5-72B-assisted category and distractor generation (Qwen et al., 2024). It contributes 300 instances, including 229 No Extreme instances. We report this component explicitly because model-assisted labeling can induce source-specific preferences. Its contribution is evaluated separately in the source-stratified analysis.

B.3 Spatial-overlap deduplication

Candidates are normalized to a shared schema containing event metadata, time, location, coordinates, options, and answer. They must contain exactly four distinct options including No Extreme, with valid coordinate values. Within each calendar date and semantic class, the overlap ratio for a new box B_i and a previously retained box B_j is

$$r(B_i, B_j) = \frac{\mathrm{area}(B_i \cap B_j)}{\min\{\mathrm{area}(B_i), \mathrm{area}(B_j)\}}. \tag{3}$$

A later candidate is removed when $r \geq 0.60$. Processing order therefore affects which representative survives. This rule handles some within-source and cross-source duplicates, but it does not establish independence for adjacent dates, related regions with smaller overlap, or differently labeled manifestations of a compound event. We preserve the distinction between these 1,661 rule-based removals and the separate event-family audit in Section L.

B.4 TARGET-TIME PRECISION AND REGION EXTENT

Source metadata with a time of day retains that time. Date-only records use 12:00 UTC on the recorded date. Of the final instances, 4,301 have time-of-day metadata and 4,197 use the date-only noon convention. In the EM-DAT component, final target-time provenance comprises 2,766 exact-source dates, 528 dates corrected using an ERA5 peak, 84 weak-peak dates, 71 mid-month fallback dates, and four script-completed dates. These choices define reference targets; they do not guarantee that the target equals physical event onset.

Regions inherit their source bounding boxes. Median width and height are each 5°, with the fifth to ninety-fifth percentile spanning 2° to 10° for both dimensions. Consequently, a location name is a convenient descriptor rather than a statement that all report statistics refer to a point station or to an administrative polygon. Longitude conventions and box boundaries should be retained when reconstructing a case.

C METEOROLOGICAL EVIDENCE AND REPORT CONSTRUCTION

C.1 ONE CUTOFF SHARED BY EVERY EVIDENCE WINDOW

The meteorological fields come from ERA5 reanalysis (Hersbach et al., 2020). The windowing, aggregation, and reporting thresholds below are EXConan-specific construction choices, rather than definitions prescribed by the ERA5 data product. For a target time T, the evidence cutoff is $C = T - 24\,\mathrm{h}$. Every observational or reanalysis field used to render the report must have valid time at or before C. The report combines a short dynamic window, a seven-day background, and a thirty-day background:

$$\mathcal{W}_{\mathrm{dyn}} = [T - 72\,\mathrm{h}, T - 24\,\mathrm{h}], \tag{4}$$

$$\mathcal{W}_7 = (C - 7\,\mathrm{d}, C], \qquad \mathcal{W}_{30} = (C - 30\,\mathrm{d}, C]. \tag{5}$$

The dynamic window contains 49 hourly state timestamps including both endpoints. Nine anchors, A0–A8, sample these states at six-hour intervals. Eight nonoverlapping precipitation periods cover $(A_{j-1}, A_j]$, with exactly six hourly precipitation increments per period. This interval convention avoids counting an increment twice at a shared boundary.

The shared cutoff is central to comparability: recent evolution and longer antecedent context obey the same temporal boundary. It remains a *valid-time* boundary. ERA5 is a retrospective reanalysis, and the benchmark does not claim that the rendered report or all of its source fields were operationally available at the corresponding historical issuance time. Likewise, a forecast from cutoff fields is a model prediction and must remain distinguishable from later observed weather.

C.2 VARIABLES, UNITS, AND TRANSFORMATIONS

Table S2: Meteorological fields and reported transformations.

Field	**Level**	**Report quantity / unit**
Total precipitation	Surface	Hourly increments and 6/24/48-hour, 7/30-day accumulation; mm
Mean sea-level pressure	Surface	Pressure and sector gradients; hPa
Two-metre temperature	Surface	Temperature quantiles, extrema, and changes; °C
u_{10}, v_{10}	10 m	Wind speed, quantiles, and coverage; m/s
Air temperature	850, 500 hPa	Temperature and $T_{850} - T_{500}$; °C
Relative humidity	850, 700 hPa	Percentiles and threshold coverage; %
u, v wind	850 hPa	Wind speed, quantiles, and coverage; m/s
Geopotential	500 hPa	Geopotential height after dividing by 9.80665; gpm
Boundary-layer height	Surface diagnostic	Height distribution; m

Total precipitation is converted from metres to millimetres before interval accumulation. Surface and pressure-level temperatures are converted from kelvin to degrees Celsius; mean sea-level pressure is

converted from pascals to hectopascals. Horizontal wind speed is computed from its components as $s = \sqrt{u^2 + v^2}$. Geopotential is divided by $9.80665\,\mathrm{m\,s^{-2}}$ to obtain geopotential height. The reported $T_{850} - T_{500}$ difference is a thermal-structure diagnostic; it is not a substitute for a measured CAPE profile or a complete convective sounding.

C.3 Spatial aggregation and threshold coverage

The report is generated from the complete source box and uses cosine-latitude weighting for area-weighted quantities. A generic weighted mean is

$$\bar{x}(t) = \frac{\sum_{g \in B} \cos(\phi_g)\, x_g(t)}{\sum_{g \in B} \cos(\phi_g)}, \tag{6}$$

where g indexes grid cells and ϕ_g is latitude. Regional means, spatial quantiles, extrema, threshold coverage, and 3×3 sector summaries provide complementary views. A small grid-cell maximum may identify a localized feature while covering little of the region; a quantile or coverage statistic describes a different property. The report retains these distinctions so that a model can compare spatial extent with intensity.

Fixed report thresholds are 10 mm per six hours for precipitation, 15 m/s for ten-metre wind, 12 m/s for 850-hPa wind, and 80% for relative humidity at 850 and 700 hPa. A wet day is defined by at least 1 mm of area-mean daily precipitation. These are *reporting thresholds*: crossing the 850-hPa wind threshold does not, by itself, establish a surface gale; failing a rainfall threshold does not exclude a cold wave or drought. The success cases in Section J illustrate why this semantic distinction matters.

C.4 Temporal evolution and persistence

The report supplies recent six-hour, same-UTC-phase twenty-four-hour, and full forty-eight-hour changes. Same-phase comparisons help separate an evolving signal from a comparison made at different hours of the daily cycle. Antecedent summaries report rainfall accumulation, wet-day counts, dry-day fractions, consecutive dry days, temperature quantiles, moisture, and wind statistics. Taken together, these views support questions about strengthening circulation, an expanding precipitation footprint, persistence of cold conditions, and accumulated dry conditions.

These summaries are class-independent. A rainfall case and a drought case receive the same report structure rather than a label-specific selection of fields. The framework's specialists provide different interpretations of the same evidence, rather than receiving privileged observations selected using the correct label.

C.5 Report validation

Automated checks verify the 49-hour dynamic time axis, precipitation/state alignment, six increments per precipitation interval, sufficient grid coverage, finite core values, and the cutoff constraint. Additional checks reject unresolved template fields and prohibited answer-bearing content. The release records input file provenance and rendering validation information separately from the prompt. Validation establishes consistency with the intended input contract; it does not prove the correctness of every heterogeneous reference label or the causal adequacy of a report for a future event.

D Question Generation, Release Composition, and Quality Assurance

D.1 Four-option construction

Every instance contains four unique categories and includes No Extreme. For an extreme-event reference label, the candidate set contains that label, No Extreme, and two labels sampled without replacement from the remaining seven hazards. For a No Extreme reference, three distractors are sampled from the eight hazards. The four labels are shuffled to A–D. The instance-specific seed is derived from the first eight bytes of the SHA-256 digest of the string formed by `20260820:`

followed by the release identifier. This makes question reconstruction deterministic given the same identifier and taxonomy.

Answer letters are approximately balanced: A has 2,142 instances, B has 2,103, C has 2,125, and D has 2,128. A uniform random choice among the four presented candidates has an expected accuracy of 25%. The task is not a nine-way open-label classifier, because candidates restrict the decision for each example. Always selecting No Extreme would achieve 18.65% on the full release, reflecting that category's support. Candidate-only and order-control results are reported in Section H.

D.2 Visible and offline fields

The benchmark exposes target time, source location, bounding box, four candidate labels, and the class-independent weather report. Direct models return a JSON object with an answer letter and a concise explanation. The top-level answer, answer text, ground-truth object, and sampling metadata are offline fields. The EXConan inference adapter further removes the release identifier and the random A–D mapping, sorts semantic labels, and exposes internal candidate identifiers W1–W4. The offline mapping is restored after the final semantic decision.

The adapter separates a `BlindCase` from an `OfflineTruth` record and checks that prompt options agree with the top-level option mapping. This separation reduces accidental label or answer-position leakage within the inference pipeline. It does not establish that location, season, candidate composition, or report-generation artifacts contain no predictive shortcuts; those questions require empirical controls.

D.3 Final composition

Table S3: Canonical category supports in the final release.

Category	**Instances**	**Share (%)**
No Extreme	1585	18.65
Tropical Cyclone / Typhoon	1532	18.03
Drought / Dry Spell	1153	13.57
Severe Convection Storm	947	11.14
Extreme Rainfall	849	9.99
Heatwave	713	8.39
Winter Storm	712	8.38
Cold Wave	601	7.07
Strong Wind / Gale	406	4.78
Total	8498	100.00

The release spans 2000–2023, with 4,214 unique target timestamps and 3,753 calendar dates. Years are not uniformly represented: 2021 and 2022 contribute 954 and 933 instances, respectively, while 2023 contributes 105. The taxonomy is also imbalanced: No Extreme and Tropical Cyclone / Typhoon together account for more than one-third of the release. We therefore pair strict accuracy with per-class recall, macro-F1, and weighted-F1 rather than treating one metric as a complete characterization.

Geographic coverage is summarized using a documented proxy: the bounding-box center is mapped to the nearest city and its country code, then to a continental group; points south of 60°S are assigned to Antarctica. The resulting counts are Asia 4,174, North America 1,378, Europe 1,242, Africa 817, South America 340, Antarctica 316, and Oceania 231. These are approximate group assignments, especially for offshore or cross-border regions. They do not establish a precise count of independently verified countries of impact.

D.4 Development membership

The 950-instance development subset is the exact union of disjoint 100- and 850-instance subsets. All 950 identifiers appear in the full release. After canonical label normalization, each of the nine

classes has 105 or 106 development instances. The 100-instance component was used for early debugging and the larger development pool for framework and protocol development. Full-release results include these examples and are reported as full-benchmark evaluations, not as untouched held-out tests.

D.5 Automated and human quality assurance

The automated pipeline checks metadata structure, valid coordinates, option uniqueness, reference-answer consistency, input-window completeness, and prohibited leakage fields. Source harmonization and deterministic rendering make these checks reproducible, and release-level manifests allow failed records to be distinguished from model failures.

The available human-review materials describe a staged protocol with three student reviewers and expert escalation. The archived blind-review assignment contains 950 distinct cases, including 190 cases assigned to all three students; this yields 1,330 student assignments. Student packets withhold the reference answer and sampling label. The intended escalation categories include label/report disagreement, compound-event ambiguity, uncertain time boundaries, and unresolved student judgments.

E EXConan Implementation and Structured Output Contracts

E.1 Versioned implementations

E.2 Inference record and evidence identifiers

For each instance, the adapter retains the target metadata and weather report, removes the release identifier and original answer-letter ordering, and numbers nonempty evidence lines as E001, E002, and so on. It sorts the semantic candidates case-insensitively and assigns W1–W4. Evidence references are checked against the current case. The offline reference answer is attached only after inference for scoring.

Numbered evidence provides a concrete interface between observations and decisions. A specialist may claim support from one line, another may cite a contradiction, and the arbiter can inspect both references. A syntactically valid reference is not equivalent to a correct meteorological inference: the cited line can still be a threshold definition, an irrelevant quantity, or a field interpreted with the wrong spatial or temporal scope. The delivered traces preserve this distinction.

E.3 Deterministic diagnostic tools

The pressure-dynamics tool extracts minimum pressure, low-pressure sector tracks, robust sector contrasts, and six-, twenty-four-, and forty-eight-hour changes. The hydro-convection tool extracts the eight rainfall periods, their coverage statistics, the first and last twenty-four-hour accumulations, and moisture/thermal-structure lines. The persistence tool organizes seven- and thirty-day context and marks absent climatology.

Tool-augmented reasoning is related to the reasoning/action interface of ReAct (Yao et al., 2022). Here the tools are deterministic report organizers, rather than an open-ended environment interaction loop. These tools operate on the supplied report; they do not add observations or access the offline label. Their purpose is to make related evidence available in a compact structured form. In particular, a persistent-hazard specialist can examine multi-day conditions without equating a low short-period rainfall total with the absence of all hazards. The reliable-interface ablations in Section H test the end-to-end contribution of this organization.

E.4 Independent specialist roles

The use of multiple model perspectives is related to multiagent debate (Du et al., 2024); EXConan instead uses one independent pass per specialist before arbitration, without inter-specialist debate rounds. Three role-conditioned calls receive the same numbered report, diagnostic output, and semantic candidates. They execute in parallel with no access to one another's messages.

Synoptic-pattern specialist. This role identifies the weather-system mechanism before assessing disaster intensity. It compares cyclone, frontal/baroclinic, heat, drought, convection, and normal-weather explanations using circulation, pressure, moisture, and thermal structure. Weak impacts at the cutoff are not automatically evidence against a developing parent system.

Hazard-severity specialist. This role evaluates spatial coverage, robust extrema, rainfall, winds, moisture, and temperature. It separates the identity of a system from the severity of its individual effects. The intended comparison is between the evidence for each candidate, not between every event and one universal threshold.

Temporal-evolution specialist. This role examines strengthening, weakening, and persistence across six-, twenty-four-, and forty-eight-hour changes and seven-/thirty-day context. It can identify a developing event or a persistent regime that a single cutoff value would obscure. Missing climatology reduces certainty rather than proving normal weather.

E.5 Shared prompt principles and output validation

The archived current implementation makes the following instructions explicit: use only numbered evidence and deterministic tool output; treat missing evidence as uncertainty; avoid event-ID and answer-position inference; prefer spatial coverage, persistence, gradients, and evolution to one-grid extrema; distinguish tropical-cyclone systems from wind/rain impacts; require more than rainfall alone for convection; require persistence evidence for persistent hazards; actively test all alternatives before selecting No Extreme; and do not invent thresholds or observations.

Each specialist must return exactly one assessment for each of the four candidates. The fields are `candidate_id`, support in $[0, 1]$, supporting E-IDs, contradicting E-IDs, and a short list of missing evidence. A recommended candidate and a concise rationale accompany the ranking. In the newer interface, evidence references are restricted to the valid E-ID set and list/text lengths are bounded. If the declared recommendation is inconsistent with the largest support score, the validator normalizes it to the first maximum-support entry and records that normalization.

Why structured output matters. An agent pipeline has several opportunities to fail before making a final classification. Truncated JSON or invalid evidence identifiers can prevent a role from contributing, independently of whether the model has useful weather knowledge. The EXConan experiments therefore use stage-specific JSON schemas and bounded outputs. Their clean end-to-end coverage makes the new component comparisons more informative than older ablations dominated by schema failures.

E.6 Candidate evidence competition

Only valid, nonfallback specialist outputs contribute to the competition score. For candidate c and valid specialist set $\mathcal{V}$,

$$S(c) = \frac{1}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} s_k(c) - 0.15 \frac{1}{|\mathcal{V}|} \sum_{k \in \mathcal{V}} \mathbb{I}[\mathcal{E}_k^-(c) \neq \varnothing]. \tag{7}$$

The second term is a contradiction penalty per specialist, not a penalty proportional to the number of cited lines. Supporting roles require support of at least 0.5 and at least one supporting reference. Missing evidence remains available in the specialist records but is not assigned an additional numerical penalty. Complete ties are resolved deterministically by the descending tuple of net score, mean support, and candidate identifier. Since identifiers follow semantic sorting, this rule should be retained when reproducing a run.

E.7 Arbitration and fallback

The arbiter receives the original evidence, all specialist assessments, and the competition table. It returns a chosen candidate, runner-up, confidence, margin, decisive E-IDs, uncertainties, and a concise explanation. It is instructed to compare mechanisms and their impacts rather than apply a

majority vote mechanically. Its confidence and margin are self-reported quantities, not calibrated posterior probabilities.

If structured arbitration fails, the implementation falls back to a deterministic vote among valid specialist recommendations, with reduced reported confidence and zero margin. A failed specialist is represented by a neutral fallback and excluded from evidence aggregation and unanimous-No-Extreme checks. If no specialist produces a valid output, the instance is a terminal failure. The primary run records ten terminal failures; they count as incorrect in strict scoring.

E.8 Decision stress testing

The stress module is a deterministic inspection of the existing decision record. It constructs views of absolute extrema, recent changes, and antecedent context, then flags any of the following: specialist disagreement; at most one decisive evidence reference; decisive evidence from at most one report section; confidence below 0.70; margin below 0.15; an expert fallback; or a No Extreme decision for which some competing hazard lacks an explicit contradiction from a valid specialist.

It does not perturb ERA5 fields, generate synthetic alternative observations, or rerun the specialists under interventions. Its operational role is to identify a fragile decision for possible review. Consequently, a successful case should not be described as surviving an adversarial physical perturbation unless a separate perturbation experiment has actually been performed.

E.9 Conditional counterfactual review

The reviewer receives the original evidence, compressed specialist records, preliminary decision, competition table, stress diagnostics, and any available forecast summary. It identifies a strong alternative explanation and checks shared threshold bias and evidence consistency. A valid output either keeps the initial candidate or switches to a distinct challenger. For a keep action, the chosen candidate must equal the preliminary choice; for a switch action, the chosen candidate must equal the challenger. The current validator normalizes inconsistent action/challenger fields according to the selected semantic candidate.

The review stage is conceptually related to reflection-based language agents (Shinn et al., 2023), but does not implement episodic reinforcement, persistent reflection memory, or repeated environment trials. There is one review stage, with up to two structured-output attempts, rather than an unbounded iterative debate. A failed review preserves the valid preliminary answer. The current explicit triggers include disagreement, low confidence, small margin, No Extreme, persistent hazards with unavailable climatology, stress instability, and expert fallback. The primary Exconan traces preserve whether review was required but do not retain every trigger reason in the newer format.

E.10 Conditional Aurora countercheck

Aurora is an Earth-system foundation model (Bodnar et al., 2025). The primary configuration attempts Aurora-small only when all three valid specialists recommend No Extreme and capability checks allow the forecast path. Initial states are at $C - 6\,\mathrm{h}$ and C, and four six-hour steps advance the forecast to $T = C + 24\,\mathrm{h}$. Input fields include surface temperature, ten-metre wind, MSLP, static fields, and pressure-level temperature, wind, geopotential, and specific humidity.

The reviewer receives regional summaries of forecast temperature, pressure, winds, humidity, and thermal structure. This configuration does not provide predicted precipitation or CAPE. Forecast output is supplied only to the reviewer; it does not rerun the specialists or independently determine the hazard label. Missing initial fields, unavailable resources, or forecast failure are recorded as missing supplementary evidence rather than support for No Extreme.

E.11 Generation and transport configuration

Table S4: Primary settings.

Setting	Primary
Cohort	Full release, 8,498
Backbone	Qwen3-4B
Random seed	42
Specialist temperature	0.1
Arbiter / reviewer temperature	0.0 / 0.0
Maximum output tokens	1,800
Thinking mode	Disabled
Response constraint	JSON + validation
Stage output attempts	2
Confidence / margin gates	0.70 / 0.15
Contradiction penalty	0.15
Primary specialists	Three fixed roles
Aurora-small	Conditional, unanimous normal
Case concurrency	Two shards, one worker each
Code / model provenance	Incomplete historical revision

The contradiction penalty, confidence threshold, and margin threshold are fixed development heuristics. The identified archive contains no systematic sweep demonstrating that these values are optimal. Stage attempts are limited to two; the network layer uses a 600-second timeout and up to three HTTP retries with a two-second backoff. These controls describe the implementation and should not be confused with a measured guarantee of latency or reliability under a different backend.

F Evaluation Cohorts, Metrics, and Statistical Conventions

F.1 Distinct evaluation cohorts

We distinguish between four scenarios for evaluation and result recording. First, the 8,498-sample release set is used to compare the performance of all models and the proposed method, including local models, public API models, and a comparison between the proposed method and direct inference baselines. Second, Bench-950 is used exclusively for ablation studies to analyze the contribution of individual components to the method's performance.

F.2 Strict classification metrics

Let $\mathcal{I}$ be the selected identifier set, $N = |\mathcal{I}|$, y_i the canonical reference category, and $\hat{y}_i$ the predicted category or an invalid-output symbol. Strict accuracy is

$$\mathrm{Acc}_{\mathrm{strict}} = \frac{1}{N} \sum_{i \in \mathcal{I}} \mathbb{I}[\hat{y}_i = y_i]. \tag{8}$$

Failed, missing, and invalid predictions remain in N and contribute zero to the numerator. For each of the nine canonical categories, we compute precision, recall, and F1. Invalid predictions produce a false negative for their reference class but are not assigned as false positives to a valid category. A zero-denominator F1 is set to zero. Macro-F1 averages the nine class F1 values equally; weighted-F1 weights them by reference support. All percentage-point differences are calculated before rounding. Cohen's κ (Cohen, 1960), where tabulated, measures agreement beyond the agreement expected from the two marginal label distributions; undefined denominators are reported as inapplicable.

More explicitly, let $C = 9$ be the number of canonical categories, n_c the number of reference instances in category c, P_c its class-wise precision, and $F_{1,c}$ its class-wise F1 score. Precision is the fraction of predictions assigned to a category that match the reference label. Macro precision gives

every category equal weight, whereas weighted-F1 weights each category by its reference support:

$$\text{MacroPrecision} = \frac{1}{C}\sum_{c=1}^{C} P_c, \tag{9}$$

$$\text{WeightedF1} = \sum_{c=1}^{C} \frac{n_c}{N} F_{1,c}. \tag{10}$$

For Cohen's κ, let m_c denote the number of predictions assigned to category c. The observed agreement p_o and the agreement expected from the reference and prediction marginals p_e give

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \text{Acc}_{\text{strict}}, \qquad p_e = \sum_{c=1}^{C} \frac{n_c}{N}\frac{m_c}{N}. \tag{11}$$

Invalid outputs remain in N but are not counted in any m_c; the invalid-output category has zero reference support and thus contributes zero to p_e. When $p_e = 1$, κ is undefined. Accuracy, macro precision, and weighted-F1 are reported as percentages, while κ is reported on its original scale.

Success-only accuracy uses only valid predictions and is supplementary. It is always accompanied by the valid count. For instance, the primary agent has 3,396 correct predictions among 8,488 valid outputs, giving 40.01% success-only accuracy, while its strict accuracy is 39.96% over 8,498 selected cases. The latter is the primary number.

F.3 Binary extreme-event measures

Binary extreme recall counts an example as detected if its reference is any extreme category and its prediction is any extreme category. It does not require the correct hazard subtype. Invalid outputs count as missed detections. The false-alarm rate is the fraction of true No Extreme instances assigned an extreme category. These two measures describe a different tradeoff from nine-category accuracy and should be reported together. A system can increase extreme recall by predicting more hazards while also increasing false alarms.

For strata containing no true negative controls, false-alarm rate is undefined; for strata with no extreme examples, extreme recall is undefined. Such quantities are marked inapplicable rather than reported as zero. This prevents source composition from creating artificial differences.

F.4 Paired comparisons and denominators

Within a matched comparison, let b be the number of direct-wrong/agent-correct instances and c the number of direct-correct/agent-wrong instances. The strict paired change is $(b - c)/N$. We use the paired-discordance framework of McNemar (McNemar, 1947) with a two-sided exact test, equivalent to a binomial test on b successes among $b + c$ discordant pairs under probability $1/2$.

For compatibility with the existing main-text analysis, we also retain the common-valid-pair comparison, which discards pairs where either method is invalid. These two denominators must remain visible. On the primary run, strict scoring yields $b = 1{,}437$, $c = 1{,}139$, and a net gain of 298 over 8,498 cases. Restricting to 8,488 common-valid pairs yields $b = 1{,}437$, $c = 1{,}134$, and a net gain of 303. The five-prediction difference is explained by direct-correct cases on which the agent failed.

F.5 Intervals, seeds, and exploratory comparisons

We supply Wilson score intervals (Wilson, 1927) for individual accuracies and paired nonparametric bootstrap intervals (Efron, 1979) for differences. The additional paired intervals use 10,000 resamples with a fixed analysis seed of 20260925, treating an instance as the resampling unit. These intervals describe instance-level variation within the archive. They do not account for correlation among records from the same physical weather process because verified event-family identifiers are unavailable.

The three seeds measure generation variability under the frozen suite. Their sample standard deviation is not a confidence interval over future weather processes. Component and slice analyses are exploratory unless otherwise specified. We retain exact per-class paired tests and Holm adjustment (Holm, 1979) across the nine class comparisons in the machine-readable tables; broader collections of ablation p values are reported without a claim of confirmatory family-wise significance.

G Full-Benchmark and Matched-Backbone Results

G.1 Benchmark difficulty across direct models

Table S5: Completed local direct-model results on all 8,498 instances. Identifiers follow archived evaluation records.

Model	Correct	Fail	Acc.	Macro-F1
Qwen3-30B-A3B-Instruct-2507	4192	0	49.33	49.16
Mistral-Small-24B-Instruct-2501	4124	4	48.53	48.85
Phi-4	4012	0	47.21	46.62
Qwen3.8-27B-Instruct	3774	0	44.41	43.93
Qwen3.8-27B-Thinking	3699	0	43.53	43.21
Qwen3-30B-A3B-Thinking-2507	3439	0	40.47	37.15
Qwen3-4B-Thinking-2507	3403	0	40.04	–
Qwen3-8B	3198	0	37.63	31.74
Qwen3-4B	3098	0	36.46	34.41
Qwen3-14B	3094	0	36.41	32.72

The Qwen3-4B-Thinking-2507 row is a reported aggregate; its restricted raw file prevented class-metric recomputation. A dash denotes unavailable evidence, not zero.

The strongest completed, auditable full-release local direct result is 49.33% accuracy for Qwen3-30B-A3B-Instruct-2507. This leaves substantial room for improving the conversion of structured weather evidence into event decisions. The broad variation across model configurations also argues for measuring performance directly rather than assuming that model size or a reasoning designation determines weather-recognition quality.

Table S6: Archived API results on the same cohort.

Model	Correct	Fail	Acc.	Macro-F1
Gemini-3.6-Flash	4833	0	56.53	59.83
GPT-5.5	3789	0	44.32	47.02
Claude-Sonnet-5	3591	0	42.00	46.25
Claude-Opus-4.8	3565	0	41.68	45.97
GPT-5.4	3398	0	39.68	43.54

Both API models and local models were evaluated on the same set of 8,498 samples, using a consistent sample denominator and retaining archived model identifiers. Among the API models, Gemini-3.6-Flash achieved the highest accuracy at 56.53%. For the other local reasoning model, only aggregated results have been reported so far.

G.2 Primary same-backbone improvement

Table S7: Full-benchmark Qwen3-4B comparison. All 8,498 selected instances enter strict metrics.

Condition	Valid	Correct	Acc.	Macro-F1	W-F1
Direct	8498	3098	36.46	34.41	34.74
EXConan	8488	3396	39.96	33.61	36.79

EXConan produces 3,396 correct predictions compared with 3,098 for direct Qwen3-4B on the same 8,498 instances. The resulting 3.51-point strict accuracy gain corresponds to 298 additional

correct decisions. Its exploratory paired 95% bootstrap interval is $[2.34, 4.67]$ points, with a strict-pair exact McNemar $p = 4.68 \times 10^{-9}$. On the 8,488 common-valid pairs, the gain is 3.57 points and $p = 2.48 \times 10^{-9}$.

This improvement is useful at the event-decision level: the model backbone and task instances are held fixed, and the framework makes the intermediate interpretations inspectable. It is not a same-compute comparison. The full class profile is essential because accuracy gains concentrate in some hazards and strict macro-F1 decreases from 34.41% to 33.61%.

G.3 Where the correction capacity is strongest

Table S8: Complete hazard-level results on the 8,498-instance benchmark. D and E denote Direct and EXConan.

Hazard	n	D rec.	E rec.	Δ pp	D F1	E F1
Cold wave	601	18.14	31.61	+13.48	25.32	36.61
Drought	1153	9.80	5.20	-4.60	16.92	9.43
Extreme rainfall	849	88.93	88.81	-0.12	53.02	54.58
Heatwave	713	36.04	4.91	-31.14	44.23	9.19
No Extreme	1585	52.87	58.68	+5.80	35.66	34.92
Severe convection	947	22.07	33.16	+11.09	28.91	39.85
Strong wind	406	49.75	7.14	-42.61	25.88	11.05
Tropical cyclone	1532	25.78	55.94	+30.16	40.02	64.90
Winter storm	712	30.90	31.88	+0.98	39.71	41.96

The largest recall gains occur for Tropical Cyclone / Typhoon, Cold Wave, and Severe Convection Storm: +30.16, +13.48, and +11.09 percentage points, respectively. These are categories for which a weather-system interpretation or temporal context can matter beyond one impact threshold. The selected examples later in this supplement show concrete corrections involving those distinctions.

No Extreme recall also improves from 52.87% to 58.68%, demonstrating that the framework can correct some false alarms as well as recover selected missed events. The profile does not support uniform improvement: heatwave, drought, and strong-wind recall deteriorate, and these failure modes are discussed in Section L. The figure below and the complete confusion matrices retain both directions of change.

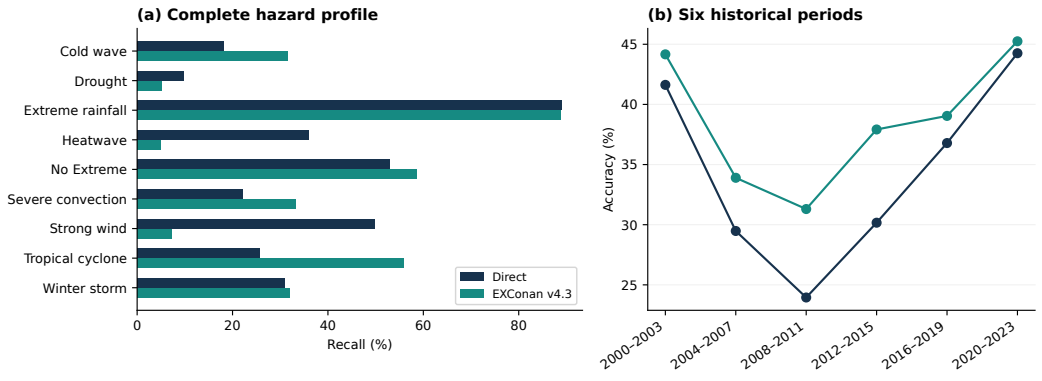


Figure S1: Full-release Qwen3-4B analysis. Left: every hazard category is shown, including deteriorations. Right: accuracy differences are positive in each of six four-year periods. These are retrospective within-benchmark strata, not held-out temporal tests.

G.4 Stronger-backbone corroboration on a matched cohort

The completed Qwen3-30B-A3B-Instruct-2507 experiment contains 8498 selected cases. Direct inference is correct on 4211, and EXConan with conditional Aurora-small is correct on 4672. Strict accuracy therefore increases from 49.26% to 54.63%, or +5.37 points. The agent has 8298 valid outputs and 200 failures. Strict macro-F1 increases from 50.47% to 53.39%.

Among the 8298 common-valid pairs, EXConan corrects 1476 direct errors and introduces 927 errors, for a net gain of 549 and exact McNemar $p = 2.27 \times 10^{-4}$. The strict comparison additionally retains the 200 failures. This provides a second matched-backbone result in favor of the framework, while the broader cross-backbone record in Section L defines the limits of generalization.

G.5 Distinguishing final results from intermediate snapshots

The primary comparison results for all models are based on the complete release set of 8,498 samples. The Qwen3-4B agent achieved a final strict accuracy of 39.96%. On the same evaluation set, Phi-4's accuracy rose from 27.73% with direct inference to 38.44% using the agent-based approach—an improvement of 10.71 percentage points.

H Frozen EXConan Experiments and Controlled Comparisons

H.1 A fixed interface for mechanism analysis

The newer suite freezes Qwen3-4B, the Bench-950 identifier set, the input hash, schema-constrained output. It includes 21 completed conditions: three full-system seeds, four component ablations, three additional stress/review factorial cells, four inference strategies including direct, and seven additional input/order controls. All comparisons below preserve the suite's own direct prompt rather than importing the older 37.68% direct result.

The full-system runs have no terminal invalid predictions. They record one, two, and one cases with an internal stage error for seeds 42, 43, and 44, respectively, with no expert fallback. This materially changes the interpretability of ablations relative to older experiments in which removing a component also produced hundreds of invalid outputs. The archived manifests and compact per-instance predictions accompany every condition.

H.2 Stability over three generation seeds

Strict accuracy is 36.95%, 37.16%, and 38.32% for seeds 42, 43, and 44. The mean and sample standard deviation are $37.47 \pm 0.74\%$, compared with 31.68% for the frozen direct protocol. All three seeds improve on that comparator, and the mean gain is 5.79 percentage points. The corresponding macro-F1 values are 35.68%, 35.47%, and 37.12%, compared with 24.26% for the frozen direct condition.

These results provide a more specific form of robustness evidence than a single favorable generation: the direction of the accuracy improvement is consistent across three seeds under a documented configuration. This demonstrates, to a certain extent, the system's invariance with respect to different prompts, backbones, data distributions, or weather process samplings.

H.3 Inference strategy and budget controls

Table S9: Frozen EXConan Bench-950 comparisons with measured inference budgets. Aurora is disabled.

Condition	Acc.	Macro-F1	Fail	Calls	Tokens
Direct (frozen prompt)	31.68	24.26	0	1.00	4.9k
Self-consistency, 4 samples	30.74	23.90	1	4.00	19.8k
Three-role voting	33.58	27.42	0	3.00	15.0k
Single-agent tools, 4 votes	30.00	21.93	0	4.00	32.4k

Condition	Acc.	Macro-F1	Fail	Calls	Tokens
EXConan (seed 42)	36.95	35.68	0	4.73	37.4k
EXConan (seed 43)	37.16	35.47	0	4.73	37.4k
EXConan (seed 44)	38.32	37.12	0	4.73	37.4k

The self-consistency control follows the principle of sampling multiple reasoning paths and aggregating their answers (Wang et al., 2023), with four samples and the benchmark-specific output contract used here. Four-sample self-consistency reaches 30.74%, three-role voting reaches 33.58%, and a single-agent strategy with the same deterministic tools and four votes reaches 30.00%. The full seed-42 system reaches 36.95%. Thus, simply making several calls or adding the diagnostic tools to a voting baseline does not recover the full system's accuracy in this suite.

The budget table also shows why the comparison should be described carefully. Full EXConan uses approximately 4.73 successful completion calls and 37.43 thousand total tokens per valid case. The four-vote single-agent tool baseline uses 4.00 calls and 32.39 thousand tokens. The full system therefore uses about 15.6% more total tokens than this closest tool-enabled comparator. Four-sample self-consistency uses about 19.76 thousand tokens, and one direct call about 4.94 thousand.

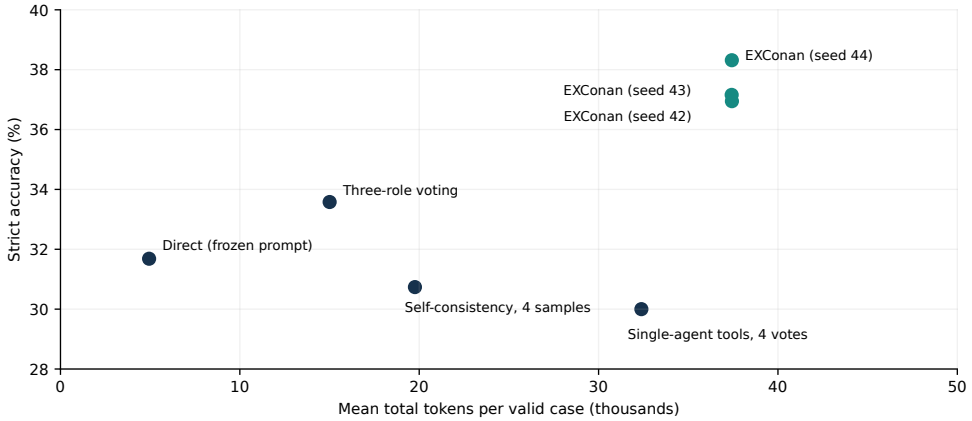


Figure S2: Measured total-token use and strict accuracy for the frozen Bench-950 inference strategies. Total tokens include prompts and completions. Nearby full-system points are distinct seeds. The plot makes the extra budget visible alongside the accuracy gain.

Table S10: Paired improvements of EXConan (seed 42) over each comparator. Invalid outputs count as wrong. Intervals are exploratory instance-level bootstrap intervals.

Comparator	Δ pp	95% CI	Repair	Harm	p
Direct (frozen prompt)	+5.26	[1.68, 8.84]	177	127	0.00487
Self-consistency, 4 samples	+6.21	[2.63, 9.68]	179	120	0.000768
Three-role voting	+3.37	[-0.32, 6.95]	171	139	0.0781
Single-agent tools, 4 votes	+6.95	[3.37, 10.53]	186	120	0.000192
Without arbiter	+6.21	[3.37, 9.16]	133	74	4.98e-05
Single generalist	+4.42	[1.37, 7.47]	135	93	0.00649
Without competition	+2.42	[-0.11, 4.95]	88	65	0.075
Without diagnostics	+4.63	[1.68, 7.68]	130	86	0.00334

The full system exceeds the four-sample self-consistency and four-vote single-agent-tool controls with exact paired $p = 7.68 \times 10^{-4}$ and $p = 1.92 \times 10^{-4}$, respectively. Its 3.37-point advantage over three-role voting is directionally positive but has $p = 0.0781$ and an exploratory interval crossing zero.

H.4 SPECIALISTS, TOOLS, ARBITRATION, AND COMPETITION

Table S11: Reliable-interface EXConan component ablations on the same 950 cases.

Condition	Acc.	Macro-F1	Fail	Calls	Tokens
EXConan (seed 42)	36.95	35.68	0	4.73	37.4k
Without arbiter	30.74	28.30	0	4.00	32.2k
Single generalist	32.53	30.54	0	2.59	18.0k
Without competition	34.53	34.09	0	4.83	37.3k
Without diagnostics	32.32	29.28	0	4.77	28.4k

Replacing the specialist panel with one generalist reduces accuracy by 4.42 points, while removing diagnostic tools reduces it by 4.63 points. Because both conditions retain complete valid coverage, these differences are not explained by the large terminal-failure imbalance seen in the older ablations. They support the combined value of role differentiation and explicit organization of the supplied weather evidence in this configuration.

Removing the LLM arbiter reduces accuracy by 6.21 points, from 36.95% to 30.74%, with exact paired $p = 4.98 \times 10^{-5}$. This provides useful end-to-end evidence for the arbitration stage under the reliable EXConan interface. The removal also changes the preliminary decision presented to downstream review, so the effect belongs to the complete pipeline rather than an isolated claim about the arbiter's intrinsic reasoning.

Removing evidence competition reduces accuracy by 2.42 points, but its exact paired $p = 0.0750$ does not establish a conventional 0.05-level difference. The candidate-level table remains useful for trace inspection, while its independent accuracy contribution is less strongly supported than those of diagnostics, specialists, and the arbiter. The opposite arbiter direction in older interfaces is discussed explicitly in Section L.

H.5 INPUT EVIDENCE AND OPTION-ORDER CONTROLS

Table S12: Input-view and answer-order controls using the frozen direct protocol.

Condition	Acc.	Macro-F1	Fail	Calls	Tokens
Direct (frozen prompt)	31.68	24.26	0	1.00	4.9k
Candidates only	22.11	15.24	0	1.00	0.4k
Metadata + candidates	32.00	25.24	0	1.00	0.8k
Evidence + candidates	27.58	18.70	0	1.00	4.7k
Within-section shuffled	30.95	23.39	0	1.00	5.0k
Option rotation 1	31.47	24.17	0	1.00	4.9k
Option rotation 2	31.26	23.99	0	1.00	4.9k
Option rotation 3	31.79	24.45	0	1.00	4.9k

Candidates alone yield 22.11%, while complete direct input yields 31.68%. Within-section evidence shuffling yields 30.95%. Three fixed, nonidentity option rotations yield 31.47%, 31.26%, and 31.79%, a narrow 0.53-point range. These controls probe different issues: candidate priors, the use of report information, and sensitivity to the mapping between semantic choices and output letters.

The option controls do not establish that all choices are semantically invariant, nor do they replace a nine-label evaluation. The metadata-plus-candidates condition reaches 32.00%, and evidence-plus-candidates without location/time reaches 27.58%. This is an important diagnostic outcome: the direct model can exploit information in metadata, and the experiment does not establish that all of its full-input accuracy is grounded in the atmospheric report. These controls test the frozen direct protocol, not a full input-ablation matrix of EXConan itself.

I Regional, Temporal, and Source-Stratified Analyses

I.1 Analysis scope

All strata in this section use the same full-release Qwen3-4B direct and EXConan predictions. No model is rerun, no stratum receives a different prompt, and invalid outputs remain incorrect. These analyses identify where the aggregate gain occurs and expose differences that would be lost in one score. They describe the existing benchmark rather than new held-out evaluations.

I.2 Geographic groups

Table S13: Full-benchmark accuracy stratified by region.

Group	n	Direct	EXConan	Δ pp
Africa	817	19.71	18.85	-0.86
Antarctica	316	38.29	50.00	+11.71
Asia	4174	40.30	45.30	+5.01
Europe	1242	43.08	40.74	-2.33
North America	1378	34.25	41.80	+7.55
Oceania	231	35.93	29.00	-6.93
South America	340	12.94	12.94	+0.00

Accuracy improves in Asia, North America, and Antarctica by 5.01, 7.55, and 11.71 points, respectively. Asia contributes the largest number of instances and 209 additional correct predictions; North America contributes 104. These results show that the gain is not confined to one of these geographic groups. Nevertheless, Africa, Europe, and Oceania decline, while South America is unchanged. The complete table is retained because global coverage and uniformly positive regional transfer are different claims.

The groups use the geographic proxies described in Section D.3. Their class and source composition differ. For example, a region containing more sampled negative controls can have a different error profile from a region dominated by documented cyclones. A formal geographic-generalization study would require a frozen region-disjoint split and appropriate event-family and spatial buffers.

I.3 Six historical periods

Table S14: Full-benchmark accuracy stratified by period.

Group	n	Direct	EXConan	Δ pp
2000–2003	1182	41.62	44.16	+2.54
2004–2007	882	29.48	33.90	+4.42
2008–2011	952	23.95	31.30	+7.35
2012–2015	1200	30.17	37.92	+7.75
2016–2019	1862	36.79	39.04	+2.26
2020–2023	2420	44.26	45.25	+0.99

The accuracy difference is positive in all six consecutive four-year periods spanning 2000–2023. Gains range from 0.99 points in 2020–2023 to 7.75 points in 2012–2015. The latter period increases from 30.17% to 37.92%; 2008–2011 increases from 23.95% to 31.30%. The result demonstrates broad temporal coverage of the observed aggregate benefit rather than reliance on one short date range.

The direction in every bin should not be confused with statistical significance in every bin. Some period-specific exploratory intervals include zero. Moreover, no chronological training/development boundary was imposed for this analysis, and source composition changes over time. It is therefore

a retrospective consistency check, not a demonstration of forecasting generalization to previously unseen years.

I.4 SOURCE COMPONENTS

Table S15: Full-benchmark accuracy stratified by source.

Group	n	Direct	EXConan	Δ pp
EM-DAT disaster records	3453	26.44	35.77	+9.33
ERA5 hard-negative mining (xhc)	548	60.77	62.41	+1.64
ERA5 heatwave threshold mining	210	31.43	5.24	-26.19
ERA5 high-latitude negative mining (yjh)	799	47.18	53.44	+6.26
ERA5 sampling + Qwen-assisted labels (zll)	300	47.33	66.00	+18.67
SPEI drought threshold mining	679	9.72	7.51	-2.21
baseline	2509	47.87	45.12	-2.75

The EM-DAT component increases from 26.44% to 35.77%, a 9.33-point improvement over 3,453 instances. This component supplies 322 net additional correct predictions. The high-latitude negative-sampling component improves by 6.26 points, and model-assisted expansion improves by 18.67 points over 300 instances. The smaller difficult-negative component improves by 1.64 points.

These differences sharpen the interpretation of the system's strengths: gains are substantial in the historical disaster component and some negative-control sources. They also expose source dependence. Dedicated heatwave mining, SPEI drought mining, and the legacy component decline. Source and hazard are strongly coupled, so these observations are not independent causal effects of source provenance. Complete source-by-label counts are included in the reproducibility materials.

I.5 CONFUSION STRUCTURE

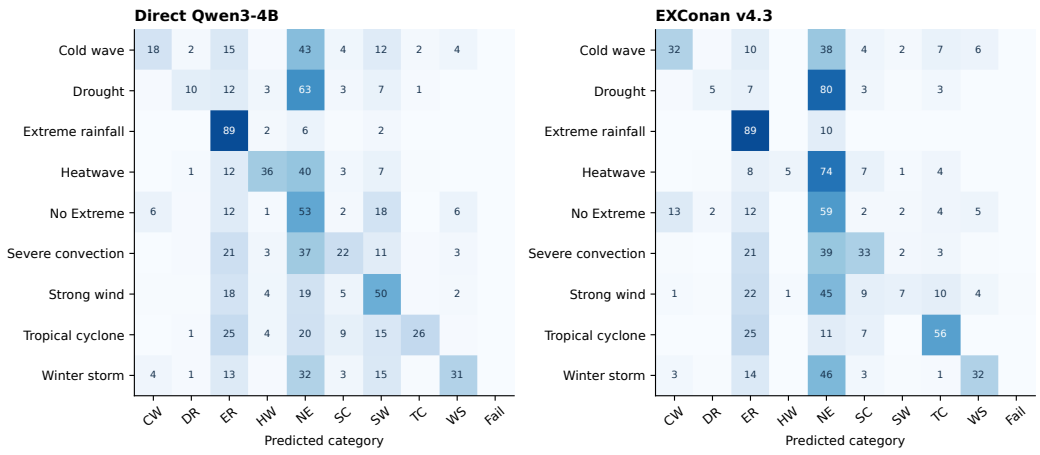


Figure S3: Row-normalized full-release confusion matrices. Rows are reference categories; columns are predictions, including a terminal-failure column. CW: cold wave; DR: drought; ER: extreme rainfall; HW: heatwave; NE: No Extreme; SC: severe convection; SW: strong wind; TC: tropical cyclone; WS: winter storm. Values below 1% are omitted from annotations but remain in the supplied counts.

The matrices complement the six successful examples by retaining the complete distribution of mistakes. Improved cyclone and convection recognition coexists with increased confusion toward No Extreme for some persistent hazards. Both the count matrices and per-instance transitions are supplied, enabling alternative cost-sensitive analyses without rerunning the models.

J Six Verified Corrections of Direct Inference

J.1 Selection and interpretation

Every case below meets four verifiable conditions: the archived direct Qwen3-4B prediction is wrong; EXConan predicts the reference label; no stage error is recorded; and no neutral expert fallback is used. The cases come from the same full-benchmark run as Table S7. We select six categories to illustrate different decision errors, rather than estimate a success rate from a handpicked sample.

There are 692 corrections satisfying the clean-trace filter in the archived run: 319 tropical-cyclone cases, 94 severe-convection cases, 79 cold-wave cases, 57 No Extreme cases, 54 winter-storm cases, 46 extreme-rainfall cases, 32 drought cases, ten heatwave cases, and one strong-wind case. Six are selected for exposition. These counts characterize the selection pool, not the number of corrections attributable to an individual module. The original report, direct output, semantic option mapping, specialist outputs, preliminary decision, and final decision are preserved for each selected identifier.

The evidence discussion below is an author interpretation checked against the report. It is not presented as a verbatim model reasoning trace or as proof that every generated sentence is scientifically correct. "Success" means that the final answer matches the archived reference category under the same supplied candidates.

Table S16: Six purposively selected corrections from the full-benchmark Qwen3-4B run. All cases have valid expert traces with no recorded stage errors or expert fallback.

ID	Reference / EXConan	Direct error	Aurora
event_03348	Tropical cyclone	Strong wind	No
event_04095	Severe convection	No Extreme	No
event_00331	Cold wave	Strong wind	No
event_00182	Winter storm	Strong wind	No
event_01241	Extreme rainfall	No Extreme	No
event_08400	No Extreme	Strong wind	Yes

J.2 Case 1: Recognizing a tropical-cyclone system beyond its wind impact

Instance and outcome. `event_03348` targets South Jeolla Province, Republic of Korea, at 2019-10-02 12:00 UTC. Its cutoff is 2019-10-01 12:00 UTC and its source box is 29.89–39.89°N, 122.00–132.00°E. The options are No Extreme, Winter Storm, Strong Wind / Gale, and Tropical Cyclone / Typhoon. Direct inference selects Strong Wind / Gale; EXConan selects the correct Tropical Cyclone / Typhoon category.

Observed pre-cutoff evidence. The report contains a 992.40-hPa absolute MSLP minimum. Across the forty-eight-hour anchor interval, the lowest-sector pressure decreases by 8.22 hPa, ten-metre wind p95 increases by 10.60 m/s, and 850-hPa wind p95 increases by 19.18 m/s. At the final anchor, the wind p95 values are 15.28 and 24.74 m/s at the surface and 850 hPa, respectively; 31.46% of the region exceeds the reporting threshold of 12 m/s at 850 hPa. The final six-hour rainfall period has a 78.04-mm grid maximum and 15.98% coverage above 10 mm.

Why direct inference fails. The direct explanation focuses on high wind speeds and treats the wind-impact category as the answer. That accounts for a visible feature but does not fully compare the parent-system alternative. Some quantities are also conflated in its prose, making it important to reconstruct values from the source report instead of repeating all numbers in the generated explanation.

How EXConan resolves the comparison. The synoptic specialist also prefers Strong Wind / Gale, while the severity and temporal specialists prefer Tropical Cyclone / Typhoon. Their disagreement is retained in the record. The arbiter chooses the cyclone category with reported confidence 0.95 and margin 0.20, drawing on pressure, winds, and evolving conditions. Review is invoked and keeps that choice; Aurora is not used. Thus, the correction is present at arbitration, before review.

Demonstrated strength. The case illustrates the value of comparing a weather system with one of its impacts. The useful behavior is the integration of pressure evolution, wind strengthening, and the precipitation footprint into the offered cyclone hypothesis. It does not establish that a reviewer repaired the answer or that a forecast model was necessary.

J.3 Case 2: Recovering severe convection from an inconsistent No Extreme judgment

Instance and outcome. `event_04095` targets Opanayaka, Sri Lanka, at 2011-11-24 12:00 UTC, with cutoff 2011-11-23 12:00 UTC. The candidates are Winter Storm, Severe Convection Storm, No Extreme, and Tropical Cyclone / Typhoon. Direct inference selects No Extreme; EXConan recovers the reference Severe Convection Storm label.

Observed pre-cutoff evidence. Seven-day area-mean rainfall is 70.93 mm with seven wet days; the thirty-day total is 288.92 mm. Seven-day median relative humidity is 85.89% at 850 hPa and 75.71% at 700 hPa. Rainfall increases from 4.67 mm in the first twenty-four hours of the dynamic window to 19.98 mm in the last twenty-four hours. The final six-hour period has 7.90 mm of area-mean precipitation, a 34.11-mm grid maximum, and 32.03% coverage above 10 mm. At the last anchor, RH850 and RH700 medians are 90.66% and 93.22%.

Why direct inference fails. The baseline states that the six-hour threshold was not met while citing 32.03% coverage above that same threshold. This is a directly inspectable inconsistency. It additionally treats the lack of surface winds above 15 m/s as support for dismissing the event, without a sufficiently specific comparison of the convection and cyclone alternatives.

How EXConan resolves the comparison. The synoptic and severity specialists choose Severe Convection Storm; the temporal specialist proposes a tropical cyclone. Arbitration selects severe convection with confidence 0.80 and margin 0.15, using the moisture-rich environment and accelerating rainfall while comparing the alternative circulation interpretation. The reviewer retains the answer, and Aurora is not used.

Demonstrated strength. The case shows a correction of a premature normal-weather decision and an inconsistent reading of threshold coverage. It also illustrates explicit comparison among plausible active-weather hypotheses. Rainfall and moisture alone do not constitute a complete physical validation of severe convection, and the report does not provide a CAPE profile; the success claim is the verified benchmark classification, supported by the available evidence and candidate comparison.

J.4 Case 3: Distinguishing sustained cold from an upper-air wind signal

Instance and outcome. `event_00331` targets Ottawa at 2018-01-01 12:00 UTC, with cutoff 2017-12-31 12:00 UTC. The choices are Heatwave, Strong Wind / Gale, Cold Wave, and No Extreme. The baseline predicts Strong Wind / Gale; all three EXConan specialists and the final system choose Cold Wave, matching the reference.

Observed pre-cutoff evidence. Seven-day two-metre temperature p05, median, and p95 are -28.05, -19.17, and $-5.05°\mathrm{C}$, respectively. The corresponding thirty-day median is $-8.88°\mathrm{C}$. The dynamic-window median is $-21.84°\mathrm{C}$, and its minimum is $-32.01°\mathrm{C}$. Surface wind p95 is only 4.59 m/s over the forty-eight-hour summary, with a 10.95-m/s maximum and zero coverage above 15 m/s. By contrast, 850-hPa wind reaches a 19.59-m/s maximum and broad coverage above the upper-air reporting threshold.

Why direct inference fails. The direct model emphasizes the 850-hPa wind and interprets it as decisive support for a gale category. This conflates an upper-air circulation feature with a surface-impact judgment and underweights the sustained cold regime. The evidence for low surface temperatures is not a single isolated grid minimum: it appears in the multi-day distribution and recent window.

How EXConan resolves the comparison. The three specialists converge on Cold Wave. The preliminary reported confidence is 0.8333 with margin 0.6667. Review is invoked and keeps the cold-wave label; no forecast is used. The temporal specialist's attention to seven-/thirty-day context supplies a useful complementary view to the wind-focused baseline.

Demonstrated strength. This example illustrates the combination of vertical-level distinctions and persistence evidence. Its defensible interpretation is sustained absolute cold relative to the other supplied hypotheses. The report has no local-seasonal anomaly calculation, and some generated trace sentences overstate anomaly evidence or repeat statistics. We retain that limitation rather than equating a correct answer with a perfectly grounded explanation.

J.5 Case 4: Identifying a winter-storm system before the target time

Instance and outcome. `event_00182` targets New York City at 2017-02-09 12:00 UTC, using a cutoff of 2017-02-08 12:00 UTC. The options are Strong Wind / Gale, Winter Storm, No Extreme, and Cold Wave. Direct inference selects Strong Wind / Gale. EXConan correctly selects Winter Storm without invoking either reviewer or Aurora.

Observed pre-cutoff evidence. The report contains a 993.38-hPa absolute minimum MSLP. The lowest-sector mean falls by 16.83 hPa across the forty-eight-hour anchor interval and by 12.32 hPa over the same-phase twenty-four-hour comparison. The 850-hPa wind p95 is 27.88 m/s in the dynamic-window summary, with a maximum of 33.81 m/s and complete regional coverage above 12 m/s at some times. Surface wind p95 is 12.14 m/s, and maximum coverage above 15 m/s is only 0.25%. Seven-day two-metre temperature p05 is $-7.28^{\circ}\mathrm{C}$, while the recent regional temperature distribution contains both cold and mild values.

Why direct inference fails. The baseline identifies the strong wind signal but selects the generic impact category. It does not give comparable weight to the evolving pressure system and thermal/moisture context. The intended task is a pre-target category decision, so the system can be relevant before all target-time impacts are present in the observations.

How EXConan resolves the comparison. All three roles select Winter Storm. The arbiter chooses it with confidence 0.80 and margin 0.50, citing pressure and circulation evolution among its decisive evidence. The valid preliminary choice is returned directly because the archived path does not require review. This distinguishes a correction due to the specialist/arbitration path from one produced by a later reviewer.

Demonstrated strength. The useful behavior is joint interpretation of circulation and context instead of equating the strongest upper-air wind with the final category. This case does not establish observed snowfall, precipitation phase, or a falling surface temperature at the target; the report includes pre-cutoff warming as well. The reference match is therefore reported as a successful event-category decision, not as an independently verified snow forecast.

J.6 Case 5: Correctly reading an expanding heavy-rainfall footprint

Instance and outcome. `event_01241` targets Qinzhou City at 2021-06-30 12:00 UTC, with cutoff 2021-06-29 12:00 UTC. The candidates are Strong Wind / Gale, Extreme Rainfall, No Extreme, and Cold Wave. Direct inference returns No Extreme, while all specialists and EXConan return the reference Extreme Rainfall label.

Observed pre-cutoff evidence. The seven-day area-mean rainfall is 57.94 mm with seven wet days, and the thirty-day total is 222.86 mm. Rainfall increases from 4.54 mm in the first twenty-four hours to 10.49 mm in the last twenty-four hours. The final six-hour period has an area mean of 3.51 mm, spatial p95 of 12.24 mm, maximum of 23.64 mm, and 10.71% regional coverage above 10 mm. The report also records maximum moisture coverage of 69.52% at 850 hPa and 91.43% at 700 hPa.

Why direct inference fails. The direct explanation states that no region exceeded the six-hour rainfall threshold in the final period, despite the explicit 10.71% coverage and a spatial p95 above

10 mm. It appears to treat a low regional mean as equivalent to an absent localized heavy-rainfall footprint. The error is inspectable against the report rather than inferred only from the wrong answer.

How EXConan resolves the comparison. All three specialists choose Extreme Rainfall. The arbiter returns the same category with confidence 0.9833 and margin 0.6167, citing the rainfall evolution and supporting report lines. The decision follows the stable path with no reviewer and no Aurora. The system therefore corrects the baseline while avoiding an unnecessary additional review stage in this case.

Demonstrated strength. The case illustrates consistency between spatial coverage, quantiles, antecedent wetness, and temporal growth. It is a successful recovery of the archived rainfall category, while the aggregate rainfall recall remains almost unchanged between methods. We include it to explain one concrete correction, not to claim that rainfall is the largest source of the system-wide gain.

J.7 Case 6: Rejecting a false gale alarm with consistent forecast counterevidence

Instance and outcome. `event_08400` targets the Madagascar Highland at 2018-07-10 00:00 UTC, with cutoff 2018-07-09 00:00 UTC. The candidates are No Extreme, Cold Wave, Strong Wind / Gale, and Heatwave. Direct inference selects Strong Wind / Gale. All three specialists, the arbiter, and the final reviewer select the reference No Extreme answer.

Observed pre-cutoff evidence. The baseline emphasizes upper-air winds: the dynamic-window 850-hPa wind p95 is 16.89 m/s with substantial coverage above 12 m/s. Surface impacts are weaker: ten-metre wind p95 is 6.56 m/s with a maximum of 9.35 m/s. All eight six-hour rainfall periods have zero coverage above 10 mm, and the largest six-hour area-mean accumulation is 0.61 mm. At the cutoff, ten-metre and 850-hPa wind p95 are 5.09 and 15.16 m/s, respectively.

How EXConan resolves the comparison. The specialists distinguish the upper-air wind feature from strong surface-wind impacts and unanimously select No Extreme. Their valid unanimity activates the conditional Aurora-small path. The arbiter's preliminary choice is already No Extreme, with reported confidence and margin both 1.0. These self-reported values are not a calibration guarantee.

Supplementary forecast evidence. The archived rollout predicts ten-metre wind p95 declining from 5.09 to 2.32 m/s and 850-hPa wind p95 declining from 15.16 to 5.60 m/s by the target time. Minimum MSLP rises from 1019.36 to 1020.28 hPa. These values are model forecasts initialized from pre-cutoff fields, not observed target-time measurements. The reviewer receives the weakening-wind and rising-pressure summaries and retains No Extreme.

Demonstrated strength. This is a verified correction of a false alarm and an example of conditional forecast corroboration. Aurora does not change the answer: the preliminary decision was already correct. The case therefore supports the inspectability of the countercheck pathway and its consistency with the final decision, while the aggregate causal contribution of Aurora remains a separate experimental question.

K Reviewer Transitions, Forecast Availability, and Cost

K.1 Primary EXConan execution record

Of 8,498 selected cases, the primary agent returns 8,488 valid outputs and ten terminal failures. Among the valid cases, 6,738 require review, equivalent to 79.38%. Of these requested reviews, 2,961 produce valid structured outputs and 3,777 fail validation. None of the valid reviewer outputs change the preliminary candidate. Both preliminary and final predictions contain 3,396 correct answers.

This trace-level fact determines how the primary success cases can be described. When review keeps an already-correct answer, it documents a retained decision, not a reviewer repair. The primary

EXConan accuracy gain is already present before the final reviewer output. Review design can be described as part of the system, but its realized contribution in this run cannot be credited with the gain.

The archived component accounting contains 23,102 valid specialist outputs and 2,362 specialist fallbacks across successful cases, with 86 recommendation normalizations. Arbitration falls back twice. Decision stress testing marks 6,583 successful cases as unstable. These counts demonstrate that a final valid answer can coexist with internal failures, which is why terminal validity and stage reliability are reported separately.

K.2 Reviewer behavior with the reliable EXConan interface

Table S17: Within-run reviewer transitions for the three frozen EXConan seeds. These are observed trace transitions, not randomized causal effects.

Seed	Reviewed	Switched	Repair	Harm	Net
42	694	456	122	125	-3
43	692	464	124	118	6
44	693	462	132	120	12

Seed 42 requests and completes 694 reviews, producing 456 changes, 122 wrong-to-correct repairs, and 125 correct-to-wrong changes. Seeds 43 and 44 produce net repairs of +6 and +12, respectively. The sign is not consistent across seeds, even though full-system accuracy improves over the frozen direct baseline for all three.

The net transition counts answer a narrower question than the full factorial: what happened between preliminary and final answers within a particular run? They are not randomized causal effects, because reviewed cases are selected by the routing policy and the upstream states may differ between complete reruns.

K.3 Binary recall and false-alarm tradeoff

On Bench-950, the full seed-42 condition has binary extreme recall of 79.41% and a No Extreme false-alarm rate of 76.19%. With review disabled and stress enabled, the corresponding values are 63.91% and 48.57%. These differences indicate a more hazard-favoring operating point when review is active. The subtype accuracy difference is only -0.21 points for full relative to review-disabled, and is not significant.

Consequently, increased extreme recall must be evaluated together with false alarms, subtype correctness, and application-specific costs. The archive does not define a universal decision-cost matrix that would make one operating point optimal for all deployments. The delivered predictions allow such costs to be applied explicitly in future analysis.

K.4 Aurora coverage and a separate cache diagnostic

Aurora is used in 577 successful primary-run cases, or 6.80% of 8,488. The archived implementation summary separately records 346 forecast-stage errors. Forecast access is conditional and resource-dependent, not assigned randomly. Availability depends on valid initial fields, checkpoint and adapter availability, cache state, and capability checks.

An analysis was conducted on bench-950 comparing an arbiter-less configuration (without Aurora) against a configuration with access to the Aurora cache. The overall accuracy shifted from 45.15% to 45.35%—a net increase of two correct predictions—with an exact-match test yielding p = 0.860.

The Madagascar case remains useful despite this aggregate uncertainty: it shows an actual forecast initialized within the cutoff and supplied to the reviewer. Its decision is corroborated, not repaired. A stronger experiment would hold forecast availability and all other inference inputs fixed while comparing forecast evidence against a matched no-forecast condition and a single-agent condition given the identical forecast.

K.5 Computation and conditional paths

The primary run records 34,646 successful LLM calls, averaging approximately 4.082 calls per valid instance. The system uses a conditional execution pipeline that combines three specialists and an arbiter with review activated as needed, supporting flexible processing paths across instances.

With random seed 42, EXConan uses an average of 4.731 calls and approximately 37,433 total tokens per valid instance. Table S9 reports LLM usage based on measured call counts, providing a quantitative basis for computational budgeting and experimental reproducibility. Forecast-model GPU computation is treated as a separate computational category from LLM token usage.

The six examples illustrate multiple traceable execution paths, including stable paths completed without review and paths that activate review or invoke Aurora. These examples make the conditional routing mechanism explicit and allow stage-level invocations and decision processes to be inspected and analyzed.

L Limitations

Computational overhead. EXConan requires more model calls than single-pass direct inference. Conditional routing supports selective execution of some stages, but the current evaluation does not systematically characterize performance across computational budgets. Future budget-matched experiments will examine the trade-offs among accuracy, call counts, and latency.

Validation of explanation quality. Structured outputs and evidence references make reasoning traces inspectable, but format validity does not guarantee the meteorological accuracy of every explanation. The case analyses demonstrate traceable reasoning, while independent expert evaluation remains necessary to systematically assess evidence relevance, numerical consistency, and explanation faithfulness.