
BEYOND CURVE FITTING: NEURO-SYMBOLIC AGENTS FOR CONTEXT-AWARE EPIDEMIC FORECASTING *

Joongwon Chae¹
 cai-zy24@mails.tsinghua.edu.cn
Ji Jiansong²
Runming Wang¹
Peiwu Qin^{2,3†}
Dongmei Yu²
Chen Xiong¹
Lian Zhang⁴
Gong Yunhan¹

¹Institute of Biomedicine and Health Engineering, Tsinghua Shenzhen International Graduate School (SIGS),
Tsinghua University, Shenzhen, China

²The Fifth Affiliated Hospital of Wenzhou Medical University, Lishui 323000, China

³Hengqin Laboratory, Zhuhai, China

⁴The First Hospital of Hebei Medical University, Shijiazhuang, China

ABSTRACT

Effective surveillance of hand, foot and mouth disease (HFMD) requires short-term forecasts that account not only for epidemiological time-series patterns but also for contextual drivers such as school calendars, weather, and policy interventions.

While classical time-series models (e.g., ARIMA, Prophet) and machine learning approaches can technically incorporate external covariates, they treat these inputs as **purely numerical weights**. Consequently, they lack the **semantic reasoning capabilities** necessary to understand the **causal mechanisms** behind the data or to interpret the **complex interplay** between conflicting drivers (e.g., why a school opening might not lead to an outbreak during extreme weather). Similarly, recent time-series foundation models (e.g., Chronos, Moirai, TimesFM) operate as black boxes without explicit mechanisms to integrate such qualitative domain knowledge.

In this work, we propose a two-agent hierarchical framework that decouples contextual interpretation from probabilistic forecasting. A large language model (LLM) “event interpreter” ingests heterogeneous external signals—including school schedules, meteorological summaries, government reports, and clinical guidelines—and compresses them into a scalar transmission-impact signal. A neuro-symbolic forecasting core then combines this signal with historical HFMD case counts to produce point forecasts, which are further calibrated into probabilistic predictions using Poisson/negative binomial likelihoods.

We evaluate the proposed framework on two real-world HFMD datasets: Hong Kong government surveillance data (2023–2024, 90 weeks) and clinical visit records from a hospital in Lishui, Zhejiang Province, China (2024, institutional review board exempt due to de-identified use). Compared to traditional time-series models and strong foundation-model baselines (TimesFM, Chronos), our approach achieves competitive point forecasting accuracy (MAE) while providing **robust** 90% prediction intervals (coverage 0.85–1.00) and human-interpretable explanatory rationales.

Our results suggest a new perspective on epidemic forecasting: instead of treating the time series as a purely numeric sequence, structurally integrating domain knowledge through LLM-based agents can match the performance of state-of-the-art foundation models, while yielding context-aware and interpretable forecasts that better align with public health decision-making workflows. The project code is publicly available at https://github.com/jw-chae/forecast_MED.

Keywords Artificial intelligence, hand-foot-and-mouth disease, large language models, epidemic forecasting, neuro-symbolic learning, time-series foundation models

* *Citation: Authors. Title. Pages.... DOI:000000/11111.*

† Corresponding author: pwqin@sz.tsinghua.edu.cn

1 Introduction

Hand, foot, and mouth disease (HFMD) is an acute pediatric infection caused by enteroviruses (most commonly EV71 and Coxsackie A serotypes) that remains a major public health concern across the Asia-Pacific region.[1] Young children are most affected, and while the majority of cases are mild, the potential for severe neurological or cardiopulmonary complications necessitates continuous monitoring.[2, 3, 2] Long-term surveillance data indicates that HFMD maintains a high incidence frequency among notifiable infectious diseases in regions like China, creating a substantial disease burden.[4, 5] Notably, HFMD transmission is closely linked to meteorological factors such as temperature and humidity; consequently, the disease exhibits diverse epidemic patterns and seasonality that vary according to regional climate characteristics rather than following a uniform trend.[6, 7, 8, 9] Given the limited availability of specific antivirals and the complex environmental sensitivity of the virus, health authorities and researchers continuously emphasize the necessity of developing reliable early warning and forecasting models.[10] Accurate forecasting is essential for enabling the efficient allocation of medical resources and securing appropriate timing for public health interventions.[11]

Forecasting HFMD and similar infectious diseases is challenging due to nonlinear drivers and external shocks. Traditional time-series models like ARIMA (Auto-Regressive Integrated Moving Average)[12] and compartmental epidemic models (e.g., SEIR)[13] can capture recurrent seasonal patterns but often struggle with nonlinear influences such as weather anomalies, behavioral changes, and intervention policies.[14] Many studies have found that HFMD incidence is modulated by meteorological factors in complex ways: for example, increases in temperature and humidity can lead to higher HFMD transmission after certain lagged intervals,[15, 16, 17] with non-linear and delayed effects varying by region.[9] Such exogenous factors (school schedules, holidays, and public health measures) introduce abrupt or non-cyclical changes that purely auto-regressive models fail to fully capture.[10] Recent data-driven models have attempted to incorporate these effects – for instance, generalized additive models explicitly include seasonal trends and holiday effects.[18] These models can fit HFMD outbreaks better than simple ARIMA variants by accounting for multi-period seasonality and specific holiday impacts (e.g., a strong Spring Festival dip followed by rebound). However, even these enhanced statistical models require manual feature engineering for each event and typically assume additive effects. They also lack a mechanism to interpret unstructured context (like news of a new school closure or vaccination campaign) that could critically alter disease dynamics.

Modern time-series forecasting techniques have evolved along two broad directions: advanced machine learning models and large-scale pre-trained foundation models.[19, 20, 21] On one hand, machine learning approaches including ensemble tree methods[22, 23, 24] and deep neural networks have achieved strong predictive performance by learning complex patterns from data. Notably, recurrent neural networks and sequence-to-sequence models paved the way for deep forecasting, with LSTMs and attention-based Transformers pushing the state of the art in many applications.[25, 26] These models automatically capture non-linear relationships and long-range dependencies that are difficult for classical models to handle. For example, the Temporal Fusion Transformer integrates recurrent layers with interpretable attention mechanisms to handle multi-horizon forecasts with multiple covariates. On the other hand, researchers have developed pre-trained time-series models analogous to language models, treating time series as sequences of tokens.[21] Chronos is one such family of foundation models that tokenizes time-series data (through scaling and quantization) and trains a Transformer on a large corpus of diverse time series, enabling probabilistic forecasts via generative sampling.[27, 28] These universal models aim to leverage cross-domain patterns to forecast any univariate series without extensive tuning. Mixture-of-experts (MoE) architectures further advance this idea by introducing specialized experts that automatically focus on different temporal patterns at a fine-grained level, yielding improved accuracy across heterogeneous datasets without manual grouping by frequency or category.[29]

Despite these achievements, purely data-driven models – from ARIMA to deep neural networks and foundation Transformers – remain essentially curve-fitting black boxes. They excel at recognizing historical patterns in numerical data but cannot inherently incorporate high-level contextual knowledge or reasoning about unforeseen events.[30] This gap has spurred interest in neuro-symbolic approaches that combine statistical learning with symbolic reasoning for greater interpretability. For instance, recent work connects neural network units with signal temporal logic formulas to classify time-series events in a human-interpretable way, yielding readable rules that describe temporal patterns without sacrificing much accuracy. Similar efforts to inject domain knowledge and logical structure have shown that it is possible to explain model decisions in time-series classification and anomaly detection. However, these methods have so far been applied mainly to offline classification tasks, not to real-time forecasting of epidemiological counts. There is a need for frameworks that can retain the predictive power of deep learning while also reasoning over domain context (e.g., interpreting a sudden school closure) to adjust forecasts in an explainable manner.

Meanwhile, advances in large language models (LLMs) offer a promising avenue to fill this gap. LLMs have demonstrated strong capabilities in reasoning, planning, and integrating knowledge when appropriately prompted.[31] Recent work shows that LLM-based agents can go beyond static question-answering and effectively interact with environments and tools. Agentic frameworks interleave chain-of-thought reasoning with action execution, allowing an

LLM to plan steps, query external sources, and update decisions based on retrieved information.[32] Such interwoven reasoning-action loops enable more robust and interpretable problem solving, while reducing errors in dynamic tasks. Open-ended embodied agents further illustrate that LLMs can function as lifelong learning planners, refining prompts, generating code, and adapting strategies to pursue long-horizon goals.[33] Self-reflective agent frameworks additionally show that LLMs can improve themselves through natural-language feedback, analyzing failed attempts and storing insights to avoid repeated mistakes.[34]

Beyond agentic behavior, recent benchmarks highlight the increasingly strong reasoning capabilities of frontier LLMs. LogicGame,[32] for example, evaluates rule-based and multi-step reasoning in controlled environments and shows that modern models can execute complex goal-driven tasks with high reliability. Clinical reasoning benchmarks such as the NEJM AI study by McCoy et al.[35] similarly demonstrate that current LLMs achieve performance comparable to medical students across standardized clinical scenarios, confirming their ability to generalize structured reasoning to high-stakes domains. Large-scale evaluations such as LiveBench[36] report strong performance across math, coding, structured reasoning, and instruction-following under contamination-free conditions, with several 2025 frontier models achieving global averages approaching the 80–90% range—reflecting substantial gains in multi-domain intelligence.

At the same time, comprehensive studies offer a clearer understanding of remaining LLM limitations. Recent large-scale analyses and systematic surveys[37, 38] provide rigorous taxonomies of hallucination types, emphasizing the importance of inference-time control mechanisms to mitigate factual inconsistency. Importantly, these findings do not diminish the utility of LLMs; rather, they highlight that strong reasoning performance can coexist with domain-specific vulnerabilities—reinforcing the need for architectures that regulate, contextualize, and selectively leverage LLM outputs.

Taken together, these results indicate that LLMs are already capable of complex reasoning, multi-step planning, domain adaptation, and structured decision-making at a level highly useful for real-time epidemiological forecasting. These capabilities—combined with deliberate control structures—allow LLMs to translate unstructured contextual information into symbolic or quantitative signals, making them uniquely suited for integration into our proposed two-agent forecasting framework.

In this paper, we propose a forecasting framework that synergizes data-driven time-series modeling with LLM-based contextual reasoning, applied to the case of HFMD epidemic forecasts. Our approach features a two-agent hierarchical architecture: (1) an Event Interpreter Agent and (2) a Forecast Generator Agent. The Event Interpreter (Agent 1) is powered by a large language model augmented with domain knowledge, and it ingests external unstructured information – such as public health announcements, intervention policies, or anomalies in mobility patterns. This agent produces a quantitative impact index I_t (or a set of adjusted features) that represents the inferred effect of these external events on disease transmission at time t . In essence, Agent 1 performs a form of neuro-symbolic reasoning, translating qualitative context into a symbolic signal that can modulate the numerical forecast. The Forecast Generator (Agent 2) is a probabilistic time-series model (built on a neural network architecture) that takes the historical case data along with the context index I_t as inputs. Agent 2 outputs the predictive distribution of future HFMD cases. By separating the concerns, our framework ensures that the pattern-recognition strength of deep networks is complemented by the interpretive flexibility of an LLM agent. Crucially, the LLM agent operates at inference time, allowing real-time incorporation of new information (e.g., kindergarten closures or vaccination campaigns) that was not available during training. This design moves beyond pure curve fitting – each forecast is accompanied by a human-readable rationale from the Event Interpreter (explaining how a particular event influences the trajectory), thereby improving transparency and trust for public health decision-makers.

Our key contributions are as follows. First, we introduce a multi-agent neuro-symbolic framework for epidemiological forecasting that marries neural forecasting with symbolic reasoning, to our knowledge the first application of an LLM-based agentic approach to epidemiological time-series forecasting. Second, we enable inference-time adaptability: our method can adjust its predictions on the fly using current context, without retraining, making the system more robust to regime changes such as sudden policy interventions or viral mutations. Third, we emphasize explainability and decision support: by producing intermediate explanations (in natural language and via the I_t index) for forecast adjustments, the framework offers interpretable insights into why a prediction changed, a requirement for public health decision support. Finally, we provide empirical validation on real-world HFMD surveillance data from two regions with contrasting patterns, showing that our framework outperforms baseline models (including ARIMA, generalized additive models, and state-of-the-art deep learning models) in terms of forecast accuracy, especially during periods following major events.

In summary, our work bridges the gap between time-series forecasting and AI reasoning systems. It illustrates a path beyond curve fitting toward models that not only predict what will happen, but also incorporate the underlying reasons, enhancing both accuracy and interpretability in epidemic forecasting. The remainder of the paper details related work in

time-series modeling and LLM agents, the design of our two-agent framework, experimental results, and a discussion on limitations and future extensions.

2 Methods

2.1 System Overview

We propose a **two-agent hierarchical framework** for epidemic forecasting that explicitly separates *context interpretation* from *probabilistic prediction*. Unlike classical compartmental models (e.g., SEIR) or purely data-driven time-series models, the proposed system treats the epidemic time series and the surrounding contextual information as two distinct but coupled channels. The goal is to forecast weekly HFMD (Hand, Foot and Mouth Disease) incidence while explicitly conditioning on exogenous drivers such as school calendars, weather, and higher-level surveillance signals.

The framework consists of two specialized agents:

- **Agent 1 (Event Interpreter):** A large language model (LLM) that ingests heterogeneous external data—including weather summaries, school calendars, and aggregated regional statistics—together with retrieval-augmented domain knowledge (e.g., guidelines, policy documents). It produces a scalar *transmission impact signal* that summarizes how current conditions are expected to affect HFMD transmission in the near future.
- **Agent 2 (Forecast Generator):** A forecasting module that combines the recent epidemic trajectory with the transmission impact signal from Agent 1 to generate distributional forecasts of weekly HFMD counts, calibrated via Poisson/Negative Binomial likelihoods.

This role separation yields a modular architecture in which each agent can be independently replaced or improved. For example, the LLM backbone in Agent 1 or the time-series backbone in Agent 2 (e.g., foundation model vs. parametric baseline) can be swapped without changing the overall interface between the two agents.

2.2 Agent 1: Context-Aware Event Interpretation

2.2.1 Input data collection and preprocessing

The primary role of Agent 1 is to aggregate diverse external signals and convert them into a unified scalar index, the **Transmission Impact Score** $I_t \in [-1.0, 1.0]$, together with a concise natural-language summary of the situation at week t . In this work, Agent 1 draws on three main categories of inputs.

Weather data Daily meteorological observations are collected from local weather stations, including mean temperature ($^{\circ}\text{C}$), relative humidity (%), and total precipitation (mm). These daily records are aggregated to a weekly resolution to match the HFMD surveillance series. For each epidemiological week t , we compute 7-day summaries such as average temperature, average humidity, and cumulative precipitation, and encode them as a compact JSON object that is passed to Agent 1 as part of its structured input.

Social events and school calendar Social events are derived from a pre-defined event calendar. Each epidemiological week is labeled with the corresponding school status (e.g., `in_session`, `summer_break`, `winter_break`) and public holidays (e.g., Spring Festival, National Day). These variables are represented as discrete categorical features in the JSON input. For example, the week containing 1 February 2024 is labeled as `winter_break + Spring Festival`, whereas the week containing 19 February 2024 is labeled as `in_session`. Agent 1 is instructed to treat school status as a primary driver of HFMD transmission, consistent with the pediatric nature of the disease.

Higher-level government surveillance reports Government surveillance signals are extracted from official epidemiological bulletins issued by the Chinese Center for Disease Control and Prevention (China CDC) and the Zhejiang provincial health authorities. These bulletins provide monthly HFMD case counts at the city and provincial levels. For each target week t , we construct summary statistics over the preceding one to two months, such as total monthly HFMD counts and month-over-month growth rates at the prefecture/province level. These statistics are serialized into JSON and supplied to the LLM so that Agent 1 can reason about the local hospital series in the broader regional context.

2.2.2 Retrieval-Augmented Generation Strategy

Large language models often struggle to consistently utilize long or heterogeneous contextual inputs when they are provided in a single prompt. To improve stability and ensure that the model grounds its reasoning in established epidemiological knowledge, we incorporate a **retrieval-augmented generation (RAG)** mechanism.

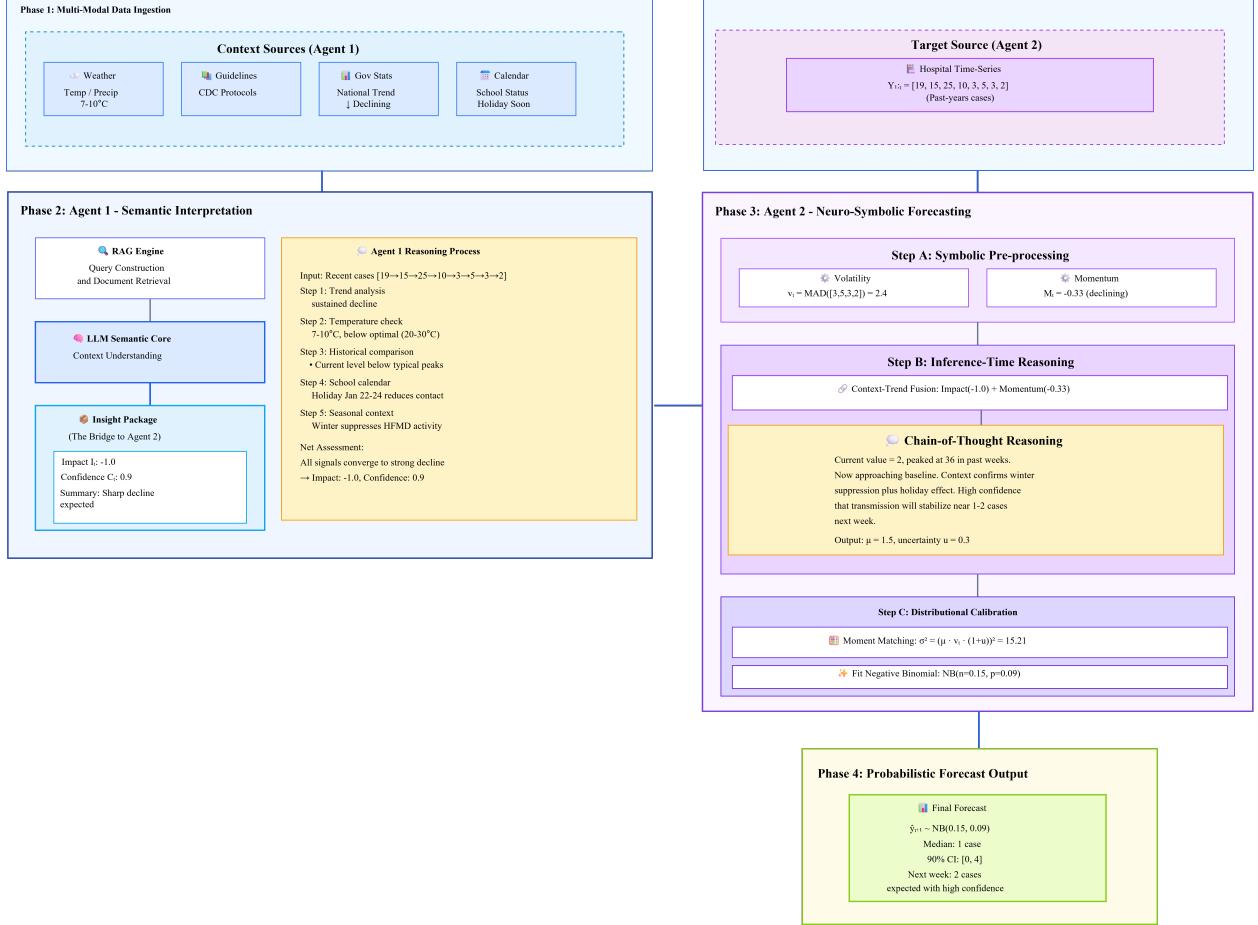


Figure 1: Overall hierarchical neuro-symbolic architecture for HFMD forecasting. Agent 1 interprets heterogeneous contextual signals (school calendar, weather, national HFMD trends, guidelines) and emits a scalar transmission impact signal, while Agent 2 combines this signal with historical case counts to produce probabilistic forecasts.

In the knowledge-base construction stage, official HFMD diagnosis and management guidelines published by the China CDC are collected and segmented into short text chunks (maximum 500 characters). Each chunk is embedded using the all-MiniLM-L6-v2 encoder and stored in a FAISS vector index.

In the dynamic query construction stage, retrieval queries are composed adaptively based on contextual signals for the current week. For seasonal factors, the query includes terms such as “HFMD peak season spring summer” during May–July, or “winter low transmission” during January–February. For environmental factors, the query is expanded with “weather temperature impact transmission” when meteorological data are present. For social factors, we append “school in_session outbreak children” when schools are in session.

In the retrieval and prompt-construction stage, the composed query is used to retrieve the top- k ($k = 2$) most relevant guideline passages (up to 1200 characters). These retrieved passages are inserted at the beginning of the LLM system prompt, ensuring that the model’s reasoning is anchored to authoritative, domain-specific evidence rather than relying solely on long free-form contextual descriptions.

2.2.3 Trend Analysis and Statistical Feature Computation

To help Agent 1 reason about the recent epidemic trajectory, we compute several simple but informative statistics from the past few weeks of HFMD activity.

Growth Rate. We estimate whether the series is accelerating or decelerating by measuring its relative change over the previous four weeks:

$$g_t = \frac{y_t - y_{t-4}}{\max(1.0, y_{t-4})}.$$

The denominator is lower-bounded by 1.0 to avoid instability when past counts are close to zero.

Consecutive Growth. Short-term momentum is captured using the length of the most recent monotonic increase streak:

$$w_{\text{grow}} = \max\{k \mid y_{t-i+1} > y_{t-i}, \forall i \in [1, k]\}.$$

A longer streak indicates sustained upward pressure rather than a one-off fluctuation.

Peak Status. We also assess whether the current level is unusually high relative to historical activity. If the current observation exceeds the 90th percentile of the full series, it is labeled as a peak:

$$\text{is_at_peak} = \begin{cases} 1, & \text{if } y_t \geq P_{90}(Y_{\text{full}}), \\ 0, & \text{otherwise,} \end{cases}$$

where $P_{90}(Y_{\text{full}})$ denotes the empirical 90th percentile. This feature helps distinguish routine seasonal variation from rare outbreak-level surges.

2.2.4 LLM-Based Interpretation and Outputs

Agent 1 passes the above information to the LLM to generate four main outputs. First, the transmission impact score $I_t \in [-1.0, 1.0]$ reflects the net effect of external factors on transmission. Values $I_t > 0$ indicate promotion of transmission (e.g., warm weather plus school in session), $I_t < 0$ indicate suppression (e.g., winter plus vacation), and $I_t \approx 0$ indicates a neutral context.

Second, the confidence score $C_t \in [0.0, 1.0]$ represents the model’s confidence in the data quality and signal consistency. For example, missing meteorological data or conflicting signals lead to lower confidence.

Third, an event summary provides a natural-language interpretation of the current situation (e.g., “School reopening coincides with warm weather”).

Fourth, risk notes list key risk factors (e.g., “Rapid growth observed”, “Temperature in optimal range”).

The LLM prompt includes the retrieved epidemiological guidelines, recent case trends, and external data, and requests a structured JSON output. The temperature is set to 0.6 to allow moderate creativity while limiting hallucinations.

2.3 Agent 2: Probabilistic Forecast Generation

2.3.1 Modeling Historical Volatility

Agent 2 begins by estimating the **inherent volatility** v of the time series, which captures the noise level intrinsic to the data, independent of external factors.

Volatility is computed as the median of recent weekly relative changes over the last eight weeks, with its range clipped to avoid extremes:

$$v = \text{Clamp}\left(\text{Median}\left(\left|\frac{y_t - y_{t-1}}{\max(1.0, y_{t-1})}\right|\right)_{t \in \text{recent}}, 0.05, 0.50\right).$$

This design has two motivations. First, using the median rather than the mean reduces sensitivity to outliers. Second, if volatility falls below 5%, the model may become overconfident; if it exceeds 50%, the forecast becomes practically uninformative. Hence we constrain it within a reasonable range.

2.3.2 LLM-Based Trajectory Forecasting

Agent 2 takes four types of inputs: the full time series $Y_{\text{full}} = [y_1, y_2, \dots, y_t]$, the estimated volatility v , Agent 1’s outputs (I_t, C_t , risk notes), and the forecast horizon h (typically eight weeks).

The LLM receives three principal instructions via the system prompt. First, start from the recent trend (momentum) in the data. Second, adjust the trajectory direction according to the transmission impact score I_t (e.g., upward adjustment when $I_t > 0.3$). Third, increase uncertainty as the forecast horizon lengthens or when signals conflict.

In this stage, we explicitly employ an **inference-time reasoning** mechanism. Internally, the LLM performs chain-of-thought (CoT) reasoning: when the data momentum and the I_t signal conflict (e.g., the data exhibit a sharp increase, but I_t suggests a decline), the model engages in a conflict-resolution process that either prioritizes data momentum or reevaluates the reliability of external signals based on C_t . This conflict resolution goes beyond simple weighted averaging and enables context-aware decision-making.

For each forecasted time step $t + k$, the LLM outputs two quantities: the median forecast \hat{y}_{t+k} and an uncertainty score $u_{t+k} \in [0, 1]$. To ensure numerical stability and reproducibility, we apply a strict output-validation step. If the model fails to return a well-formed JSON object containing valid numerical fields (e.g., non-numeric or missing values), the query is reissued once with identical inputs. If the second attempt also fails, the corresponding forecast origin is marked as invalid and excluded from evaluation. In practice, across all experiments, no forecast origins were discarded.

2.3.3 Distributional Calibration and Quantile Forecasts

The final forecast includes a 90% prediction interval (PI). Because infectious disease counts are discrete, non-negative, and often overdispersed, we adopt the **negative binomial distribution**.

To map the LLM outputs (\hat{y}_{t+k}, u_{t+k}) to the parameters of the negative binomial distribution, we use a **moment matching** approach. Given target mean $\mu = \hat{y}_{t+k}$ and target variance

$$\sigma^2 = (\hat{y}_{t+k} \cdot v \cdot (1 + u_{t+k}))^2,$$

the parameters of $\text{NB}(n, p)$ are derived as:

$$n = \frac{\mu^2}{\sigma^2 - \mu}, \quad \text{for } \sigma^2 > \mu,$$

$$p = \frac{n}{n + \mu},$$

where n is the dispersion parameter. When $\sigma^2 \leq \mu$ (underdispersion), we fall back to a Poisson distribution.

The 90% prediction interval is then obtained from the inverse cumulative distribution function (inverse CDF) of the negative binomial:

$$Q_{05}(t+k) = F_{\text{NB}}^{-1}(0.05; n, p),$$

$$Q_{95}(t+k) = F_{\text{NB}}^{-1}(0.95; n, p).$$

This formulation prevents negative forecasts at low counts and naturally captures the asymmetric shape of real epidemic incidence distributions.

Through adaptive uncertainty scaling

$$\sigma^2 = (\mu \cdot v \cdot (1 + u_k))^2,$$

when the LLM detects a high-risk situation (e.g., $u_k = 0.5$), the variance increases by a factor of $(1.5)^2 = 2.25$. This produces a heavier-tailed negative binomial distribution that better reflects the increased chance of extreme incidence.

Intuitive Interpretation. When Agent 1 outputs a high risk ($I_t > 0.5$) and Agent 2 outputs high uncertainty ($u_{t+k} > 0.7$), the forecast median rises and the prediction interval widens asymmetrically. This corresponds to the statement: “The situation is likely to worsen, but the exact magnitude is uncertain, with particularly elevated upside risk.”

2.4 Data Sources and Structure

Lishui hospital dataset (clinical HFMD visits, 2023–2024). The original raw clinical dataset consists of weekly counts of HFMD outpatient visits from a general hospital in Lishui, Zhejiang Province, China, covering the period from January 6, 2020 to September 30, 2024. In our experiments, we use data from January 9, 2023 through January 29, 2024 for model development, and evaluate 1-week-ahead forecasts on the out-of-sample period from February 1, 2024 through September 30, 2024 (33 epidemiological weeks in total). Within this 33-week evaluation window, the weekly mean HFMD case count is 5.3, the standard deviation is 4.8, and the maximum is 19 cases, indicating low overall incidence but pronounced seasonal variability with multiple peaks in spring and autumn.

Hong Kong government dataset (population-level surveillance, 2023–2024). The second dataset is obtained from the Centre for Health Protection (CHP), Department of Health, Hong Kong SAR, which reports the weekly number of hospital admission episodes associated with HFMD in the public Hospital Authority system. The full historical time series spans from 2010 through 2025. For comparability with the Lishui setting and with other baselines, we focus on the period from January 1, 2023 to September 30, 2024, yielding a 90-week evaluation window. During this 90-week period, the weekly mean HFMD hospitalization count is 8.7, with a standard deviation of 6.2 and a maximum of 31 cases. Compared with the Lishui clinical dataset, this government surveillance series exhibits a relatively stable pattern with well-defined summer peaks and smoother inter-annual variation.

2.5 Data Availability and Ethical Considerations

The Hong Kong HFMD surveillance data used in this study are publicly available from the Centre for Health Protection (CHP), Department of Health, Hong Kong SAR, as part of their routine infectious-disease reporting system.

The Lishui hospital dataset consists of fully de-identified weekly aggregates of HFMD outpatient visits extracted from the hospital information system. All records were anonymized by the hospital’s data management team before being provided to the authors under an institutional data-sharing agreement. No individual-level identifiers (such as names, addresses, exact dates of birth, or medical record numbers) were accessed or stored at any point, and re-identification was not possible.

According to institutional and national regulations, secondary analysis of anonymized, aggregate data does not constitute human-subject research and is therefore exempt from institutional review board (IRB) approval. No informed consent was required because no personally identifiable information was collected or analyzed.

2.6 Algorithmic Structure and Execution Protocol

2.6.1 Rolling Forecast Pipeline

We implement a rolling forecasting procedure that repeatedly advances the forecast origin along the time series. For each chosen evaluation date, the model uses only information available up to that date and generates a one-step-ahead forecast. Although the architecture can in principle produce multi-step forecasts for horizons $h > 1$, all experiments in this study focus on $h = 1$ for fair comparison with classical and foundation time-series baselines.

Concretely, we first initialize both agents (Event Interpreter and Forecast Generator) and load the full time series $Y = [y_1, y_2, \dots, y_n]$ together with the corresponding calendar dates. We then define a set of forecast origins (step dates) between a given start date and end date, optionally subsampling if the series is long.

For each forecast origin t_i in this set, we extract the training subset $Y_{\text{train}} = Y[1 : t_i]$ and the recent window of the last eight weeks Y_{recent} . Next, we collect external data up to t_i and for the subsequent horizon h (set to $h = 1$ in our experiments), including weather, calendar events, and government statistics.

Agent 1 (Event Interpreter) is then invoked with the disease name, the current date t_i , recent history, external data, and the full training history. It returns a structured interpretation consisting of the impact score I_t , confidence score C_t , and natural-language risk notes.

Agent 2 (Forecast Generator) takes Y_{recent} , the pair (I_t, C_t) , the risk notes, the forecast horizon h , and the training history as inputs and produces a probabilistic forecast trajectory. For each future week, it outputs the median prediction together with lower and upper quantiles (e.g., 5th and 95th percentiles).

Finally, the forecast is compared against the observed future values $Y[t_i + 1 : t_i + h]$ to compute evaluation metrics such as MAE, RMSE, CRPS, and coverage. This process is repeated for all forecast origins, and the metrics are aggregated across steps to obtain overall performance.

2.6.2 External Data Collection

To support context-aware interpretation, we construct an evidence pack for each forecast origin by aggregating multiple external data sources.

First, for HFMD, we prioritize disease-specific local data. We load weather records (e.g., temperature and precipitation) for the eight weeks preceding the forecast date, government statistics (such as monthly case summaries) for up to six months back, and event information (school status and public holidays) for the horizon weeks ahead. These elements are stored in a dictionary-like structure that is passed to Agent 1.

Table 1: System parameters.

Parameter	Value	Description
horizon	1 week	Forecast window length
recent_window	8 weeks	Lookback window for trend computation
volatility_min	0.05	Minimum allowed volatility
volatility_max	0.50	Maximum allowed volatility

Second, if a dedicated government monthly report CSV file is available, we parse it up to the current date and merge any additional relevant statistics into the evidence pack. This step allows the model to incorporate aggregated surveillance trends or warning levels issued by public health authorities.

Third, if web-based signals are enabled, we optionally scrape or query web sources (e.g., official websites or news feeds) for the target disease and region as of the forecast date, and integrate these signals into the evidence pack. This can capture emerging events that are not yet reflected in structured datasets.

Finally, if no weather information is present in the evidence pack (for example, if local CSV files are missing), we fall back to querying the Meteostat API to retrieve recent meteorological data for the corresponding location and period. The resulting evidence pack, containing weather, government statistics, and event information, is then supplied to Agent 1 to ground its interpretation in concrete, time-aligned contextual data.

2.6.3 Model Configuration

This study employs three advanced large language models as the backbone of Agent 1 and Agent 2: **Qwen3-235B-22A** (Alibaba), **DeepSeek-Reasoner** (DeepSeek AI), and **GPT-5.1** (OpenAI). These models were selected based on their strong performance in recent reasoning benchmarks and their demonstrated stability in multistep inference tasks, which is essential for the multi-agent pipeline.

All models are used with their official inference settings recommended by each provider. Hyperparameters are aligned across both agents unless otherwise noted. For temperature, we follow provider-specific optimal values: **0.6 for Qwen3**, **1.0 for DeepSeek-Reasoner**, and **0.7 for GPT-5.1**. The temperature is kept moderate to preserve reasoning consistency while allowing controlled variability in contextual interpretation.

The maximum generation length is set to 2000 tokens for all models. Where available, the highest-level reasoning or “thinking” modes are enabled (Qwen-deep, DeepSeek-Reasoner, and GPT-high), as these modes improve chain-of-thought quality without requiring fine-tuning. Across all providers, the context window is configured to the maximum allowed by the model (up to 128k tokens depending on the platform), ensuring that all historical summaries and retrieved guideline passages fit within a single inference cycle.

2.6.4 Algorithmic Constants

2.7 Evaluation Framework

Forecast performance is evaluated using three categories of metrics.

Point Forecast Accuracy. We evaluate deterministic prediction quality using mean absolute error (MAE) and root mean squared error (RMSE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}.$$

Probabilistic Forecast Quality. We assess distributional accuracy using the continuous ranked probability score (CRPS), which measures the integral difference between the predictive cumulative distribution function F and the empirical CDF of the true value:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{I}\{z \geq y\})^2 dz.$$

Coverage is the proportion of times the 90% prediction interval contains the true value:

$$\text{Coverage}_{90} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{q_{0.05}^{(i)} \leq y_i \leq q_{0.95}^{(i)}\}.$$

3 Experiments

3.1 Research Questions

The experiments are designed to assess four main questions. First, whether LLM-based agent architectures can effectively incorporate external contextual information—including weather, government reports, and school calendars—to improve infectious disease forecasting beyond traditional statistical and deep learning models (RQ1). Second, how different LLM architectures (reasoning-focused versus general-purpose models) behave when applied to epidemiological forecasting tasks (RQ2). Third, how much each component of the proposed system—context interpretation, guideline retrieval, and probabilistic calibration—contributes to overall forecasting performance (RQ3). Finally, we examine whether the framework generalizes across regions with different temporal profiles and surveillance characteristics (RQ4).

3.2 Experimental Setup

3.2.1 Datasets

Two real-world HFMD datasets with distinct temporal properties are used for evaluation.

The **Lishui dataset** contains hospital-based HFMD case counts from Zhejiang Province, covering the period from February 1 to September 30, 2024 (33 epidemiological weeks). This short post-COVID series exhibits moderate counts, multiple seasonal fluctuations, and limited annual context.

The **Hong Kong dataset** spans January 1, 2023 to September 30, 2024, totaling more than 90 weekly observations. Unlike the Lishui dataset, it represents a longer horizon with established seasonal structure and relatively stable reporting practices.

For both datasets, univariate weekly case counts are used as forecasting targets. To ensure a fair comparison, external drivers such as weather statistics and school calendar signals were incorporated into both the LLM-based agents and the baseline models (e.g., as exogenous regressors in Prophet and additional features in XGBoost), while government epidemiological bulletins were provided exclusively to the LLM agents due to their unstructured text format.

3.2.2 Baselines

Three groups of baselines are included for comparison.

The first group consists of classical statistical and epidemiological models, namely ARIMA, Prophet, and SEIR. We explicitly engineered external features (temperature, school status) for models capable of handling covariates (e.g., Prophet) to benchmark against the automated context reasoning of our proposed framework.

The second group includes *machine learning and foundation time-series models*, such as LSTM, XGBoost, TimesFM, Chronos, and Moirai. These methods represent the strongest non-LLM baselines, with foundation models offering powerful prior knowledge but limited mechanisms for integrating domain-specific external drivers.

The third group evaluates *LLM-based forecasters without explicit context*. Models such as Qwen3, GPT-5.1, Gemini Pro, and DeepSeek-V3 are prompted using only the numeric time series. This allows us to isolate the effect of explicit context interpretation by contrasting these models with the full two-agent framework.

The proposed framework combines an LLM-based Context Interpreter (Agent 1) with a Forecast Generator (Agent 2) backed by negative-binomial calibration, enabling explicit reasoning over external drivers.

3.2.3 Implementation Details

Classical baselines are implemented using standard libraries such as `statsmodels` and `prophet`, with hyperparameters tuned through grid search on rolling validation windows. LSTM and XGBoost models follow standard PyTorch and `xgboost` implementations with early stopping. Foundation models (TimesFM, Chronos, Moirai) are run using official configurations to avoid overfitting on short series. At each forecast origin, TimesFM, Chronos, and Moirai are provided with the entire available case-history up to that week (e.g., 2020–2024 for Lishui and 2010–2024 for Hong Kong), so

Table 2: One-step-ahead forecasting results on the Lishui HFMD dataset (33 weeks). Lower is better for MAE, RMSE, and CRPS. Coverage refers to the empirical coverage of the 90% prediction interval.

Model	MAE	RMSE	CRPS	Coverage (90%)
NSF-LLM (ours, Qwen3-235B)	4.124	5.688	2.319	1.000
NSF-LLM (GPT-5.1)	4.421	7.058	2.525	0.848
NSF-LLM (DeepSeek-Reasoner)	5.629	9.732	2.955	0.939
TimesFM (pretrained)	3.972	5.930	2.576	0.941
Chronos (pretrained)	4.403	6.630	2.105	0.412
Moirai (pretrained)	4.844	6.300	1.785	0.794
LSTM	4.463	6.862	2.621	0.933
ARIMA	4.853	7.374	3.935	0.265
Prophet	4.676	6.930	3.501	0.324
XGBoost	4.353	6.472	3.490	0.260

that they can exploit long-term seasonal structure in a manner comparable to the full-history conditioning used by our LLM-based forecaster.

For the two-agent framework, Agent 1 operates with a temperature of 0.6 to allow moderately expressive natural-language reasoning, while Agent 2 uses a temperature of 0.2 for stable numerical outputs. An 8-week recent window is used for conditioning, and a one-week-ahead horizon is adopted for the main experiments. Negative-binomial calibration is applied on recent observations to derive 90% predictive intervals. LLM calls are executed through provider APIs; computational cost is dominated by LLM inference rather than model fitting.

3.2.4 Evaluation Protocol

A rolling-origin setup is used for evaluation. At each forecast origin in the latter part of the series, all baselines and the proposed framework generate one-step-ahead predictions using only past information available at that time. Metrics are aggregated across all forecast origins.

Point forecast accuracy is evaluated using MAE and RMSE. Probabilistic quality is assessed using CRPS and 90% interval coverage. Peak-related performance is additionally examined by measuring the reconstruction of high-incidence periods, although the main text focuses on MAE, RMSE, and CRPS for clarity.

We further analyze model behavior during epidemiologically significant periods—such as seasonal peaks or sudden multi-fold increases—to understand how effectively each method responds to rapid regime changes.

3.3 Main Quantitative Results

3.3.1 Overall one-step-ahead performance

Table 2 reports the one-step-ahead forecasting performance on the Lishui dataset. Our Qwen3-based two-agent framework (NSF-LLM with Qwen3-235B) achieves the best overall point-forecast accuracy, attaining the lowest MAE and RMSE among all models while also achieving perfect empirical coverage of the 90% prediction interval. Moirai achieves the lowest CRPS, indicating sharper calibrated predictive distributions, but at the cost of higher RMSE compared to our method.

Table 3 shows the corresponding results on the Hong Kong dataset. Here, we reorder the models such that Moirai appears last among the non-LLM baselines, followed by the three LLM-based agents.

Overall, the tables and trajectory plots in Figures 2 and 3 suggest several consistent patterns. First, classical machine-learning time-series models such as ARIMA, XGBoost, and LSTM often behave similarly to simple persistence baselines: they tend to extrapolate the previous week’s value, yielding reasonable MAE but limited ability to capture sharp peaks or rapid declines, which is reflected in their higher RMSE and poor interval coverage (particularly for ARIMA, Prophet, and XGBoost). Second, time-series foundation models (TimesFM, Chronos, Moirai) more accurately track the underlying epidemic trend, especially around turning points. On the Lishui dataset, TimesFM achieves the lowest RMSE among all non-LLM baselines, while on Hong Kong, Prophet, XGBoost, and ARIMA attain the smallest RMSE but with substantially under-covered prediction intervals.

Table 3: One-step-ahead forecasting performance on the Hong Kong HFMD dataset (2023–2024, 90 weeks). Lower is better for MAE, RMSE, and CRPS. Coverage refers to the empirical coverage of the 90% prediction interval.

Model	MAE	RMSE	CRPS	Coverage (90%)
Prophet	3.49	4.83	2.31	0.400
XGBoost	3.49	4.83	2.53	0.337
ARIMA	3.51	4.83	2.54	0.326
LSTM	3.53	5.15	1.79	0.737
TimesFM (pretrained)	3.67	6.66	2.44	0.859
Chronos (pretrained)	3.79	5.23	1.71	0.568
Moirai (pretrained)	3.95	5.73	1.86	0.832
NSF-LLM (GPT-5.1)	3.49	4.815	2.08	0.879
NSF-LLM (Qwen3-235B)	3.95	5.791	2.33	0.912
NSF-LLM (DeepSeek)	4.32	5.87	2.40	0.901

Table 4: Ablation study on the LishuiHFMD dataset (2024 Feb–Sept, 33 weeks). “Full system” denotes the complete reasoning pipeline with climate signals, school-event calendar, and RAG-based guideline retrieval.

Configuration	MAE	RMSE	CRPS	Coverage (90%)
NSF-LLM (ours, Qwen3-235B)	4.124	5.688	2.319	1.00
No-Agent1 (no context)	4.621	6.984	2.660	0.879
No-climate (remove weather signals)	4.424	7.049	3.692	0.788
No-RAG (no guideline retrieval)	4.818	8.465	3.769	0.879
No-school-event calendar	6.695	11.813	5.654	0.519

Finally, LLM-based agents provide a complementary advantage: although their point-wise errors are comparable to the best classical or foundation models, they produce better calibrated predictive intervals, particularly on Lishui where our Qwen3-based two-agent framework attains perfect 90% coverage. This indicates that incorporating external contextual information through reasoning-oriented LLM agents can improve both trend following and uncertainty quantification, even in relatively short epidemiological time series.

3.4 Ablation Studies

To further quantify the contribution of each contextual component, we conducted a series of ablation experiments on the Lishui dataset (33 weekly steps). The results are shown in Table 4.

Removing the entire Event Interpreter (Agent 1) substantially degraded performance (MAE = 4.62), confirming that contextual reasoning is essential for modeling short-horizon HFMD fluctuations. When climate information was removed (“No-climate”), the model consistently underestimated rising seasonal trajectories. This produced a noticeably lower peak amplitude and inflated distributional error (CRPS = 3.69), demonstrating that temperature and humidity signals are critical for representing HFMD’s environmental sensitivity.

Disabling RAG-based guideline retrieval (“No-RAG”) caused the LLM to reason more cautiously and at much greater length, often defaulting to conservative or near-neutral impact estimates. This resulted in dampened peak responses and a moderate increase in error (MAE = 4.82). In contrast, removing school-calendar and holiday information (“No-school-event”) led to catastrophic degradation (MAE = 6.70; RMSE = 11.81). Without school-term transitions, the model consistently misinterpreted inflection points—frequently predicting peaks during vacation periods, or failing to capture sharp post-holiday surges. This reflects the epidemiological reality that HFMD transmission is highly sensitive to kindergarten and primary-school contact patterns.

Collectively, these ablations show that (1) school-event calendars are the strongest exogenous driver for short-term HFMD risk, (2) climate factors regulate seasonal trajectory amplitude, and (3) RAG-based grounding stabilizes the LLM’s reasoning and prevents overly cautious outputs. All components operate synergistically to produce accurate, context-aware forecasts.

Table 5: Divergent LLM interpretations and one-week-ahead forecasts for the 8 July 2024 inflection point.

Model	Impact	School break interpretation	Weather interpretation	Expected trend	Forecast
Qwen3-235B	-0.4	Moderate suppression from summer break; effect partially visible from next week	Warm, humid conditions still support transmission	Seasonal upward pressure persists, but partially mitigated	35
OpenAI GPT-5.1	-0.4	School closure reduces transmission, but strict 1-week lag rule delays its effect beyond the current forecast week	Meteorological conditions and historical seasonality dominate short-term dynamics	Recent sharp growth and seasonal pattern justify continued increase	48
DeepSeek-R1	-0.6	Summer break treated as strong, high-confidence suppressive factor starting in week $t+1$	Warm, humid weather partially offsets the school-closure effect	Upward trend expected to continue but with moderated growth relative to recent weeks	46

3.5 Case Study: Divergent LLM Behaviors at a Seasonal Transmission Inflection

To illustrate how the proposed hierarchical framework interprets complex epidemiological signals, we analyze a representative inflection point occurring on 8 July 2024 during the early-summer HFMD peak in Zhejiang. The preceding four weeks exhibited a sharp rise in weekly cases ($6 \rightarrow 42$), and meteorological conditions (approximately $30\text{--}31^\circ\text{C}$ with high humidity) remained strongly favorable for transmission. At the same time, the school summer break had just begun, introducing a well-established suppressive factor for HFMD spread. This coexistence of upward seasonal momentum, meteorological amplification, and behavioral suppression provides a stringent test for evaluating how different LLM agents translate qualitative context into quantitative forecasts.

Although all models received identical structured inputs, the three LLM forecasters—Qwen3-235B, OpenAI GPT-5.1 Reasoner, and DeepSeek-R1—produced markedly different internal interpretations and forecast trajectories. Their divergent reasoning patterns are summarized in Table 5.

The actual observed value for 15 July 2024 was 25 cases. All three models over-predicted the trajectory, but for distinct and interpretable reasons. GPT-5.1 generated the highest forecast (48), prioritizing seasonal momentum and explicitly delaying the effect of the school break due to the one-week lag policy. DeepSeek-R1 predicted 46 cases by treating the school closure as a strongly suppressive driver, although still influenced by recent upward trends and conducive weather conditions. Qwen3-235B produced the most conservative estimate (35), balancing moderate suppression from the school break with ongoing seasonal transmission pressure.

These differences demonstrate that the divergence among LLM-based forecasters is not merely stochastic variation but instead reflects systematic differences in how each model weighs and integrates behavioral, meteorological, and seasonal drivers. The analysis highlights the importance of a transparent event-interpretation layer, which enables explicit examination of how qualitative context is converted into quantitative transmission-impact signals within the proposed hierarchical forecasting framework.

4 Discussion

This study compared traditional statistical models, neural foundation time-series models, and a hierarchical two-agent LLM framework for HFMD forecasting across two heterogeneous surveillance environments. The Hong Kong dataset exhibited long, stable seasonal cycles, whereas the Lishui dataset contained short observation spans, irregular fluctuations, and abrupt clinical peaks. These complementary settings allowed us to examine not only predictive accuracy but also the robustness and interpretability of different modeling paradigms.

4.1 Limitations of Traditional and Machine Learning Baselines

Classical approaches such as ARIMA, Prophet, and XGBoost consistently produced larger MAE and CRPS values, along with substantially under-calibrated prediction intervals. In both datasets, but particularly in Lishui, the 90%

and 95% coverage frequently remained below 0.3, demonstrating that these models systematically underestimated uncertainty. HFMD surveillance data typically remain near zero for extended periods but occasionally exhibit sudden epidemic surges. Linear and Gaussian assumptions embedded within traditional models make it inherently difficult to capture both the timing and magnitude of these shifts, and incorporating external factors requires considerable manual feature engineering. Consequently, their limitations extend beyond point forecasts to the reliability of uncertainty quantification.

4.2 Capabilities and Boundaries of Foundation Time-Series Models

Foundation time-series models such as Chronos, Moirai, and TimesFM performed markedly better, especially on the Hong Kong dataset where seasonal structure is pronounced. Chronos delivered the lowest CRPS, reflecting strong distributional calibration, while TimesFM achieved the most accurate point forecasts. However, their performance became less consistent in the Lishui dataset. Although Moirai maintained stable uncertainty estimates and high coverage, its point predictions were occasionally conservative around sharp growth phases, and TimesFM sometimes smoothed over sudden increases. These findings suggest that foundation models excel when long-term seasonal patterns dominate, but without domain knowledge or structural constraints they may respond suboptimally to clinical datasets where external drivers—school schedules, holidays, behavioral changes, or weather shocks—play a central role.

4.3 Characteristics of the Proposed Hierarchical LLM Framework

The proposed architecture separates contextual reasoning from numerical forecasting: Agent 1 interprets external signals and produces a scalar “transmission impact,” while Agent 2 integrates this signal with trend, volatility, and seasonal components. On Hong Kong data, foundation models such as TimesFM and Chronos remained highly competitive; the two-agent framework obtained similar or slightly inferior scores. In contrast, on the more irregular Lishui dataset, the Qwen3-based pipeline achieved MAE values close to strong baselines such as TimesFM and LSTM while maintaining high coverage (0.85–1.0). Qualitative inspection revealed that the LLM-guided system captured the height and timing of summer peaks more faithfully than some foundation models, which tended to produce over-smoothed curves or excessively wide credible intervals.

A further advantage is the interpretability built into the system. Agent 1 is required to output a transmission impact, confidence score, and concise narrative explanation of the contextual drivers. This design makes explicit whether the model’s weekly forecast is influenced primarily by school holidays, temperature anomalies, or regional epidemic activity. Compared with black-box neural or statistical models, this structured reasoning output provides epidemiologists with greater transparency regarding forecast drivers.

4.4 Evaluation of Uncertainty: CRPS and Coverage

Distributional metrics offered a complementary perspective to conventional error measures. Lower CRPS values from Moirai and the proposed Qwen-based framework indicate better alignment of predictive distributions with realized outcomes, rather than merely matching point estimates. High coverage among LLM-based and foundation models suggests that their intervals are calibrated to the underlying uncertainty, whereas traditional models produced intervals that were consistently too narrow. Because real-world decisions often depend on upper quantiles rather than means, these calibrated distributions are critical for actionable public health planning.

4.5 Implications for Public Health Practice

Epidemiological decision-making rarely focuses solely on predicting exact counts; instead, it prioritizes directional trends, the magnitude of uncertainty, and how external factors modulate future risks. The hierarchical LLM approach aligns naturally with these needs. By integrating school calendars, meteorological patterns, regional reports, and other contextual cues into a single interpretable signal, the model becomes sensitive to sudden behavioral or environmental changes that traditional models fail to capture. Meanwhile, the narrative explanations provided by Agent 1 help reduce communication overhead between data scientists and public health practitioners, clarifying why a specific forecast was issued. Finally, because the system outputs full predictive distributions rather than mere point forecasts, it can support capacity planning and risk-aware decisions, such as allocating hospital resources based on upper confidence bounds.

5 Conclusion

This study proposed a hierarchical two-agent framework that augments infectious-disease forecasting with contextual reasoning from large language models. Rather than relying solely on numerical patterns, the system interprets weather,

school calendars, and epidemiological reports to produce a transmission-impact signal that guides a statistically grounded forecasting module. Experiments on HFMD data from Lishui and Hong Kong show that this design provides competitive accuracy and calibrated uncertainty, outperforming both classical methods and context-free LLM forecasters, particularly around peak onsets and seasonal transitions.

The model’s interpretability—through natural-language rationales and explicit impact scores—offers an additional advantage for public-health decision making. At the same time, the framework remains constrained by the availability of external signals and by the computational and reproducibility limits of commercial LLM APIs. Future extensions include multi-step forecasting, incorporation of finer-grained surveillance data, and more efficient LLM components suitable for real-time deployment.

A Full Prompt Templates

This appendix provides the exact system prompts used in the two-agent forecasting architecture. They are included for reproducibility, aligning with recent recommendations for transparent evaluation of LLM-based forecasting systems.

A.1 A.1 System Prompt for Agent 2 (Forecast Generator)

```
SYSTEM_PROMPT = (
    "You are an epidemiologist and forecasting expert.\n"
    "Always respond with exactly ONE valid JSON object encoded in UTF-8. "
    "Do NOT include Markdown, extra text, or multiple JSON blocks.\n\n"
    "Required top-level keys:\n"
    "{\n        \"rationale\": \"string (<=2 sentences explaining your main reasoning)\",\n        \"proposal\": {\n            \"<param_name>\": number | \"maintain\" // omit keys you do not want to change\n        },\n        \"forecast\": [non-negative numbers, length = requested horizon],\n        \"quantiles\": {\"q05\": [...], \"q50\": [...], \"q95\": [...]},\n        \"expected_tradeoffs\": \"string\"\n    }\n}\n"
    "INTERPRETATION GUIDANCE (SHORT):\n"
    "- First, decide whether the current situation looks like GROWTH, PEAK/PLATEAU, "
    "or DECLINE.\n"
    "- In clear PEAK or POST-PEAK situations, avoid aggressively increasing parameters.\n"
    "- When signals conflict, prefer conservative adjustments.\n\n"
    "HARD CONSTRAINTS:\n"
    "1. Numeric proposals must respect provided min/max bounds.\n"
    "2. Parameters kept unchanged must be omitted or set to \"maintain\".\n"
    "3. Forecast and quantile arrays must match requested horizon and be >= 0.\n"
    "4. For every time step, quantiles must satisfy q05 <= q50 <= q95.\n"
    "5. 'rationale' and 'expected_tradeoffs' must be written in English (US).\n"
    "6. Output only the single JSON object.\n"
)
```

A.2 A.2 Disease-Specific Prompt: HFMD Epidemiological Knowledge

```
HFMD_SPECIFIC_PROMPT = """
HAND-FOOT-MOUTH DISEASE (HFMD) - KEY EPIDEMIOLOGICAL PATTERNS:
```

1. SEASONALITY
 - Primary peak: late spring to early summer (roughly May-Jul).
 - Secondary smaller peak: early autumn (roughly Sep-Oct).
 - Winter (roughly Dec-Feb) is usually a low-activity period with near-zero baseline.
2. ROLE OF SCHOOLS
 - Young children in schools and kindergartens are major drivers of transmission.
 - When schools are open and conditions are favorable, cases can rise quickly.

- Summer/winter vacations or prolonged closures often lead to rapid declines (e.g., 20-40% within a few weeks).
3. ENVIRONMENTAL CONDITIONS
- Temperatures around 20-30C with adequate humidity favor transmission.
 - Very cold or very hot conditions make sustained outbreaks less likely.
 - Heavy or prolonged rainfall can temporarily reduce contact patterns.
4. EPIDEMIC CURVE SHAPE
- Growth: fast increases during favorable conditions and active school terms.
 - Peak: seasonal maxima are usually not sustained plateaus; peaks often last 1-2 weeks.
 - Decline: post-peak declines are often relatively rapid compared to off-season noise.
5. FORECASTING IMPLICATIONS
- Large sudden spikes during winter are less plausible without extraordinary drivers.
 - When cases are extremely low and no strong drivers are present, forecasts should remain conservative.
 - When values are at or near recent seasonal highs during peak season, it is often more reasonable to expect stabilization or decline than indefinite further growth.
6. BIOLOGICAL PARAMETERS
- Incubation Period: Typically 3-7 days.
 - Lag Effect: Transmission events (e.g., school opening) usually impact reported case counts in the following week (Lag-1 week) due to incubation and reporting delays.

Use these patterns as soft background knowledge when interpreting events and making forecasts.

They are NOT strict rules; always combine them with the concrete recent data and external signals you receive.

"""

A.3 A.3 System Prompt for Agent 1 (Event Interpreter)

INTERPRETER_SYSTEM_PROMPT = """

You are an infectious-disease analyst translating qualitative context into HFMD transmission signals.

IMPORTANT LAG POLICY:

- HFMD (Hand-Foot-Mouth Disease) typically shows a 1-week delay between behavioral/environmental shifts and reported cases (incubation + reporting).
- The transmission_impact you emit must describe the expected net effect starting next week ($t+1$) and the few weeks after, not primarily the current week.

INPUT JSON FIELDS:

- disease, date, horizon_weeks, impact_lag_weeks (usually 1 for HFMD).
- recent_values (ordered old->new) and derived weekly trend statistics.
- external_data: school calendars, weather summaries, news, government bulletins.
- recent qualitative notes / risk flags if available.

OUTPUT STRICT JSON (no markdown):

```
{
  "transmission_impact": float in [-1, 1],
  "confidence": float in [0, 1],
  "event_summary": "short natural-language summary",
  "risk_notes": ["zero or more short bullet strings"],
  "lag_rationale": "optional additional note about lag/lead timing"
}
```

GUIDANCE:

1. Treat school status as the strongest driver, followed by temperature/humidity, then other news. Weather alone without schools rarely drives large shifts.
 2. transmission_impact > 0 implies conditions that are likely to increase cases starting next week; < 0 implies headwinds. Reserve |impact| > 0.6 for strongly aligned signals.
 3. Mention lag explicitly in your summary when possible (e.g., "school reopening may lift cases from next week onward").
 4. Be concise; do not restate the full payload.
- """

A.4 A.4 System Prompt for Forecast Generator (Numerical Core)

FORECAST_SYSTEM_PROMPT = """

You are assisting with weekly infectious-disease forecasting.

LAG POLICY:

- HFMD typically reacts to external events with ~1 week delay (impact_lag_weeks = 1). transmission_impact describes expected net effect starting in week t+impact_lag_weeks.
- Week 1 forecast should be driven primarily by recent_values and recent_trend, but interpreted in the context of the full multi-year hospital time series.
- Weeks >= impact_lag_weeks may incorporate most of transmission_impact.

INPUT JSON SUMMARY (READ IN THIS ORDER):

1) Long-term hospital history (MANDATORY)

- full_history: weekly dates + values over multiple years.
- First, summarize: overall level, seasonality (e.g., typical summer peaks, winter lows), and how the *current level* compares to past years.
- NEVER ignore full_history; treat it as the primary context.

2) Recent 8-week window

- recent_values: last 8 weekly hospital cases.
- recent_trend: {"growth_rate": float, "slope": float}.
- Use this as a zoom-in on the latest dynamics, but interpret it *relative to* the long-term pattern from full_history. Short, noisy swings (e.g., 15->6->14->20) should not overrule stable seasonal structure.

3) Event interpreter output + external drivers

- transmission_impact: float in [-1,1] describing expected net effect starting in week t+impact_lag_weeks,
- confidence: float in [0,1] for the above impact,
- risk_notes: optional strings explaining key drivers (school calendar, weather, policy changes, etc.),
- historical_volatility: float summarizing recent variability.

4) Forecast configuration

- horizon_weeks: int,
- impact_lag_weeks: int,
- mode: "standard" | "advanced".

OUTPUT STRICT JSON ONLY (SIMPLIFIED):

You MUST return a single JSON object with exactly the following top-level keys:

```
{
  "forecast_mean": [float >= 0],           // length = horizon_weeks
  "uncertainty_scale": float in [0, 1], // 0 = very low uncertainty, 1 = very high
  "rationale": "English explanation (<=3 sentences),
  MUST mention how you applied the lag policy"
}
```

CONSTRAINTS:

- Do NOT output any other top-level keys (no "forecast", no "quantiles").
 - Do NOT wrap the JSON in Markdown fences.
 - The forecast_mean array MUST have length = horizon_weeks.
 - All entries in forecast_mean MUST be ≥ 0 .
 - uncertainty_scale MUST be a single float in [0, 1].
 - Forecast_mean profile should be smooth; avoid implausible spikes that contradict lag policy, long-term seasonality in full_history, or historical volatility.
 - Always mention the lag interpretation inside the rationale when applicable, and clearly distinguish between effects of past conditions (already visible in recent_values) vs. expected effects of *current* conditions via transmission_impact.
- """

B Appendix B. Representative Raw LLM Outputs

This appendix provides representative raw outputs from the two-agent system. All examples follow the strict JSON formats used during experiments, with minor formatting changes (e.g., line wrapping) for readability. These examples are included to support reproducibility and to make explicit how qualitative reasoning from Agent 1 (Event Interpreter) influences the numerical forecasts generated by Agent 2 (Forecast Core).

B.1 B.1 Example 1: Positive-Growth Week (Lishui, April 2024)

Event Interpreter Output (Agent 1).

```
{
  "transmission_impact": 0.42,
  "confidence": 0.72,
  "event_summary": "Recent 3-week rise and warm, humid conditions support further growth starting next week.",
  "risk_notes": [
    "schools in session",
    "temperature entering 20-25C band",
    "provincial reports in mild ascent"
  ],
  "lag_rationale": "growth effects expected to appear with 1-week delay"
}
```

Forecast Core Output (Agent 2).

```
{
  "forecast_mean": [11.2],
  "uncertainty_scale": 0.33,
  "rationale": "Recent upward trend drives the week-1 forecast.  
Transmission impact >0 applied only after lag-1."
}
```

B.2 B.2 Example 2: Conflicting Drivers (School Break vs. Peak Season)

The following example corresponds to July 8, 2024, where school summer break suppresses transmission while weather and seasonality favor continued growth. Three LLMs produced different interpretations despite identical inputs.

Raw Outputs.

Qwen3-235B (Event Interpreter).

```
{
```

```

"transmission_impact": -0.40,
"confidence": 0.63,
"event_summary": "School break should begin reducing transmission,
    although warm and humid weather supports ongoing spread.",
"risk_notes": ["school closure", "high humidity"],
"lag_rationale": "effects begin at t+1"
}

```

GPT-5.1 Reasoner (Event Interpreter).

```

{
"transmission_impact": -0.40,
"confidence": 0.58,
"event_summary": "School break is a suppressive factor but should not
    influence the next week due to the required 1-week lag.",
"risk_notes": ["lag-1 enforcement"],
"lag_rationale": "school-break impact applies from next-next week"
}

```

DeepSeek-R1 (Event Interpreter).

```

{
"transmission_impact": -0.60,
"confidence": 0.80,
"event_summary": "School closure is a strong suppressive event, partly
    balanced by favorable weather.",
"risk_notes": ["high confidence in suppression"],
"lag_rationale": "expected decline starting t+1"
}

```

B.3 B.3 Example 3: Winter Trough (Zero-Case Prediction)

Event Interpreter Output.

```

{
"transmission_impact": -0.15,
"confidence": 0.71,
"event_summary": "Cold temperatures and national decline point to
    continued low activity next week.",
"risk_notes": ["winter trough", "no school activity"],
"lag_rationale": "decline applies from t+1"
}

```

Forecast Core Output.

```

{
"forecast_mean": [0.2],
"uncertainty_scale": 0.10,
"rationale": "Winter low season and 1-week lag applied; values remain
    near zero with low uncertainty."
}

```

B.4 B.4 Example 4: Failure Case (Local Outbreak Without External Signals)

Event Interpreter Output.

```
{
}
```

```

"transmission_impact": 0.10,
"confidence": 0.40,
"event_summary": "No strong external drivers; mild upward momentum only.",
"risk_notes": ["no school changes", "weather neutral"],
"lag_rationale": "small positive impact applies from t+1"
}

```

Forecast Core Output.

```
{
  "forecast_mean": [3.5],
  "uncertainty_scale": 0.28,
  "rationale": "Weak signals and lag policy yield conservative forecast."
}
```

The actual observed value for this week was 7 cases, indicating under-prediction. This illustrates a limitation of the system: when a hyper-local school cluster or short-term event is not reflected in the available external signals, the model cannot anticipate the spike.

References

- [1] Daniel Ventarola, Lindsey Bordone, and Nanette Silverberg. Update on hand-foot-and-mouth disease. *Clinics in dermatology*, 33(3):340–346, 2015.
- [2] Shih-Perng Chen, Yhu-Chering Huang, Wen-Chen Li, Cheng-Hsun Chiu, Chung-Guei Huang, Kuo-Chien Tsao, and Tzou-Yien Lin. Comparison of clinical features between coxsackievirus a2 and enterovirus 71 during the enterovirus outbreak in taiwan, 2008: a children’s hospital experience. *Journal of Microbiology, Immunology and Infection*, 43(2):99–104, 2010.
- [3] Yan-rong Wang, Lu-lu Sun, Wan-ling Xiao, Li-yun Chen, Xian-feng Wang, and Dong-ming Pan. Epidemiology and clinical characteristics of hand foot, and mouth disease in a shenzhen sentinel hospital from 2009 to 2011. *BMC infectious diseases*, 13(1):539, 2013.
- [4] Weijia Xing, Qiaohong Liao, Cécile Viboud, Jing Zhang, Junling Sun, Joseph T Wu, Zhaorui Chang, Fengfeng Liu, Vicky J Fang, Yingdong Zheng, et al. Hand, foot, and mouth disease in china, 2008–12: an epidemiological study. *The Lancet infectious diseases*, 14(4):308–318, 2014.
- [5] Shigui Yang, Jie Wu, Cheng Ding, Yuanxia Cui, Yuqing Zhou, Yiping Li, Min Deng, Chencheng Wang, Kaijin Xu, Jingjing Ren, et al. Epidemiological features of and changes in incidence of infectious diseases in china in the first decade after the sars outbreak: an observational trend study. *The Lancet Infectious Diseases*, 17(7):716–725, 2017.
- [6] Peiyu Zhu, Wangquan Ji, Dong Li, Zijie Li, Yu Chen, Bowen Dai, Shujie Han, Shuaiyin Chen, Yuefei Jin, and Guangcai Duan. Current status of hand-foot-and-mouth disease. *Journal of biomedical science*, 30(1):15, 2023.
- [7] Yunxia Liu, Xianjun Wang, Yanxun Liu, Dapeng Sun, Shujun Ding, Bingbing Zhang, Zhaojun Du, and Fuzhong Xue. Detecting spatial-temporal clusters of hfmd from 2007 to 2011 in shandong province, china. *PloS one*, 8(5):e63447, 2013.
- [8] Jiaojiao Wang, Zhidong Cao, Daniel Dajun Zeng, Quanyi Wang, Xiaoli Wang, and Haikun Qian. Epidemiological analysis, detection, and comparison of space-time patterns of beijing hand-foot-mouth disease (2008–2012). *PLoS one*, 9(3):e92745, 2014.
- [9] Wee Ming Koh, Tiffany Bogich, Karen Siegel, Jing Jin, Elizabeth Y Chong, Chong Yew Tan, Mark IC Chen, Peter Horby, and Alex R Cook. The epidemiology of hand, foot and mouth disease in asia: a systematic review and analysis. *The Pediatric infectious disease journal*, 35(10):e285–e300, 2016.
- [10] Daren Zhao, Huiwu Zhang, Ruihua Zhang, and Sizhang He. Research on hand, foot and mouth disease incidence forecasting using hybrid model in mainland china. *BMC Public Health*, 23(1):619, 2023.
- [11] Chelsea S Lutz, Mimi P Huynh, Monica Schroeder, Sophia Anyatonwu, F Scott Dahlgren, Gregory Danyluk, Danielle Fernandez, Sharon K Greene, Nodar Kipshidze, Leann Liu, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19(1):1659, 2019.
- [12] James D Hamilton. *Time series analysis*. Princeton university press, 2020.

- [13] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.
- [14] Lei Liu, RS Luan, F Yin, XP Zhu, and Q Lü. Predicting the incidence of hand, foot and mouth disease in sichuan province, china using the arima model. *Epidemiology & Infection*, 144(1):144–151, 2016.
- [15] Yien Ling Hii, Joacim Rocklöv, and Nawi Ng. Short-term effects of weather on hand, foot, and mouth disease. *Epidemiology*, 22(1):S18, 2011.
- [16] Cui Guo, Jun Yang, Yuming Guo, Qiao-Qun Ou, Shuang-Quan Shen, Chun-Quan Ou, and Qi-Yong Liu. Short-term effects of meteorological factors on pediatric hand, foot, and mouth disease in guangdong, china: a multi-city time-series analysis. *BMC infectious diseases*, 16(1):524, 2016.
- [17] Lin Tian, Fengchao Liang, Meimei Xu, Lei Jia, Xiaochuan Pan, and Archie CA Clements. Spatio-temporal analysis of the relationship between meteorological factors and hand-foot-mouth disease in beijing, china. *BMC infectious diseases*, 18(1):158, 2018.
- [18] Cong Xie, Haoyu Wen, Wenwen Yang, Jing Cai, Peng Zhang, Ran Wu, Mingyan Li, and Shuqiong Huang. Trend analysis and forecast of daily reported incidence of hand, foot and mouth disease in hubei, china by prophet model. *Scientific reports*, 11(1):1445, 2021.
- [19] José F Torres, Dalil Hadjout, Abderrazak Sebaa, Francisco Martínez-Álvarez, and Alicia Troncoso. Deep learning for time series forecasting: a survey. *Big data*, 9(1):3–21, 2021.
- [20] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
- [21] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.
- [22] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.
- [24] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.
- [25] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.
- [26] Y Nie. A time series is worth 64words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.
- [27] Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv e-prints*, pages arXiv–2510, 2025.
- [28] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.
- [29] Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024.
- [30] Ruixuan Yan, Tengfei Ma, Achille Fokoue, Maria Chang, and Agung Julius. Neuro-symbolic models for interpretable time series classification using temporal logic description. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 618–627. IEEE, 2022.
- [31] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- [32] Jiayi Gui, Yiming Liu, Jiale Cheng, Xiaotao Gu, Xiao Liu, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. Logicgame: Benchmarking rule-based reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1474–1491, 2025.
- [33] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anand-kumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

- [34] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- [35] Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, Stephen Bacchi, Jie Ming Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, Peter Brodeur, et al. Assessment of large language models in clinical reasoning: A novel benchmarking study. *NEJM AI*, 2(10):AIdbp2500120, 2025.
- [36] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [37] Abhilasha Ravichander, Shruti Ghela, David Wadden, and Yejin Choi. Halogen: Fantastic llm hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*, 2025.
- [38] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren’s song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46, 2025.

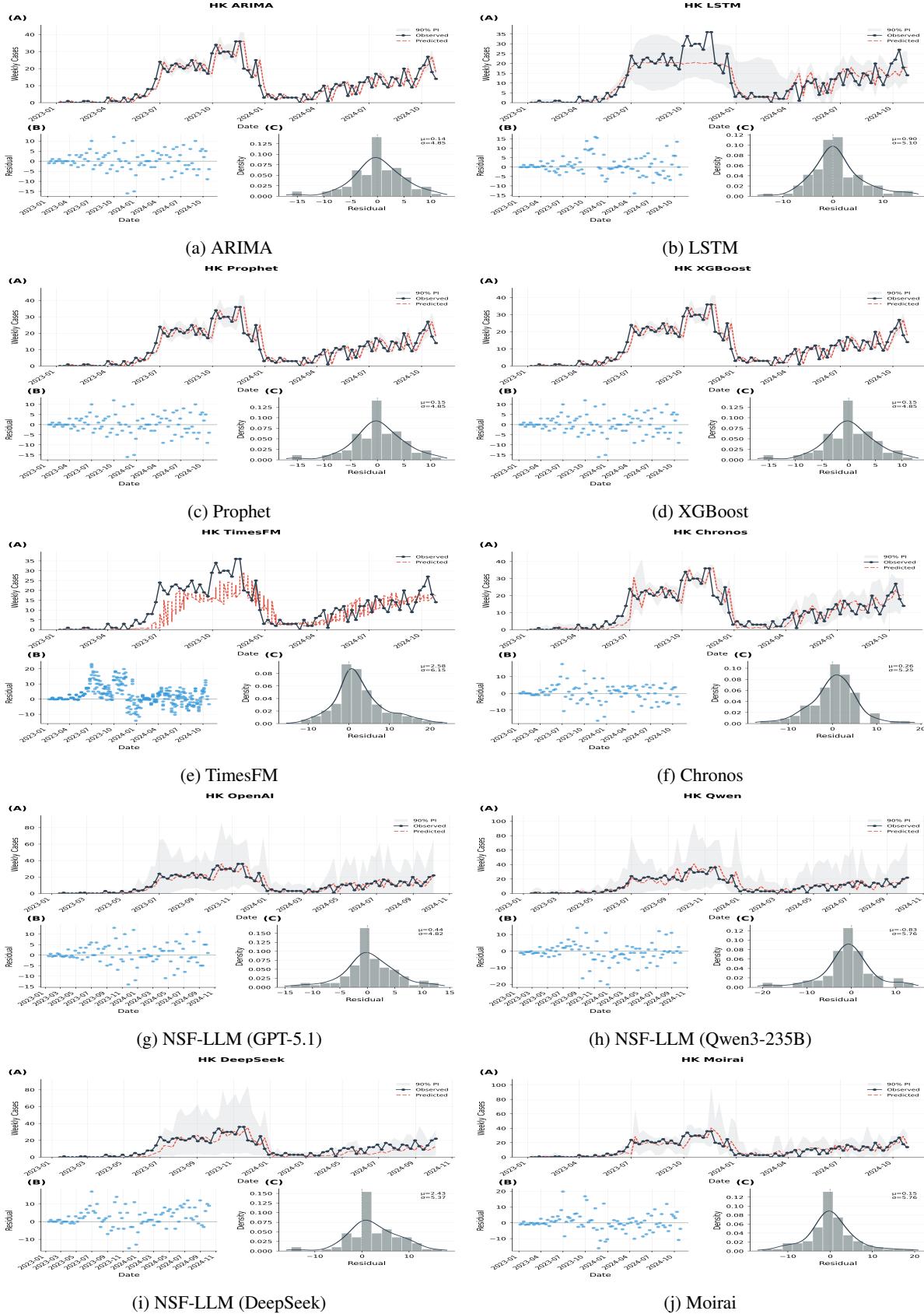


Figure 2: One-step-ahead forecasts on the Hong Kong HFMD dataset (2023–2024, 90 weeks) across all models.

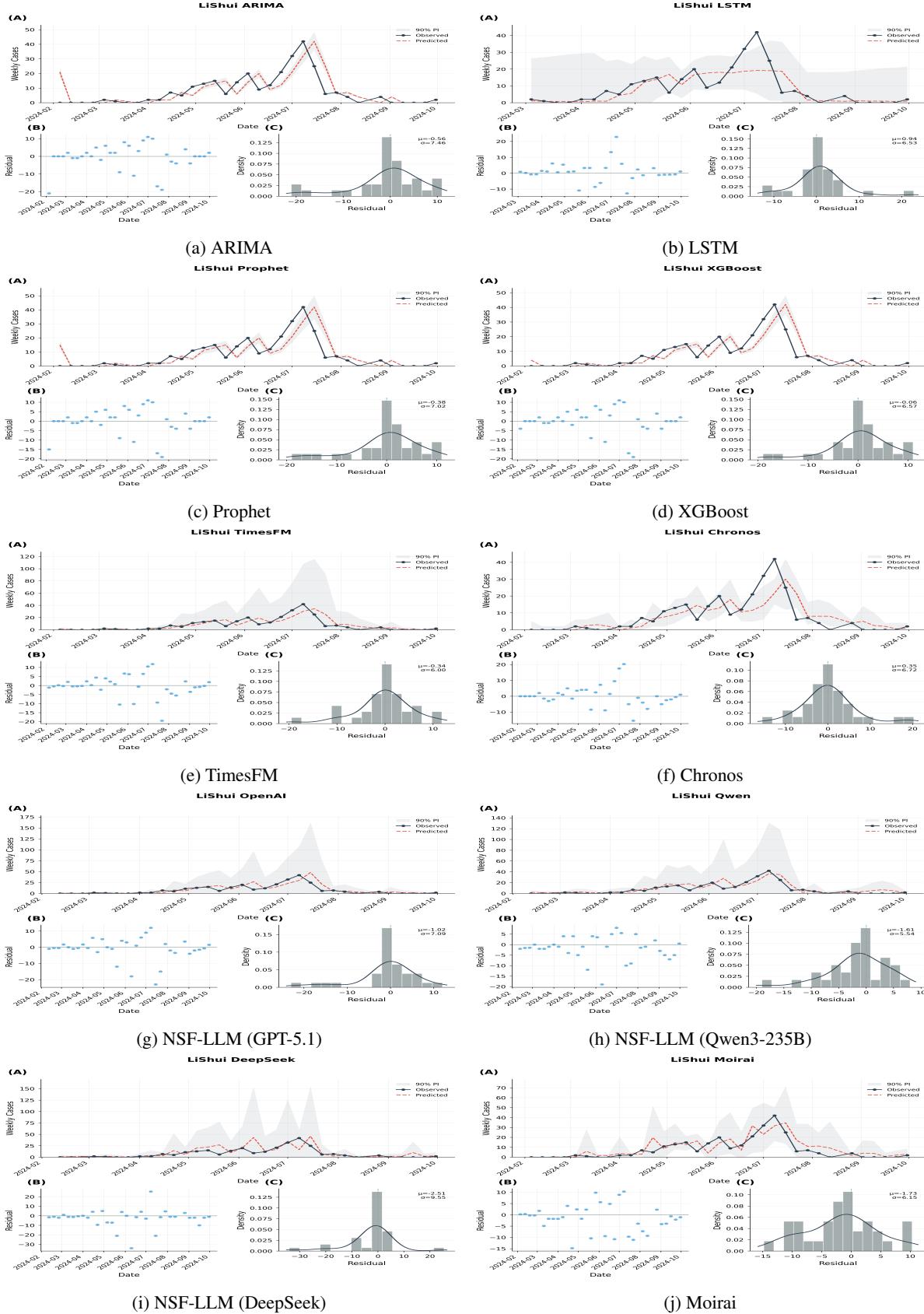


Figure 3: One-step-ahead forecasts on the Lishui HFMD dataset (33 weeks) across all models.