nature computational science

Article

https://doi.org/10.1038/s43588-025-00798-6

Advancing real-time infectious disease forecasting using large language models

Received: 15 August 2024

Accepted: 1 April 2025

Published online: 6 June 2025

Check for updates

Hongru Du[1,2,9], Yang Zhao[1,2,9], Jianan Zhao[3,4,9], Shaochong Xu[1,2], Xihong Lin[5,6], Yiran Chen[7] ✉, Lauren M. Gardner[1,2,8] ✉ & Hao 'Frank' Yang[1,2,7] ✉

Forecasting the short-term spread of an ongoing disease outbreak poses a challenge owing to the complexity of contributing factors, some of which can be characterized through interlinked, multi-modality variables, and the intersection of public policy and human behavior. Here we introduce PandemicLLM, a framework with multi-modal large language models (LLMs) that reformulates real-time forecasting of disease spread as a text-reasoning problem, with the ability to incorporate real-time, complex, non-numerical information. This approach, through an artificial intelligence–human cooperative prompt design and time-series representation learning, encodes multi-modal data for LLMs. The model is applied to the COVID-19 pandemic, and trained to utilize textual public health policies, genomic surveillance, spatial and epidemiological time-series data, and is tested across all 50 states of the United States for a duration of 19 months. PandemicLLM opens avenues for incorporating various pandemic-related data in heterogeneous formats and shows performance benefits over existing models.

Pandemic forecasting is essential for providing situational awareness and supporting decision-making for policymakers during public health emergencies. The ability to forecast short-term disease outcomes is crucial for informing resource allocation and risk mitigation strategies in near-term timeframes, ultimately aiming to minimize the burden of diseases. Predominantly, the existing forecasting models used in research and practice can be categorized into two types: mechanistic models, which simulate transmission dynamics in the population through compartmental models such as the susceptible, infectious, and recovered (SIR) model and its derivatives[1], and statistical models, which adopt data-driven approaches to forecast disease trends using historical data patterns[2,3]. Different models are suited for distinct forecasting needs: mechanistic models are beneficial for long-term projections owing to their ability to integrate scenario assumptions, whereas statistical models are helpful for short-term forecasting because they can adapt to immediate trends effectively[4]. Despite the pivotal role of statistical models in short-term forecasting, they are limited in their ability to (1) adapt to multi-modal data in real time[5], (2) respond to rapidly changing policies[6], (3) account for the emergence of new variants[7], and (4) translate the predicted results into useful decision-support guidance with trustworthiness[8]. Consequently, current pandemic forecasting models have often struggled to identify and predict critical turning points in the pandemic's trajectory[9]. Moreover, the absence of a transparent interpretation of these results can diminish public trust in these models and even potentially undermine the efficacy of public health responses.

The COVID-19 pandemic highlighted each of these deficiencies in the existing set of disease forecasting tools, which, as a result, struggled to accurately forecast disease spreading patterns[10]. In particular, the complexities of the COVID-19 pandemic arose from several interrelated factors that were difficult to incorporate into predictive models: (1) the virological characteristics of the virus[11], such as relative

[1]Center for Systems Science and Engineering, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD, USA. [3]Mila - Quebec AI Institute, Montreal, Quebec, Canada. [4]Department of Computer Science, Universite de Montréal, Montreal, Quebec, Canada. [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [6]Department of Statistics, Harvard University, Cambridge, MA, USA. [7]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. [8]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [9]These authors contributed equally: Hongru Du, Yang Zhao, Jianan Zhao. ✉e-mail: yiran.chen@duke.edu; l.gardner@jhu.edu; haofrankyang@jhu.edu

transmissibility, severity and impact on immunity; (2) the diverse demographic profiles of affected populations, including their evolving immunity[12]; and (3) the dynamic relationship between public policy and human behavior[13]. The virological dynamics, including factors such as the characteristics of viruses and their current prevalence, represent a biological 'language' with its own syntax and semantics, ruling how a virus behaves and evolves[14]. Simultaneously, human dynamics, including demographics, behavioral patterns and responses to policy changes, are akin to a socio-cultural 'language', rich with subtle cognitive processes. Recognizing these complexities, we aim to enhance disease forecasting by exploiting the richness of timely, complex data made available in multiple structures and formats, diverging from the current research trend of simplifying information into solely numerical forms for the purposes of mathematical modeling.

Recently, the large language models (LLMs) stand out as a substantial advancement in artificial intelligence (AI), which have demonstrated proficiency in multi-modal contextual feature learning[15–17]. Their strength in text generation and reasoning suggests potential for understanding the complex dynamics of disease spread[18,19]. However, LLMs are designed for processing information in natural language, which presents the following distinct challenges when applying them to pandemic forecasting. (1) Incorporating multi-modality information: the data used for forecasting pandemics include epidemiological time series, public health policy, genomic surveillance and demographics profiles, requiring these multi-modality inputs to be organized and tailored into suitable learned representations and structured contextual prompts. (2) Incapability in modeling temporal dynamics: LLMs typically struggle with time-series data owing to the tokenization of continuous numbers[20], requiring new methodologies that strategically embed temporal information into prompts, enabling LLMs to effectively utilize these information for reasoning.

In light of the limitations of existing models and the well-established text-reasoning capability of LLMs, we propose PandemicLLM, an LLM-based framework for pandemic forecasting (Fig. 1). We reformulate pandemic forecasting as a text-reasoning task, allowing the integration of new data streams that were not previously utilized in pandemic forecasting models. The proposed framework is designed for state-level COVID-19 hospitalization forecasting across the United States, targeting 1-week and 3-week prediction horizons. We tackle the challenge of multi-modality data input in LLMs by employing an AI–human cooperative approach for prompt generation, coupled with the use of a recurrent neural network (RNN) for temporal representation learning of epidemiological time series. This strategy enables an LLM to effectively process complex pandemic-related information, encompassing spatial, epidemiological time series, public health policy and genomic surveillance aspects. Furthermore, by recasting forecasting as an ordinal classification of hospitalization trends, we align the model's output with the needs of public health decision-makers, also adhering to Centers for Disease Control and Prevention (CDC) guidance[21]. We present three versions of PandemicLLM, each trained on progressively larger datasets, and rigorously evaluate their performance over extended periods (4 to 19 months) without retraining. Comprehensive testing across all 50 US states demonstrates the following key advantages. PandemicLLM delivers robust and trustworthy forecasts, offering categorical predictions with confidence levels that support public health policymaking. Moreover, it exhibits timely and accurate responses to emerging variants, leveraging the textual analysis of timely genomic surveillance data. These findings highlight the potential of leveraging LLMs in public health emergency response settings.

Results

The PandemicLLM framework

To provide actionable and interpretable forecasts, we formulate pandemic forecasting as an ordinal classification problem. Leveraging the heterogeneous data types discussed in 'Multi-modality pandemic data', we adapt these data into a textual format amenable to LLM learning facilitated by an AI–human cooperative prompt design. Subsequently, the PandemicLLM undergoes supervised fine-tuning using designed prompts and targets. To ensure the accuracy and reliability of our predictions, we undertake an extensive performance evaluation, focusing particularly on the results' trustworthiness. Figure 1b illustrates the entire construction pipeline of PandemicLLM, with subsequent sections providing detailed explanations of each segment of the pipeline.

In our effort to build PandemicLLM, particular attention was given to the design of the prediction targets, aiming to refine the targets used for COVID-19 forecasting. We argue that continuous targets, such as reported cases or hospitalizations, are prone to reporting errors and hinder clear uncertainty communication for stakeholders. The frequent ambiguity observed in the conflicting predicted trends of 50% and 95% confidence intervals from the COVID-19 Forecast Hub[22] highlights this problem. To address this, we introduce a straightforward yet informative target: hospitalization trend category (HTC). This target categorizes future hospitalization trends into five levels: substantial decrease, moderate decrease, stable, moderate increase and substantial increase. Extended Data Fig. 1 shows a visualization of the designed targets for 1 week and 3 weeks ahead. Detailed definitions for the HTCs can be found in 'Epidemiological time series data' in Methods.

The cornerstone of the PandemicLLM framework is the AI–human cooperative prompt design, a process that blends human insight with AI efficiency. This framework involves the embedding of multi-modality data into textual formats, and each resulting prompt has around 300 words (see Supplementary Section 6 for examples of completed prompts). Visualizations of the data transformation into text are provided in Fig. 2, with detailed descriptions available in 'Multi-modality data textualization' in Methods.

We fine-tuned LLaMA2[23]—a publicly accessible LLM by Meta—to predict the HTC for the upcoming 1 week and 3 weeks for each state. As illustrated in Fig. 1b, the model was fine-tuned to maximize the probability of predicted tokens corresponding to the target category. We developed three versions of PandemicLLM, each trained and validated on different data periods (Extended Data Fig. 1). Specifically:

- PandemicLLM-1: trained and validated on data up to June 2021 (5 months) with an 80/20 partition ratio, then tested on subsequent data for 19 months without any retraining.
- PandemicLLM-2: trained and validated on data up to December 2021 (11 months) with an 80/20 partition ratio, then tested on the following 13 months without any retraining.
- PandemicLLM-3: trained and validated on data up to September 2022 (20 months) with an 80/20 partition ratio, then tested on the subsequent 4 months without any retraining.

Evaluation metrics: as our proposed framework reformulates pandemic forecasting as an ordinal classification problem, we adopt the widely used accuracy and mean-squared error (MSE) for evaluation. Nevertheless, these two metrics evaluate only the prediction with the largest probability, overlooking the prediction distribution. For a fine-grained evaluation, we propose to evaluate using weighted MSE (WMSE) along with the Brier score[24], and the ranked probability score (RPS)[25]. Specifically, the Brier score quantifies the precision of probability estimations. The RPS penalizes predictions where the predicted probabilities deviate from the target category. The WMSE escalates its penalization in proportion to the divergence of the predicted probabilities from the target category. The detailed definition for each error metric is documented in 'Evaluation' in Methods.

Reference model: we selected the CDC Ensemble Model[26] as the reference for performance comparison owing to its status as the state-of-the-art model for short-term forecasting during the COVID-19 pandemic. This model produces point predictions accompanied by quantile values, which represent the confidence interval of the

https://doi.org/10.1038/s43588-025-00798-6

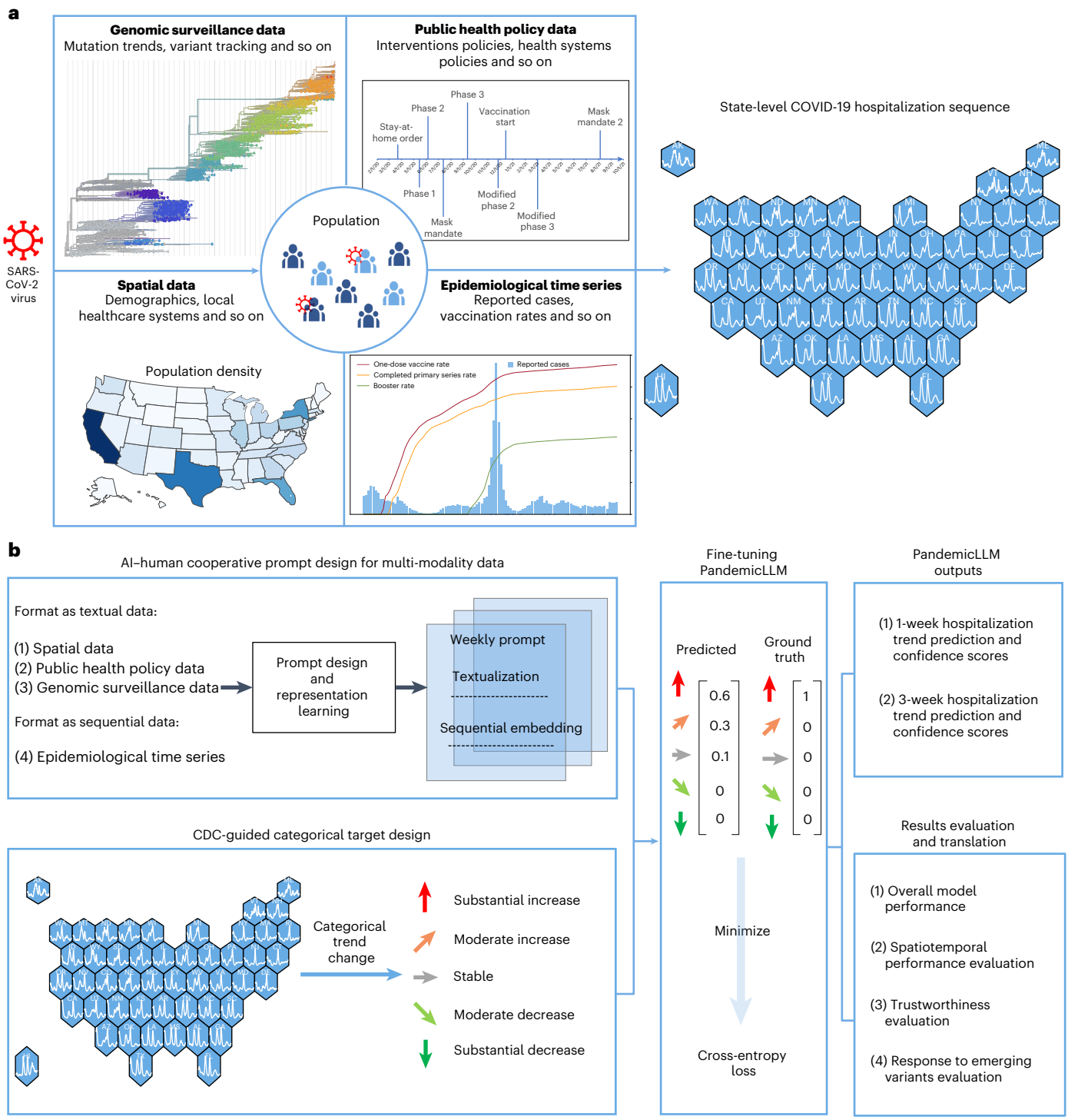


Fig. 1 | Overview of PandemicLLM's pandemic data streams and pipeline. **a**, Multi-modality data insights into the pandemic. Our multi-modality dataset integrates four types of pandemic data sources: spatial, epidemiological time-series, public health policy and genomic surveillance data. Spatial data include demographic and healthcare indicators, whereas the epidemiological time-series aspect covers reported cases, hospitalizations and vaccination rates. Data about policy detail governmental interventions in a textual format (dates are in the format month-day-year), and the genomic surveillance data integrate textual descriptions of variants with weekly sequences regarding their prevalence. The data comprise 5,200 records, covering all 50 US states over 104 weeks. The phylogenetic tree of SARS-CoV-2 was generated using Nextstrain[53]. **b**, PandemicLLM's construction pipeline. To forecast pandemic hospitalization trends, we formulate the problem as an ordinal classification task. We define five categories following CDC guidance[21]: substantial decrease, moderate decrease, stable, moderate increase and substantial increase. By converting multi-modality data into a text format through AI–human cooperative prompt design, PandemicLLM is fine-tuned with these prompts and targets for 1-week and 3-week forecasts. We emphasize rigorous performance assessment to verify the accuracy and trustworthiness of our predictions.

https://doi.org/10.1038/s43588-025-00798-6

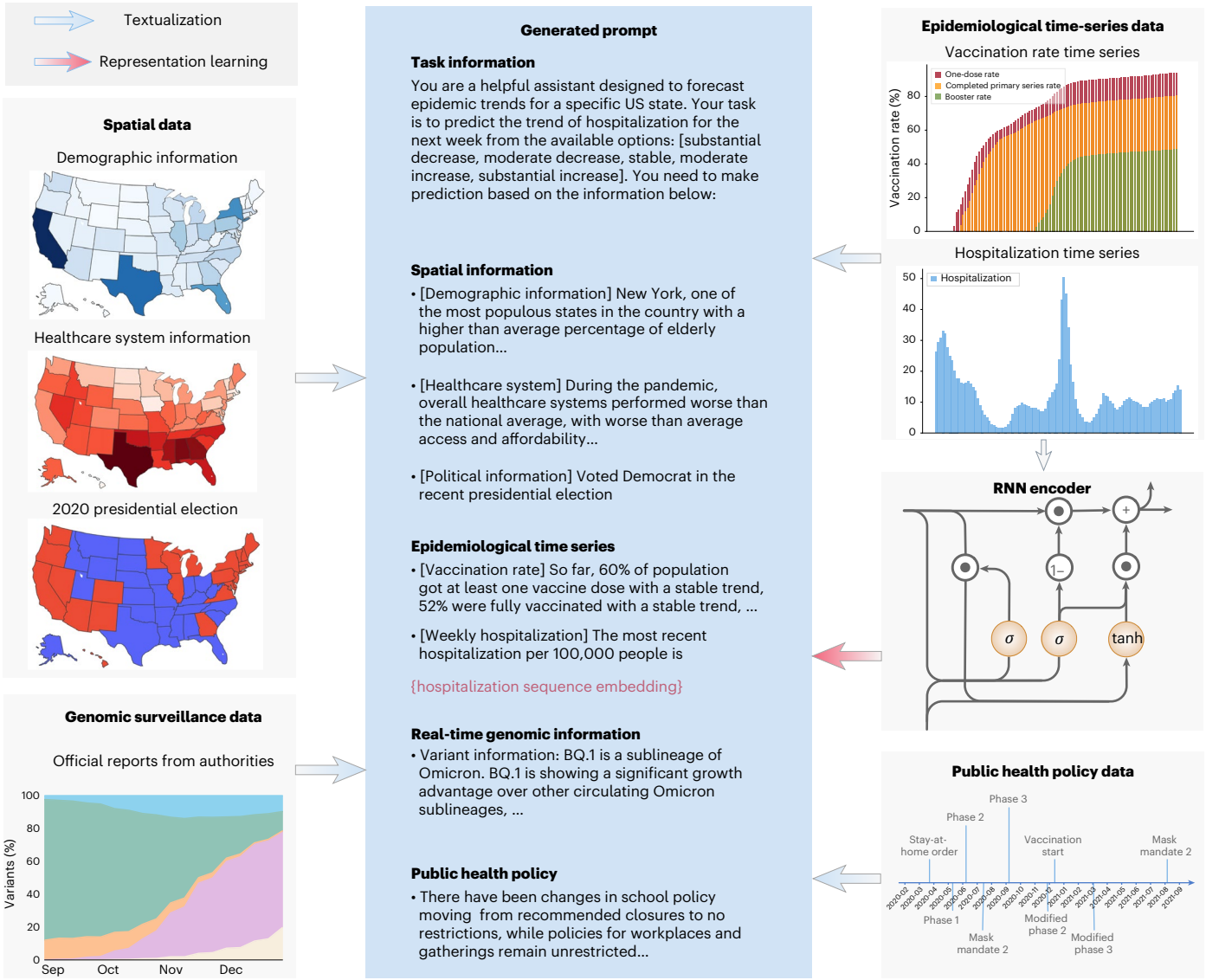


Fig. 2 | Summary of the AI–human cooperative prompt design. Spatial data for all 50 US states are converted into descriptions to reflect their rankings; the policy data include stringency levels and changes from week-to-week. Epidemiological time-series data use both narrative generation and representation learning. The representation learning component uses an RNN encoder, and the figure shows a GRU-based encoder where σ represents the sigmoid activation function. Genomic surveillance data combine textual summaries of variant characteristics with recent prevalence. The blue arrow indicates the information textualization and the red arrow indicates the sequence representation learning. Each designed prompt has 296 to 322 words. Dates are in the format year-month.

forecast. To ensure comparability with our PandemicLLMs, we converted these continuous quantile values into categorical outcomes. Specifically, each quantile value was assigned to one of the five HTC categories based on predefined thresholds. For each set of quantile predictions, we then calculated the frequency distribution of these categorical outcomes across the five categories. The category with the highest frequency was selected as the final prediction of the CDC Ensemble Model.

Multi-modality pandemic data

In addressing the intricate dynamics of COVID-19 transmission, which involves data of heterogeneous formats, the proposed PandemicLLM was built on four data categories: spatial, epidemiological time-series, public health policy and genomic surveillance data (Fig. 1a). This categorization reflects the distinct nature and representation of each data type. Spatial data, sourced at the state level, comprised numeric and static variables, including demographic information, healthcare system scores and political affiliations. Epidemiological time-series data were collected at weekly resolution for each state, encompassing numerical and sequential data, such as reported COVID-19 cases, hospitalization and vaccination rates. Public health policy data, also state specific, were captured in a textual format, detailing the stringency level and types of government policy. Genomic surveillance data are a hybrid of textual and sequential formats, with textual data sourcing from the authoritative reports of variants' virological characteristics and sequential data reflecting the current prevalence of these variants.

Our multi-modality pandemic dataset encompasses information spanning 50 states across the United States and 104 weeks from January 2021 to January 2023, culminating in a total of 5,200 data records. This dataset includes five types of public health policy, three variations in vaccine rates and genomic data on five Omicron sublineages. Utilizing the AI–human cooperative prompt design ('Multi-modality data textualization' in Methods), these data records have been transformed into a set of prompts, aggregating approximately 1.51 million words.

https://doi.org/10.1038/s43588-025-00798-6

Table 1 | A summary of models' performances for the three PandemicLLMs and CDC Ensemble Model by different periods

Prediction period	Target	Model	Evaluation metric (model rank)					Average rank
			Accuracy ↑	MSE ↓	WMSE ↓	Brier score ↓	RPS↓	
June 2021–December 2021	1 week	CDC Ensemble	0.337 (2)	2.958 (2)	5.425 (2)	1.121 (2)	0.327 (2)	2
		PandemicLLM-1	**0.568** (1)	**1.082** (1)	**1.177** (1)	**0.653** (1)	**0.124** (1)	1
	3 week	CDC Ensemble	0.211 (2)	5.156 (2)	5.603 (2)	1.127 (2)	0.370 (2)	2
		PandemicLLM-1	**0.379** (1)	**2.325** (1)	**2.238** (1)	**0.956** (1)	**0.222** (1)	1
December 2021–September 2022	1 week	CDC Ensemble	0.376 (3)	2.385 (3)	5.524 (3)	1.114 (3)	0.334 (3)	3
		PandemicLLM-1	0.598 (2)	0.862 (2)	0.925 (2)	0.579 (2)	0.102 (2)	2
		PandemicLLM-2	**0.644** (1)	**0.775** (1)	**0.842** (1)	**0.525** (1)	**0.092** (1)	1
	3 week	CDC Ensemble	0.245 (3)	4.392 (3)	5.400 (3)	1.154 (3)	0.351 (3)	3
		PandemicLLM-1	0.427 (2)	1.815 (2)	1.823 (2)	0.862 (2)	0.182 (2)	2
		PandemicLLM-2	**0.567** (1)	**1.389** (1)	**1.498** (1)	**0.587** (1)	**0.125** (1)	1
September 2022–February 2023	1 week	CDC Ensemble	0.239 (4)	2.266 (4)	3.492 (4)	0.919 (4)	0.197 (4)	4
		PandemicLLM-1	0.540 (2)	0.905 (3)	0.968 (3)	0.621 (2)	0.109 (3)	2.6
		PandemicLLM-2	0.526 (3)	0.742 (2)	0.807 (2)	**0.575** (1)	**0.095** (1)	1.8
		PandemicLLM-3	**0.560** (1)	**0.628** (1)	**0.767** (1)	0.634 (3)	0.098 (2)	1.6
	3 week	CDC Ensemble	0.213 (4)	2.754 (4)	3.259 (4)	0.871 (4)	0.194 (4)	4
		PandemicLLM-1	0.409 (3)	1.760 (3)	1.777 (3)	0.876 (3)	0.183 (3)	3
		PandemicLLM-2	**0.468** (1)	1.220 (2)	1.354 (2)	**0.621** (1)	0.121 (2)	1.6
		PandemicLLM-3	0.464 (2)	**0.760** (1)	**0.908** (1)	0.695 (2)	**0.106** (1)	1.4

The up (down) arrows indicate that higher (lower) metric values signify better performance. The numbers in parentheses represent the relative ranking of each model. The bold numbers represent better performance.

Further details regarding the data sources and the methods used in data preprocessing are documented in 'PandemicLLM datasets' in Methods.

PandemicLLM performance evaluation

In this section, we provide a comprehensive evaluation of the three PandemicLLMs, comparing their performance both among themselves and against the CDC Ensemble Model. Table 1 provides a detailed summary of the models' performance across various periods and Fig. 3 visually illustrates their performance trajectories over time. Each bar plot in Fig. 3 represents the WMSE distribution across all states for a specific week, where a lower bar indicates a better performance. Equivalent evaluations with alternative error metrics are included in Supplementary Section 10. See Supplementary Section 9 for state-by-state predictions visualization against the CDC Ensemble Model and the ground-truth data.

Our analysis reveals several key findings regarding the performance of the three PandemicLLMs. First, all three versions consistently outperform the CDC Ensemble model in forecasting both 1-week and 3-week hospitalization trends, with the exception of the 3-week predictions for December 2022 (Fig. 3). This temporary dip in performance for models trained before the emergence of the Omicron variant highlights the need for dynamic retraining strategies in response to shifts in disease transmission dynamics (see Supplementary Section 14 for a detailed analysis of when retraining is necessary). Notably, even PandemicLLM-1, trained on only 5 months of data, shows robust predictive capability over a one-and-a-half-year period. This finding underscores the value of adapting LLMs for infectious disease forecasting. While LLMs may require longer initial training times, their ability to maintain accuracy and reliability over extended periods, far exceeding the lifespan of many traditional models, makes them an invaluable tool in this domain. Second, the PandemicLLMs show a notable performance advantage when the CDC Ensemble model produces larger errors, particularly around inflection points such as September 2021 and early 2022, when hospitalization trends transitioned from increasing to decreasing (Extended Data Fig. 1). This highlights the models' ability to adapt to shifts in disease dynamics, maintaining accuracy across various stages of an outbreak.

Table 1 compares the forecasting performance of the three PandemicLLMs and the CDC Ensemble model across three distinct time periods. The results show that all three PandemicLLMs outperform the CDC Ensemble model in each evaluated period, with a clear margin across all metrics. We observe that increasing the amount of training data leads to more substantial performance improvements in 3-week forecasting compared with 1-week forecasting. This is particularly evident from December 2021 to September 2022, where increasing the training data from 5 months (PandemicLLM-1) to 11 months (PandemicLLM-2) yields substantial enhancements in performance across both forecasting windows. However, from September 2022 to February 2023, PandemicLLM-2 (11 months of training data) and PandemicLLM-3 (20 months) show comparable performance in 1-week-ahead forecasting. For 3-week forecasts, PandemicLLM-3 showcases notable improvements in metrics that penalize more severe errors, such as WMSE and RPS, with 61% and 49% reductions, respectively. However, it shows a slightly lower percentage of correct predictions than PandemicLLM-2. These results suggest that PandemicLLMs achieve stable predictive accuracy with less than 1 year of fine-tuning data. Furthermore, the inclusion of more data reduces the occurrence of large forecasting errors.

The proposed PandemicLLM framework is enhanced by the integrated sequential GRU encoder for hospitalization time series, leading to an accuracy improvement of 17–24% over the model without the encoder (Supplementary Table 2). This design allows for easy future extensions to accommodate other time-series encoder architectures (such as attention mechanisms), potentially providing better suitability for different disease types and propagation patterns. The AI–human collaborative prompt design is aimed to enable automation in real time, allowing human expertise to be strategically focused on selecting the most relevant and timely information from authoritative sources, such as policy updates and variant reports.

https://doi.org/10.1038/s43588-025-00798-6

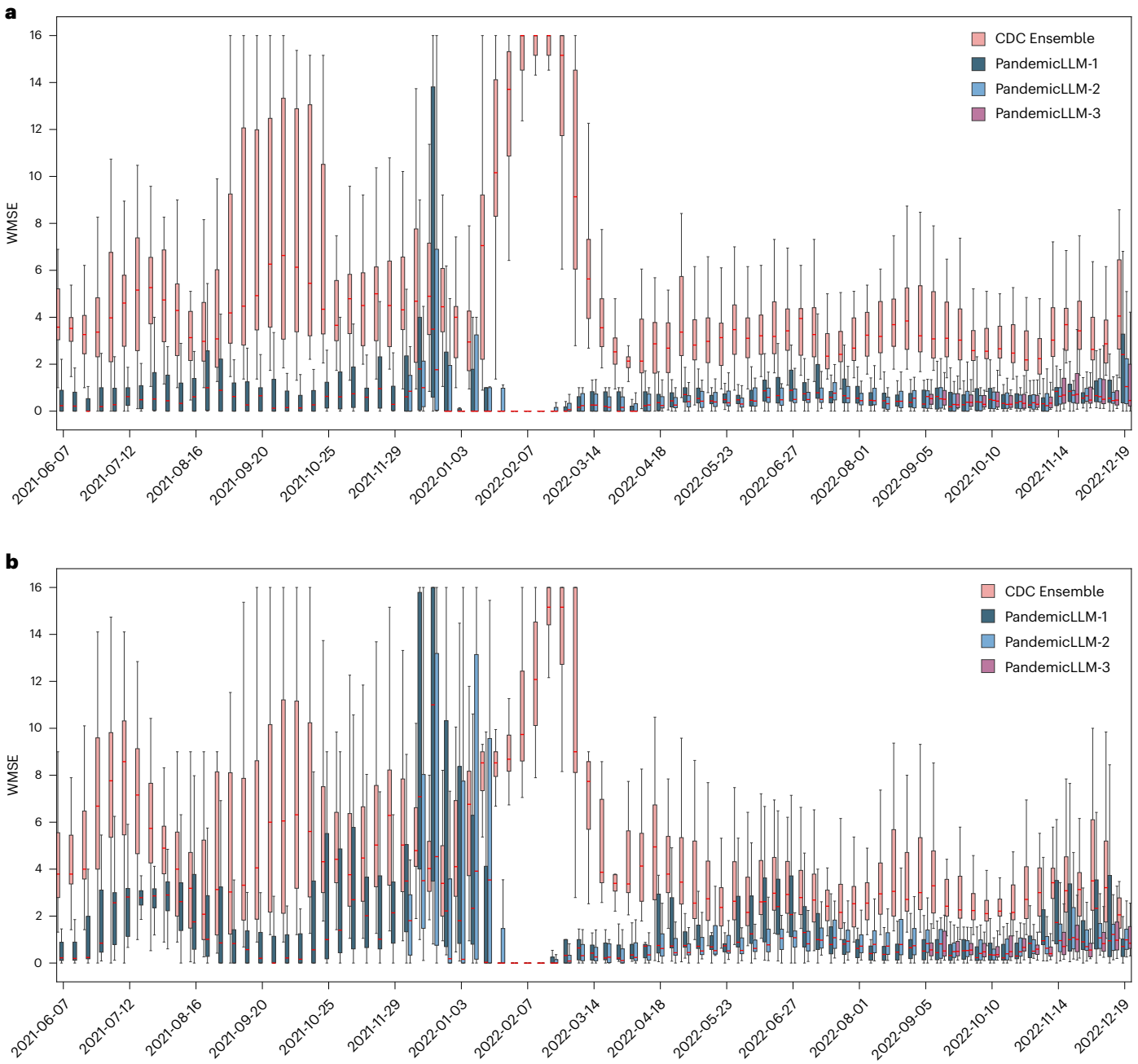


Fig. 3 | PandemicLLM performance evaluation. Performance comparison of the three PandemicLLMs with the CDC Ensemble Model across time. The *y* axis represents the scale of WMSE. Each set of bar graphs in the figure represents the distribution of WMSE for all 50 states during a specific week ($n = 50$, where each state represents an independent sample). Data are presented as median values, with the box indicating the interquartile range (IQR: 25th percentile (Q1) to 75th percentile (Q3)). The whiskers extend to the most extreme data points within 1.5 times the IQR from the lower and upper quartiles, while outliers (beyond this range) are not displayed. The color bars represent the error distributions for different models. **a**, The 1-week forecasting performance. **b**, The 3-week forecasting performance. See Supplementary Section 9 for predictions visualization and Supplementary Section 10 for alternative evaluation metrics. Dates are in the format year-month-day.

Trustworthy and robust results

To ensure the integrity of our predictions, we implemented a time-anonymized prompt design that excludes specific dates and historical outcomes, focusing solely on the information provided in the prompts. Zero-shot tests using LLaMA2 (Supplementary Table 8) confirmed that performance was only slightly better than random guessing without fine-tuning. Furthermore, fully anonymized experiments, removing all disease-specific terms and state identifiers, showed that PandemicLLM maintained robust performance (Supplementary Table 9). These results validate that prior knowledge does not influence PandemicLLM's forecasts, ensuring trustworthy predictions.

The standard for models designed to aid public health decision-making is rigorous, necessitating a reliable model that provides clear guidance on its utility. In response, we show that the confidence level of PandemicLLM, the highest probability assigned to the predicted category, is a robust indicator of reliability. Specifically, we present the models' accuracy with respect to different confidence thresholds in Fig. 4a,b. As we elevate the confidence threshold, wherein only predictions surpassing this criterion are considered, notable enhancements

https://doi.org/10.1038/s43588-025-00798-6

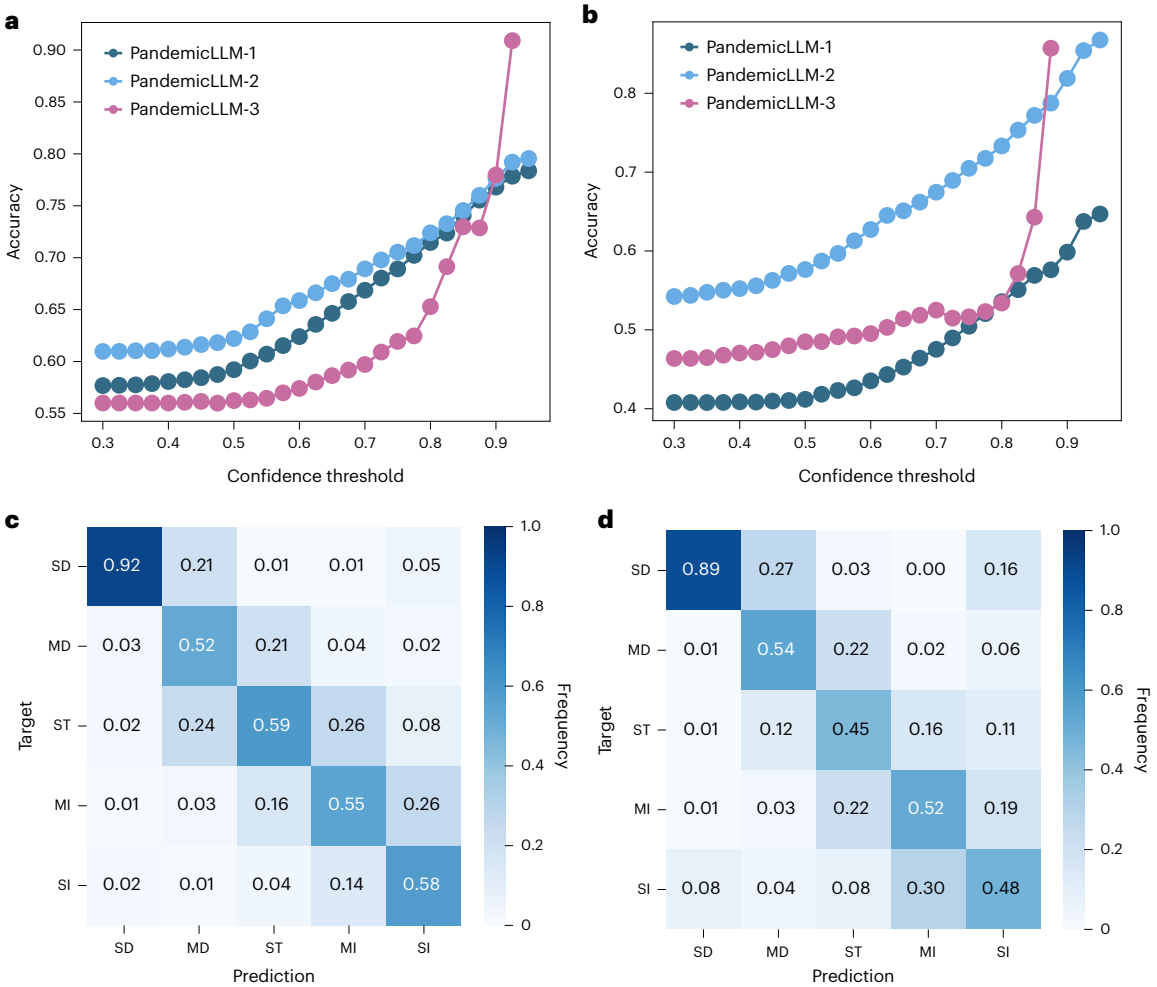


Fig. 4 | Trustworthiness for PandemicLLMs. a,**b**, The accuracy of 1-week (**a**) and 3-week (**b**) predictions varied across various levels of prediction confidence. **c**,**d**, The 1-week (**c**) and 3-week (**d**) confusion matrix for PandemicLLM-2. The color bar indicates the frequency (proportion of samples) from 0 to 1. Confusion matrices for other models are presented in Supplementary Section 11. SD, substantial decrease; MD, moderate decrease; ST, stable; MI, moderate increase; SI, substantial increase.

in prediction accuracy are observed for all three PandemicLLMs. For instance, setting the confidence threshold at 0.85, PandemicLLM-2 attains an accuracy of 75% for 1-week forecasts and 77% for 3-week forecasts. This characteristic of our model is invaluable for public health policy decision-making, offering decision-makers the flexibility to devise strategies grounded in the predictive confidence provided by the model.

The stakes of making incorrect public health decisions are exceedingly high, such as prematurely easing restrictions when future hospital admissions are rising. A reliable forecasting model is paramount to prevent substantial prediction errors, for instance, misidentifying a 'substantial increase' as a 'substantial decrease,' which could severely undermine public health strategies. To evaluate the reliability of PandemicLLM under different pandemic phases, we analyze the confusion matrices of PandemicLLM-2 shown in Fig. 4c,d. The findings underscore two pivotal insights. (1) PandemicLLM shows strong predictive capability during the decreasing phases, particularly in identifying substantial decreases. This attribute is invaluable for informing policies on reopening and easing restrictions safely. (2) Errors in forecasting predominantly occur between neighboring categories. For instance, when PandemicLLM-2 forecasts 'substantial increase' predictions, the actual situation reflected at least a 'moderate increase' in 84% of 1-week forecasts and in 67% for 3-week forecasts. This level of reliability suggests that PandemicLLM can offer valuable insights for making informed operational decisions across various phases of a pandemic. (3) Predicting upward trends is more challenging than downward trends, often influenced by behavior changes and emerging variants[27]. Early waves such as Alpha and Delta faced delays in genomic-sequence reporting[28], impacting timely surveillance and forecasting accuracy. The following section examines how incorporating real-time genomic information during the Omicron wave improves performance during increasing phases.

Integrating real-time genomic surveillance information for timely response

One unique ability of PandemicLLM is that it can incorporate previously unseen information through text reasoning. One notable example is the emergence of new variants that pose challenges for traditional real-time forecasting. In this section, we present the capability of our model to mitigate such challenges by incorporating timely information on new variants. This analysis focuses on 3-week forecasting, aiming to provide an extended lead time for informed decision-making. We concentrate on the later period of the Omicron wave owing to the shorter lag between sample collection and reporting, ensuring more reliable variant proportion estimates.

Specifically, the hospitalizations surge during the testing period can be attributed to the rise of the SARS-COV-2 BQ.1 variant starting in October 2022, which became the predominant strain by December 2022, as depicted in Fig. 5a. The initial authoritative report detailing the virological properties of the BQ.1 variant was released on 27 October 2022[29].

https://doi.org/10.1038/s43588-025-00798-6

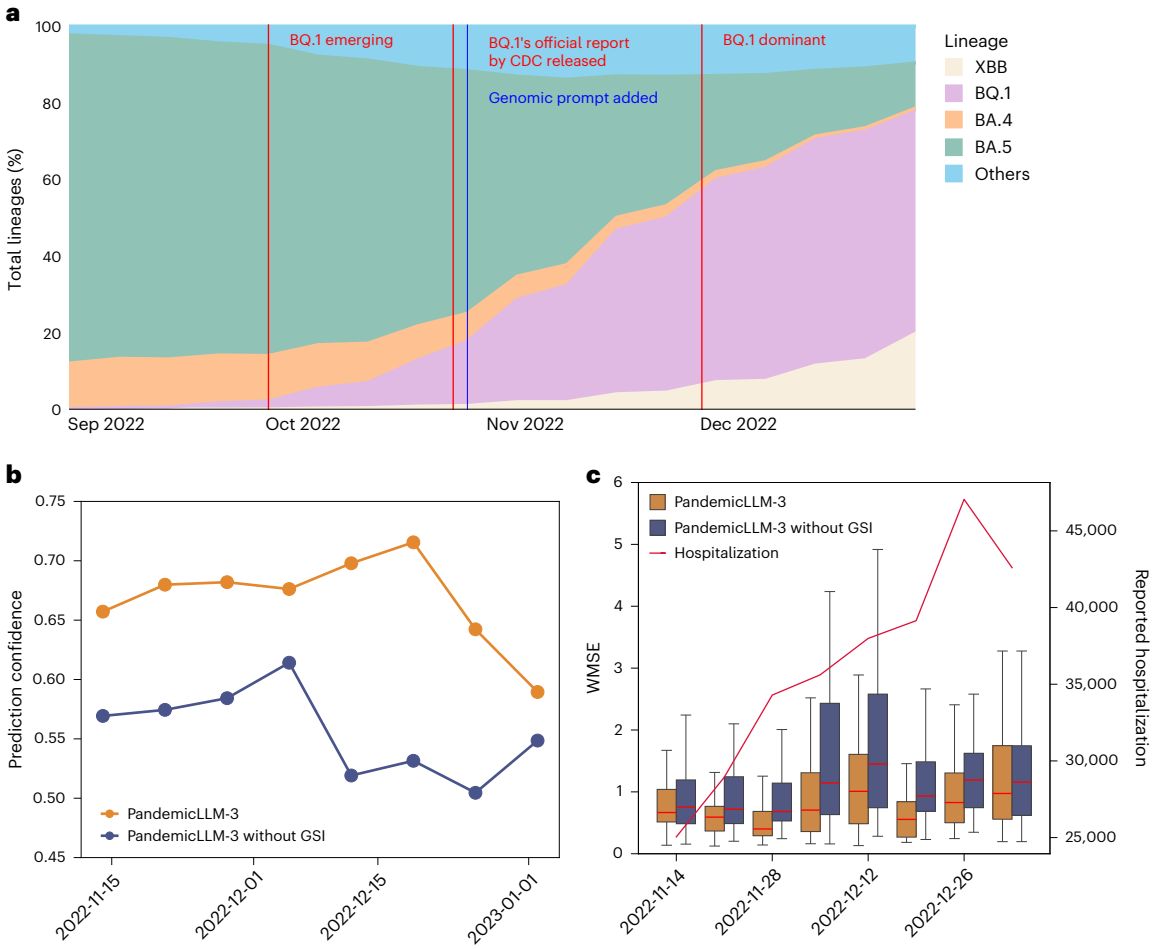


Fig. 5 | A comparative analysis with and without the real-time genomic surveillance information. a, National estimates of weekly proportions of SARS-CoV-2 variants from September 2022 to January 2023. **b**, Prediction confidence of PandemicLLMs with and without real-time genomic surveillance information (without GSI). The blue line represents the models without GSI. Dates are in the format year-month-day. **c**, Comparison of model performance with and without GSI. The blue bars represent the models without GSI. Each set of bar graphs in the figure represents the distribution of WMSE for all 50 states during a specific week ($n = 50$, where each state represents an independent sample). Data are presented as median values, with the box indicating the interquartile range (IQR, 25th percentile (Q1) to 75th percentile (Q3)). The whiskers extend to the most extreme data points within 1.5 times the IQR from the lower and upper quartiles, while outliers (beyond this range) are not displayed. Dates are in the format year-month-day.

Subsequently, within the same week, we incorporated these specific characteristics of the BQ.1 variant (infectiousness, severity and resistance to immunity), together with the latest variant proportion estimates from the CDC[30], into the PandemicLLMs (see Supplementary Section 6 for example genomic prompts).

The results indicate that PandemicLLM responds to the real-time genomic surveillance information. Figure 5c shows a comparison of two sets of distinct predictions: (1) predictions generated by PandemicLLM-3 and (2) predictions generated by PandemicLLM-3 without genomic surveillance information. Figure 5b highlights that the model's prediction confidence changes when variant data are included. The findings reveal that the integration of variant information enhances both the performance and confidence of the models, particularly in the case of PandemicLLM-3. Specifically, introducing variant information in PandemicLLM-3 leads to an average increase in prediction confidence of 20.1% and an average improvement of 28.2% in WMSE. This enhancement in performance is particularly evident during the transition of the dominant variant from BA.5 to the BQ.1 lineage.

Discussion

By reformulating pandemic forecasting as a text-reasoning problem, PandemicLLMs push the boundaries of traditional forecasting methods. Traditional disease forecasting models heavily depend on structured numerical data, overlooking the wealth of information hidden within diverse disease-relevant sources. For instance, public health policies and real-time reports on emerging variants offer crucial information in textual formats that traditional models cannot access. To address this need, our proposed model unlocks the full potential of relevant information for pandemic forecasting, by reformulating it as a text-reasoning problem and adapting LLMs. This approach inclusively converts all information into text, allowing PandemicLLM to process and reason from data inaccessible to traditional frameworks. Our information contribution analysis (Supplementary Section 13) suggests that the relative importance of different data streams varies across the different phases of an outbreak, which aligns with existing literature[3]. For instance, epidemiological time-series information (through ChatGPT embedding) is most impactful during the decreasing and stable phases, while policy, demographic and political information are more influential during the increasing phases. We aim to enhance our model by integrating a broader range of disease-relevant data, including local comorbidities, wastewater epidemiology and human behavior, to improve predictive accuracy. In addition, we will incorporate a variable to model population-level waning immunity, accounting for the decline in both vaccine-induced and infection-induced immunity over time.

For a model to effectively support public health decision-making, it needs to be reliable, trustworthy and accurate[15,17]. Moreover, pandemic forecasts are inherently uncertain, and effectively communicating forecast uncertainty to decision-makers and the public is critical for transparency, yet remains challenging[22,31]. In consideration of these needs, PandemicLLM was fine-tuned to predict future pandemic trends by generating probabilities within defined categories, where the probability of the predicted category indicates the model's confidence in its predictions. As revealed through the 'Trustworthy and robust results' section, the performance of PandemicLLM improves consistently as confidence increases. This observation is consistent with the capabilities of Flan-PaLM 540B[15], an LLM with encoded clinical knowledge, further emphasizing the proficiency of PandemicLLM in representing uncertainties with the COVID-19-related knowledge. Consequently, the confidence level can effectively function as an indicator of prediction reliability, offering model users a definitive guide to gauge the trustworthiness of the forecasts.

The sustained performance without retraining and the ability to adapt to emerging variants demonstrate the potential to generalize our framework to other diseases with similar transmission mechanisms and sufficient historical data (for example, influenza and respiratory syncytial virus) and adapt to emergency public health scenarios requiring rapid decision-making. The success of PandemicLLM in capturing disease dynamics at the state level offers a promising foundation for extending its forecasts to more localized levels, such as counties or even hospitals. Our preliminary exploration into county-level predictions supports the potential for localized forecasting (Supplementary Section 15). Initial results using zero-shot inference on ten populous US counties showed promising accuracy rates of 42% at 1 week and 41% at 3 weeks. However, the lack of granular data limited performance, particularly during the rising spread of disease (Supplementary Fig. 43). Future efforts will focus on integrating more detailed local data and fine-tuning the model to improve county-level forecasts and better support local decision-making. As we look toward the future, the insights and methodologies developed through PandemicLLM offer a glimpse of what might be possible for the next generation of public health forecasting models.

This study has several limitations. First, while the proposed framework demonstrates the effectiveness of LLM-based categorical forecasting for pandemics over a short-term horizon of 1 to 3 weeks, it does not currently support precise numerical hospitalization predictions or long-term projections (beyond 1 month). Another potential limiting factor of the proposed model is the computational cost of employing LLMs. In efforts to address this concern, we empirically show that freezing the pre-trained LLaMA2 parameters achieves much better efficiency without sacrificing performance. However, PandemicLLM still demands a substantial amount of computational resources as the gradients of the LLM parameters are involved in the optimization of trainable RNNs and input token embeddings. This could be a limiting factor in scenarios where resources are scarce or in situations demanding rapid model deployment. Moreover, the PandemicLLM framework is inherently data driven, necessitating a sufficient amount of training data for effective fine-tuning. Further research is needed to explore how LLMs can be adapted to handle new or emerging diseases with limited or no historical data. Finally, LLMs lack theoretical transparency. To address this issue, we explicitly evaluate the empirical reliability and trustworthiness of PandemicLLMs. However, there remains a need to enhance the interpretability of LLMs. This is particularly important for public health-related applications of LLMs, where elucidating the reasoning behind model predictions is pivotal for fostering user confidence and trust.

Methods

PandemicLLM datasets

The proposed PandemicLLM is fine-tuned using multiple disparate categories of data including spatial, epidemiological time-series, genomic surveillance and public health policy data. Our data covers all 50 states in the United States, ensuring a comprehensive nationwide scope for our study. All the spatial data are available at the state resolution and all the time-varying data are available at the weekly resolution. In this section, we comprehensively discuss the sources of data and the preprocessing process.

Epidemiological time-series data. *Hospitalizations*. Our study focuses on refining the targets used for COVID-19 forecasting models. The choice of hospitalization data over cases and deaths data stems from its capability to reflect the disease's spread and overall harmful impacts on healthcare systems, making it a more comprehensive measure.

Shifting from the conventional reliance on numerical targets, our study embraces categorical targets due to two main considerations. First, the unreliability of data reporting poses a challenge[32]. The inconsistencies in reported cases, deaths and hospitalizations undermine the effectiveness of traditional disease forecasting models that rely heavily on these metrics. Second, the experience with the COVID-19 pandemic has demonstrated that the performance of numerical target predictions often falls short of expectations[8,10].

In light of these limitations, we advocate for the use of HTCs as our predictive target instead of relying solely on numerical values. These categories, crafted from weekly COVID-19 hospitalization time series, offer a more robust indicator of both the disease spread and its impact on healthcare systems. We categorized the hospitalization numbers into five distinct trends as follows:

$$\mathrm{HTC}_1^{i,t} = \begin{cases} \text{Substantial increase} & \text{if } \mathrm{HT}_1^{i,t} > 3 \\ \text{Moderate increase} & \text{if } 3 > \mathrm{HT}_1^{i,t} > 1 \\ \text{Stable} & \text{if } 1 > \mathrm{HT}_1 i^{i,t} > -1 \\ \text{Moderate decrease} & \text{if } -1 > \mathrm{HT}_1^{i,t} > -3 \\ \text{Substantial decrease} & \text{if } -3 > \mathrm{HT}_1^{i,t}, \end{cases} \quad (1)$$

$$\mathrm{HTC}_3^{i,t} = \begin{cases} \text{Substantial increase} & \text{if } \mathrm{HT}_3^{i,t} > 4.5 \\ \text{Moderate increase} & \text{if } 4.5 > \mathrm{HT}_3^{i,t} > 1.5 \\ \text{Stable} & \text{if } 1.5 > \mathrm{HT}_3^{i,t} > -1.5 \\ \text{Moderate decrease} & \text{if } -1.5 > \mathrm{HT}_3^{i,t} > -4.5 \\ \text{Substantial decrease} & \text{if } -4.5 > \mathrm{HT}_3^{i,t}, \end{cases} \quad (2)$$

where $\mathrm{HTC}_1^{i,t}$ and $\mathrm{HTC}_3^{i,t}$ represents the 1-week- and 3-weeks-later HTC for state i at week t. The 1-week and 3-weeks hospitalization trends (HTs) for state i at week t ($\mathrm{HT}_1^{i,t}$ and $\mathrm{HT}_3^{i,t}$) are defined as follow:

$$\mathrm{HT}_1^{i,t} = \mathrm{HR}^{i,t} - \overline{\mathrm{HR}}^{i,t-1}, \quad (3)$$

$$\mathrm{HT}_3^{i,t} = \mathrm{HR}^{i,t} - \overline{\mathrm{HR}}^{i,t-3}, \quad (4)$$

where $\mathrm{HR}^{i,t}$ represents the hospitalization rate per 100,000 people for state i at week t. $\overline{\mathrm{HR}}^{i,t}$ indicates the smoothed 3-week average hospitalization rate per 100,000 people for state i at week t. The decision to define HT by contrasting current reported hospitalization values with lagged, smoothed data was to diminish the influence of data irregularities and reporting errors. This approach is intended to yield a robust and more representative target, thereby providing a more reliable reflection of the actual progression of the pandemic.

While the HTC is utilized as the target, HT and the HR are employed as input epidemiological time-series data. The state-level COVID-19 hospitalizations data were collected from the US Department of Health and Human Services[33].

Reported cases. Our study utilized the state-level, daily reported COVID-19 case from the Johns Hopkins University Center for Systems Science and Engineering[34]. The raw case data are aggregated at the weekly level as input data C_i^t, where C_i^t represents the number of reported cases for state i at week t.

Previous infections. Numerous research findings have highlighted the protective role of previous infections in guarding against reinfection and severe outcomes of COVID-19[35]. To incorporate the impact of immunity acquired from previous infections in mitigating severe disease upon subsequent infections, we created a variable that represents the total number of reported infections over a 3-month period. The mathematical formulation of previous infection (PI) is defined as follows:

$$\mathrm{PI}_i^t = \frac{\sum_{j=t-16}^{t-4} C_i^j}{\mathrm{pop}_i}, \tag{5}$$

where PI_i^t represents the cumulative reported infection rate for state i from 16 to 4 weeks before to week t, C_i^j is the number of reported cases for state i at week j, and pop_i is the population for state i. The numerator represents the aggregate of C_i^j from the period spanning weeks $t-16$ to $t-4$ preceding week t.

Vaccination. The vaccine-induced immunity is widely regarded as a crucial approach to reducing the harmful impact of COVID-19. Our model incorporated state-level vaccination data from the US CDC Vaccine Tracker[36], leveraging this information to enhance our analysis. We included state-level cumulative vaccination rates normalized by population. The vaccination data cover three types of time series: the partial vaccination rate, the completed primary series rate and the booster vaccination rate. Our approach offers a detailed view of the vaccination landscape across different states by including these distinct metrics.

Spatial data. *Demographics*. Data on the state population, including age group and race-related data, was sourced from the American Community Survey by the US Census Bureau[37]. Recognizing that COVID-19 affects different racial and age groups unevenly, as established by existing studies[38], we specifically focused on the population aged over 65 and vulnerable racial groups. The demographic information utilized in our research is derived from the 2022 National Census Survey.

Healthcare systems. In our study, we incorporated healthcare system scores from the Commonwealth Fund[39], an annually updated and comprehensive dataset that evaluates the performance of healthcare systems in every US state. This score on State Health System Performance offers a unique perspective on how state healthcare systems have coped with and managed the challenges posed by the COVID-19 pandemic. For each state, this resource provides not only an overall healthcare system performance ranking but also offers detailed rankings in specific categories such as COVID-19 response, access and affordability, prevention and treatment, avoidable hospital use and cost, healthy lives, income disparity, and racial and ethnic equity.

2020 presidential election results. Previous research has identified a notable association between political affiliation and COVID-19 health outcomes at the state level[40]. In light of this, we collected data for the 2020 US Presidential election results from the Federal Elections Commission[41]. On the basis of the state-level voting results, we labeled each state as either Democrat or Republican.

Public health policy data. The impact of the COVID-19 epidemic is influenced by the stringency and timing of government-implemented policies[42]. These policies include measures such as closing schools, canceling public events and protecting the elderly vulnerable populations. We collected policy data from the Oxford COVID-19 Government Response Tracker[43]. However, in contrast to the majority of studies that utilize the generated policy index, we integrated the textual descriptions of selected policies as input for our prompt design, providing a unique approach to analyzing policy impact. To effectively assess the impact of policies on the spread of the disease and the strain on healthcare systems, we specifically chose policies from two categories: 'C' for containment and closure policies and 'H' for health system policies. The complete list of policies included in our study from these categories is provided in Supplementary Fig. 1.

Our prompt utilized the summary of the stringency levels for each policy as input. These summaries are derived directly from the original descriptions provided by the data source. Owing to restrictions in token size, we have limited our input to the specified policy categories despite the potential impact of other policies on the pandemic's dynamics.

Genomic surveillance data. In the constantly changing COVID-19 pandemic, the emergence of new variants has a pivotal role in shaping pandemic trends and potential new waves. Genomic surveillance data are indispensable in this context, as they enable the timely detection and tracking of these variants, providing essential insights into their characteristics and spread[44]. Existing forecasting models that attempt to incorporate genomic surveillance data frequently encounter limitations owing to reporting delays and the challenge of accurately encoding the virological characteristics of different variants[3]. PandemicLLM offers an opportunity to include this data more effectively. Through zero-shot inference, LLMs can process and interpret previously unseen variant data without training, providing a fresh and timely perspective in dealing with the emergence of new variants. In our study, we utilized two primary types of genomic surveillance data to forecast the COVID-19 HTC:

- Virological characteristics: our study integrated official updates from authoritative organizations, including the World Health Organization, the CDC, and the European Centre for Disease Prevention and Control. These updates, presented in text format, offer information on the virological characteristics of newly emerging variants. We then summarized the information from three aspects: (1) infectiousness, (2) severity and (3) resistance to immunity. More detailed data sources are documented in Supplementary Section 3.
- Weighted estimates of variant proportions: we also included the weighted estimates of variant proportions from the CDC[30] in our analysis. Weighted estimates refer to the proportions of circulating variants derived from empirical genomic sequencing data that has been observed and recorded. These data are vital for assessing the prevalence and distribution of these variants, offering an overview of the pandemic's status and aiding in more accurate forecasting.

Preliminaries and notations

Large language models. In recent years, LLMs have catalyzed a pivotal transformation in the field of natural language processing. These models are transformer decoders[45] autoregressively pre-trained on extensive text corpus such as GitHub, Wikipedia, CommonCrawl and BooksCorpus. Consequently, LLMs enjoy a broad knowledge base and demonstrate proficiency across a spectrum of tasks, ranging from conversational agents[46] to complex reasoning[47] and decision-making applications[48]. Their performance often rivals, and occasionally surpasses, that of humans in these domains[46]. Despite the remarkable performance of LLMs, the state-of-the-art LLMs are often closed-source LLMs, for example, ChatGPT and GPT-4[46], hampering the development and application of LLMs for the research community. To bridge this gap, LLaMA and LLaMA2[49] have been proposed. They are open-source LLMs trained on publicly available datasets and have achieved comparable or better performance than GPT-3. In this study, we used LLaMA2 as the backbone for its robust performance in language understanding.

Autoregressive language modeling. Most existing LLMs are trained in an autoregressive manner, where the model predicts the next token in a text sequence based on its predecessors. This process begins with tokenization, for example, byte pair encoding[50], where a tokenizer segments a string into discrete tokens. Then the tokenized sequence for data sample i, denoted as T_i is reconstructed autoregressively:

$$p_\theta(T_i) = \prod_{j=1}^{|T_i|} p_\theta(t_j^{(i)} | t_1^{(i)}, \cdots, t_{j-1}^{(i)}), \tag{6}$$

where θ is the set of trainable parameters, p_θ is the LLM model, $t_j^{(i)}$ denotes the jth token in T_i. By maximizing the likelihood $p_\theta(T) = \prod_{i=1}^{N} p_\theta(T_i)$ for all data samples $T_i \in T$, the LLM's parameters are learned.

Challenges of LLMs in encoding disease-related data. Despite the remarkable abilities of LLMs, pandemic forecasting integrates various data sources, including demographic profiles, public health policy and epidemiological time series. These diverse sources constitute a composite, multi-modal input for the pandemic prediction model. However, conventional LLMs primarily handle discrete text inputs, which can not directly leverage these multi-modality data. Besides, for continuous sequential data, tokenizers such as byte pair encoding segment continuous sequences into discrete tokens, breaking numbers into awkward chunks that make learning basic numerical operations challenging[20]. To leverage the cross-domain knowledge of LLMs while preserving sequential information effectively, it is essential to innovate how multi-modal pandemic data are assimilated as input for LLMs.

Notations. Before we formally start to introduce the PandemicLLM, we establish the following notations.

- T_i denotes the prompt text of the ith item, including task description, spatial, epidemiological time-series, genomic surveillance and public health policy information. An example can be found in Fig. 2.
- X_i denotes the time series data of the ith item in the dataset, for example, hospitalization time-series data.
- y_i denotes the token for the HTC of the ith item in the dataset, representing one of the following trends {<Substantial decrease>, <Moderate decrease>, <Stable>, <Moderate increase>, <Substantial increase>}.
- f_θ denotes an LLM with parameters θ. θ_{in} and θ_{out} denote the parameters for the input and output layer.
- g_ϕ denotes the RNN encoder with parameters ϕ.
- $\mathbf{H}_i$ denotes the embedding matrix of text tokens T_i that serves as the input for the transformer layers.
- $\mathbf{z}_i$ denotes the embedding vector of time series data X_i.

Methodology

Overview of PandemicLLM. We propose to formalize the pandemic prediction problem as a multi-modality ordinal classification problem: For a sample i (for example, New York state at a given week), given the prompt with multi-modality text information, denoted as T_i, and the sequential data X_i, the goal is to predict the HTC y_i. Specifically, as illustrated in Supplementary Section 2, PandemicLLM models $p(y_i|T_i, X_i)$ using two models: an RNN-based sequential encoder g_ϕ that projects sequential data to the representations in the text space; and a transformer-based LLM f_θ that forecasts the categorical targets distribution using the encoded text representations containing multi-modality information and the encoded time series information. In this way, PandemicLLM seamlessly integrates text reasoning and numerical time-series learning using one unified framework.

Multi-modality data textualization. PandemicLLM utilizes an LLM to address the challenge of processing multi-modal disease-relevant data. For each item in the dataset, we construct a composite prompt to effectively parse and integrate the multi-modality data. Further details of the textualization process are shown in Supplementary Sections 3–5, and examples are shown in Supplementary Section 6:

- Spatial data textualization: in spatial data textualization, we processed numerical state demographic data and healthcare system statistics. Those data for all 50 US states were numerically ranked and subsequently transformed into categorical descriptions reflecting their relative positions. States ranking in the top 5 were labeled 'One of the best'. Those between 6th and 20th were categorized as 'Higher than the national average', while rankings from 21st to 30th were described as 'Close to the national average'. States falling between 31st and 45th were labeled 'Lower than the national average', and those in the bottom 5 were described as 'One of the lowest'. Regarding presidential election outcomes, each state was characterized based on its voting percentage, being labeled as either predominantly voting for Democrats or Republicans.
- Epidemiological time series data textualization: in the phase of textualizing epidemiological time series data, an AI model (ChatGPT-3.5[46]) is employed to convert sequential data into narrative summaries. This method leverages the sophisticated capabilities of AI to transform numerical sequences into detailed textual narratives, thereby reducing the need for laborious manual annotation. For instance, the AI-facilitated textualization process for the weekly vaccination rate time series is illustrated in Supplementary Section 4.
- Epidemiological time-series representation learning: as time series plays a critical role in pandemic forecasting, we leverage an additional RNN encoder to effectively distill the useful information into LLM's input space. Specifically, the prompt for time-series data is initialized by a special token <time-series-special-token> whose embedding $\mathbf{z}_i$ is encoded by an RNN encoder g_ϕ:

$$\mathbf{z}_i = g_\phi(X_i) \in \mathbb{R}^d, \tag{7}$$

- where X_i denotes the time-series data, for example, hospitalization time series, for data sample i. As shown in Supplementary Table 2, the RNN encoder can improve the model's performance. We choose GRU as the implementation of g_ϕ in our experiments.
- Policy textualization: our study incorporated six policy categories: school and workplace closures, public events, gathering restrictions, facial coverings, and elderly protection. We designed the prompt's weekly policy stringency level descriptions, highlighting policy shifts (Supplementary Section 5). For instance, if New York shifted its school closure policy from 'recommended closures' to 'No measures' while maintaining 'No restriction' on gatherings, the description would be: 'There have been changes in school policy moving from recommended closures to no restrictions while gathering policy remains unrestricted'.
- Genomic surveillance data textualization: the genomic surveillance data consisted of two main categories: reports about new variants from authoritative sources and sequential data derived from the CDC's weighted estimates of variant proportions. Key information, which includes the variant's relative transmissibility, severity and the impact on immunity, was incorporated directly into the prompt design. The weighted variant proportion estimates were processed using the same method outlined in the temporal data textualization section (for further details, refer to Supplementary Sections 3 and 5).

Now we describe how we organize the multi-modality data into the input space of an LLM: we first provide the essential task information at the beginning of the prompt, then we use expert knowledge combined with AI assistance to describe the spatial and temporal information for pandemic forecasting. Then the sequential data are followed, which will be further encoded by the RNN encoder. The input for the LLM transformer decoder is a mixture of text information and encoded sequential information.

Specifically, for data sample i, the tokenized text T_i, containing both spatial and temporal information, is encoded into the text embeddings $\mathbf{H}_i$:

$$\mathbf{H}_i = f_{\theta_{in}}(T_i) \in \mathbb{R}^{|T_i| \times d}, \tag{8}$$

where $f_{\theta_{in}}$ is the input embedding layer for LLM, $|T_i|$ denotes the number of tokens for T_i, and d is the embedding dimension of the LLM. Then the encoded information $\mathbf{z}_i$ is used to replace the embedding of the time-series special token, that is <time-series-special-token>, in $\mathbf{H}_i$ to generate the final representation for the transformer's input $\mathbf{H}'$:

$$\mathbf{H}'[j,:] = \begin{cases} \mathbf{H}_i[j,:] & \text{if } j \neq s_i, \\ \mathbf{z}_i & \text{if } j = s_i, \end{cases} \tag{9}$$

where s_i denotes the index for the time-series special token. In this way, PandemicLLM seamlessly fuses the sequential information and textual information into the input space of an LLM, enabling it to perform reasoning with both types of information.

LLM for pandemic prediction. With $\mathbf{H}'$ encoding both textual and sequential information, PandemicLLM leverages an LLM (that is, transformer decoder) to perform pandemic prediction as a text-generation problem. As introduced in 'Preliminaries and notations', an autoregressive LLM generates one token at a time conditioned on the previously generated tokens. To make a prediction, we extend the original vocabulary of LLM with 'class tokens': {<Substantial decrease>, <Moderate decrease>, <Stable>, <Moderate increase>, <Substantial increase>} that represent the trend for future hospitalization. Then we use the predicted distribution of 'class tokens' after the output prompt 'The answer is' as the target distribution for the pandemic forecasting. In this way, PandemicLLM formulates pandemic forecasting as a text-reasoning problem using an LLM.

The model is optimized by the autoregressive loss defined in equation (6), maximizing the likelihood of the ground-truth text. For better scalability and efficiency, we freeze the transformer parameters for LLMs[51] and train only the vocabulary embeddings $f_{\theta_{in}}$ and output prediction layer $f_{\theta_{out}}$, as well as the GRU encoders g_ϕ. The detailed training and testing time is documented in Supplementary Section 8.

Evaluation. We comprehensively evaluate our model using five error metrics: (1) accuracy, (2) MSE, (3) WMSE, (4) Brier score and (5) RPS.

Accuracy offers a direct method to assess the performance of a model's classification, defined by the following equation:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N}[y_i = \hat{y}_i]}{N}, \tag{10}$$

where y_i represents the actual class, $\hat{y}_i$ denotes the predicted class and N signifies the total number of samples. However, despite being widely used in traditional classification problems, it equally treats all errors and neglects the inherent ordering between the HTC classes. For example, if the ground-truth trend is 'Substantial increase', accuracy will treat the prediction error of 'Moderate increase' and 'Substantial decrease' equally.

To model the order information of HTC classes, we map the classes {substantial decrease, moderate decreasing, stable, moderate increasing, substantial increase} into a numeric scale {1, 2, 3, 4, 5}, and use the mapped value to compute numerical error metrics. To start with, the MSE is a promising way to evaluate the ordinal classification, which is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i - \bar{y}_i)^2, \tag{11}$$

where $\bar{y}_i$ and $\bar{y}_i$ denote the numerical values of the ground-truth class and the predicted class. To further evaluate the predicted distribution, we use WMSE, which introduces a probability weighting into the MSE, defined as:

$$\text{WMSE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} P(\bar{y}_i = k)(k - \hat{y}_i)^2, \tag{12}$$

where K denotes the number of classes ($K = 5$ for HTC prediction) and $P(\bar{y}_i = k)$ denotes the predicted probability for the kth class for the ith data sample.

The Brier score is used to measure the accuracy of probabilistic predictions. It is calculated as the mean-squared difference between the predicted probability assigned to the possible outcomes and the actual outcome:

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(P(\bar{y}_i = k) - o_k^{(i)}\right)^2, \tag{13}$$

where $P(\bar{y}_i = k)$ is the forecast probability for the ith item in the kth class, and $o_k^{(i)}$ is the one-hot encoded actual outcome, with a value of 1 for the ground-truth class and 0 for all other classes.

We also utilize the RPS[25], which is particularly useful for evaluating the accuracy of the entire predicted probability distribution across categorical outcomes. The RPS is defined as:

$$\text{RPS} = \frac{1}{N-1}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(q(\bar{y}_i = k) - \bar{q}_k^{(i)}\right)^2, \tag{14}$$

where $q_k^{(i)}$ and $\bar{q}_k^{(i)}$ are the ground-truth and predicted cumulative probability distributions for the ith case in the kth class. The ground-truth cumulative probability is defined as a step function that increases from 0 to 1 at the class of the actual observed outcome. The predicted cumulative probability distribution is defined as the sum of the predicted probabilities for all classes up to and including the kth class.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data utilized in this study derive from publicly accessible sources. Hospitalization data were collected from the US Department of Health and Human Services[33] and reported case data were sourced from the Johns Hopkins University CSSE COVID-19 Dashboard[34]. Vaccination data were obtained from the US CDC Vaccine Tracker[36]. For spatial data, demographic information was sourced from the US Census Bureau[37], healthcare system data from the Commonwealth Fund[39], and presidential election results from the Federal Elections Commission[41]. Public health policy data were acquired from the Oxford COVID-19 Government Response Tracker[43]. Official reports on variants were collected from the World Health Organization (WHO) and the European Centre for Disease Prevention and Control (ECDC), with estimated variant proportions sourced from the CDC[30]. The data are available at an archived repository[52] and at https://github.com/miemieyanga/PandemicLLM. Source data for Figs. 3–5 and Extended Data Fig. 1 are provided with this paper.

Code availability

All codes are written using Python 3.11.5. Codes are publicly accessible at an archived repository[52] and at https://github.com/miemieyanga/PandemicLLM.

References

1. Giordano, G. et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* **26**, 855–860 (2020).
2. Li, X. et al. Wastewater-based epidemiology predicts COVID-19-induced weekly new hospital admissions in over 150 USA counties. *Nat. Commun.* **14**, 4548 (2023).
3. Du, H. et al. Incorporating variant frequencies data into short-term forecasting for COVID-19 cases and deaths in the USA: a deep learning approach. *EBioMedicine* **89**, 104482 (2023).
4. Reich, N. G. et al. Collaborative hubs: making the most of predictive epidemic modeling. *Am. J. Public Health* **112**, 839–842 (2022).
5. Rosenfeld, R. & Tibshirani, R. J. Epidemic tracking and forecasting: lessons learned from a tumultuous year. *Proc. Natl Acad. Sci. USA* **118**, e2111456118 (2021).
6. Hsiang, S. et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**, 262–267 (2020).
7. Li, J., Lai, S., Gao, G. F. & Shi, W. The emergence, genomic diversity and global spread of SARS-CoV-2. *Nature* **600**, 408–418 (2021).
8. Nixon, K. et al. An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation. *Lancet Digit. Health* **4**, e738–e747 (2022).
9. Castro, M., Ares, S., Cuesta, J. A. & Manrubia, S. The turning point and end of an expanding epidemic cannot be precisely forecast. *Proc. Natl Acad. Sci. USA* **117**, 26190–26196 (2020).
10. Ioannidis, J. P., Cripps, S. & Tanner, M. A. Forecasting for COVID-19 has failed. *Int. J. Forecast.* **38**, 423–438 (2022).
11. Telenti, A. et al. After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* **596**, 495–504 (2021).
12. Nepomuceno, M. R. et al. Besides population age structure, health and other demographic factors can contribute to understanding the COVID-19 burden. *Proc. Natl Acad. Sci. USA* **117**, 13881–13883 (2020).
13. Ruggeri, K. et al. A synthesis of evidence for policy from behavioural science during COVID-19. *Nature* **625**, 134–147 (2023).
14. Searls, D. B. The language of genes. *Nature* **420**, 211–217 (2002).
15. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
16. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
17. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
18. Bzdok, D. et al. Data science opportunities of large language models for neuroscience and biomedicine. *Neuron* **112**, 698–717 (2024).
19. Williams, R., Hosseinichimeh, N., Majumdar, A. & Ghaffarzadegan, N. Epidemic modeling with generative agents. Preprint at https://arxiv.org/abs/2307.04986 (2023).
20. Gruver, N., Finzi, M., Qiu, S. & Wilson, A. G. Large language models are zero-shot time series forecasters. *Adv. Neural Inf. Process. Syst.* **36**, 19622–19635 (2023).
21. Covid Data Tracker. *CDC* https://covid.cdc.gov/covid-data-tracker/#maps_new-admissions-percent-change-state (2024).
22. Nixon, K. et al. Real-time COVID-19 forecasting: challenges and opportunities of model performance and translation. *Lancet Digit. Health* **4**, e699–e701 (2022).
23. Touvron, H. et al. LLaMA 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
24. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **63**, 938–939 (2010).
25. Leung, K., Wu, J. T. & Leung, G. M. Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing. *Nat. Commun.* **12**, 1501 (2021).
26. Cramer, E. Y. et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl Acad. Sci. USA* **119**, e2113561119 (2022).
27. Lopez, V. K. et al. Challenges of COVID-19 case forecasting in the US, 2020–2021. *PLoS Comput. Biol.* **20**, e1011200 (2024).
28. Kalia, K., Saberwal, G. & Sharma, G. The lag in SARS-CoV-2 genome submissions to GISAID. *Nat. Biotechnol.* **39**, 1058–1060 (2021).
29. TAG-VE statement on Omicron sublineages BQ.1 and XBB. *WHO* https://www.who.int/news/item/27-10-2022-tag-ve-statement-on-omicron-sublineages-bq.1-and-xbb (2024).
30. Ma, K. C. Genomic surveillance for SARS-CoV-2 variants: circulation of Omicron lineages—United States, January 2022–May 2023. *MMWR Morb. Mortal. Wkly Rep.* **72**, 651–656 (2023).
31. Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B. & Sledge, D. The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl Acad. Sci. USA* **117**, 16732–16738 (2020).
32. Dong, E. et al. The Johns Hopkins University Center for Systems Science and Engineering COVID-19 dashboard: data collection process, challenges faced, and lessons learned. *Lancet Infect. Dis.* **22**, e370–e376 (2022).
33. COVID-19 reported patient impact and hospital capacity by facility. *Department of Health and Human Services* https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u/about_data (2024).
34. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
35. Du, H., Saiyed, S. & Gardner, L. M. Association between vaccination rates and COVID-19 health outcomes in the United States: a population-level statistical analysis. *BMC Public Health* **24**, 220 (2024).
36. COVID Data Tracker. *CDC* https://covid.cdc.gov/covid-data-tracker/#vaccine-delivery-coverage (2024).
37. State population totals and components of change: 2020–2023. *US Census Bureau* https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html#v2022 (2024).
38. Dowd, J. B. et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc. Natl Acad. Sci. USA* **117**, 9696–9698 (2020).
39. Radley, D., Collins, S. & Hayes, S. *The Commonwealth Fund 2019 Scorecard on State Health System Performance* (The Commonwealth Fund, 2022).
40. Bollyky, T. J. et al. Assessing COVID-19 pandemic policies and behaviours and their economic and educational trade-offs across US states from Jan 1, 2020, to July 31, 2022: an observational analysis. *Lancet* **401**, 1341–1360 (2023).
41. Federal Elections 2020. *FEC* https://www.fec.gov/introduction-campaign-finance/election-results-and-voting-information/federal-elections-2020/ (2024).
42. Haug, N. et al. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **4**, 1303–1312 (2020).
43. Hale, T. et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5**, 529–538 (2021).
44. Stockdale, J. E., Liu, P. & Colijn, C. The potential of genomics for infectious disease forecasting. *Nat. Microbiol.* **7**, 1736–1743 (2022).

45. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
46. OpenAI et al. *GPT-4 Technical Report*. Preprint at https://arxiv.org/abs/2303.08774 (2024).
47. Fu, Y., Peng, H., Sabharwal, A., Clark, P. & Khot, T. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations* (2023).
48. Liang, J. et al. Code as policies: language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* 9493–9500 (IEEE, 2023).
49. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
50. Gage, P. A new algorithm for data compression. *C Users J.* **12**, 23–38 (1994).
51. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, 34892–34916 (2023).
52. Du, H. & Zhao, Y. Advancing real-time infectious disease forecasting using large language models. *Zenodo* https://doi.org/10.5281/zenodo.14788491 (2025).
53. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

Acknowledgements

This work was supported by NSF Award ID 2229996 (L.M.G.), cooperative agreement CDC-RFA-FT-23-0069 (L.M.G.), NOA: 6 NU38FT000012-01, from the CDC Center for Forecasting and Outbreak Analytics (L.M.G.), Merck KGaA Future Insight Prize (L.M.G.), NSF Award ID 2112562 (Y.C.) and ARO W911NF-23-2-0224 (Y.C.). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies.

Author contributions

H.D., Y.Z., J.Z. and H.F.Y. conceptualized and designed the study. H.D. and S.X. collected the data. H.D. processed the data and designed prompts. H.D., J.Z. and Y.Z. performed the experiments. S.X. ran the baseline models. H.D., J.Z., Y.Z., S.X. and H.F.Y. prepared the figures. H.D., J.Z., Y.Z. and H.F.Y analyzed the results. H.D., J.Z., Y.Z. and H.F.Y. wrote the initial draft. L.M.G., Y.C., X.L. and H.F.Y. provided guidance and feedback for the study. L.M.G., X.L. and H.F.Y. revised the paper. L.M.G. and Y.C. acquired the funding. Y.C. provided computational resources. All authors prepared the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43588-025-00798-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00798-6.

Correspondence and requests for materials should be addressed to Yiran Chen, Lauren M. Gardner or Hao 'Frank' Yang.

Peer review information *Nature Computational Science* thanks Rafael de Andrade Moral and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team.

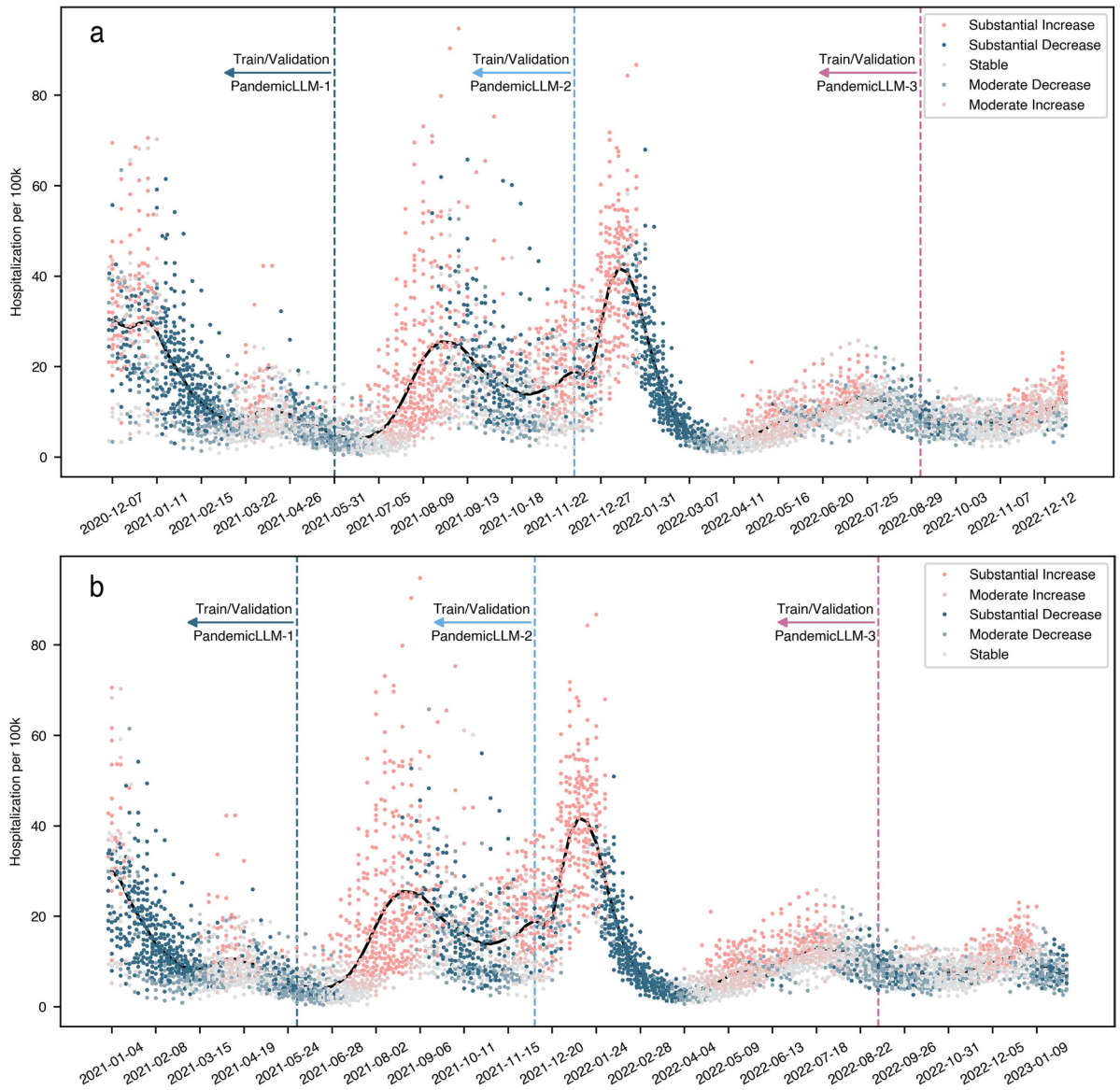
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2025

https://doi.org/10.1038/s43588-025-00798-6



Extended Data Fig. 1 | Overview of 1 and 3-week Targets Design and Experimental Setups for the PandemicLLMs. The black line shows the average COVID-19 hospitalization rate per 100,000 people in the U.S. plotted over time. Each data point (dot) represents the hospitalization rate for a specific state at a particular week. The color of each dot corresponds to the state's resulting HTC category for that week. Three PandemicLLMs were developed, each trained and validated on data ending in June 2021 (PandemicLLM-1), December 2021 (PandemicLLM-2), and September 2022 (PandemicLLM-3), respectively. All models were subsequently evaluated on a dataset spanning from their respective end dates to February 2023, without retraining. Panel (**a**) visualizes 1-week targets, while panel (**b**) shows 3-week targets. Dates are in the format year-month-day.

nature portfolio

Corresponding author(s): Hao Frank Yang

Last updated by author(s): Jan 29, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All the code used to collect the data and generate prompts are available at: https:/ /github.com/miemieyanga/PandemicLLM.
Data analysis	All codes are written in Python 3.11.5. All the code used to analysis the data and perform the experiments are available at: https:// github.com/miemieyanga/PandemicLLM.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data utilized in this study derive from publicly accessible sources. Hospitalization data were collected from the U.S. Department of Health & Human Services (HHS) (https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u/about_data), , while reported case data were sourced from the Johns Hopkins University CSSE COVID-19 Dashboard (https://github.com/CSSEGISandData/COVID-19). Vaccination data were obtained from the U.S. CDC

Vaccine Tracker (https://covid.cdc.gov/covid-data-tracker/#vaccine-delivery-coverage). For spatial data, demographic information was sourced from the U.S. Census Bureau (https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html#v2022), healthcare system data from the Commonwealth Fund (https://2019scorecard.commonwealthfund.org/), and presidential election results from the Federal Elections Commission (https://www.fec.gov/introduction-campaign-finance/election-results-and-voting-information/federal-elections-2020/). Public policy data from the Oxford COVID-19 Government Response Tracker (https://github.com/OxCGRT/covid-policy-tracker). Official reports on variants were collected from WHO (https://www.who.int/news/item/27-10-2022-tag-ve-statement-on-omicron-sublineages-bq.1-and-xbb) and ECDC (https://www.ecdc.europa.eu/en/news-events/epidemiological-update-sars-cov-2-omicron-sub-lineages-ba4-and-ba5), with estimated variant proportions sourced from the CDC (https://data.cdc.gov/Laboratory-Surveillance/SARS-CoV-2-Variant-Proportions/jr58-6ysp/about_data).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study does not include any human data.
Reporting on race, ethnicity, or other socially relevant groupings	This study does not include any human data.
Population characteristics	This study does not include any human data.
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal sample size calculation was performed. The sample size (n = 50) was determined based on the number of U.S. states, as each state represents an independent unit of study. This choice ensures comprehensive geographic coverage and accounts for variations in pandemic forecasting performance across diverse populations and regions. Given that the analysis includes all available states, the sample size is inherently representative and sufficient for evaluating model performance across different locations.
Data exclusions	No data exclusions were performed. All available data points from the 50 U.S. states were included in the analysis.
Replication	Replication in this study is based on independent samples from 50 U.S. states, each serving as a distinct unit of analysis. No technical replicates were used, and all results are derived from independent state-level predictions over multiple weeks.
Randomization	The dataset was partitioned based on three temporal cutoffs: June 2021, December 2021, and September 2022. Data preceding these dates underwent random allocation into training (80%) and validation (20%) subsets. All data subsequent to these dates was designated for testing.
Blinding	No experimental groups are used in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A