AFDBench: A Benchmark for Evaluating AI-Generated National Weather Service Forecast Discussions

Anonymous Authors[1]

Abstract

Area Forecast Discussions (AFDs) are critical scientific texts produced by the U.S. National Weather Service. Despite progress in AI weather prediction, no benchmark evaluates whether language models can generate professional meteorological text. We introduce **AFDBench**, comprising 7,732 expert-written AFDs from 13 NWS offices, paired with structured AI weather forecasts from Google's WeatherNext 2. AFDBench defines three evaluation metrics: *Met-Align* (numerical accuracy), *Style-Align* (professional vocabulary), and *Input-Grounding* (fidelity to source data). Zero-shot baselines with open-source LLMs reveal a significant capability gap, highlighting the challenge of domain-specific scientific text generation.

1. Introduction

The National Weather Service (NWS) issues Area Forecast Discussions (AFDs) to explain the scientific reasoning behind weather forecasts. While models like GraphCast (Lam et al., 2023) and Pangu-Weather (Bi et al., 2023) excel at predicting gridded numerical data, the "last mile" of translating these grids into expert natural language remains untested. LLMs face challenges here requiring numerical precision, domain-specific vocabulary, and strict structural conventions.

To address this, we introduce **AFDBench**, a dataset and evaluation suite. It provides 7,732 professional AFDs paired with WeatherNext 2 structured JSON forecasts. We define three metrics to measure numerical grounding, stylistic adherence, and input data fidelity.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI Scientists – Tools, Co-authors, or Founders? workshop (ICML 2026). Do not distribute.

2. Dataset Construction

We collected 7,732 AFDs from 13 NWS Weather Forecast Offices spanning diverse U.S. climate regions between Dec 31, 2025, and Apr 6, 2026. Each AFD text (mean 951 words) is paired with a corresponding single-timestep forecast from Google's WeatherNext 2 model, capturing surface conditions, comfort indices, and upper-air fields.

Samples follow a 4-field JSONL format: an **Instruction** specifying the office, the **Input** WeatherNext 2 JSON, a **Thinking** field containing synoptic reasoning extracted from the human AFD, and the **Output** full AFD text. We apply a geographic hold-out split: 11 offices (6,699 samples) are used for training, and 2 unseen offices (BOX and MRX, 1,033 samples) are reserved for testing generalizable reasoning.

3. Evaluation Metrics

AFDBench employs three automated metrics:

Met-Align (%): Measures the intersection of 2–3 digit numbers between the AI generated text and the human reference text. It serves as a proxy for numerical faithfulness.

Style-Align (0–1): Measures adherence to the NWS professional register via the overlap of 12 canonical meteorological terms (e.g., SYNOPSIS, CONVECTION, ADVECTION) present in both the generated and reference text.

Input-Grounding (0–1): Measures fidelity to the input data independently of the human reference by checking three conditions: temperature accuracy (within 3°F of input), correct cardinal wind direction, and correct pressure regime (high/ridge vs. low/trough).

4. Baseline Evaluation

We evaluated zero-shot open-source LLMs (Qwen2.5-7B, Mistral-7B, Hermes-3-8B) on a test subset. All models achieved $\sim$13% Met-Align, restricted by the single-timestep nature of the input versus the multi-timestep references in human AFDs. Style-Align scores averaged 0.33, demonstrating that standard models struggle with NWS vocabulary. In contrast, an AI-Met-v1 model trained with domain-

specific GRPO nearly doubled Style-Align (0.619) and improved Input-Grounding to 0.940, proving the benchmark's utility in differentiating domain-adapted models.

5. Conclusion

AFDBench is the first benchmark for AI-generated meteorological text. By combining expert-written discussions with real AI forecast data, it provides a rigorous testbed for evaluating numerical precision and domain adaptation in scientific language models.

References

Bi, K., Xie, L., Zhang, H., et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023.

Lam, R., Sanchez-Gonzalez, A., Willson, M., et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.