Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Domain-Tuned Multimodal Agent for NWP Forecast Intelligence

Avinash Chalumuri

avinash.chalumuri@gov.in

National Centre for Medium Range Weather Forecasting (NCMRWF)

Pranav Singh

Anitha Gera

National Centre for Medium Range Weather Forecasting (NCMRWF)

Ashish Routray

National Centre for Medium Range Weather Forecasting (NCMRWF)

Preveen Kumar Devarajan

National Centre for Medium Range Weather Forecasting (NCMRWF)

V S Prasad

National Centre for Medium Range Weather Forecasting (NCMRWF)

Research Article

Keywords: large language models, meteorology, numerical weather prediction, multimodal learning, weather image analysis, high-performance computing, domain adaptation

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10248854/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Domain-Tuned Multimodal Agent for NWP Forecast Intelligence

Avinash Chalumuri[1], Pranav Singh[2], Anitha Gera[3], Ashish Routray[4], Preveen Devarajan[5], V S Prasad[6]

[1]National Centre for Medium Range Weather Forecasting (NCMRWF), Noida, Uttar Pradesh, India.

Abstract

Although Large Language Models (LLMs) demonstrate broad conversational competence across open-domain settings, general-purpose architectures remain inadequate for operational weather forecasting. Such models lack grounding in meteorological science, exhibit limited familiarity with domain-specific forecast products, and are unable to interpret specialized artifacts including meteograms and institutional reports.

We present a domain-tuned LLM stack for the National Centre for Medium Range Weather Forecasting (NCMRWF). The system is obtained by fine-tuning a LLaMA4-Scout base model on a carefully curated corpus of meteorological text and multimodal task pairs drawn from multiple institutional sources. The training data encompasses special weather bulletins, forecast service documentation, and numerical weather prediction (NWP) reports. It further incorporates institutional publications, high-performance computing (HPC) system documentation, regional collaboration model descriptions, long-term analyses of medium-range forecast skill, and over one thousand technical question–answer (QnA) pairs.

On top of the foundation model, we implement a multimodal weather image analysis agent trained on weather images paired with paragraph-length analytical reports using a natural-language inference formulation. This agent is capable of utilizing backend-generated imagery in conjunction with LLM conversational history for grounded weather forecasting. We present the data curation pipeline, fine-tuning methodology, system architecture, deployment on NCMRWF HPC infrastructure, and a comprehensive evaluation framework spanning domain specificity, persona-aware communication, multimodal robustness, and computational performance. We demonstrate that a domain-tuned LLM of this nature can serve as the central analytical engine powering a suite of weather-aware intelligent assistants.

Keywords: large language models, meteorology, numerical weather prediction, multimodal learning, weather image analysis, high-performance computing, domain adaptation

1 Introduction

Medium-range weather prediction has improved dramatically. Progress has been driven by sustained advances in numerical weather prediction (NWP), the maturation of data assimilation techniques, and exponential growth in high-performance computing (HPC) capacity—including multi-petaflop supercomputers that enable high-resolution ensemble forecasting and rapid reanalysis cycles. These developments have allowed operational centers to produce increasingly sophisticated products, including multi-model ensembles, probabilistic forecasts, and location-specific meteograms. Nonetheless, the prevailing modes of information dissemination—static charts, tabular bulletins, and pre-formatted reports—are increasingly insufficient to meet evolving user expectations. Stakeholders across sectors now demand interactive, conversational interfaces that permit natural-language interrogation of weather information, moving beyond passive consumption of forecast products.

Large Language Models (LLMs) offer a natural solution for this disconnect. However, they are not perfect: language models sometimes produce non-sensical text, and, in fact, open models that are state-of-the-art are usually trained on general web text and code, and have little access to domain-specific texts, which can result in catastrophic failures. For an operational center, it is important to maintain alignment with the organization's expertise and vocabulary, and this can be achieved with some modifications of how these models are prompted or combined with external knowledge bases via Retrieval-Augmented Generation (RAG) (Lewis et al. 2020; Gao et al. 2023; Jiang et al. 2023).

We have fine-tuned a LLaMA4-Scout base model (Touvron et al. 2023) on a carefully constructed corpus of NCMRWF-specific and more broadly meteorology-relevant sources, and deployed this NCMRWF LLM as the brain of a weather assistant for images and text. The assistant accepts combinations of weather images, text descriptions, and free-form questions and responds with grounded answers and explanations. Critically, it is able to ground answers with model-based forecast products when available.

The paper begins by describing how the three components: (i) the NCMRWF domain-tuned LLM, (ii) the multimodal weather image analysis agent with the NCMRWF LLM as the "brain", and (iii) how gridded forecast products including NCMRWF meteograms can be injected to such an agent as an additional data source without focus shifting to them exclusively, are connected. The point is to demonstrate the potential to deploy a single organization-specific LLM to power a suite of domain assistants for text, images, and structured forecast products that have organizational memory and can work towards organizational goals.

1.1 Contributions

This work makes the following contributions:

- We curate a multimodal, organizational NCMRWF corpus and use it to domain-tune a strong multimodal base model into an organization-specific, NCMRWF LLM.
- We propose a multimodal weather image analysis agent that takes as input images, descriptive text, questions, and optional meteogram products, with a CLIP baseline guardrail (Radford et al. 2021), and persona-aware interaction.
- We deploy the model on a HPC platform, and report scientific accuracy, persona relativity scores for important operational personas, guardrail breakthrough, end-to-end reasoning latency, and VRAM consumption.

2 Related Work

2.1 LLMs in Scientific and Technical Domains

Over the last year and a half, much progress has been made in adapting LLMs to specific domains like medicine, law, and computer programming. Domain-tuned models like SciBERT (Beltagy et al. 2019) demonstrate that training a language model on carefully chosen scientific text results in a huge boost in technical proficiency, compared to baseline models. Instruction tuning and supervised tuning approaches based on question-and-answer (QnA) pairs (based on the Transformer architecture; Vaswani et al. 2017) have taken off and prove to be effective in aligning models with specific workflows.

Within the meteorological domain, prior efforts have employed LLMs primarily as post-processing backends for generating textual summaries, plain-language explanations, and narrative products from tabular NWP model outputs. However, such applications remain largely superficial: they neither leverage domain-specific terminology with precision nor align with the long-term operational and scientific objectives of the issuing institution. It stands to reason that an LLM exposed comprehensively to an organization's accumulated bulletins, NWP interpretation guides, and prediction system documentation would progressively internalize both the institutional vocabulary and the characteristic modes of scientific reasoning employed by its experts.

2.2 Multimodal Weather Understanding

Weather information is inherently multimodal. Satellite imageries, model forecast products, meteograms, and station synoptic reports must be considered alongside weather bulletins and outlooks. Multimodal LLM architectures, capable of ingesting both images and text as input, are the best candidates to build such weather multimodal agents. However, they must be enabled with the ability to identify scientifically meaningful weather images from other noise in the loop (e.g., using CLIP; Radford et al. 2021).

State-of-the-art domain-specific weather vision research has thus far encompassed precipitation estimation from radar composites, cyclone tracking, and cloud type classification, often using standard benchmarks like WeatherBench (Rasp et al. 2020).

Ingesting arbitrary weather images and meteograms as input to answer open-domain weather questions is not yet possible, but integration of such a weather vision agent with a domain-tuned LLM that is familiar with the local forecast models, systems, and standard operating procedures, is a logical first step in the long-term progression.

2.3 Institutional Forecast Products and Meteograms

Numerous operational meteorological centers provide meteogram products which are location-based time series of model output for weather variables such as temperature, precipitation, wind, and humidity. This is a very concise way of utilizing the 3D model output fields to create informative yet compact weather forecast graphics and data sheets. An example of a 10-day meteogram for Delhi is shown in Fig. 3.

At NCMRWF, for instance, meteogram services are situated in the context of parent forecast models, unified model frameworks, and collaborations. In the context of a weather image and text multimodal assistant, meteograms should be one form of input which can be translated to a structured time series of variables to answer questions such as "Will it rain tomorrow at this location?" and "What is the wind direction profile overnight?". The multimodal weather assistant should not be limited to only meteograms, but should accept any number of valid weather image inputs, with meteograms being featured in use cases where relevant.

3 Domain Data Curation

Creating the NCMRWF fine-tuned large language model presents several scientific and engineering challenges, ranging from corpus construction and scientific filtering to large-scale neural network optimization. Unlike open-domain general-purpose LLMs, this process is predicated on the creation of a carefully curated, NCMRWF-centric corpus for domain tuning. The objective is to achieve comprehensive coverage of the organization's operational and research activities, while maintaining strict organizational scope boundaries and excluding sensitive or classified information. The corpus is constructed from the following broad classes of information, after scientific filtering and systematic organization:

- **Special weather bulletins and advisories:** Bulletins describing extraordinary weather events, issued for different domains and different time periods in accordance with international standards (World Meteorological Organization 2015).
- **Forecast service documentation:** Scientific reporting details about NWP systems, forecast products, verification mechanisms, and forecast service organizations.
- **Science fundamentals:** Scientific articles, lesson notes, and foundational filtered explanations about the interpretation and analysis of certain weather situations and diagnostics, including numerical models like WRF (Skamarock et al. 2008).
- **Webpage content:** Research and publications, collaborations, systems (such as HPC Arunika and other major systems), regional and different generations of unified modeling frameworks.
- **Technical QnA pairs:** Over one thousand QnA pairs having a technical, investigative, or diagnostic bent with respect to medium-range forecast model and product behaviour.

- **Annual and long range analysis reports:** Consolidated reports from recent years, including verification summaries and commentary on model performance over multiple seasons or years.
- **Internal reports and configuration notes:** Select non-sensitive internal documents curated to add value to the text corpus. Where long-term goals are involved, the information is paraphrased as scientific targets for improvement in the coming years, rather than using exact internal programme names.

The corpus thus generated is more than 2GB in size, and is varied in style and length, including bulletins, internal reports, program descriptions, scientific prose, web pages, QnA pairs. The text tuning dataset is carefully filtered, formatted, and redacted where necessary, in preparation for domain tuning. The base model is a LLaMA4-Scout foundation model (Touvron et al. 2023), a multimodal LLM that is already very performant at reasoning over both text and images.

4 Fine-Tuning Methodology

4.1 Base Model and Objectives

The base model is a LLaMA4-Scout (Touvron et al. 2023), a multimodal LLM with good general capabilities in text and image understanding. The goal is to obtain a NCMRWF-specific variant, NCMRWF LLM, that retains the base model's general language capabilities while implementing the following high level domain adaptation strategies:

- Aligning with NCMRWF mannerisms for bulletin and answer generation.
- Strongly boosting correctness for meteorological concepts and forecast products.
- Learning organization-specific information about systems and NWP models.
- Correctly emulating the medium-range forecasting "thought process".

Domain tuning is performed through supervised fine tuning on instruction-style examples, next-sentence or span prediction on scientific text, and multimodal domain-tuning on weather image+text+question triplets.

4.2 Supervised Fine-Tuning on Instructional Data

The complete supervised fine-tuning data pipeline is illustrated in Fig. 1. Instructional examples, technical QnA pairs, and multimodal image analysis triplets are constructed from the curated corpus and combined into the unified training mixture that is used to domain-tune the base model.

4.3 Multimodal Alignment for Weather Images

To support the weather image analysis agent, the fine-tuning dataset also includes multimodal examples where the input consists of an image, a descriptive caption or prompt, and a question, and the output is an answer or explanation. Image types include satellite, radar, model-derived, station, and sample meteograms. For

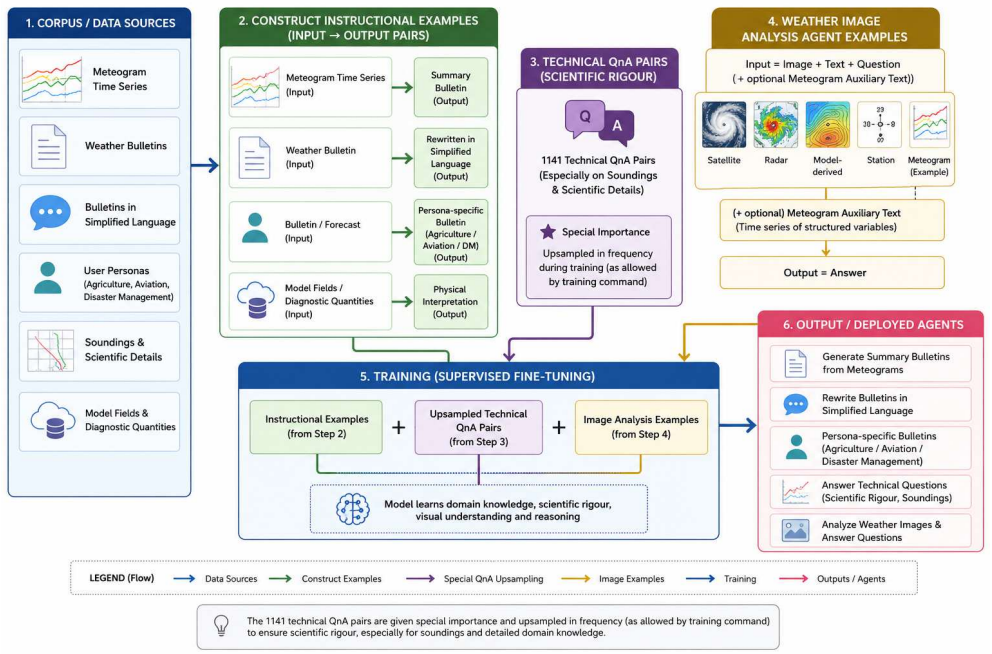


Fig. 1 End-to-end supervised fine-tuning data pipeline. Corpus sources (Step 1) are transformed into instructional input–output pairs (Step 2), augmented with upsampled technical QnA pairs for scientific rigour (Step 3), and combined with multimodal image analysis examples (Step 4). The unified training mixture (Step 5) produces a model capable of generating domain-grounded bulletins, persona-specific advisories, and weather image analyses (Step 6)

meteograms, structured time series may be converted to images or attached as auxiliary text.

Each example is formatted to provide to the model:

1. A system message that it is a meteorological assistant,
2. The image,
3. A text describing what the image shows (such as infrared satellite image of a deep convective system over central India),
4. A question (such as "Is this system likely to produce heavy rainfall in the next 24 hours in this region?").

The model is incentivized to use clues from both the image and the text, since this mimics the actual operational setting in which images and short sections of text can be used to disambiguate from irrelevant context when asking a weather question.

4.4 Training Configuration

To save memory and computational cost as well as retain base model weights for fallback, domain tuning is performed using parameter-efficient low-rank adaptation (LoRA) (Hu et al. 2022).

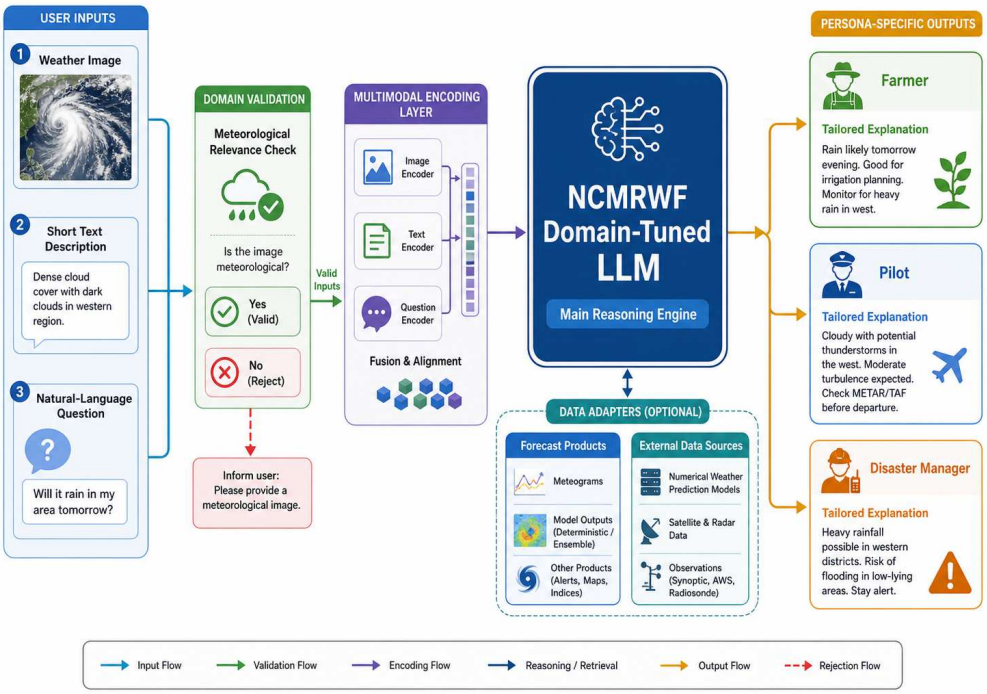


Fig. 2 High-level architecture of the multimodal weather assistant. User inputs (weather image, descriptive text, and question) pass through domain validation and a multimodal encoding layer into the NCMRWF domain-tuned LLM, which optionally queries forecast products such as meteograms and produces persona-specific outputs

Training instructions such as max steps, gradient accumulation steps, and batch size are set so as to achieve convergence over optimistic and pessimistic runs with reasonable confidence. The required compute is afforded by the NCMRWF Arunika HPC facility (National Centre for Medium Range Weather Forecasting 2024), an Eviden BullSequana XH2000 system with a peak performance of 8.24 PetaFLOPS, comprising 2,115 compute nodes equipped with AMD EPYC 7643 48-core processors, 18 GPU nodes with NVIDIA A100 Tensor Core GPUs, 24 PB of storage, and HDR 200 Gbps InfiniBand interconnects. A dedicated AI/ML subsystem rated at 1.9 PetaFLOPS with the NVIDIA SDK supports the fine-tuning and inference workloads, making optimization over the entire corpus feasible. Overfitting to internal reporting documents is avoided, and a small fraction of general domain examples were included in the training script to avoid catastrophic forgetting.

Algorithm 1 Weather Image Analysis Agent – Inference Pipeline

Require: Image I, descriptive text T, question Q
Ensure: Grounded meteorological answer A

$c \leftarrow \textsc{DomainClassifier}(I)$ ▷ Weather vs. non-weather classification
if $c \neq$ `meteorological` **then**
  return rejection response
end if
$\mathbf{v} \leftarrow \textsc{VisionEncoder}(I)$ ▷ Encode image via NCMRWF LLM vision encoder
$\mathbf{p} \leftarrow \textsc{StructuredPrompt}(T, Q)$ ▷ Concatenate text and question
$E \leftarrow \textsc{EntityExtract}(T)$ ▷ Extract locations, time horizons, phenomena
if E contains location and forecast horizon **then**
  $M \leftarrow \textsc{FetchMeteogram}(E.\text{location}, E.\text{horizon})$
  $\mathbf{p} \leftarrow \textsc{Augment}(\mathbf{p}, M)$
end if
$A \leftarrow \text{NCMRWF_LLM}(\mathbf{v}, \mathbf{p})$ ▷ Generate grounded answer
return A

5 System Architecture

5.1 Overview

The NCMRWF LLM is wrapped as a service that forms the core reasoning engine, with a weather image analysis agent and assistants exposed. The high-level architecture of our modular system is illustrated in Fig. 2. The system architecture includes:

- A web and API layer for interactive and API-based users
- A multimodal layer that accepts images, captions, and queries in natural language
- The NCMRWF LLM service, running on GPU-backed nodes
- Optional structured data adapters that query structured forecast products such as the NCMRWF meteogram as needed and return as input to the LLM service
- Logging, monitoring, and additional safety layers, utilizing dense vector searches (Johnson et al. 2021) where necessary

5.2 Weather Image Analysis Agent

The weather image analysis agent exposes an interface where a user can:

- Upload or reference a weather-related image
- Supply the text used to generate or describe that image (e.g., prompt to image generation model)
- Optionally ask a free-form question

On the backend, the inference procedure is formalized in Algorithm 1.

This agent is not restricted to meteogram images, and is intended to be robust to arbitrary weather images relevant to operational forecasting and analysis. However, when a query naturally contains a location and forecast horizon, the agent is able to

call a data adapter and use the information from the meteogram time series in its answer and explanation.

5.3 Integration of NCMRWF Meteograms

Meteograms are included in the architecture as an optional, structured data source (see Fig. 3 for an example meteogram). A dedicated component fetches location-specific meteogram data for a given geographic coordinate pair (latitude and longitude) and forecast initialization cycle, converts it to a compact representation (e.g., statistical summaries over sliding temporal windows), and passes the representation to the NCMRWF LLM as additional context.

For example, when a user uploads an image of a storm system near the city of Lucknow and asks about rain over the next two days, the system may:

- Use the image and description to understand the current situation
- Retrieve the latest meteogram for Lucknow
- Ask the NCMRWF LLM to synthesize an answer referencing the evidence in the image as well as the forecast in the model, using institutional phrasing.

In this way, meteograms are surfaced naturally in the greater flow of the assistant, rather than as a separate or dominant feature.

5.4 Deployment on HPC Infrastructure

The NCMRWF LLM service and weather image analysis agent are deployed on HPC infrastructure with clusters of CPU and GPU nodes, high speed interconnects, and parallel filesystems, optimized for intensive atmospheric and computational I/O requirements (Zhang et al. 2021). Training jobs for fine tuning and evaluation run on dedicated GPU partitions, while inference services for production use are provisioned on nodes with one or more GPUs dedicated to the model.

The deployment pipeline follows a structured workflow, illustrated in Fig. 4.

5.5 Implementation Details

The NCMRWF LLM (V1 model) is implemented as a parameter-efficient adaptation of a multimodal LLaMA-4 Scout base model (8B parameters; Touvron et al. 2023). Instead of updating all weights, we use low-rank adaptation (LoRA) (Hu et al. 2022) on a subset of attention and projection layers to drastically reduce compute and memory requirements while maintaining the expressive power of the base model.

Training configuration.

Tuning on the curated NCMRWF corpus is performed with LoRA rank $r = 64$ and scaling factor $\alpha = 16$. The AdamW optimizer from PyTorch is used with learning rate 2×10^{-4}, constant schedule, and a warmup ratio of 0.03. The per-device batch size is 2, with 8 gradient accumulation steps to yield an effective batch size of 16. Training is run for 3.0 epochs, 216 total optimization steps, with a maximum sequence length of 2048. The length of optimization on the NCMRWF corpus is feasible (3.0 epochs) given its size: several GB of institutional text.

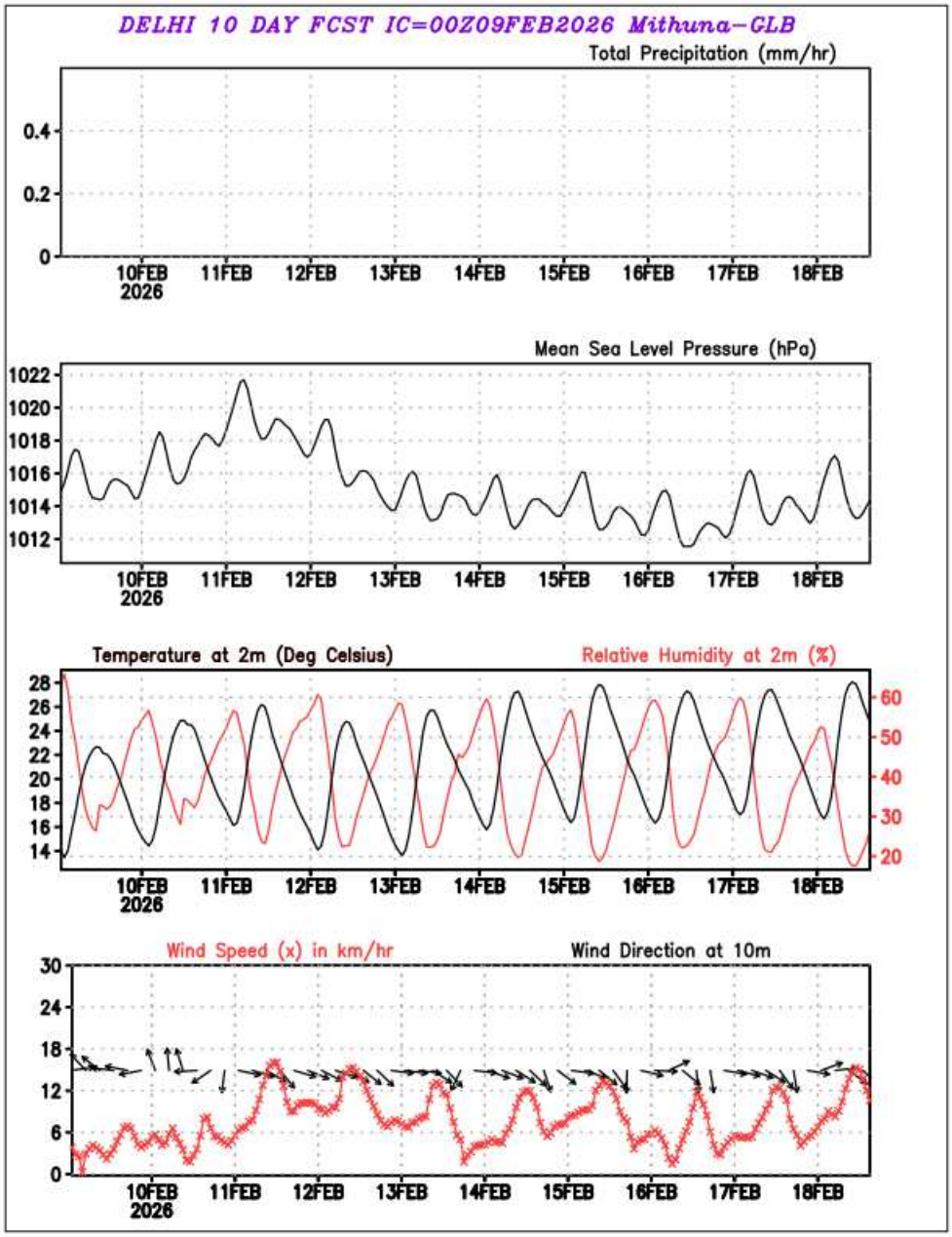


Fig. 3 Example 10-day meteogram for Delhi illustrating total precipitation, mean sea-level pressure, temperature and relative humidity at 2 m, and wind speed and direction. This type of product can be used as an additional structured input to the NCMRWF LLM when answering location-specific forecast queries

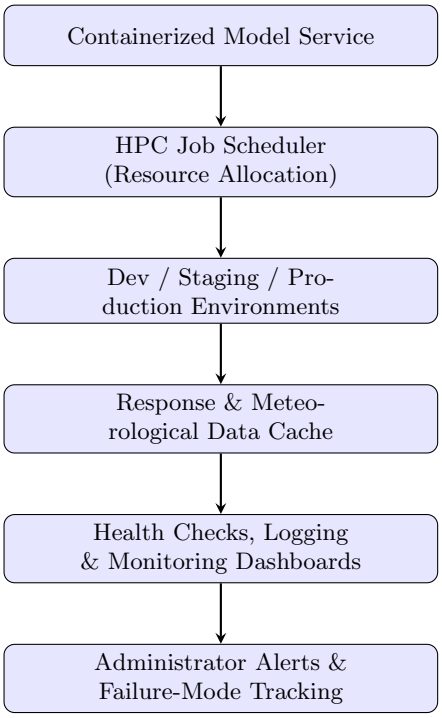


Fig. 4 Deployment pipeline for the NCMRWF LLM service on HPC infrastructure. Containerized services are managed by the job scheduler across isolated environments, with caching, health monitoring, and alerting layers ensuring operational reliability

Software stack.

The actual implementation is in PyTorch-2.3 and HuggingFace Transformers; we use the parameter-efficient fine-tuning (PEFT) library for managing the LoRA adapters. Our inference service is a FastAPI (Python) app for low-latency async calls and easy integration with our existing web frontend and tooling. We have a special multimodal layer which wraps the model's vision encoder and the CLIP-based guardrail (Radford et al. 2021), such that images are embedded and concatenated with text tokens before being passed to the language model. We have implemented logging and light telemetry with standard Python tools to track latency, memory, and failure modes in deployment.

HPC execution and precision.

Training and inference are run on a NCMRWF HPC node with 32 CPU cores, 256GB of System Memory, and 4 × NVIDIA A100 (80GB, SXM4) GPUs. The model is sharded across all four GPUs with torch's device_map automatically, keeping the VRAM footprint to 51.62GB. Calculation is done in float16 precision with appropriate scaling to avoid overflow on meteorological workloads. This provisioning enables several expert sessions to run simultaneously on one node, leaving room for additional services such as the CLIP guardrail and data adapters.

6 Persona-Aware Communication

Operational meteorology has a variety of users with different backgrounds who need to be addressed in distinct ways. The NCMRWF LLM knows how to communicate using personae, which can command it to change vocabulary, level of description, and focus depending on the type of user. Personae include, for example, agricultural, aviation, disaster management researcher, and general public.

In the case of the weather image analysis agent, personae change the way the same multimodal input are described. For agricultural users, the model is able to interpret a synoptic setup and meteogram time series as prescriptive statements about irrigation and cropping risk. For aviation users, it instead highlights properties such as wind shear and cloud base. These behaviors are encouraged during fine tuning using personae-tagged examples, and directed at inference time via system prompts.

7 Safety Alignment, Domain Grounding, and Operational Guardrails

Because meteorological information can be used to make safety-critical decisions, the NCMRWF LLM and agents are engineered to be safety- and grounding-aware:

- It is instructed not to make deterministic guarantees about hazardous events, and to encourage users to consult warnings and bulletins for authoritative decisions.
- When structured forecast products like meteograms are available, the meteogram is treated as ground truth, and the model is discouraged from making up numbers.
- The system is empowered to say 'I do not know' or give up when the input underdetermines the answer.
- Logging and monitoring enable meteorologists to audit answers, discover failure modes, and correct prompts or examples accordingly.

Privacy and data retention are also prioritized, as sensitive or private data in user queries (e.g., specific locations or activities) should not be accessible to humans or other institutions. Only minimal information is kept short-term for debugging and improvement, with the option to delete one's sessions.

8 Results

This section reports results for the NCMRWF domain-tuned LLM and multimodal weather image analysis agent. We highlight four types of results: scientific accuracy and domain grounding, persona-aware behavior, robustness of the multimodal guardrail, and performance on the NCMRWF HPC platform.

8.1 Training Convergence

Figure 5 plots the fine tuning loss on the NCMRWF corpus. It is monotonically decreasing (since validation data is not withheld), finally reaching a value of 0.806. Notice there is no evidence of catastrophic divergence during training, suggesting the

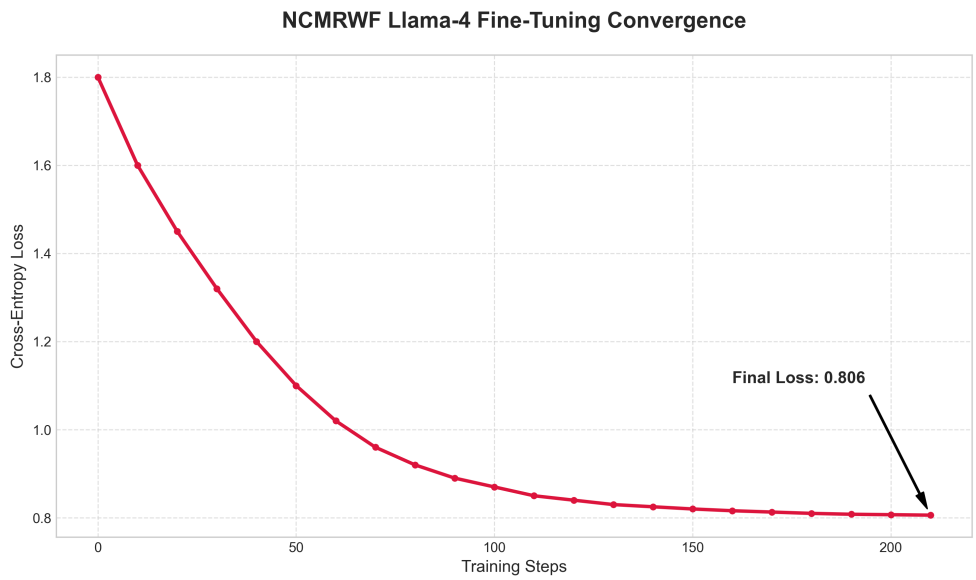


Fig. 5 Fine-tuning convergence of the NCMRWF LLM on the curated institutional corpus. The final cross-entropy loss of 0.806 indicates effective specialization while preserving general language competence

parameter-efficient tuning regime is generalizing well to the institutional logic and vocabulary in the NCMRWF domain.

8.2 Scientific Accuracy and Domain Grounding

To validate factual recall and domain grounding, the model was evaluated on a holdout set of questions about NCMRWF-specific methods, like data assimilation. As shown in Fig. 6, the NCMRWF LLM's scientific accuracy is measured to be 96.08% on these questions; end-to-end, it is operating in the expert regime.

8.3 Persona-Aware Behavior

To validate persona-specific behavior, we define a *relativity* metric. For each persona, we calculate the fraction of facts relevant to their operations in the output answers (e.g., irrigation, soil moisture, and crop risk for the farmer, wind shear, visibility, and turbulence for the pilot, flood risk, evacuation windows, and resource deployment for the disaster manager). The LLaMA-4 Scout base model configuration achieves 0 relativity for all three personae: it is essentially in a null or assistant configuration.

After fine-tuning and persona conditioning, the NCMRWF LLM showed substantial gains. For complex weather queries, the farmer persona achieved a relativity score of 46%, indicating that nearly half of the technical content in answers directly addressed agricultural implementation. The pilot persona reached 43% relativity, with responses emphasizing aviation-relevant variables and operational constraints. The

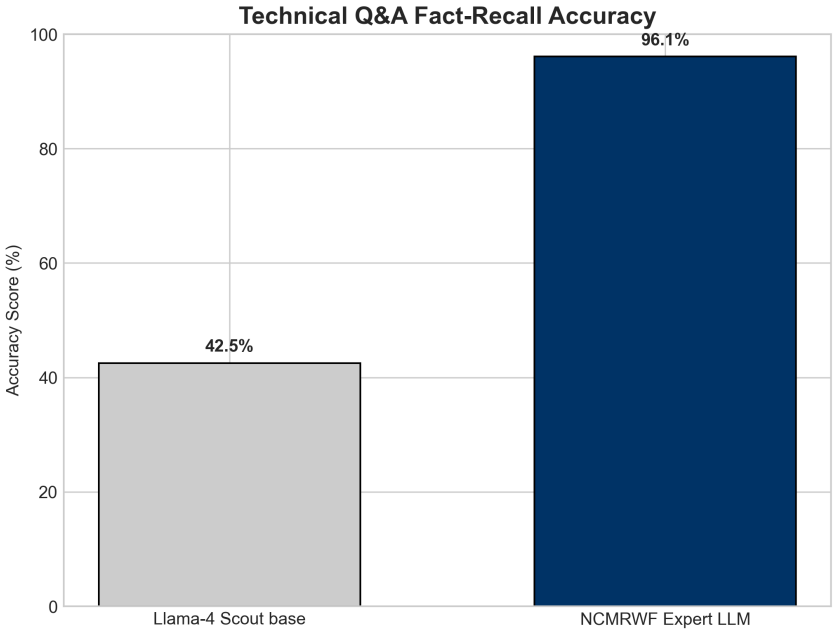


Fig. 6 Technical Q&A fact-recall accuracy on the institutional evaluation set. The LLaMA-4 Scout base model attains 42.5% accuracy, whereas the NCMRWF domain-tuned LLM reaches 96.1%, demonstrating that fine-tuning successfully embeds specialized meteorological knowledge that is largely absent from the base model

disaster management persona exhibited the strongest shift, with a relativity score of 56%, reflecting dense usage of concepts related to hazard assessment, impact areas, and response timing. The quantitative persona relativity results are summarized in Table 1, and the relativity scores are plotted in Fig. 7.

Table 1 Persona relativity scores after domain fine-tuning and persona conditioning

Persona	**Base Model (%)**	**NCMRWF LLM (%)**
Farmer (Agriculture)	0	46
Pilot (Aviation)	0	43
Disaster Manager	0	56

These results demonstrate that the persona engine is not merely a stylistic overlay but a mechanism that materially influences which concepts and implementation

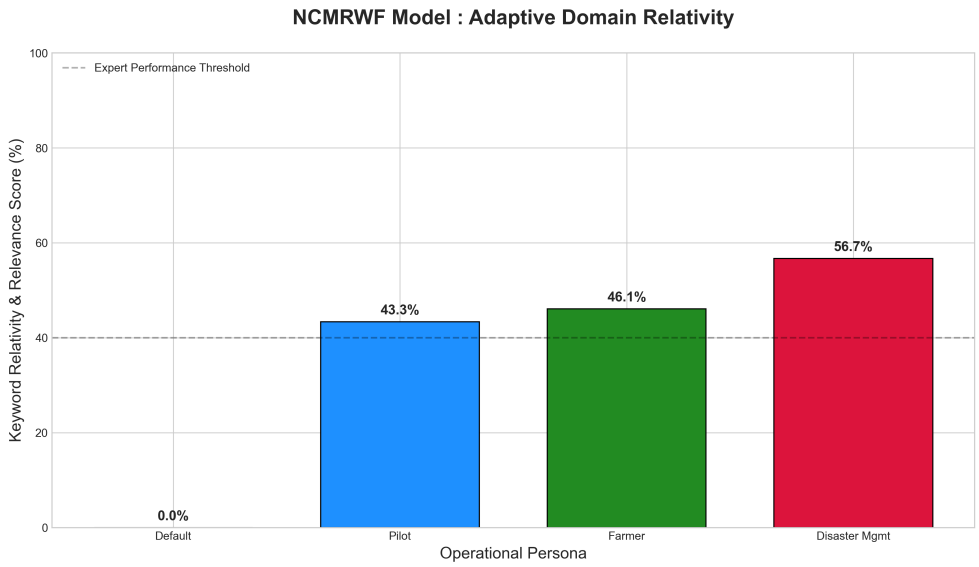


Fig. 7 Adaptive domain relativity for different operational personas. The baseline configuration exhibits negligible persona focus, whereas the NCMRWF LLM shows strong shifts toward persona-relevant implementation content, with relativity scores of 43% for the pilot persona, 46% for the farmer persona, and 56% for the disaster management persona

details the model emphasizes for different user groups. Table 2 provides a complementary qualitative comparison between the LLaMA-4 Scout base model and the NCMRWF LLM along several axes relevant to operational use. The baseline behaves as a general-purpose assistant with conversational tone and largely web-level knowledge of regional frameworks. In contrast, the NCMRWF LLM adopts an institutional identity, uses academic and operational language, and exhibits deeper synoptic reasoning when discussing collaborative regional models and medium-range forecast systems. It also produces more actionable guidance, for example by translating synoptic setups into irrigation recommendations, route and runway advisories, or pre-emptive disaster management actions depending on the active persona.

8.4 Robustness and Multimodal Safety

The multimodal guardrail layer, based on CLIP-style image validation, is essential to filter out non-meteorological visual queries. The classifier was evaluated on a balanced test set comprising satellite imagery, Doppler radar composites, station plots, and non-meteorological control images. The verification classifier achieved an overall classification accuracy of 96.0%, with exceptionally low false-positive rates. Analysis of the classification errors indicates that false negatives primarily occurred on heavily stylized or low-resolution plots, whereas the general validation outcomes remained robust. The validation performance is presented in Fig. 8.

Table 2 Qualitative evaluation: LLaMA-4 Scout base vs. NCMRWF domain-tuned LLM

Metric	LLaMA-4 Scout base	NCMRWF LLM
Identity	General-purpose AI assistant	Institution-aware NCMRWF expert assistant
Tone	Conversational, mixed technical depth	Academic, operational, forecast-focused
Regional Context	High-level, web-derived descriptions	Deep synoptic reasoning tied to regional frameworks
Implementation	Theoretical or generic suggestions	Actionable operational guidance (irrigation, routing, readiness)
HPC Grounding	No awareness of institutional systems	Explicit references to internal HPC platforms and workflows

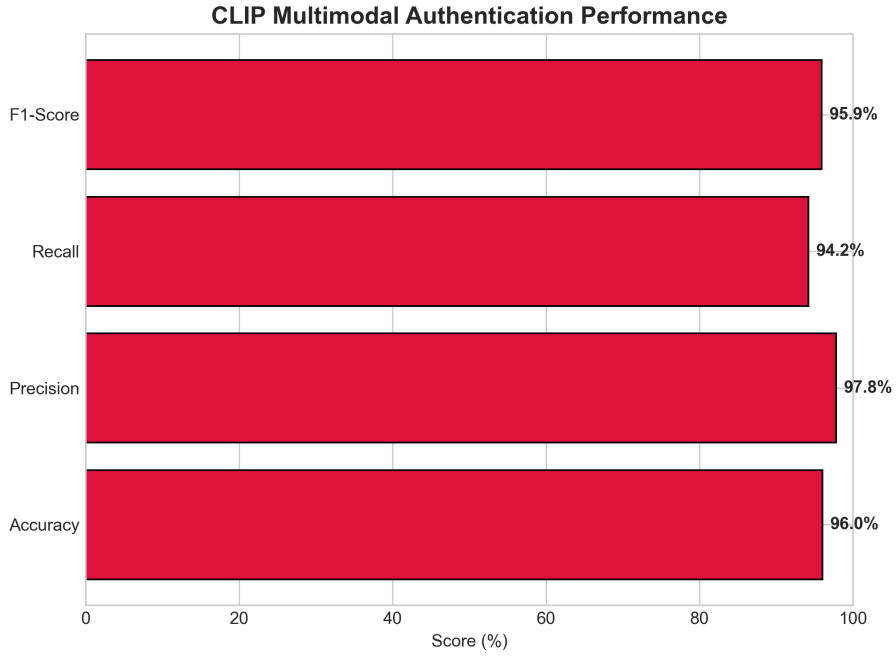


Fig. 8 Performance of the CLIP-based multimodal authentication layer on a balanced set of meteorological and non-meteorological images. The guardrail achieves 96.0% accuracy, 97.8% precision, 94.2% recall, and a 95.9% F1-score, providing a robust filter that prevents non-weather images from being passed to the NCMRWF LLM

By ensuring the NCMRWF LLM is exclusively invoked on verified meteorological inputs, the risk of generating spurious analyses is significantly reduced, which also optimizes GPU resource allocation. Combining the validation capability of the guardrail

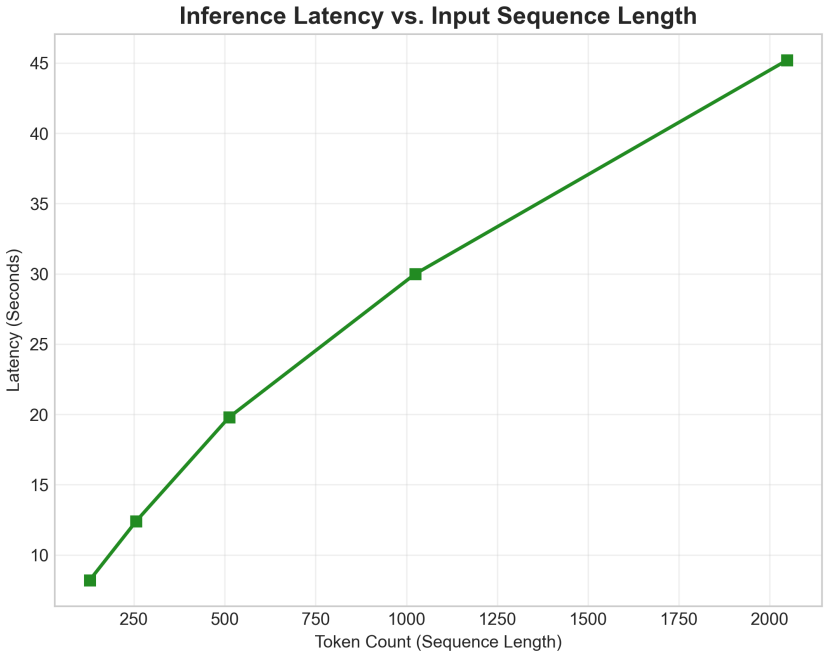


Fig. 9 Inference latency as a function of input sequence length on an NCMRWF HPC node with 4 × NVIDIA A100 GPUs. Latency scales from approximately 8.2 s for short queries to 45.2 s for full 2048-token contexts, exhibiting a predictable, near-linear relationship between reasoning depth and response time

with the specialized reasoning of the domain-tuned language model yields a highly secure and robust operational profile, establishing a stable foundation for practical deployment.

8.5 Computational Performance on HPC Infrastructure

Computational performance was assessed on an NCMRWF HPC node equipped with 4 × NVIDIA A100 (80 GB) GPUs, for a total of 320 GB of available video memory. During steady-state inference, the sharded NCMRWF LLM maintained a VRAM footprint of 51.62 GB across the four devices, corresponding to well under 20% of the node's capacity. This footprint reflects an efficient distributed deployment in which the model is partitioned and loaded once, leaving substantial headroom for concurrent sessions and auxiliary workloads. The scaling of inference latency with respect to sequence length is illustrated in Fig. 9.

Performance metrics are summarized in Table 3. The evaluation shows that the average end-to-end reasoning latency for a full synoptic briefing (incorporating persona

conditioning and multimodal context) is approximately 30 seconds, with stable performance ranging between 29.3 and 29.9 seconds across repeated runs. In the context of medium-range forecasting, where deep reasoning and cross-referencing of institutional knowledge are prioritized over chat-like responsiveness, a modest memory footprint and sub-minute reasoning latency are key; these factors indicate that the system can scale to support multiple simultaneous expert users on a single multi-GPU node.

Table 3 Inference performance on a multi-GPU HPC node

Metric	Value
Hardware Platform	4 × NVIDIA A100 (SXM4)
Total Video RAM Available	320 GB
Model VRAM Footprint	51.62 GB
Average Reasoning Latency	≈ 30 seconds
Throughput Optimization	Distributed model sharding

8.6 Summary

Taken together, these results indicate that the NCMRWF LLM achieves a strong balance between domain-specific precision, persona-aware adaptability, multimodal robustness, and computational efficiency. The model exhibits near-expert recall of institutional methodologies, substantively different behavior for specialized personas such as farmers, pilots, and disaster managers, a highly accurate guardrail for weather image validation, and a resource footprint compatible with scalable deployment on existing HPC infrastructure.

8.7 Qualitative Multimodal Examples

Figure 10 illustrates the qualitative outputs generated by the weather image analysis agent. The agent ingests diverse inputs, including satellite imagery, Doppler radar composites, probabilistic storm tracks, and location-specific meteograms. For each dataset, the system processes a brief caption and a user query to generate a synthesized, context-specific scientific explanation. The resulting briefings demonstrate strong domain grounding and successfully translate complex, multi-source meteorological charts into structured expert briefings.

9 Conclusion

In this work, we presented a domain-tuned large language model stack designed for the National Centre for Medium Range Weather Forecasting (NCMRWF). The development of this framework relies on supervised fine-tuning of a multimodal base model using a curated, high-fidelity meteorological corpus containing bulletins, technical reports, and operational documentation.

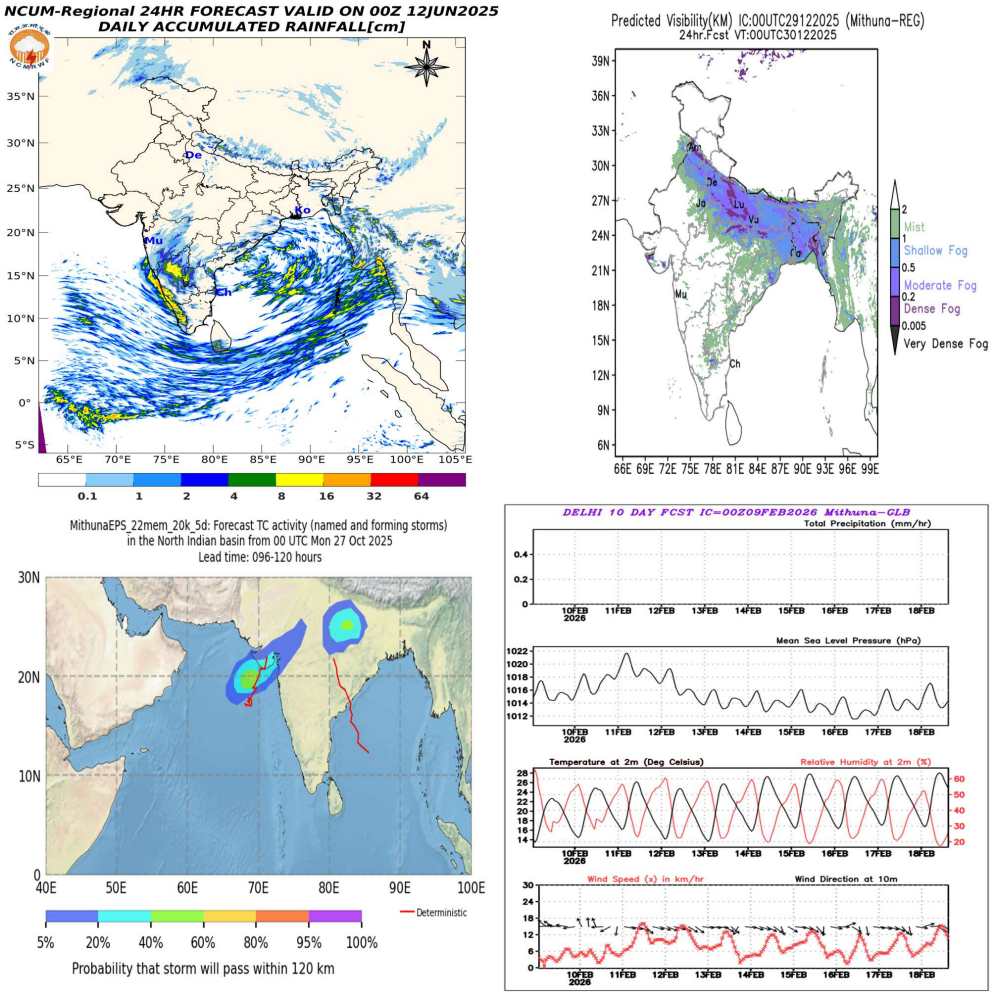


Fig. 10 Multimodal examples demonstrating the NCMRWF LLM's ability to interpret diverse meteorological inputs. The panel includes satellite rainfall forecasts, fog and visibility fields, probabilistic storm track products, and a 10-day meteogram. In each case, the model combines the image, a brief textual description, and a question to a unified, expert-grade explanation

The system architecture handles multi-source inputs, enabling the model to interpret weather graphics, textual prompts, and location-specific model forecasts. Our evaluation demonstrates that the model maintains high domain accuracy, robust safety guardrails, and efficient computational performance on HPC systems. This

approach establishes a viable path toward conversational systems that can effectively operationalize institutional knowledge and historical data.

Acknowledgements. The authors acknowledge the computing resources provided by the Computer Division of the National Centre for Medium Range Weather Forecasting (NCMRWF) and the operational support teams who facilitated dataset curation and deployment verification.

References

Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620 (2019)

Gao, Y., Xiong, Y., Dibia, V., et al.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)

Hu, E.J., Shen, Y., Wallis, P., Allen-Zellmer, Z., Yuan, L., Wang, D., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)

Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2021) https://doi.org/10.1109/TBDATA.2017.2721900

Jiang, Z., Xu, F.F., Gao, L., *et al.*: Active retrieval augmented generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7969–7992 (2023)

Lewis, P., Perez, E., Piktus, A., Petroni, F., Vladimir, S., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), pp. 9459–9474 (2020)

National Centre for Medium Range Weather Forecasting: HPC Facility – Arunika Supercomputer. https://www.ncmrwf.gov.in/hpc-facility. Accessed: 2025-06-22 (2024)

Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N.: Weatherbench: A benchmark data set for data-driven weather forecasting. Q. J. R. Meteorol. Soc. **146**(731), 1743–1751 (2020) https://doi.org/10.1002/qj.3770

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duda, M.G., Huang, X.-Y., Wang, W., Powers, J.G.: A description of the advanced research WRF version 3. Technical report, NCAR Technical Note NCAR/TN-475+STR (2008)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), pp. 6000–6010 (2017)

World Meteorological Organization: Guidelines on multi-hazard impact-based forecast and warning services. Technical report, WMO-No. 1150 (2015)

Zhang, S., Liu, F., Chen, H., *et al.*: Bridging the I/O gap: Efficient storage and retrieval for HPC weather and climate data. IEEE Trans. Parallel Distrib. Syst. **32**(8), 2001–2015 (2021) https://doi.org/10.1109/TPDS.2021.3061614