Structured LLM-Augmented Forecasting for Non-Stationary Public Health Time Series Data

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

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

Abstract

Artificial intelligence systems have been increasingly proposed as part of medical and public health resource management workflows for tasks such as forecasting bed capacity during large-scale healthcare interruptions, such as operational failures, epidemics, or pandemics. These healthcare forecasting tasks have unique properties, such as non-stationary, noisy, multivariate data that exhibit unique types of measurement noise. This makes it challenging to develop and deploy such forecasting models in practice. Foundation models, in particular Large Language Models(LLMs), can address some structural challenges in forecasting tasks by incorporating richer forms of context into numerical forecasting, such as demographic, geographic, and population-level features. Accordingly, 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. We evaluate this system with bias and lead–lag alignment metrics 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 forecast performance depends not only on the availability of contextual information, but also on how that information is integrated into forecasting systems. Furthermore, these findings suggest that, in non-stationary healthcare resource forecasting, LLMs are most useful when embedded within structured hybrid models rather than used as standalone forecasters.

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 [11, 14]. 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 [14] and widespread disruption of essential services such as chemotherapy and rehabilitation [2]. As a result, healthcare systems, including hospitals and public health agencies, require reliable decision support to (1) guide resource allocation during large-scale healthcare disruptions and (2) respond effectively to emerging system strain [8, 9].Beyond their operational importance, hospitalization forecasting presents a challenging and unique forecasting problem because it is situated within a rapidly evolving system that combines non-stationarity, noisy exogenous signals, and heterogeneous data sources. Furthermore, contextual signals such as public health response measures, mobility patterns, and hospital capacity provide significant predictive value when assessing hospitalization demand; however, this predictive value can change over time and vary across different population densities. This motivates forecasting approaches that remain stable under noisy and rapidly changing conditions while still incorporating contextual information relevant to emerging system strain.

Aggregate forecast accuracy can mask operationally important failures, and while simple persistence models achieve strong error metrics, they offer limited decision value. Furthermore, autoregressive models can lag behind regime shifts under non-stationary conditions. This motivates forecasting systems that incorporate contextual signals while remaining stable enough for real-world resource planning.

Large Language Models (LLMs), with their advanced contextual reasoning capabilities using operational and public health knowledge(heterogeneous contextual information), could potentially provide a mechanism for incorporating heterogeneous contextual information into forecasting tasks under non-stationary conditions. We introduce a two-stage hybrid framework, HybridARX, that uses LLMs as contextual reasoning agents within a structured time-series forecasting pipeline. In this approach, an LLM first predicts the next-week 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 Pennsylvania hospitalization data against prompt-only LLM forecasting and classical time-series baselines, including persistence, autoregressive, exponential smoothing, and ARX models.

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) Structured integration of LLM-generated outputs as inputs to classical models improves robustness to noisy contextual signals while preserving the interpretability of classical forecasting models.

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 [11]. In these settings, forecasting utility depends not only on aggregate accuracy but also on calibration and reliability, which directly affect planning.

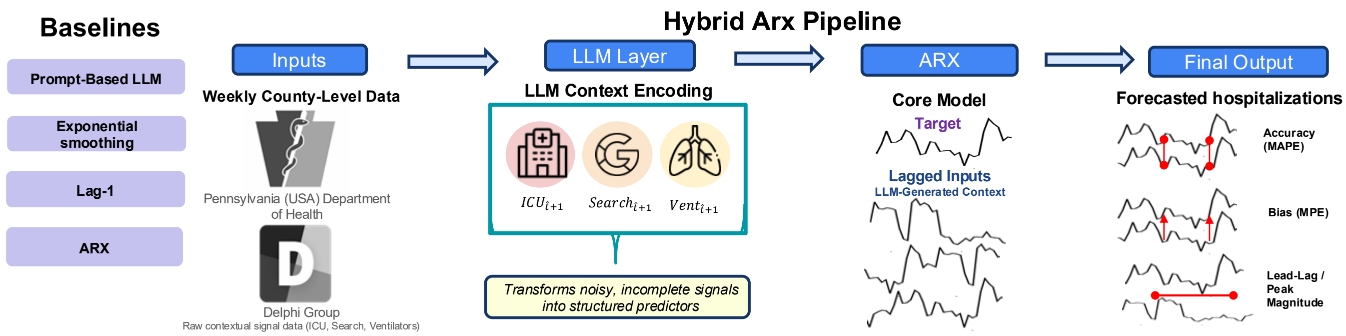

Figure 1: HybridARX Pipeline. Data is pulled from multiple data sources with varying granularity and is processed into standardized input. Then, it is evaluated on metrics relevant to decision-making tasks.

Prior work has applied machine learning and time series methods such as random forests, gradient-boosted trees, support vector machines, and elastic nets to forecast hospital admissions [1]. CDC forecasting efforts, such as FluSight [4] and the COVID-19 Forecast Hub [12], include multiple models, which are generally trend-extension and mechanistic hybrid approaches. Leading indicators such as hospital capacity, symptom search, mobility, and outpatient activity may precede hospitalization changes, but their relationship to hospitalizations can shift over time and across regions. This creates a need for forecasting systems that utilize contextual information.

Recent work has explored LLMs when forecasting the short-term spread of disease outbreaks by reformulating real-time forecasting as a text reasoning problem [5], as well as for monitoring time-series models to flag implausible or inaccurate predictions in large-scale retail settings [3]. 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 [10].[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.

Leading indicators: We used the Delphi COVIDcast API [13] 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 [6], 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 [15]. Counties were ranked by their mean weekly hospitalization counts over March 2020–January 2022 and categorized into lower, middle, and upper tertiles.

[1] Seven 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.

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, and there were strong associations between hospitalizations and measures of hospital capacity (e.g., ventilator utilization (0.98) and ICU beds (0.76)), as well as anosmia/ageusia search activity (0.717), and are used to guide our evaluation of how contextual signals impact forecasting performance.

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

3.4 Baseline Models

We compare HybridARX against a prompt-only LLM baseline and four classical baselines: Lag-1 persistence, AR(1), exponential smoothing, and ARX which uses the same exogenous variables as HybridARX.

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.

3.5 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 metrics:

1. MAPE Analysis. Forecasting accuracy is assessed using mean absolute percent error (MAPE), defined as $100 \times \frac{|\hat{y}_{t+1} - y_{t+1}|}{\max(y_{t+1}, 1)}$, 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 as $100 \times \frac{\hat{y}_{t+1} - y_{t+1}}{\max(y_{t+1}, 1)}$ Positive values indicate systematic overestimation, while negative values indicate systematic underestimation.

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 $\ell \in [-\ell_{\max}, \ell_{\max}]$: $\rho(\ell) = \mathrm{corr}(\Delta \hat{y}_{t-\ell}, \Delta y_t)$, computed over all overlapping evaluation weeks.

We then define the lead–lag estimate as $\ell^\star = \arg\max_{\ell \in [-\ell_{\max}, \ell_{\max}]} \rho(\ell)$. Positive values of $\ell^\star$ 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 $\ell^\star$ across counties. We additionally report the peak correlation value $\rho^\star = \max_{\ell \in [-\ell_{\max}, \ell_{\max}]} \rho(\ell)$, 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 $\rho^\star = 1$) mathematically guaranteed.

4 Results

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, despite using the same exogenous information. This suggests that forecast performance critically depends on how contextual signals are integrated rather than the contextual information itself. 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

Model	MAPE: Hospitalization Intensity		
	Low	Mid	High
Baselines			
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*
Proposed Method			
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.

Model	Lead–Lag Analysis			
	$\ell^\star$ Mean	$\ell^\star$ SD	$\rho^\star$ Mean	$\rho^\star$ SD
Baselines				
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*
Proposed Method				
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 ($\rho^\star$). Italics indicate values with overlapping intervals relative to the best-performing model.

Model	MPE: Hospitalization Intensity			
	Overall	Low	Mid	High
Baselines				
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*
Proposed Method				
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.

that incorporating noisy exogenous variables destabilized the regression, negatively impacted forecasting accuracy, and amplified instability under non-stationary conditions. The prompt-only LLM shows relatively consistent performance across county types and is competitive with 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.

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 under noisy forecasting conditions 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.

Because most methods yield mean lead–lag estimates close to $\ell^\star = -1$, we primarily focus on $\rho^\star$, 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. The prompt-only LLM forecasts, in particular, achieve competitive MAPE and calibration despite consistently lagging observed hospitalization changes by approximately one week. This suggests that LLMs have the ability to smooth temporal structure rather than capture broader directional trends. Furthermore, 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 ($\rho^\star = 0.80$), and ARX achieves zero-lag alignment in Philadelphia County ($\rho^\star = 0.65$). 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 and Conclusion

To better gain an understanding of the capabilities of HybridARX, we examine where it succeeds and where it falls short in comparison to other forecasting approaches and the merits of the approach itself. Compared to classical ARX, HybridARX consistently achieved lower MAPE, lower bias (MPE), and stronger trend alignment, even while using the same exogenous variables. This suggests that HybridARX reduces the degradation observed when noisy contextual signals are incorporated directly into ARX models. However, HybridARX did not consistently outperform the strongest statistical baselines, particularly Lag-1 and Exponential Smoothing, which often achieved lower forecasting error. However, hospitalization forecasting is not solely determined by point forecasting accuracy, but also encompasses calibration, robustness, and the ability to effectively leverage contextual information. While Exponential Smoothing generally produced better MAPE, HybridARX exhibited lower overall bias and improved calibration in the mid- and high-intensity counties. These findings suggest that HybridARX's primary value lies not necessarily in maximizing forecasting accuracy, but in better leveraging contextual information under non-stationary conditions while reducing instability associated with noisy exogenous signals. By transforming noisy contextual signals into more structured predictors, HybridARX allow forecasting models to effectively use contextual information while mitigating the degradation in ARX-based approaches

One explanation is that LLM-generated contextual signals help smooth or regularize noisy external indicators before they are incorporated into the forecasting model. Many public health indicators are inconsistently reported and partially missing, which can reduce their usefulness when directly incorporated into regression-based models as exogenous inputs. Therefore, since leading indicators anticipate shifts in hospitalization demand, improving the stability and temporal consistency of these signals is particularly important under non-stationary conditions.

Prompt-only LLM forecasts were still competitive – they had low MAPE 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 forecasting models [7].

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. This suggests LLMs are more effective as contextual reasoning components within structured forecasting pipelines than as standalone forecasters.

Our evaluations demonstrate that LLM-generated contextual representations improve forecast stability, robustness, and calibration under noisy and non-stationary conditions by reducing the degradation observed when exogenous signals are incorporated directly into classical ARX models. These findings suggest that forecast performance depends not only on the availability of contextual information, but also on how that information is integrated into forecasting systems. While this work focuses on LLMs within structured pipelines, 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] 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

[3] 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]

[4] 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.

[5] 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

[6] Google LLC. 2022. COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/. Accessed January 31, 2026.

[7] 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).

[8] 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

[9] 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

[10] 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.

[11] 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

[12] 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

[13] 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

[14] 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.

[15] U.S. Census Bureau. 2022. Pennsylvania — Census Bureau Profile. https://data.census.gov/profile/Pennsylvania?g=040XX00US42. data.census.gov. Accessed January 31, 2026.

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-Oracle:** AR(1) model with exogenous inputs X_B, X_V, and s_t. For one-step-ahead forecasting, observed contextual variables from the prediction week are supplied as exogenous inputs.

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 $\rho^\star$ and lag $\ell^\star$.

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

Implementation AR(1) and ARX models were estimated using ordinary least squares (OLS) as implemented in `statsmodels.tsa.ar`$_{m}odel.AutoR$
$wisemeans.AllLLMcallswereexecutedwithtemperature1anddefaultsettings.Nor$
$basedmethodswererepeatedforthreeindependentruns.$