

Context-Aware Hospitalization Forecasting Evaluations for Decision Support using LLMs

Rhea Makkuni
Johns Hopkins University
Baltimore, Maryland, USA
rmakkun1@jhu.edu

Ananya Joshi
Johns Hopkins University
Baltimore, Maryland, USA
aajoshi@jhu.edu

Abstract

Medical and public health experts must make real-time resource decisions, such as expanding hospital bed capacity, based on projected hospitalization trends during large-scale healthcare disruptions (e.g., operational failures or pandemics). Forecasting models can assist in this task by analyzing large volumes of resource-related data at the facility level, but they must be reliable for decision-making under real-world data conditions. Recent work shows that large language models (LLMs) can incorporate richer forms of context into numerical forecasting. Whereas traditional models rely primarily on temporal context (i.e., past observations), LLMs can also leverage non-temporal public health context such as demographic, geographic, and population-level features. However, it remains unclear how these capabilities should be used to produce stable or decision-relevant predictions in real-world healthcare settings. To address this gap, we introduce HYBRIDARX, a context-augmented hybrid framework that integrates LLM-derived contextual signals into structured forecasting models, and evaluate it against direct LLM-based forecasting and classical time-series methods across 60 U.S. counties with low-, mid-, and high-intensity hospitalization patterns. Because the goal is operational decision-making rather than error minimization alone, we evaluate performance with bias and lead-lag alignment in addition to standard forecasting metrics. Our results show that HYBRIDARX improves over classical ARX by yielding more stable and better-calibrated forecasts, particularly when incorporating noisy contextual signals into structured time-series models. These findings suggest that, in non-stationary healthcare resource forecasting, LLMs are most useful when embedded within structured hybrid models.

CCS Concepts

• **Computing methodologies** → **Machine learning**: *Time series analysis*; • **Applied computing** → *Health informatics*.

Keywords

hospitalization forecasting, time series forecasting, large language models, HybridARX, decision support, healthcare operations

ACM Reference Format:

Rhea Makkuni and Ananya Joshi. 2026. Context-Aware Hospitalization Forecasting Evaluations for Decision Support using LLMs. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Accurate hospitalization forecasting is vital to healthcare systems because it enables timely resource allocation, staffing, and capacity planning to maintain quality of care during periods of fluctuating demand [15, 19]. As demonstrated by the COVID-19 surges from 2020–2023, unexpected demand can overwhelm the ability of healthcare systems to allocate resources in real time, leading to strained intensive care units [19] and widespread disruption of essential services such as chemotherapy and rehabilitation [3]. As a result, healthcare systems, including hospitals and public health agencies, require reliable decision support to both (1) guide resource allocation during large-scale healthcare disruptions and (2) respond effectively to emerging system strain [11, 13].

This means that models that perform well in aggregate may still fail in practice if they do not capture regional variation, thereby compounding disparities between resource-constrained and well-resourced hospitals. Accordingly, many general forecasting models that are used as the backbone for resource allocation algorithms do not reliably support decision-making. Often, Lag-1 models, or simply reusing the previous measurement as the forecast, have the highest accuracy values. However, these are not necessarily meaningful for decision-making. Other models, such as autoregressive (AR) models, rely heavily on recent observations and are therefore unable to anticipate regime shifts or rapid changes in disease dynamics under non-stationary conditions [6]. This effect is pronounced across different population levels, where the operational cost of forecasting error can vary substantially, which motivates the need for methods that can incorporate additional non-temporal contextual information.

Large Language Models (LLMs), with their advanced contextual reasoning capabilities using operational and public health knowledge, could potentially provide a mechanism for incorporating heterogeneous contextual information into forecasting tasks under non-stationary conditions. However, it remains unclear how these contextual capabilities should be integrated into forecasting systems to produce stable decision-making forecasts in real world healthcare settings.

Rather than treating LLMs as standalone forecasters, we introduce a two-stage hybrid framework, HybridARX, that uses LLMs as contextual reasoning agents within a structured forecasting pipeline. In this approach, an LLM first predicts the next-week

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '26, Jeju Island, Republic of Korea

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

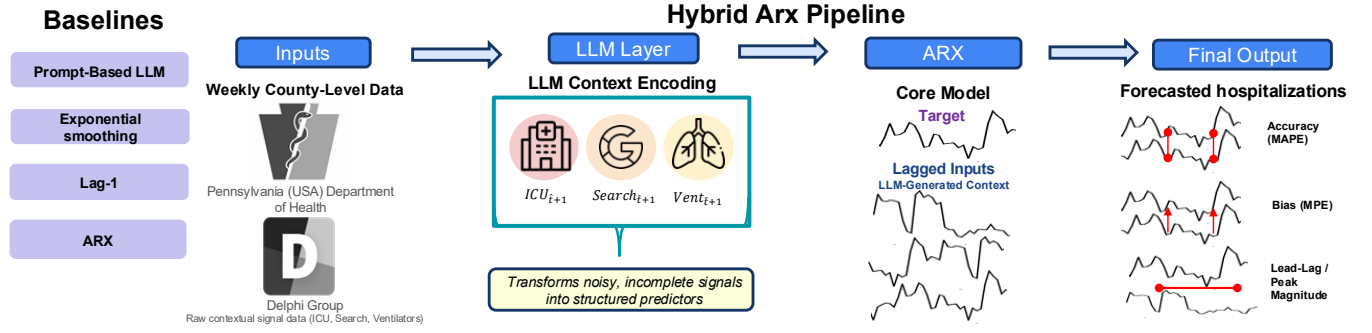


Figure 1: HybridARX Pipeline. Data is pulled from multiple data sources with varying granularity and is processed into standardized input. Then, it is evaluated on metrics relevant to decision-making tasks.

values of the leading indicators (i.e., observable signals that tend to precede hospitalizations), which are then supplied as exogenous inputs to classical time-series models for the final hospitalization forecast.

We evaluate HybridARX on county-level hospitalization data from the Pennsylvania Department of Health against a prompt-only LLM baseline and classical time-series baselines, including univariate and autoregressive models (Lag-1, AR(1), exponential smoothing), as well as multivariate autoregressive models with exogenous inputs (ARX).

Our method highlights three novel insights under the context of non-stationary data:

- 1) LLMs are most useful not as standalone forecasters but as contextual reasoning agents within hybrid pipelines that transform heterogeneous contextual inputs into structured predictors for classical time-series models.

- 2) Incorporating non-temporal contextual information alone does not guarantee improved forecasts; instead, performance depends on how contextual signals are integrated into forecasting pipelines.

- 3) LLM-generated outputs as inputs to classical models improve forecast stability and interpretability.

We perform a county-level analysis across 60 counties within a single state, using a shared reporting and policy environment that enables controlled comparison across varying low-, mid- and high-intensity regions (defined using tertiles of mean weekly hospitalizations from Pennsylvania Department of Health data).

Our results indicate that HybridARX reduces large forecast errors relative to classical ARX in several settings and produces more stable forecasts when noisy contextual signals are incorporated into structured time-series models, though it does not uniformly outperform simpler methods. HybridARX demonstrates that LLMs are valuable as contextual signal encoders that convert heterogeneous inputs into structured predictors within forecasting pipelines.

2 Related Work

Real-time hospitalization forecasting is essential for decision-making, enabling hospitals to maintain high-quality care, especially when resources such as beds are limited [15]. In these settings, forecasting utility depends not only on aggregate accuracy but also on calibration and reliability, which directly affect planning.

Prior work has applied machine learning methods such as random forests, gradient-boosted trees (XGBoost), support vector machines (SVM), and elastic nets to forecast hospital admissions [1]. Time-series models such as exponential smoothing state space models (ETS), TBATS, and additive models with nonlinear trends and seasonal effects (Prophet) have been benchmarked against these approaches in disrupted settings, such as psychiatric admissions during the pandemic [23]. More broadly, large-scale epidemiological forecasting efforts often include hospital resource targets such as bed occupancy [12]. CDC forecasting efforts, such as FluSight [5] and the COVID-19 Forecast Hub [16], include multiple models, which are generally trend-extension and mechanistic hybrid approaches. Variants of these methods, such as ARIMAX or mechanistic models (e.g., SEIR-based models), also exhibit high accuracy [20]. Additionally, prior work has incorporated non-temporal context through coarse, bucketed representations, such as stratifying populations by age, geography, or risk group, or including aggregated demographic covariates in regression-based models. Newer foundation time-series models, such as MOMENT, extend this idea by training transformers on large collections of time-series data [9], including epidemiological data, but require a large number of historical samples (e.g. 512) that may not be reliable or relevant in a nonstationary setting, where there may be between 8-10 reliable previous measurements.

While promising, many of these approaches were evaluated solely from an accuracy perspective instead of a decision-making lens, masking challenges with real world data forecasting. First, as a result of the structure of the models, they still rely on stable historical relationships and require frequent refitting. Second, many leading indicators can be inconsistent because their conversion rates (the probability that a leading indicator results in hospitalization) are influenced by contextual factors, making resource allocation during surges particularly challenging [17]. In addition, any bucket-based approaches that treat additional context as fixed, discretized inputs may disproportionately impact regions with less representative buckets (e.g. lower hospitalization intensity). Many such methods also rely on expert-selected clinical characteristics, hand-crafted preprocessing, and tuned hyperparameters, and they rarely operate directly on unstructured sources or without significant maintenance [2]. This requirement may be incompatible with the budgets and real-time constraints faced by many public health

departments. As a result, resource allocation decisions frequently rely on ad hoc adjustments or expert judgment layered on top of trend-based forecasts, reducing reproducibility and scalability across regions and time [6]

LLMs offer a different form of flexibility. LLMs have been explored for forecasting the short-term spread of disease outbreaks by reformulating real-time forecasting as a text reasoning problem [7], as well as for monitoring time-series models to flag implausible or inaccurate predictions in large-scale retail settings [4]. Despite these advances, the role of LLMs for operational hospital forecasting and, more generally, as primary forecast generators remains unclear and unexplored. This work seeks to address this gap by introducing HYBRIDARX, a context-augmented hybrid framework that leverages LLMs as contextual signal encoders within classical forecasting pipelines.

3 Methods

For each county, we consider a weekly time-series of COVID-19 hospitalizations. Let y_t denote the observed hospitalization level at week t , and let z_t denote a vector of auxiliary contextual indicators. The forecasting task is to produce a one-week-ahead prediction:

$$\hat{y}_{t+1} = f(y_{t-L:t}, z_{t-L:t})$$

where f is a forecasting model that could be either statistical, LLM-based, or the proposed HybridARX framework. All models are evaluated using a rolling-origin one-step-ahead forecasting protocol to prevent data leakage.

We consider three forecasting model classes: (1) the HYBRIDARX approach, which integrates LLM-generated contextual signals with classical time-series models to improve stability, calibration, and temporal alignment in forecasting (2) classical statistical baselines, including the lag-1 model, autoregressive (AR) models, and exponential smoothing, which serve as interpretable reference methods without explicit contextual inputs, and (3) a prompt-only LLM approach that generates predictions using recent observations expressed as structured text.

For each county and forecast time $t+1$, the model is provided with the most recent $L = 8$ weekly observations of the hospitalization series.

3.1 Data Selection and Extraction

We constructed a weekly panel using the following public data sources:

Hospitalization and hospital capacity data: Obtained from the Pennsylvania Department of Health [14].¹ The dataset includes weekly reported values derived from daily data, including 14-day averaged daily counts of COVID-19 hospitalizations, as well as adult and pediatric ICU and ward bed availability and occupancy. These reported counts may be affected by testing practices, reporting delays, backlogs, and routine data cleaning. County assignment was based on patient or provider address and may be imperfect for ZIP codes that cross county boundaries.

¹Seven counties (Forest, Juniata, Perry, Pike, Snyder, Cameron, and Sullivan) were excluded from per-county evaluation due to data missingness and reporting irregularities across the study period.

Leading indicators: We used the Delphi COVIDcast API [18] to obtain leading indicators, including Facebook symptom survey measures (mask wearing, vaccine acceptance, and COVID-like symptoms in individuals and in the community), Google symptom search intensity for anosmia and ageusia, and COVID-related outpatient and emergency department visit rates, which are reported at mixed geographic resolutions. We aggregated daily values to weekly sums. When signals were unavailable at the county level for a given week, the corresponding Pennsylvania state-level signal was used.

We also used mobility indicators derived from Google COVID-19 Community Mobility Reports [8], based on percentage changes in visits to retail and recreation and residential locations. These categories were selected because they reflect high-contact public activity and voluntary risk avoidance.

Population estimates: Population estimates from the 2020 U.S. Census were used to stratify counties [21]. Counties were ranked by their mean weekly hospitalization counts over March 2020–January 2022 and categorized into lower, middle, and upper tertiles. The final panel includes these features (x) and the target hospitalization outcome (y), defined as the 14-day average number of hospitalized COVID-19 patients.

3.2 Leading Indicators and Correlation Analysis

To identify the most informative leading indicators, we analyzed associations between hospitalization outcomes and a broad set of candidate signals from the panel data. Consistent with the forecasting setup, all indicators were aligned to a weekly resolution, with hospitalization outcomes defined as the 14-day average number of hospitalized COVID-19 patients reported each week.

We calculated Pearson correlation coefficients between each indicator and the hospitalization outcome using only weeks in which both variables were observed. Indicators were ranked by the magnitude of their correlations to assess their potential relevance as leading signals. The resulting coefficients, shown in Fig. 2, highlight strong associations between hospitalizations and measures of hospital capacity (e.g., ventilator utilization and ICU beds), as well as anosmia/ageusia search activity, and are used to guide our evaluation of how contextual signals impact forecasting performance.

3.3 Context-Augmented Hybrid Model [Multivariate] (HYBRIDARX)

To address the challenge of incorporating contextual signals into forecasting under non-stationary conditions, we propose HybridARX, a context-augmented hybrid framework that separates context generation from final prediction. In the first stage, a LLM is used to predict next-week contextual variables:

- $\hat{X}_{B,t+1}$: adult ICU beds,
- $\hat{X}_{V,t+1}$: ventilator utilization,
- $\hat{s}_{t+1}$: anosmia/ageusia search volume

In the second stage, the LLM-predicted contextual variables

$$(\hat{X}_{B,t+1}, \hat{X}_{V,t+1}, \hat{s}_{t+1})$$

are supplied to classical forecasting models (ARX and linear regression) to produce the final hospitalization forecast $\hat{y}_{t+1}$. We consider

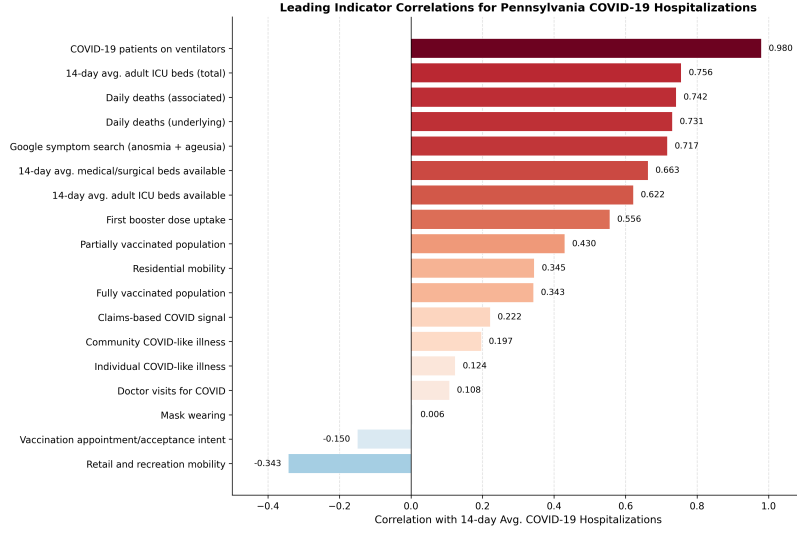


Figure 2: Pearson correlation coefficients between candidate leading indicators and county-level 14-day average COVID-19 hospitalizations. Hospital capacity measures and anosmia/ageusia search volume exhibit the strongest positive associations.

ARX to test the impact of temporal dependence assumptions. Additional implementation details, including prompt templates, pre-processing steps, and hyperparameter settings, are provided in the supplement.

3.4 Baseline Models

3.4.1 Prompt-Only LLM [Univariate]. Serving as a baseline comparison, in the prompt-only LLM setting, the model is used as a univariate forecaster. Each prompt consists of (1) the county identifier, (2) a chronologically ordered list of the previous eight weeks of numeric hospitalization values, where each value corresponds to the reported 14-day average number of hospitalized COVID-19 patients for that week, and (3) an instruction to return a single numerical prediction for the subsequent week. The model output is then parsed to a numeric value corresponding to y .

3.4.2 Classical Time-Series Baselines [Univariate and Multivariate]. As a baseline comparison, we evaluate several classical statistical baselines, each fitted independently for each county using a rolling window of $L = 8$ weekly observations.

- **Lag-1 baseline:** A naïve persistence model predicting the next week’s hospitalization level as the most recently observed value.
- **Autoregressive model (AR(1)):** A first-order autoregressive model with an intercept, trained on the previous eight weeks of hospitalization data.
- **Exponential smoothing:** An additive exponential smoothing model with trend is used. Seasonal components are omitted because our evaluation provides insufficient historical context to estimate seasonality.
- **Autoregressive model with exogenous inputs (ARX):** An AR(1) model augmented with weekly exogenous variables, including adult ICU beds (X_B), ventilator utilization (X_V), and anosmia/ageusia search volume (s_t).

3.5 Evaluation Framework

Operational decision-making depends not only on accuracy but also on timing, stability, and systematic bias. We evaluate HybridARX’s performance using accuracy, bias, and lead-lag behavior (stratified by hospitalization intensity) to capture error magnitude, direction, temporal alignment, and run-to-run variability in LLM-based forecasts.

All methods are evaluated under a consistent rolling-origin, one-step-ahead forecasting protocol at a weekly resolution, using the most recent $L = 8$ observations at each forecast time. To quantify variability in LLM-based methods, the full rolling-window forecasting procedure is repeated $n = 3$ times to quantify variability in LLM-based forecasts. For each county-week-model combination, we record the predicted hospitalization level and the corresponding percent error for every run, using the following performance metrics:

1. MAPE Analysis. Forecasting accuracy is assessed using mean absolute percent error (MAPE), defined as

$$\text{MAPE} = 100 \times \frac{|\hat{y}_{t+1} - y_{t+1}|}{\max(y_{t+1}, 1)},$$

where y_{t+1} denotes the observed 14-day average hospitalization level in the subsequent week, and $\hat{y}_{t+1}$ denotes the corresponding forecast. MAPE is computed at each forecast point and then averaged across counties and evaluation weeks.

2. MPE Analysis (Bias). To quantify systematic over- or under-prediction, we compute mean percent error (MPE):

$$\text{MPE}(t) = 100 \times \frac{\hat{y}_{t+1} - y_{t+1}}{\max(y_{t+1}, 1)}.$$

Positive values indicate systematic overestimation, while negative values indicate systematic underestimation.

Model	MAPE: Hospitalization Intensity		
	Low	Mid	High
<i>Baselines</i>			
Lag-1	23.80 ± 3.97	22.66 ± 3.49	17.85 ± 2.68
AR(1)	32.44 ± 7.66	26.93 ± 5.83	20.10 ± 3.87
Exp. Smoothing	26.47 ± 4.94	22.54 ± 4.44	15.49 ± 4.17
ARX	35.74 ± 7.75	31.17 ± 7.13	26.06 ± 14.91
LLM (Prompt-Only)	25.44 ± 4.15	22.59 ± 4.18	19.59 ± 3.30
<i>Proposed Method</i>			
HYBRIDARX	31.16 ± 7.20	24.94 ± 5.52	18.84 ± 3.87

Table 1: MAPE (%) by forecasting model across hospitalization-intensity tertiles (mean ± SD across counties). Bold indicates the best mean value in each column. Italics indicate values with overlapping intervals relative to the best mean.

3. Lead-Lag Analysis. In addition to pointwise error, we evaluate temporal alignment between predicted and realized hospitalization trajectories. For each county, we compute differences $\Delta y_t = y_t - y_{t-1}$ and $\Delta \hat{y}_t = \hat{y}_t - \hat{y}_{t-1}$, and define a lagged trend correlation for integer offsets $\ell \in [-\ell_{\max}, \ell_{\max}]$: $\rho(\ell) = \text{corr}(\Delta \hat{y}_{t-\ell}, \Delta y_t)$, computed over all overlapping evaluation weeks. We then define the lead-lag estimate as

$$\ell^* = \arg \max_{\ell \in [-\ell_{\max}, \ell_{\max}]} \rho(\ell).$$

Positive values of ℓ^* indicate that the forecasted trend leads the realized trend (i.e., anticipates changes), while negative values indicate lagging behavior. We report the mean and standard deviation of ℓ^* across counties. We additionally report the peak correlation value $\rho^* = \max_{\ell \in [-\ell_{\max}, \ell_{\max}]} \rho(\ell)$, which quantifies the strength of trend alignment regardless of whether the model leads or lags. For this specific metric, we omit the Lag-1 baseline from lead-lag evaluation because, under $\hat{y}_t = y_{t-1}$, we have $\Delta \hat{y}_t = \Delta y_{t-1}$. This makes perfect alignment at $\ell = -1$ (and thus $\rho^* = 1$) mathematically guaranteed.

4 Results

4.1 Accuracy: MAPE

Across all models, MAPE (Tab 1) is highest in low-intensity counties and lowest in high-intensity counties, supporting that low-volume counties are harder to forecast because small absolute fluctuations produce disproportionately large relative errors. The HYBRIDARX model improves substantially on ARX across the board, and is competitive in both mid-intensity and high-intensity counties. This suggests that forecast performance critically depends on how contextual signals are integrated. Among the classical baselines, exponential smoothing and the Lag-1 model perform competitively in high-intensity counties. In contrast, AR models show substantially higher variance, particularly when exogenous inputs are included. Classical ARX performs poorly in high-intensity counties and mid-intensity counties, which suggests that incorporating noisy exogenous variables destabilized the regression and negatively impacted forecasting accuracy. The prompt-only LLM shows relatively consistent performance across county types and is competitive with

Model	MPE: Hospitalization Intensity			
	Overall	Low	Mid	High
<i>Baselines</i>				
Lag-1	+0.7 ± 1.5	+1.1 ± 1.4	+0.4 ± 1.9	+0.6 ± 1.2
AR(1)	+3.2 ± 6.5	+9.5 ± 6.5	+0.9 ± 4.1	-0.9 ± 2.6
Exp. Smoothing	-3.4 ± 3.5	-0.6 ± 3.7	-4.4 ± 2.7	-5.4 ± 1.9
ARX	+5.5 ± 10.8	+8.6 ± 8.1	+3.6 ± 5.9	+4.1 ± 15.7
LLM (Prompt-Only)	+0.6 ± 3.1	+2.5 ± 2.3	+1.0 ± 3.0	-1.7 ± 2.4
<i>Proposed Method</i>				
HYBRIDARX	+0.7 ± 6.1	+5.2 ± 7.7	-0.4 ± 4.3	-2.7 ± 2.1

Table 2: Mean percent error (MPE, %) by forecasting model across hospitalization-intensity strata (mean ± SD across counties). Bold indicates the value closest to zero in each column. Italics indicate values with overlapping intervals relative to the best-performing model in that column. Positive values indicate overprediction, while negative values indicate underprediction.

Model	Lead-Lag Analysis			
	ℓ^* Mean	ℓ^* SD	ρ^* Mean	ρ^* SD
<i>Baselines</i>				
AR(1)	-0.97	0.26	0.802	0.139
ARX	-0.84	0.76	0.563	0.133
Exp. Smoothing	-1.00	0.00	0.872	0.036
LLM (Direct)	-1.00	0.00	0.859	0.057
<i>Proposed Method</i>				
HYBRIDARX	-0.97	0.18	0.771	0.123

Table 3: Lead-lag summary across forecasting models (mean ± SD across counties). Bold indicates the highest trend alignment (ρ^*). Italics indicate values with overlapping intervals relative to the best-performing model.

exponential smoothing, which once again supports that, even without exogenous variables, LLMs can capture short-term temporal structure and smooth random fluctuations in hospitalization trends.

4.2 Directional Bias (MPE)

Mean Percent Error (MPE) captures systematic over- and under-prediction through positive and negative values, respectively (Tab 2). Once again, HYBRIDARX performs better than ARX, supporting the conclusion that LLM generated contextual signals may improve calibration and reduce uncertainty when used as part of hybrid pipelines in decision-making models. However, in most cases, lag-1 was the most calibrated model, with MPE very close to zero across all hospitalization intensities. Interestingly, the prompt-only LLM model also exhibited bias close to zero overall (+0.6 ± 3.1), with relatively small deviations across hospitalization intensities.

4.3 Lead-Lag Analysis

Because most methods yield mean lead-lag estimates close to $\ell^* = -1$, we primarily focus on ρ^* , the peak trend-correlation

metric, when assessing temporal alignment. HybridARX improves trend alignment relative to classical ARX (0.771 ± 0.123). Among the baselines, exponential smoothing and the prompt-only LLM exhibit the strongest trend alignment, while AR(1) and ARX achieve moderate to weaker alignment. These lead-lag patterns help reconcile the MAPE and MPE results and reinforce a consistent finding throughout the analysis: effective forecasting under non-stationary conditions depends not only on the contextual information itself, but on how contextual signals are integrated. Although most forecasts lag observed changes, there are rare cases in which exogenous information eliminates this delay. In rare cases, HybridARX achieves zero-lag alignment in Berks County ($\rho^* = 0.80$), and ARX achieves zero-lag alignment in Philadelphia County ($\rho^* = 0.65$). Both counties belong to the high-intensity tertile. While these cases are uncommon, they highlight the best-case boundary of short-horizon forecasting where structured exogenous signals are reducing temporal delay under favorable data conditions.

5 Discussion

This study examined whether large language models (LLMs) can improve hospitalization forecasting through the incorporation of contextual signals compared to classical time-series methods or through direct prompting. We identify a key mechanism by which LLMs improve forecasting within traditional pipelines. While LLMs may not reliably function as standalone “black-box” forecasters, they are effective as contextual signal extractors within structured forecasting pipelines. Classical ARX models performed poorly across several evaluation metrics; however, HYBRIDARX consistently outperformed classical ARX across all metrics.

One possible explanation is that LLM-generated contextual signals help smooth or regularize noisy external indicators before they are incorporated into the forecasting model. Because leading indicators anticipate shifts in hospitalization demand, improving the stability of these signals is particularly important. Many public health indicators are inconsistently reported and partially missing, which can reduce their usefulness when directly incorporated into regression-based models. Transforming noisy leading signals into more structured predictors may allow forecasting models to better leverage this information. These results suggest that LLMs are currently most valuable in this role. This is further supported by the improvements of HYBRIDARX relative to classical ARX in lower-intensity settings, where signals are inherently sparse.

Prompt-only LLM forecasts were competitive in terms of MAPE, which measures forecast accuracy, and exhibited minimal bias, indicating that the model did not consistently overpredict or underpredict hospitalization levels. Although the LLM forecasts typically lagged observed hospitalization changes by approximately one week, they were still able to capture the overall trend of increasing and decreasing hospitalizations, in some settings with comparable or lower variance than Lag-1 models. This supports the capability of LLMs to extract temporal structure from simple numeric sequences and track the direction and magnitude of hospitalization trends, despite not being specialized time-series models [10].

However, our results also highlight that simple baselines such as Lag-1 persistence and exponential smoothing should not be overlooked. While Lag-1 produced accurate forecasts with minimal bias,

exponential smoothing, despite its strong accuracy, exhibited consistent negative bias, systematically underestimating hospitalization levels. This highlights an important operational consideration: a model that achieves strong accuracy may still be problematic if it systematically underestimates demand or misrepresents critical turning points. These periods are often when healthcare decision-makers require the most support in planning staffing and resource allocation. If forecasts are misleading during these critical periods, they may place additional strain on healthcare systems.

Furthermore, the performance of the ARX model suggests that incorporating temporal structure alone is not sufficient. The addition of contextual signals does not automatically improve forecasting accuracy, and across the diverse data sources considered, simply including more inputs was not sufficient to improve performance [22]. However, it matters how these contextual signals are integrated into downstream forecasting systems. Advancing hospitalization forecasting therefore requires moving beyond short-term trend extrapolation toward identifying contextual signals that enable earlier detection of shifts in patient surges. LLMs can play a significant role in this transition due to their ability to reduce noise in heterogeneous contextual data and transform these signals into structured predictors suitable for forecasting pipelines.

HybridARX introduces additional computational overhead because contextual variables are given to an LLM before being incorporated into the downstream ARX model. However, unlike end-to-end LLM forecasting approaches, the language model is used only to generate a small set of contextual features, while the final prediction remains a lightweight and interpretable statistical model. This design limits the role of the LLM to generating contextual features, while the final prediction is produced by an interpretable statistical model. The purpose of this study was to use GPT-5 as a testbed for evaluating the HybridARX framework, rather than to compare the performance of different LLMs. Finally, the results further suggest that the greatest benefit of HybridARX occurs during periods of rapidly changing hospitalization dynamics, whereas simpler statistical baselines may remain appropriate under more stable conditions.

6 Conclusion

Our results show that LLMs are most effective not as standalone forecasters, but as components within hybrid pipelines that transform contextual information into usable predictors. HYBRIDARX embeds LLMs within constrained pipelines rather than relying on unconstrained black-box predictions, enabling the use of contextual information while maintaining reliable behavior in high-stakes decision environments. We find that simply incorporating contextual signals is insufficient: performance depends on how they are integrated within structured forecasting frameworks. In this setting, LLM-generated contextual representations improve forecast stability and consistency under noisy and non-stationary conditions. The HybridARX framework generalizes to real-world settings, where forecasting occurs under heterogeneous capacity constraints and fragmented data, and highlights that effective hospitalization forecasting depends not only on accuracy, but also on stability, reduced bias, and strong temporal alignment. While this work focuses on predictive performance, an important direction for future work

is identifying which exogenous signals directly influence hospitalization outcomes. Understanding which signals meaningfully precede changes in demand can help narrow the factors that hospitals should prioritize, enabling more actionable and decision-aware forecasting for epidemic response.

References

- [1] M. A. Al Aawar and A. Srivastava. 2025. Random forest of epidemiological models for influenza forecasting. *Epidemics* 53 (2025), 100862. doi:10.1016/j.epidem.2025.100862
- [2] S. M. Baik, M. Lee, K. S. Hong, and D. J. Park. 2022. Development of Machine-Learning Model to Predict COVID-19 Mortality: Application of Ensemble Model and Regarding Feature Impacts. *Diagnostics* 12, 6 (2022), 1464. doi:10.3390/diagnostics12061464
- [3] E. Berger, J. Winkelmann, H. Eckhardt, U. Nimptsch, D. Panteli, C. Reichebner, et al. 2022. A country-level analysis comparing hospital capacity and utilisation during the first COVID-19 wave across Europe. *Health Policy* 126, 5 (2022), 373–381. doi:10.1016/j.healthpol.2021.11.009
- [4] Luke Bhan, Hanyu Zhang, Andrew Gordon Wilson, Michael W. Mahoney, and Chuck Arvin. 2025. The Forecast Critic: Leveraging Large Language Models for Poor Forecast Identification. *arXiv* (2025). arXiv:2512.12059 [cs.AI]
- [5] Centers for Disease Control and Prevention. 2025. FluSight: Influenza forecasting. <https://www.cdc.gov/flu-forecasting/index.html>. U.S. Department of Health and Human Services. Accessed January 31, 2026.
- [6] Estee Y. Cramer, Evan L. Ray, Velma K. Lopez, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H. House, Yuxin Huang, et al. 2022. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *Proceedings of the National Academy of Sciences* 119, 15 (2022), e2113561119. doi:10.1073/pnas.2113561119
- [7] H. Du, J. Zhao, Y. Zhao, S. Xu, X. Lin, Y. Chen, L. M. Gardner, and H. F. Yang. 2024. Advancing real-time pandemic forecasting using large language models: A COVID-19 case study. *arXiv* (2024). arXiv:2404.06962
- [8] Google LLC. 2022. COVID-19 Community Mobility Reports. <https://www.google.com/covid19/mobility/>. Accessed January 31, 2026.
- [9] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski. 2024. MOMENT: A family of open time-series foundation models. *arXiv* (2024). arXiv:2402.03885
- [10] Mingyu Jin, Hua Tang, Chong Zhang, Qinkai Yu, Chengzhi Liu, Suiyuan Zhu, Yongfeng Zhang, and Mengnan Du. 2024. Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities. *arXiv preprint arXiv:2402.10835* (2024).
- [11] Sameer S. Kadri, Jia Sun, Ayman Lawandi, Joshua R. Strich, Lyle M. Busch, Michael Keller, Atheer Babiker, Ahmed Babiker, Aaron J. Littman, John Chaitram, et al. 2021. Association Between Caseload Surge and COVID-19 Survival in 558 U.S. Hospitals. *Annals of Internal Medicine* 174, 9 (2021), 1240–1251. doi:10.7326/M21-1213
- [12] Jonathon Mellor, Maria L. Tang, Owen Jones, Thomas Ward, Steven Riley, and Sarah R. Deeny. 2025. Forecasting COVID-19, influenza, and RSV hospitalizations over winter 2023–4 in England. *International Journal of Epidemiology* 54, 3 (2025), dyaf066. doi:10.1093/ije/dyaf066
- [13] Ray Moynihan, Sally Sanders, Zoe A. Michaleff, Anna M. Scott, Justin Clark, Emily J. To, Matthew Jones, Emma Kitchener, Matthew Fox, Mattias Johansson, Eddie Lang, Michelle Duggan, Ian Scott, and Loai Albarqouni. 2021. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. *BMJ Open* 11, 3 (2021), e045343. doi:10.1136/bmjopen-2020-045343
- [14] Pennsylvania Department of Health. 2021. COVID-19 Aggregate Hospitalizations: Current Weekly County Health. <https://data.pa.gov/Covid-19/COVID-19-Aggregate-Hospitalizations-Current-Weekly/kayn-sjhx>. Accessed January 31, 2026.
- [15] A. Preiss, E. Hadley, K. Jones, M. C. D. Stoner, C. Kery, P. Baumgartner, G. Bobashev, J. Tenenbaum, C. Carter, K. Clement, and S. Rhea. 2022. Incorporation of near-real-time hospital occupancy data to improve hospitalization forecast accuracy during the COVID-19 pandemic. *Infectious Disease Modelling* 7, 1 (2022), 277–285. doi:10.1016/j.idm.2022.01.003
- [16] Nicholas G. Reich et al. 2022. The United States COVID-19 Forecast Hub dataset. *Scientific Data* 9 (2022), 409. doi:10.1038/s41597-022-01517-w
- [17] Nicholas G. Reich, Yuxin Wang, Matt Burns, Rafael Ergas, Eva Y. Cramer, and Evan L. Ray. 2023. Assessing the utility of COVID-19 case reports as a leading indicator for hospitalization forecasting in the United States. *Epidemics* 45 (2023), 100728. doi:10.1016/j.epidem.2023.100728
- [18] Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, Sumit Agrawal, Wael Al Saeed, Taylor Arnold, Amartya Basu, Jacob Bien, Ángel A. Cabrera, Andrew Chin, Eu Jing Chua, Brian Clark, Sarah Colquhoun, Nat DeFries, David C. Farrow, Jodi Forlizzi, Jed Grabman, Samuel Gratzl, Alden Green, George Haff, Robin Han, Kate Harwood, Addison J. Hu, Raphael Hyde, Sangwon Hyun, Ananya Joshi, Jimi Kim, Andrew Kuznetsov, Wichada La Motte-Kerr, Yeon Jin Lee, Kenneth Lee, Zachary C. Lipton, Michael X. Liu, Lester Mackey, Kathryn Mazaitis, Daniel J. McDonald, Phillip McGuinness, Balasubramanian Narasimhan, Michael P. O'Brien, Natalia L. Oliveira, Pratik Patil, Adam Perer, Collin A. Politsch, Samyak Rajanala, Dawn Rucker, Chris Scott, Nigam H. Shah, Vishnu Shankar, James Sharpnack, Dmitry Shemetov, Noah Simon, Benjamin Y. Smith, Vishakha Srivastava, Shuyi Tan, Robert Tibshirani, Elena Tuzhilina, Ana Karina Van Nortwick, Valérie Ventura, Larry Wasserman, Benjamin Weaver, Jeremy C. Weiss, Spencer Whitman, Kristin Williams, Roni Rosenfeld, and Ryan J. Tibshirani. 2021. An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences* 118, 51 (2021). doi:10.1073/pnas.2111452118
- [19] P. Sandhu, A. B. Shah, F. B. Ahmad, J. Kerr, H. B. Demeke, E. Graeden, S. Marks, H. Clark, J. M. Bombard, M. Bolduc, K. Hatfield-Timajchy, E. Tindall, A. Neri, K. Smith, C. Owens, T. Martin, and F. V. Strona. 2022. Emergency Department and Intensive Care Unit Overcrowding and Ventilator Shortages in US Hospitals During the COVID-19 Pandemic, 2020–2021. *Public Health Rep.* 137, 4 (Jul-Aug 2022), 796–802. doi:10.1177/00333549221091781 Epub 2022 Jun 1. PMID: 35642664; PMCID: PMC9257510.
- [20] R. Somyanonthanakul, K. Warin, W. Amasiri, K. Mairiang, C. Mingmalairak, W. Panichkitkosolkul, K. Silanun, T. Theeramunkong, S. Nitikraipot, and S. Suebnukarn. 2022. Forecasting COVID-19 cases using time series modeling and association rule mining. *BMC Medical Research Methodology* 22, 1 (2022), 281. doi:10.1186/s12874-022-01755-x
- [21] U.S. Census Bureau. 2022. Pennsylvania — Census Bureau Profile. <https://data.census.gov/profile/Pennsylvania?g=040XX00US42>. data.census.gov. Accessed January 31, 2026.
- [22] Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. 2025. Context is Key: A Benchmark for Forecasting with Essential Textual Information. In *Proceedings of the International Conference on Machine Learning (ICML)*. <https://servicenow.github.io/context-is-key-forecasting/v0> Poster.
- [23] J. Wolff, A. Klimke, M. Marschollek, and T. Kacprowski. 2022. Forecasting admissions in psychiatric hospitals before and during COVID-19: a retrospective study with routine data. *Scientific Reports* 12, 1 (2022), 15912. doi:10.1038/s41598-022-20190-y

Reproducibility Details

Overview. We evaluate hospitalization forecasting using classical time-series models, a prompt-only LLM, and the proposed HYBRIDARX pipeline. All models operate on a weekly time series with a fixed rolling window of $L = 8$ observations and produce one-step-ahead forecasts $\hat{y}_{t+1}$.

Data and Preprocessing. All models are trained on county-level weekly time series. Each observation corresponds to a 14-day average hospitalization value. Data are aggregated to weekly resolution and aligned by county and week. Missing values in contextual indicators are handled using column-wise mean imputation within the rolling window, with remaining missing values set to zero. All forecasts are clipped to be non-negative.

Forecasting Setup. All methods use a rolling-origin, one-step-ahead forecasting protocol. For each county and forecast time $t + 1$, models are trained using the previous $L = 8$ weekly observations. Classical models and hybrid models are fit independently for each county and forecast origin.

Classical Baselines. We evaluate the following baselines:

- **Lag-1:** $\hat{y}_{t+1} = y_t$.
- **AR(1):** First-order autoregressive model with intercept (lags=1, trend='c').
- **Exponential Smoothing:** Additive trend model (no seasonality).
- **ARX:** AR(1) model with exogenous inputs X_B , X_V , and s_t .

Prompt-Only LLM. We use an LLM (gpt-5-2025-08-07) to directly forecast hospitalizations from the previous $L = 8$ observations. The prompt structure is fixed across all counties and weeks.

```
Given the last 8 weekly observations for region <Geography>:
<RecentData>
Predict next week's value (week ending <PredictionDate>)
for the numeric series y.
Return exactly one line:
y: <number>
```

Predictions are parsed using regular expressions. If parsing fails, the prompt is retried once. Three independent runs are performed per county-week.

HYBRIDARX. The proposed HYBRIDARX pipeline separates contextual forecasting from hospitalization prediction.

Stage 1 (LLM Context Prediction). An LLM forecasts next-week contextual indicators:

$$(\hat{X}_{B,t+1}, \hat{X}_{V,t+1}, \hat{s}_{t+1})$$

using the previous $L = 8$ weekly observations.

```
Given the last 8 weekly observations for region <Geography>:
<RecentData>
Predict next week's values (week ending <PredictionDate>)
for three numeric series:  $X_B$ ,  $X_V$ ,  $s_t$ .
Return exactly three lines:
X_B: <number>
X_V: <number>
s_t: <number>
```

Stage 2 (Statistical Forecasting). The LLM-predicted indicators are used as exogenous inputs in a downstream ARX model to generate $\hat{y}_{t+1}$. Training uses the previous $L = 8$ weeks of observed hospitalizations and contextual indicators.

Evaluation. Performance is evaluated using:

- Mean Absolute Percent Error (MAPE),
- Mean Percent Error (MPE),
- Lead-lag analysis using peak correlation ρ^* and lag ℓ^* .

All LLM-based methods are repeated for three independent runs to characterize variability.

All model configurations, prompts, and preprocessing steps are fully specified above. The code and data processing script, along with full experimental results, will be released upon publication.