Can LLMs Forecast on Irregular Time Series?

Hanze Qin*
Xingjian Wang*
hqin68@gatech.edu
xwang3040@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

Seasonal influenza planning requires forecasts that remain reliable when surveillance data are noisy, irregular, or only partially observed. We evaluate whether large language models (LLMs) can provide such forecasts by comparing a zero-shot Gemini-2.5-Flash forecaster—using only serialized numeric inputs—with seasonal ARIMA and Gaussian Process (GP) models on ten seasons of CDC influenza-like illness (ILI) data, holding out 2024–2025 as a test year. Two settings are considered: robustness to irregular training data generated by 0–60% random missingness, and practical early-season forecasting where models observe only a thresholded portion of the current season. All methods are assessed using Mean Absolute Error (MAE). ARIMA and GP achieve the lowest and most stable errors, but the strongest LLM prompt becomes competitive with GP despite requiring no training or preprocessing. In the early-season setting, both statistical models produce usable forecasts from limited information, with GP showing greater stability. Overall, zero-shot LLMs are not yet a replacement for classical models, but they offer a practical and accessible complement for flu forecasting under irregular and incomplete data.

1 Introduction

Seasonal influenza control depends on timely, actionable forecasts. Public health agencies must decide when to expand clinic capacity, promote vaccination, or issue advisories—often before the seasonal peak is visible. Yet flu surveillance data are noisy and, outside well-resourced settings, often irregular due to reporting delays, sparse coverage, or missed weeks. These characteristics challenge many classical forecasting tools that assume clean, regularly sampled series.

Large Language Models (LLMs) offer a complementary path. Recent studies suggest that LLMs can act as *zero-shot* forecasters when numeric sequences are serialized as text, potentially enabling meaningful predictions without specialized modeling expertise. If robust, this property could lower the barrier to deploying early-warning systems in regions lacking dedicated statisticians or complex pipelines. The key question we ask is: *can LLMs make epidemiologically meaningful flu forecasts—despite noise and irregular sampling—and thus assist real-world flu control?*

To address this question, we design a study aligned with how flu forecasts are actually used in public health decision-making. We use weekly percentages of outpatient visits for influenza-like illness (ILI) from the CDC FluView/Viral Surveillance system, covering all U.S. states. To ensure fairness and avoid data leakage, we fix **September 2024–September 2025** as the held-out test year—well beyond the likely training cutoff for current LLMs and representing a complete flu season.

To represent weaker surveillance systems, we also test robustness under *missingness*. Using the same train–test setup, we randomly remove 0–60% of training data in 20% increments. Classical baselines—**seasonal ARIMA** and **Gaussian Process** models—handle missing data directly, while **LLM** inputs include explicit markers for missing weeks.

Because public health agencies seldom forecast from the very start of a season, we simulate more practical decision points. Instead of predicting from the lowest flu activity, we start forecasting once case levels begin to rise. For each state's 2024–2025 series, we locate the pre-season trough and the subsequent peak, then define several thresholds between them—representing early, middle, and late stages of the seasonal increase. When the ILI level crosses a threshold, we treat that week as the forecast starting point and predict the remaining trajectory. This approach mirrors how officials might act when flu activity shows noticeable growth and early interventions are being considered.

We evaluate forecasting performance using Mean Absolute Error (MAE), which measures the average absolute difference between predicted and true values. MAE provides a clear and interpretable measure of general accuracy, allowing us to compare how well each method captures overall flu activity levels.

In summary, this setup tests whether LLMs can provide *practical and actionable* flu forecasts under realistic conditions of irregular, incomplete, or partially observed data, and whether their zero-shot capability can complement or substitute traditional time-series models in public health applications.

2 Literature Review

The use of LLMs for time-series forecasting is a rapidly evolving research frontier. Gruver et al. [2] introduced *LLMTime*, a pioneering framework that reformulated forecasting as a text continuation task. In LLMTime, numeric time-series values were serialized into a plain-text format, and an off-the-shelf LLM (e.g., GPT-3) was prompted to continue the sequence. Despite the absence of fine-tuning, the model achieved competitive zero-shot forecasting accuracy compared to classical approaches such as ARIMA and exponential smoothing. This finding suggested that pretrained LLMs may implicitly encode sufficient inductive bias for temporal reasoning.

Building on this idea, Jin et al. [3] proposed *TIME-LLM*, which integrates structured embeddings and cross-attention adapters to bridge numerical data and language representations. By mapping time-series patches into token spaces, the model leverages frozen

*Both authors contributed equally to this research.

LLM backbones to capture both temporal and semantic dependencies. While this design improves quantitative accuracy, it introduces additional preprocessing and parameter overhead. The tradeoff between architectural simplicity and performance thus remains an open question.

Other studies have reported mixed results. Tang et al. [5] observed that LLMs excel when the data exhibit clear seasonal trends or smooth trajectories, but their accuracy declines sharply in the presence of noise or irregular sampling. Park et al. [4] further demonstrated that minor perturbations or missing segments can significantly degrade performance, often reducing LLMs below even linear autoregressive baselines. These results collectively emphasize the need for systematic evaluation under realistic data imperfections.

From the classical perspective, time-series forecasting has long relied on statistical and mechanistic models. Seasonal ARIMA (SARIMA) models capture autoregressive and moving-average dynamics given period cycles and remain among the most widely used baselines. Gaussian Processes (GPs) extend this framework by providing probabilistic forecasts with principled uncertainty quantification. Because GPs model correlations in continuous time, they naturally handle irregular sampling and missing data, offering a strong benchmark against which to compare LLMs.

Recent public-health forecasting challenges, such as the FluSight competitions, emphasize not only mean error metrics but also epidemiologically meaningful quantities—such as the predicted week of peak incidence and the peak magnitude of infection. These criteria directly inform intervention timing and vaccine distribution, making them highly relevant to this study. However, existing work has rarely explored whether LLM-based models can accurately reproduce such epidemic features. The present project aims to bridge this gap by introducing both standard and disease-aware evaluation metrics.

3 Data Processing

We use weekly influenza-like illness (ILI) percentages from the CDC Viral Surveillance System, obtained through the FluView Portal [1]. The dataset covers ten influenza seasons from September 2015 to September 2025, with one data point per epidemiological week for each U.S. state. Each record represents the proportion of outpatient visits attributed to ILI symptoms. The long temporal span, regular weekly resolution, and clear seasonal patterns make this dataset well suited for evaluating time-series forecasting models.

While the data are mostly regular, a few reporting gaps exist due to delayed or missing submissions. Across all ten years, we identify only eleven missing weeks—less than 0.1% of the total. These are linearly interpolated to form a continuous reference series, which serves as ground truth for all evaluations.

We define the training horizon as Septermber 2020–September 2024 and kept one whole year September 2024–September 2025 as the testing data, which covers a full flu season beyond the likely training data cutoff for existing LLMs.

To support consistent evaluation across U.S. states, we precompute two seasonal markers for each region and year: the week of minimum ILI (start of season) and the week of maximum ILI (peak). These features are used to design the threshold-based partial-observation experiment described in Section 4. All series are standardized by percentage units and aligned by epidemiological week for comparability.

4 Methodology

This section outlines the overall experimental design for evaluating how well large language models (LLMs) perform on flu forecasting tasks compared to classical time-series models. The experiments are designed to reflect both data quality challenges and real-world forecasting contexts. We organize the methodology into three main parts: (1) description of the two experimental tasks, (2) overview of the prediction models, and (3) explanation of evaluation metrics and their relevance to epidemic forecasting.

4.1 Experimental Design

We conduct two complementary experiments to evaluate forecasting performance under different practical and data-quality conditions.

Experiment 1: Impact of Irregular Time Series. The first experiment investigates how irregularity and missingness in surveillance data affect model performance. Although the CDC ILI data are mostly regular, we simulate varying levels of missingness to represent weaker surveillance systems or incomplete data reporting. For each U.S. state, we randomly remove between 0% and 60% of weekly observations in increments of 20%, then fit or prompt the models on the resulting irregular series. Each experiment is repeated once per region, and performance metrics are averaged across all states. This experiment quantifies how robust each forecasting approach is to missing or unevenly sampled data.

Experiment 2: Practical Early-Season Forecasting. The second experiment focuses on how models perform under realistic public health forecasting conditions. In practice, health agencies rarely issue forecasts at the very beginning of a flu season when case counts are at their lowest. Instead, forecasts are typically generated once flu activity begins to rise, allowing officials to anticipate the upcoming peak and plan interventions such as vaccination campaigns or hospital resource allocation. To simulate this decision context, we design a progressive early-season forecasting setup that controls how much of the ongoing flu season is visible to the model at the time of prediction.

We begin by identifying two reference points in each region's 2024–2025 ILI series: the **pre-season trough** (the week with the lowest ILI percentage before the seasonal rise) and the **seasonal peak** (the week with the highest ILI percentage). These two points define the amplitude of the flu season for that region. Between the trough and the peak, we construct a set of relative thresholds representing different stages of the seasonal increase:

$$\text{Threshold} = (1 - w) \times \text{Trough} + w \times \text{Peak}, \quad w \in \{0, \tfrac{1}{4}, \tfrac{1}{3}, \tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{4}\}.$$

Each threshold value corresponds to a specific level of observed flu intensity:

- $w = 0$ — the earliest stage, immediately after the trough (almost no rise observed);
- $w = \frac{1}{4}$ — early rise stage, where ILI has increased modestly;

- $w = \frac{1}{3}$ — mild rise, indicating early spread of flu activity;
- $w = \frac{1}{2}$ — mid-season growth, where the disease trend becomes clearly visible;
- $w = \frac{2}{3}$ — late growth stage, near the acceleration toward the peak;
- $w = \frac{3}{4}$ — near-peak stage, representing advanced flu activity levels.

For each threshold, we locate the first week in the 2024–2025 test year where the ILI percentage exceeds that threshold. All data up to that week are treated as *observed*, and the remaining weeks form the *forecasting horizon*. The model is trained or prompted using only the observed portion and tasked with predicting the remainder of the season. This process is repeated across all threshold levels, providing multiple observation–forecast splits that represent different practical stages of situational awareness during a flu season.

By comparing performance across these thresholds, we can evaluate how much early-season information is needed for reliable forecasting and whether LLMs, like traditional statistical models, can anticipate the upcoming peak as more of the seasonal curve becomes visible. This experiment therefore assesses the operational usefulness of LLM-based forecasts under progressive information availability.

4.2 Prediction Models

This study compares three forecasting approaches that differ in their structure and assumptions about temporal data:

Gaussian Process Regression (GP).. Gaussian Process Regression (GPR) provides a flexible, probabilistic framework for modeling smooth and seasonal time series such as influenza-like illness (ILI) activity. The GP models observations $y_t = f(t) + \epsilon_t$, where $f(t)$ is drawn from a Gaussian Process with mean zero and covariance $k(t, t')$, allowing both interpolation and uncertainty-aware forecasting.

We define time indices on a normalized scale ($t_{\text{train}} \in [0, 1]$) and standardize ILI values using the training mean and standard deviation for numerical stability. The covariance kernel combines interpretable components:

$$k(t, t') = C\big(k_{\text{trend}} + k_{\text{RQ}} + k_{\text{per}}\big) + k_{\text{white}},$$

where the `DotProduct` term models linear trends, the `RationalQuadratic` term captures smooth multi-scale variation, the `ExpSineSquared` term encodes annual periodicity, and the `WhiteKernel` accounts for observation noise.

Although the periodic kernel is initialized with a 52-week period (the expected flu cycle), all key hyperparameters—period, length scales, and noise level—are learned automatically by maximizing the marginal log-likelihood within reasonable bounds. This allows the model to adapt if the true seasonal period deviates slightly from one year. The GP is fit only on observed points, handles missing data naturally, and provides predictive means and uncertainties for both training and future weeks. Overall, it captures smooth trends, seasonality, and variability while learning all kernel parameters directly from the data.

Seasonal ARIMA.. We use seasonal ARIMA with annual seasonality $s = 52$ for weekly ILI. Candidate nonseasonal orders are screened on a compact grid ($p, q \le 3$, $d \le 2$) by AIC using only the training window (Sep 2020–Sep 2024). For stability and comparability across states, we then fix ARIMA$(2, 1, 2)$ as the default specification unless the AIC-selected model is strictly better by a nontrivial margin; in either case the seasonal structure is handled in the state–space recursion. For the missingness experiment (0–80% random removal on the training portion), ARIMA is refit on series with linearly interpolated gaps to avoid numerical issues from long missing runs. Under the early-season protocol, ARIMA is refit after each threshold reveal (from trough to peak, $w \in \{0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$) and then used to forecast the remainder of the season, reporting point paths with 95% confidence intervals. This mirrors the operational workflow where simple integrated models are re-estimated as new weeks arrive.

LLM Forecaster. The LLM we used through the study is Gemini-2.5-Flash, due to the balance of response time and cost. We design a hierarchical prompting scheme with three levels of contextual information. At **Level 1**, the model receives only a task description: it is instructed to forecast the future of a univariate time series, to treat " , , " (two consecutive commas) as an explicit indicator of a missing value, and to output a comma-separated list of exactly H predicted numbers with no accompanying text. **Level 2** augments this with domain context, noting that the input corresponds to weekly percentages of outpatient visits attributed to Influenza-Like Illness (ILI) in U.S. surveillance data, and that the historical sequence extends to approximately September 2024. **Level 3** further adds modeling guidance, asking the model to implicitly account for long-term trends, seasonal influenza patterns (including winter peaks and off-season troughs), and short-term variability.

Importantly, the *only* preprocessing applied to the input time series is a simple textual serialization step. Each numeric value is rounded to a fixed decimal precision (typically two digits), converted to a string, and appended to a comma-separated list. If a value is missing, we insert an empty token—represented as two consecutive commas (" , , "), with no placeholder, imputation, or additional symbol. No detrending, scaling, smoothing, or normalization is performed for the LLM input; the model receives exactly the raw numeric trajectory in text form, with missing weeks encoded solely through the comma structure. The prompt then ends by showing this serialized sequence and instructing the LLM to output H future values, also in comma-separated format.

4.3 Evaluation Metrics

We evaluate model performance using a single general accuracy measure that reflects overall predictive reliability.

Accuracy Metric. We report **Mean Absolute Error (MAE)** as the primary indicator of forecast accuracy. MAE measures the average absolute deviation between predicted and observed values, providing a clear and interpretable assessment of how closely each model captures the true underlying flu activity.

This metric offers a consistent basis for comparing methods and is aligned with the goal of evaluating general predictive performance.

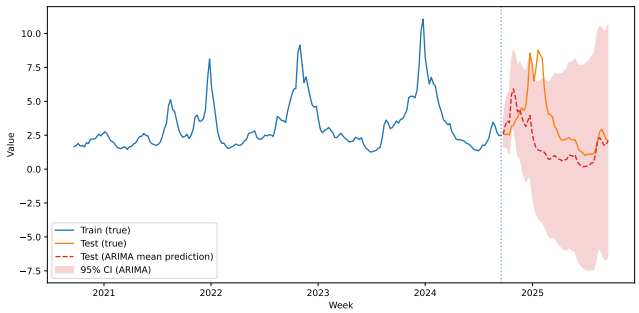

Figure 1: Example ARIMA forecast for Georgia: training trajectory, true 2024–2025 flu season, and one-year-ahead forecast with 95% confidence interval.

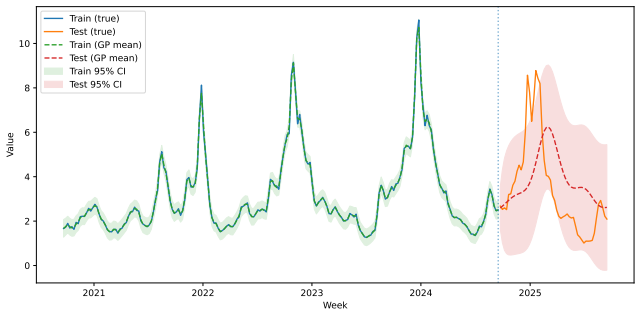

Figure 2: Example Gaussian Process forecast for Georgia: training data, true test season, and posterior mean with 95% credible intervals.

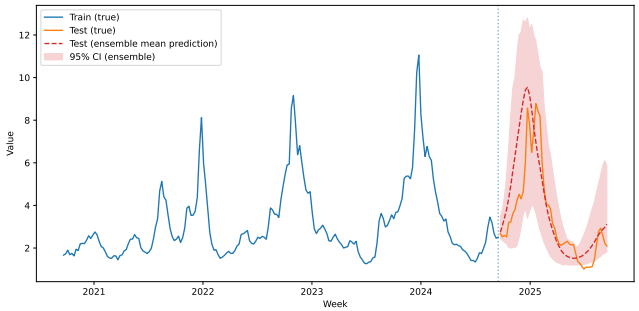

Figure 3: Example Gemini ensemble forecast for Georgia: training data, true 2024–2025 season, ensemble mean prediction, and empirical 95% interval computed from multiple prompt runs.

5 Results

We now present empirical results from the two experiments described in Section 4.1. All performance is evaluated using Mean Absolute Error (MAE) over the corresponding test horizon (52 week for experiment 1 and varying length for experiment 2).

5.1 Case Study: Georgia Forecast Trajectories

We begin with a qualitative comparison of forecasts for the state of Georgia (GA) with no missing, which illustrates typical behavior of the three model classes. Figures 1–3 show one-year-ahead forecasts for the 2024–2025 season produced by seasonal ARIMA, Gaussian Process (GP), and the Gemini LLM ensemble, along with the preceding four years of training data.

The ARIMA model (Figure 1) captures the overall rise and fall of the flu season but sometimes misaligns the timing and magnitude of the peak, especially when the 2024–2025 season deviates from the historical pattern. The 95% interval widens substantially toward the end of the horizon, reflecting accumulated uncertainty.

The GP model (Figure 2) yields a smoother mean trajectory with credible intervals that adapt to the density of training points. The periodic kernel helps align the forecast with the expected annual cycle, often tracking both the onset and decline of the season more closely than ARIMA. Residual errors primarily occur when the true peak is unusually early, late, or multi-modal relative to previous years.

The Gemini ensemble (Figure 3) reproduces the broad seasonal shape but exhibits larger discrepancies in both peak timing and peak height. The empirical 95% band, derived from multiple prompt completions, is wide, indicating substantial run-to-run variability. This qualitative pattern is consistent with the higher and more dispersed MAE values observed for Gemini in the aggregate analyses below.

5.2 Experiment 1: Impact of Irregular Time Series

We evaluate how each forecasting method responds to increasing levels of missingness in the training data. Figure 4 summarizes state-level MAE for four missing-rate conditions (0, 0.2, 0.4, 0.6), comparing ARIMA, GP, and the three Gemini prompting levels (L1–L3).

Across all panels, all five methods remain surprisingly robust: even with 60% of training weeks removed, performance degrades only marginally, and all models preserve reasonable predictive accuracy. ARIMA and GP remain broadly comparable in median MAE, with both maintaining values typically between 1.0 and 1.3 across missingness levels. Their error distributions, however, differ in characteristic ways. ARIMA occasionally produces large outliers (visible as isolated high-MAE points), while GP exhibits a noticeably wider interquartile range, indicating more variation between the 25th and 75th percentiles.

The three Gemini prompting configurations show a clear progression. L1 has the highest error and largest variance, but performance steadily improves from L1 to L2 and from L2 to L3. Under the highest prompt level, L3 becomes surprisingly competitive with GP—its median MAE under moderate missingness is very similar, though with a wider overall spread. While Gemini does not surpass GP or ARIMA, it achieves this performance in a completely zero-shot manner, requiring no statistical model fitting or hyperparameter tuning. Given this minimal setup cost, the L3 prompt remains a practical and useful option for rapid forecasting when classical modeling resources are limited or unavailable.

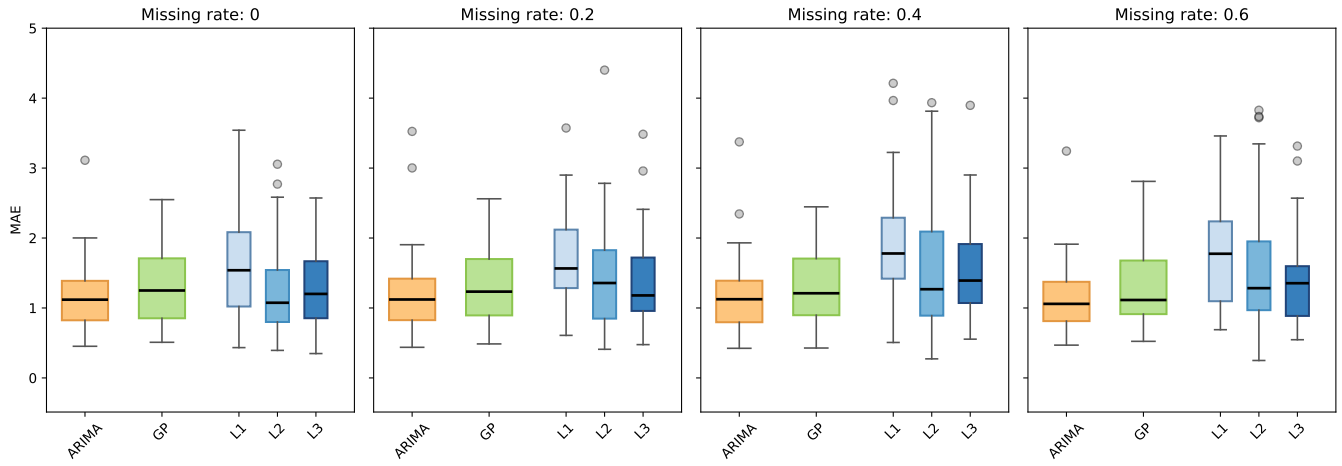

Figure 4: State-level MAE under different levels of randomly induced missingness in the training data (0, 0.2, 0.4, 0.6). Each box summarizes the distribution of MAE across states for one method: ARIMA, GP, and three Gemini prompting levels (L1–L3).

5.3 Experiment 2: Practical Early-Season Forecasting

The second experiment assesses how well models can forecast the remainder of a flu season when only an early portion of the 2024–2025 curve is visible. For each state, we construct six observation levels between the pre-season trough and the seasonal peak ($w \in \{0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$). At each level the models observe data only up to the first threshold crossing and must forecast the remaining trajectory.

Figure 5 reports the distribution of MAE across states for ARIMA and GP at each observation level. Several consistent patterns emerge from the panels. First, GP is noticeably more stable than ARIMA across all six thresholds: its interquartile range remains compact, and its median MAE changes only slightly as more of the season is revealed. Second, additional information does not necessarily improve ARIMA's performance. In fact, in this particular season the true peak occurs later than in most historical years, so later-threshold forecasts can introduce mismatches between the partially observed curve and the patterns learned from previous seasons. This leads to higher variance and, in some cases, worse accuracy for ARIMA at deeper observation levels.

GP, in contrast, remains relatively robust. Because it models temporal correlation directly through the kernel structure, partial-season observations do not substantially distort its predictions; even when the peak timing is atypical, the GP continues to generate forecasts aligned with the emerging trend. Notably, however, neither ARIMA nor GP shows systematic improvement as more of the season becomes visible. This reflects a structural limitation of both statistical models: their reliance on historical patterns means that additional partial observations help only when the ongoing season resembles the past, which is not guaranteed when the peak timing or amplitude deviates.

The result for LLM is not available so far hence no boxplot shown in the figure.

6 Conclusion

This study provides a systematic evaluation of whether large language models can serve as practical forecasters for epidemiological time series, particularly under real-world challenges such as irregular sampling, missing observations, and early-season uncertainty. Using ten years of CDC ILI data and a held-out 2024–2025 flu season, we benchmarked a zero-shot LLM forecaster against two classical statistical models—seasonal ARIMA and Gaussian Process regression— across two experimental settings designed to mimic operational flu forecasting.

Across both experiments, several consistent insights emerged. First, all methods demonstrated a surprising degree of robustness to substantial irregularity: even with 60% missingness, median state-level MAE degraded only slightly, and both ARIMA and GP maintained stable performance across states. In this setting, LLM performance improved significantly with richer prompting, and the most informative prompt (L3) achieved median accuracy comparable to GP, though with a wider spread. This highlights an important practical point: despite its simplicity and lack of parameter fitting, a carefully designed zero-shot prompt can yield forecasts competitive with established statistical models when data are sparse or incomplete.

Second, in the early-season forecasting experiment, both ARIMA and GP generated usable forecasts even when only a small fraction of the current season was visible. GP consistently produced narrower interquartile ranges and more stable error across observation thresholds, while ARIMA's accuracy sometimes deteriorated as additional partial-season data were revealed. This behavior reflects each model's underlying assumptions: GP benefits from its kernel structure, which smooths over irregularities in early-season curvature, whereas ARIMA can be more sensitive to atypical timing or amplitude relative to historical patterns. Neither model, however, showed systematic improvement as more data became available, underscoring a structural limitation of classical approaches when the ongoing season departs from learned seasonal templates.

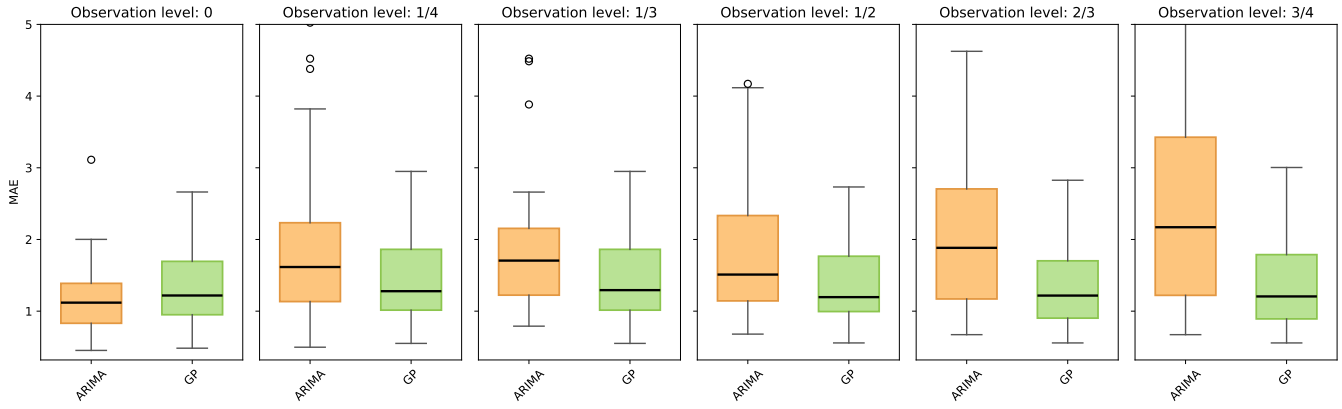

Figure 5: State-level MAE for the early-season forecasting experiment. Each panel corresponds to a different observation level ($w \in \{0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$) and compares ARIMA and GP. Boxes summarize the distribution of MAE across states for each model.

Taken together, these findings suggest that LLMs offer a promising complement—not a so far—to traditional time-series models in public health surveillance. Classical statistical models remain strong baselines, especially when historical patterns are reliable and computational resources are readily available. Yet zero-shot LLMs require no model training, no hyperparameter tuning, and only minimal preprocessing, making them attractive in settings where surveillance systems are incomplete, expertise is limited, or rapid deployment is critical. As LLMs continue to evolve, their ability to interface directly with raw and irregular data may offer new opportunities for accessible, human-in-the-loop forecasting pipelines.

7 Limitations

Although the results presented here offer useful insights into the potential of LLMs for epidemiological forecasting, several limitations remain. First, due to daily request limits and budget constraints, we relied exclusively on the free `gemini-2.5-flash` model, and were unable to include more capable but paid models such as Gemini 2.5 Pro or GPT-5-series models. Preliminary trials suggest that more expensive models do not necessarily yield better forecasting performance, but the landscape of LLM architectures is diverse and rapidly evolving. As a result, the true state-of-the-art performance of LLM forecasters on irregular epidemiological time series remains unexplored.

Second, computational limitations prevented us from completing Experiment 2 (early-season forecasting) for the LLM forecaster. While ARIMA and GP showed no systematic improvement when more partial-season information was revealed, it is possible that an LLM—unconstrained by fixed statistical assumptions—could leverage increasing context more effectively. At present, we cannot determine whether LLMs benefit from additional partial-season observations or whether their performance follows patterns similar to classical models.

Third, our study employed extremely minimal preprocessing for the LLM input. We performed only a simple serialization of numeric values with fixed decimal precision, using two consecutive commas to indicate missing observations. Prior work has emphasized the importance of sophisticated tokenization strategies and feature preprocessing for LLM forecasting. However, with more advanced models, the benefits of such preprocessing may diminish. Our results show that even under this minimal setup, LLMs can produce meaningful forecasts, but a more systematic comparison of preprocessing choices is needed to determine how much they influence performance in modern architectures.

These limitations reflect both practical constraints and open scientific questions, and they motivate several directions for future research.

8 Future Work

Several extensions of this work are natural. A primary direction is to conduct a comprehensive evaluation across a wider range of LLM architectures, including both proprietary models (e.g., GPT-5, Gemini 2.5 Pro) and emerging open-source alternatives. Such an evaluation would clarify whether the performance observed here generalizes across models or is highly model-dependent.

Completing Experiment 2 for LLMs also remains an important next step. The question of whether LLMs can effectively utilize increasing amounts of partial-season information—and potentially outperform statistical models in this regime—has direct implications for early-warning surveillance systems.

Future studies should also examine the role of data preprocessing and tokenization. While our minimal serialization approach demonstrated that advanced preprocessing is not strictly required for obtaining useful forecasts, it remains unknown whether thoughtfully engineered representations could narrow the performance gap between LLMs and classical methods or improve robustness in low-data regimes.

Finally, extending this analysis to other epidemiological indicators, multiple forecast horizons, or probabilistic LLM outputs would provide a more complete picture of the strengths and limitations of zero-shot LLM forecasting in public health contexts. As LLM capabilities continue to evolve, understanding how to integrate them effectively—and safely—into operational forecasting pipelines is a promising and impactful direction for future research.

Acknowledgments

We gratefully acknowledge Harsha Kamarthi for inspiring the original idea of this study. We also thank our teaching assistant, Shangqing Xu, for his guidance and support throughout the course, and Professor Prakash for his supervision and excellent teaching.

References

[1] Centers for Disease Control and Prevention. 2025. FluView: Weekly U.S. Influenza Surveillance Report. https://www.cdc.gov/flu/weekly/fluviewinteractive.htm. Accessed: 2025-10-20.

[2] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems* 36 (2023), 19622–19635.

[3] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[4] Junwoo Park, Hyuck Lee, Dohyun Lee, Daehoon Gwak, and Jaegul Choo. 2025. Revisiting LLMs as Zero-Shot Time-Series Forecasters: Small Noise Can Break Large Models. *arXiv preprint arXiv:2506.00457* (2025).

[5] Hua Tang, Chong Zhang, Mingyu Jin, Qinkai Yu, Zhenting Wang, Xiaobo Jin, Yongfeng Zhang, and Mengnan Du. 2025. Time series forecasting with llms: Understanding and enhancing model capabilities. *ACM SIGKDD Explorations Newsletter* 26, 2 (2025), 109–118.