AgentCaster: Reasoning-Guided Tornado Forecasting

Michael Chen
Department of Computing + Mathematical Sciences
California Institute of Technology
Pasadena, CA
mhchen@caltech.edu

Abstract

There is a growing need to evaluate Large Language Models (LLMs) on complex, high-impact, real-world tasks to assess their true readiness as reasoning agents. To address this gap, we introduce AgentCaster, a contamination-free framework employing multimodal LLMs end-to-end for the challenging, long-horizon task of tornado forecasting. Within AgentCaster, models interpret heterogeneous spatiotemporal data from a high-resolution convection-allowing forecast archive. We assess model performance over a 40-day period featuring diverse historical data, spanning several major tornado outbreaks and including over 500 tornado reports. Each day, models query interactively from a pool of 3,625 forecast maps and 40,125 forecast soundings for a forecast horizon of 12-36 hours. Probabilistic tornado-risk polygon predictions are verified against ground truths derived from geometric comparisons across disjoint risk bands in projected coordinate space. To quantify accuracy, we propose domain-specific TornadoBench and Tornado-Hallucination metrics, with TornadoBench highly challenging for both LLMs and domain expert human forecasters. Notably, human experts significantly outperform state-of-the-art models, which demonstrate a strong tendency to hallucinate and overpredict risk intensity, struggle with precise geographic placement, and exhibit poor spatiotemporal reasoning in complex, dynamically evolving systems. AgentCaster aims to advance research on improving LLM agents for challenging reasoning tasks in critical domains.

1 Introduction

LLMs have rapidly progressed from text-only pattern recognizers to general-purpose reasoning agents capable of planning, using tools, and operating in multi-turn interactions [1, 2, 8, 13, 7, 12]. As these models are increasingly envisioned for autonomous roles, evaluating their true capabilities on more challenging and higher impact problems becomes paramount [11]. This evaluation gap inhibits our understanding of both LLM limitations and progress, particularly in domains where reliable performance is critical.

Severe convective weather represents precisely such a domain. Predicting tornadoes carries immense importance; from 2010 through 2024, tornadoes in the United States caused over USD 25 billion in property damage and claimed more than 1,200 lives [10]. Human forecasters at the NWS Storm Prediction Center (SPC) must synthesize heterogeneous high-resolution numerical weather prediction (NWP) fields, examine vertical atmospheric profiles, reason across extensive geographic areas and timeframes, and ultimately produce nested probabilistic polygons that communicate risk to emergency managers and the public [3]. However, despite decades of research, tornado forecasting remains notoriously challenging.

To address this evaluation gap, we develop a framework that can rigorously test LLM capabilities in a real-world forecasting environment. We introduce AgentCaster, a novel, contamination-free

Evaluating the Evolving LLM Lifecycle, NeurIPS 2025.

evaluation framework that assesses multimodal LLM agents end-to-end on tornado forecasting. Within AgentCaster, LLMs function as AI meteorologists, interactively querying a rich archive of historical, high-resolution weather forecast data. Finally, agents synthesize their findings to produce probabilistic tornado risk predictions as geospatial polygons in standard GeoJSON format, analogous to official SPC outlooks.

Our contributions include: (1) AgentCaster, a multimodal, interactive, and contamination-free agent framework for evaluating LLM reasoning on the challenging and real-world task of tornado forecasting using daily generated high-resolution forecast data; (2) domain-specific evaluation metrics based on geometric verification against ground truths; (3) a curated 40-day benchmark dataset comprising 145,000 processed forecast maps, on-demand generation for 1,605,000 forecast soundings, SPC outlooks for baseline comparison, and processed ground truth tornado reports; (4) initial evaluation of state-of-the-art multimodal LLMs against human expert baselines; and (5) release of all code and datasets to facilitate reproducibility and further research. We hope AgentCaster will catalyze research on *high-impact, real-world reasoning tasks* and motivate progress towards agents that can meaningfully assist human experts in critical domains.

2 AgentCaster

2.1 Framework Overview

AgentCaster is an interactive environment where an LLM agent is placed in the role of an AI meteorologist tasked with issuing a tornado risk forecast for the Continental United States (CONUS). Agents make sequential requests for meteorological data products using a defined set of tools. They begin with access to a wide array of forecast maps and can subsequently request vertical atmospheric profiles for specific locations and times. The agent must predict the probability of a tornado occurring within 25 miles of any point during a 24-hour period from 12:00 UTC on the target date to 12:00 UTC the following day, aligning with operational forecasting timelines used by human meteorologists. For all experiments reported here, we freeze a contiguous 40-day benchmark window (March 1, 2025 to April 9, 2025) to ensure fair composition and reproducibility, even though the framework is designed for live daily forecasting.

AgentCaster's design enables: (1) *realistic assessment of domain expertise* by requiring reasoning similar to expert human forecasters; (2) *interactive exploration* through deliberate tool usage to analyze heterogeneous data; and (3) *contamination-free evaluation* using rolling numerical weather prediction archives. Distinct from text-based or purely simulated environments, AgentCaster dynamically integrates real-world, multimodal meteorological data (including on-demand visual sounding generation triggered by agent requests) within an interactive loop. AgentCaster is also *extensible*, allowing for the future inclusion and modification of different NWP models, prediction objectives, or prediction horizons.

2.2 Meteorological Data Sources

AgentCaster utilizes archived data from daily runs of the High-Resolution Rapid Refresh (HRRRv4) [4] model, processed into formats suitable for multimodal LLM inputs. The HRRRv4 is the state-of-the-art, 3-km resolution, convection-allowing numerical weather prediction system operated by NOAA, built on the WRF-ARW dynamical core [9]. For each day, we process the 00:00 UTC HRRR model run to extract and visualize all 145 available map products. These include convective parameters (CAPE, CIN), wind fields (shear, helicity), moisture variables, temperature profiles, and simulated radar reflectivity, among others. To access full vertical atmospheric structure near any given point, the framework provides forecast soundings derived from HRRR BUFKIT data. These are generated *on-demand* during the agent's interaction.

2.3 Agent Interaction Loop

The agent operates in a multi-turn conversational loop that mirrors an iterative forecast workflow. It begins with an initial prompt specifying the task, date, and available tools, then issues map and sounding requests as needed. Each request is processed by the backend, which returns confirmations and embeds the requested images in-line; sounding replies also report the remaining daily quota.

The agent analyzes these multimodal inputs and decides what to request next, repeating a request–receive–analyze cycle until the evidence is sufficient. When confident, it submits the final tornado-risk polygons (GeoJSON), which ends the day's interaction.

3 TornadoBench and TornadoHallucination

3.1 Ground Truth Generation

Converting discrete tornado reports into a continuous probability field requires spatial smoothing to capture the inherent uncertainty of tornado occurrences. To generate an objective verification target, we adapt and extend the Practically Perfect Forecast (PPF) methodology of [5], developing a multi-step approach to construct high-resolution ground-truth risk fields. Our modified approach transforms discrete tornado observations into a continuous probability field representing a theoretically ideal probabilistic forecast.

3.2 TornadoBench Score and TornadoHallucination Metrics

We propose TornadoBench as the primary metric for AgentCaster. It is designed to evaluate the agent's ability to accurately delineate the location, extent, and intensity of tornado risk; it addresses the limitations of standard metrics by incorporating domain-specific weighting and geometric accuracy across multiple probability thresholds. LLMs are known to hallucinate information [14, 6], and in a forecasting context, we define this as predicting risk where none exists or predicting risk in an entirely non-overlapping location on a risk day. Evaluating hallucinations is particularly important in tornado prediction, where false alarms can lead to unnecessary costs and public complacency. We introduce two metrics to quantify these behaviors.

4 Experiments and Evaluation

We evaluated a suite of reasoning and non-reasoning multimodal LLMs with knowledge cutoff dates prior to March 1st. The human expert baseline is the first official SPC Day 1 Convective Outlook issued for the 12:00 UTC cycle, processed identically to agent predictions. All LLM agents were initialized with a detailed system prompt outlining their role as an AI meteorologist, the forecasting objective, data access tools, and the GeoJSON output format requirements. The primary forecasting accuracy, hallucination metrics, and maximum risk matching for the LLM configurations and the SPC baseline are presented in Table 1. Agent interaction statistics and centroid distance errors are detailed in Table 2. The SPC baseline achieves a TornadoBench score of 18.31%, significantly outperforming all evaluated LLM agents. Among the LLM agents, performance varied, with the highest-scoring models achieving TornadoBench scores below 10%. A notable challenge for several LLMs was the consistent generation of valid GeoJSON outputs. The models with the fewest valid predictions, gemini-2.5-flash-preview:thinking (16 days), also had the lowest TornadoBench scores.

Within the GPT-5 family, increasing reasoning correlates with a monotonic drop in TornadoBench (8.51%, 7.23%, 6.28%, 3.54% for gpt-5-minimal, gpt-5-low, gpt-5-medium, and gpt-5-high, respectively). This degradation occurs despite mixed shifts in hallucination severity. Furthermore, claude-3.7-sonnet (non-thinking) marginally outperforms its thinking variant on TornadoBench (6.79% vs. 6.64%). LLM agents exhibit a strong tendency towards hallucinations. The TornadoHallucinationHard scores for LLMs were substantially higher than SPC's, with not only more frequent but also more severe hallucinations or complete misplacement of risk areas. The average centroid distance errors indicate significant challenges for LLMs in accurately placing the core of the predicted tornado threat, with most errors exceeding 400-500 km, compared to SPC's 182 km (overall) and 236 km (max risk). Agent interaction patterns varied across models. Except for one model, the number of sounding requests remained well below the daily quota of 50.

Among the three 30% risk days in our benchmark, we show March 14, 2025, the day whose SPC daily TornadoBench score is closest to the top model's score. On this day, the top LLM agent achieved a daily TornadoBench score of 9.45%, approaching SPC's 9.51% (Figure 1).

Table 1: Primary forecasting performance metrics. For TornadoHallucination metrics, lower is better. Max Risk Match shows the percentage of days the model's maximum predicted risk was Under/Match/Over the ground truth maximum risk.

Model	TornadoBench (%)	TornadoHallucination Simple	TornadoHallucination Hard	Max Risk Match (%) Under / Match / Over
SPC (Human Expert)	18.31	0.275	0.70	5.0 / 55.0 / 40.0
gpt-5-minimal	8.51	0.385	2.56	12.8 / 20.5 / 66.7
gpt-5-low	7.23	0.444	1.92	11.1 / 27.8 / 61.1
claude-3.7-sonnet	6.79	0.400	3.30	10.0 / 22.5 / 67.5
claude-3.7-sonnet:thinking	6.64	0.359	3.10	17.9 / 23.1 / 59.0
gpt-5-medium	6.28	0.484	2.65	9.7 / 22.6 / 67.7
gpt-4.1	5.63	0.444	3.64	11.1 / 19.4 / 69.4
gemini-2.5-pro-preview-03-25	4.26	0.406	4.50	15.6 / 21.9 / 62.5
grok-4	3.85	0.538	8.85	2.6 / 7.7 / 89.7
gpt-5-high	3.54	0.500	2.30	16.7 / 0.0 / 83.3
o4-mini-high	3.37	0.528	5.39	11.1 / 13.9 / 75.0
o3	3.27	0.550	5.50	10.0 / 7.5 / 82.5
gemini-2.5-flash-preview:thinking	1.57	0.625	4.50	6.3 / 6.3 / 87.5

Table 2: Agent interaction statistics and centroid distance errors.

Model	Prediction Days	Centroid Dist. (Avg. / Max Risk) (km)	Avg. Assistant Turns	Avg. Tool Calls	Sounding Requests (Avg. / Max)
SPC (Human Expert)	40	182 / 236	N/A	N/A	N/A
gpt-5-minimal	39	358 / 354	8.93	18.32	0.12 / 3
gpt-5-low	36	417 / 469	4.00	35.58	0.05 / 1
claude-3.7-sonnet	40	405 / 441	21.80	21.80	4.83 / 8
claude-3.7-sonnet:thinking	39	474 / 493	21.57	21.57	4.97 / 11
gpt-5-medium	31	398 / 447	4.45	41.27	0.05 / 1
gpt-4.1	36	361 / 377	11.32	23.07	4.47 / 13
gemini-2.5-pro-preview-03-25	32	494 / 561	5.55	18.38	2.23 / 5
grok-4	39	450 / 487	5.83	24.23	4.00 / 8
gpt-5-high	30	449 / 525	4.75	39.25	0.40 / 4
o4-mini-high	36	583 / 623	6.58	6.55	0.12 / 1
o3	40	478 / 564	13.70	13.70	0.62 / 5
gemini-2.5-flash-preview:thinking	16	601 / 595	7.05	32.38	2.70 / 50

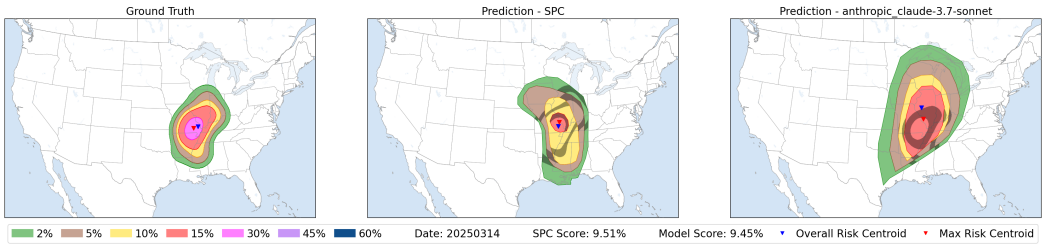

Figure 1: Evaluation of SPC and the top performing model on March 14, 2025. Overlapping solution regions are shaded.

5 Conclusion

We introduced AgentCaster, a novel framework for evaluating multimodal LLM agents on the task of tornado forecasting. Through an interactive environment utilizing high-resolution meteorological data, AgentCaster assesses agentic reasoning. Our metrics, TornadoBench and TornadoHallucination, applied over a 40-day period, revealed gaps between current LLM capabilities and human expert performance. By establishing a challenging benchmark, we aim to drive progress toward more capable and reliable AI agents while highlighting the current limitations of LLMs.

References

[1] Tom B. Brown et al. *Language Models are Few-Shot Learners*. July 22, 2020. DOI: 10.48550/arXiv.2005.14165. URL: http://arxiv.org/abs/2005.14165.

[2] Yupeng Chang et al. “A Survey on Evaluation of Large Language Models”. In: *ACM Trans. Intell. Syst. Technol.* 15.3 (Mar. 29, 2024), 39:1–39:45. ISSN: 2157-6904. DOI: 10.1145/3641289. URL: https://dl.acm.org/doi/10.1145/3641289.

[3] Stephen F. Corfidi. “The Birth and Early Years of the Storm Prediction Center”. In: (Aug. 1, 1999). Section: Weather and Forecasting. ISSN: 1520-0434. URL: https://journals.ametsoc.org/view/journals/wefo/14/4/1520-0434_1999_014_0507_tbaeyo_2_0_co_2.xml.

[4] David C. Dowell et al. “The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description”. In: (Aug. 3, 2022). Section: Weather and Forecasting. DOI: 10.1175/WAF-D-21-0151.1. URL: https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-21-0151.1.xml.

[5] Nathan M. Hitchens, Harold E. Brooks, and Michael P. Kay. “Objective Limits on Forecasting Skill of Rare Events”. In: (Apr. 1, 2013). Section: Weather and Forecasting. DOI: 10.1175/WAF-D-12-00113.1. URL: https://journals.ametsoc.org/view/journals/wefo/28/2/waf-d-12-00113_1.xml.

[6] Lei Huang et al. “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”. In: *ACM Transactions on Information Systems* 43.2 (Mar. 31, 2025), pp. 1–55. ISSN: 1046-8188, 1558-2868. DOI: 10.1145/3703155. URL: http://arxiv.org/abs/2311.05232.

[7] Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. *AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents*. Dec. 23, 2024. DOI: 10.48550/arXiv.2401.13178. URL: http://arxiv.org/abs/2401.13178.

[8] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. *A Comprehensive Overview of Large Language Models*. Oct. 17, 2024. DOI: 10.48550/arXiv.2307.06435. URL: http://arxiv.org/abs/2307.06435.

[9] Jordan G. Powers et al. “The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions”. In: (Aug. 1, 2017). Section: Bulletin of the American Meteorological Society. DOI: 10.1175/BAMS-D-15-00308.1. URL: https://journals.ametsoc.org/view/journals/bams/98/8/bams-d-15-00308.1.xml.

[10] *Storm Events Database | National Centers for Environmental Information*. URL: https://www.ncdc.noaa.gov/stormevents/.

[11] Lei Wang et al. “A survey on large language model based autonomous agents”. In: *Frontiers of Computer Science* 18.6 (Mar. 22, 2024), p. 186345. ISSN: 2095-2236. DOI: 10.1007/s11704-024-40231-1. URL: https://doi.org/10.1007/s11704-024-40231-1.

[12] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. *MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback*. Mar. 12, 2024. DOI: 10.48550/arXiv.2309.10691. URL: http://arxiv.org/abs/2309.10691.

[13] Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. *A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems*. Feb. 28, 2024. DOI: 10.48550/arXiv.2402.18013. URL: http://arxiv.org/abs/2402.18013.

[14] Yue Zhang et al. *Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. Sept. 24, 2023. DOI: 10.48550/arXiv.2309.01219. URL: http://arxiv.org/abs/2309.01219.