Supplementary Information for: Advancing Real-time Pandemics Forecasting Using Large Language Models: A COVID-19 Case Study

Hongru Du[1, 2+], **Jianan Zhao**[3, 4+], **Yang Zhao**[1, 2+], **Shaochong Xu**[1, 2], **Xihong Lin**[5, 6], **Yiran Chen**[7*], **Lauren M. Gardner**[1, 2, 8*], **and Hao (Frank) Yang**[1, 2, 7*]

[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, Montréal, QC, Canada.
[4]Department of Computer Science, Universite de Montréal, Montréal, QC, 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.
[+]The authors contributed equally.
[*]The corresponding authors information: yiran.chen@duke.edu, l.gardner@jhu.edu, haofrankyang@jhu.edu

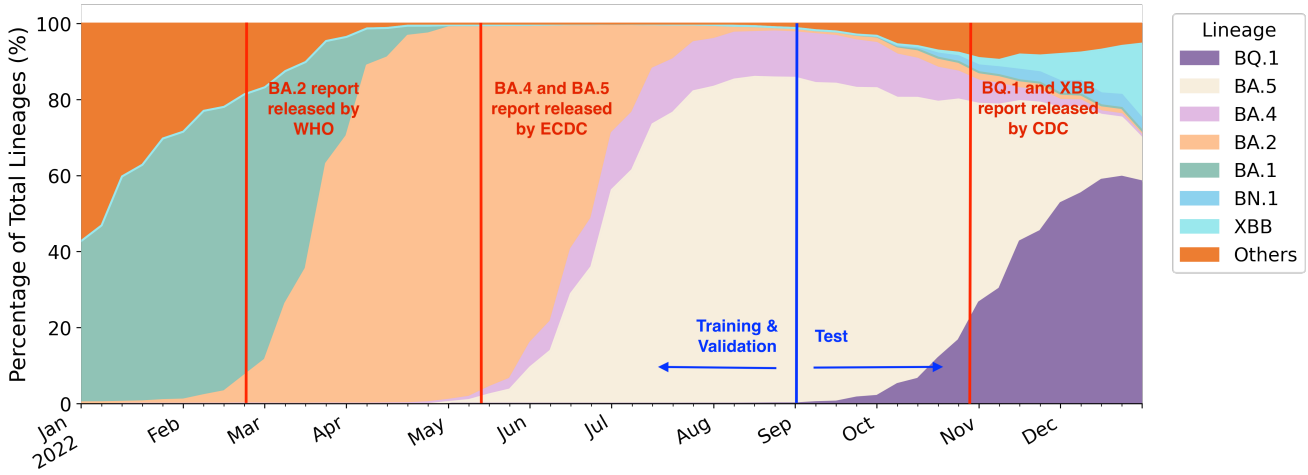
1 Genomic surveillance reports

Of particular interest in this study is the potential value of incorporating SARS-CoV-2 variant information in forecasting models with the PandmicLLMs framework. As elaborated in the main manuscript, our analysis incorporates textual summaries of the virological attributes of these variants. For comprehensive transparency, the detailed data sources and the release dates of genomic information are documented in the table below:

Supplementary table 1: Summary of genomic surveillance report

Variant	Release date	Source	Reference
BA.2	Feb 22, 2022	WHO	1
BA.4	May 13, 2022	ECDC	2
BA.5	May 13, 2022	ECDC	2
XBB	Oct 27, 2022	WHO	3
BQ.1	Oct 27, 2022	WHO	3

Supplementary figure 1 depicts a graphical representation of the chronological distribution of genomic data and variants allocated to the training and testing datasets.



Supplementary figure 1: National estimates of weekly proportions of SARS-CoV-2 variants from January, 2022 to January, 2023.

2 Human-based textualization

The integration of spatial, policy, and textual genomic surveillance information into prompts is achieved via human-based textualization.

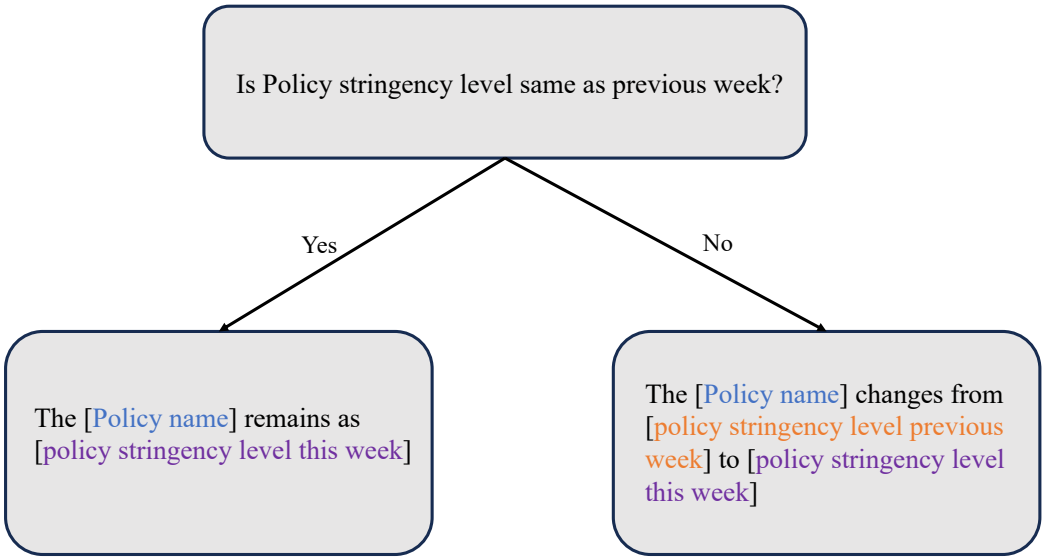
Spatial information: For the spatial data component, variables are evaluated and ranked across 50 states, creating categorical descriptions that reflect their relative standings. These descriptions are organized into five predefined ranking categories, as detailed subsequently:

> **Spatial data rank categories**
> **[Top 5]**: One of the best in country
> **[6^{th} to 20^{th}]**: Better than the national average
> **[21^{st} to 30^{th}]**: Close to the national average
> **[31^{st} to 45^{th}]**: Worse than the national average
> **[Buttom 5]**: One of the worst in country

Then, these defined categories are filled in the template to create a static prompt. Using California as an example:

> **Static prompt California**
> During the COVID-19 pandemic, overall healthcare systems performed **[Better than the national average]**, with **[Better than the national average]** in Access and Affordability, **Worse than the national average** in Prevention and Treatment, **[One of the best in country]** in population health conditions, **[Better than the national average]** in Income Disparity. California has **[Close to the national average]** ICU stress and **Worse than the national average** hospital staffing shortages.

Policy information: The textualization of policy data is performed at the weekly level for each state and policy, highlighting the difference in stringency levels across the week. This step is visualized in Supplementary figure 2.



Supplementary figure 2: Human-based policy data textualization.

Textual genomic information We summarized the virological characteristics of variants of interest as described in the referenced official genomic surveillance reports (Supplementary Information section 1). This summary includes the differences between these variants from previously circulating strains from three aspects: 1) Transmissibility, 2) Resistance to immunity, and 3) Severity. The detailed description for each variant is presented below:

Variants characteristics descriptions
[BA.2]: The new emerging variant is more transmissible than previous sublineages. No evidence of immunity escape. No reported difference in severity.
[BA.4]: The new emerging variant has high growth advantages over other sublineages. Increased immunity escape, may cause an overall increase in COVID-19 cases. No significant increase in infection severity.
[BA.5]: The new emerging variant has high growth advantages over other sublineages. Increased immunity escape, may cause an overall increase in COVID-19 cases. No significant increase in infection severity.
[XBB]: The new emerging variant has an increased impact on transmissibility. Increased impact on immunity escape. No evidence on impact on severity.
[BQ.1]: The new emerging variant has a significant growth advantage over other sublineages. No reported increase in immunity escape. No reported increase in disease severity.

3 Example prompts

Example of PandemicLLMs prompt for California:
[Task Information]: You are a helpful assistant designed to forecast epidemic trends for **California**. Your task is to predict the trend of hospitalization for the next week from the available options: [Substantial Decrease, Moderate Decrease, Stable, Moderate Increase, Substantial Increase]. You need to make prediction based on the information below:

% Spatial Information
California with one of the largest populations with close to average Black demographic, voted for Democratic in the recent Presidential election. During the pandemic, overall healthcare systems performed better than the national average, with better than average Access and Affordability, worse than Prevention and Treatment, one of the best in population health conditions, and better than the national average in Income Disparity. California has close to average ICU stress and worse than the national average hospital staffing shortages.

% Epidemiological Time Series
To date, 82% of the population got at least one vaccine dose with a Slight Increase trend, 70% were fully vaccinated with a Slight Increase trend, and 48% got booster with a Moderate Increase trend. Around 9.7% of the population reported infections over the past three months, the population immunity is Moderate Decrease.

% Public Health Policy
Recommend closing for school policy, no restrictions for workplace, no restrictions for gatherings. for elderly patients, narrow restrictions for isolation, some limitations on external visitors.

% Sequential Embedding
The sequential information is:
Hospitalization time-series: <*time-series-special-token*>

% Real-time Genomic Information
BA.2 is a sublineage of Omicron. The trend of new emerging COVID-19 variant proportion has increased from 10.0% to 16.0% in recent weeks, indicating a notable rate of change. The new emerging variant is more transmissible than previous sublineages. No evidence of immunity escape. No reported difference in severity.

% Task Prompt
Now, predict the trend of hospitalization for the one/three weeks later from the available options: [<*substantial decrease*>: substantial increase, <*moderate decrease*>: moderate decrease, <*stable*>: stable, <*moderate increase*>: moderate increase, <*substantial decrease*>: substantial decrease]

Example of PandemicLLMs prompt for New York:
[Task Information]: You are a helpful assistant designed to forecast epidemic trends for **New York**. Your task is to predict the trend of hospitalization for the next week from the available options: [Substantial Decrease, Moderate Decrease, Stable, Moderate Increase, Substantial Increase]. You need to make prediction based on the information below:

% Spatial Information
New York, one of the most populous states in the country with a higher than average percentage of Black citizens, voted Democrat in the recent Presidential election. During the pandemic, the state's healthcare system has fared better than the national standards, boasting above average Access and Affordability, Prevention and Treatment, population health conditions, ICU stress, and hospital staffing shortages. In particular, the disparity in income levels was among the best in the country.

% Epidemiological Time Series
To date, 91% of the population got at least one vaccine dose with a Slight Increase trend, 77% were fully vaccinated with a Stable trend, and 46% got booster with a Moderate Increase trend. Around 2.5% of the population reported infections over the past three months, the population immunity is Rapid Increase.

% Public Health Policy
For school policy change from recommend closing to no restrictions. workplace policy remains as no restrictions. gatherings policy remains as no restrictions. for elderly patients, narrow restrictions for isolation, some limitations on external visitors.

% Sequential Embedding
The sequential information is:
Hospitalization time-series: <*time-series-special-token*>

% Real-time Genomic Information
BA.5 is a sublineage of Omicron. The trend of new emerging COVID-19 variant proportion has increased from 21.0% to 31.0% in recent weeks, indicating a notable rate of change. The new emerging variant has high growth advantages over other sublineages. Increased immunity escape, may cause an overall increase in COVID-19 cases. No significant increase in infection severity.

% Task Prompt
Now, predict the trend of hospitalization for the one/three weeks later from the available options: [<*substantial decrease*>: substantial increase, <*moderate decrease*>: moderate decrease, <*stable*>: stable, <*moderate increase*>: moderate increase, <*substantial decrease*>: substantial decrease]

Example of PandemicLLMs prompt for Texas:
[Task Information]: You are a helpful assistant designed to forecast epidemic trends for **Texas**. Your task is to predict the trend of hospitalization for the next week from the available options: [Substantial Decrease, Moderate Decrease, Stable, Moderate Increase, Substantial Increase]. You need to make prediction based on the information below:

% Spatial Information
In the most recent Presidential election, Republicans were favored in **Texas**, the state with one of the greatest populations and a higher-than-average Black population. During the pandemic, the healthcare systems in Texas had a below-average performance in Access and Affordability, which was worse than the national average. Prevention and Treatment was close to the national average, but Population Health Conditions and Income Disparity were worse. Additionally, ICU stress and hospital staffing shortages were both worse than the national average.

% Epidemiological Time Series
To date, 64% of the population got at least one vaccine dose with a Moderate Increase trend, 55% were fully vaccinated with a Rapid Increase trend, and 21% got booster with a Rapid Increase trend. Around 1.6% of the population reported infections over the past three months, the population immunity is Rapid Decrease.

% Public Health Policy
For school policy change from require closing some to no restrictions. workplace policy remains as no restrictions. gatherings policy remains as no restrictions. for elderly patients, narrow restrictions for isolation, some limitations on external visitors.

% Sequential Embedding
The sequential information is:
Hospitalization time-series: <*time-series-special-token*>

% Real-time Genomic Information
No emerging variant for **Texas** this week.

% Task Prompt
Now, predict the trend of hospitalization for the one/three weeks later from the available options: [<*substantial decrease*>: substantial increase, <*moderate decrease*>: moderate decrease, <*stable*>: stable, <*moderate increase*>: moderate increase, <*substantial decrease*>: substantial decrease]

4 Baseline models

4.1 Heuristic-based baseline

The PrevTrend heuristic-based baseline is designed to predict future states based on historical distribution trends observed in the most recent data. For example, the 1-week prediction for **Stable** at time t can be formulated as:

$$p(HTC_1^{i,t} = \textbf{Stable} | \{HTC_1^{i,t-1} | i = 1, ..., 50\}) = \frac{|\{HTC_1^{i,t-1} = \textbf{Stable} | i = 1, ..., 50\}|}{50}, \quad (1)$$

where i is the index for each state, HTC follows the same definition as Method section 7.1, and $||$ represents the cardinality of the set. The projections for the PrevTrend model are visualized in Supplementary Fig. 6.

4.2 Machine learning baselines

Our machine learning baselines utilized hospitalization, epidemiological, numerical policy index, and vaccination data at a weekly temporal and state spatial resolution. These models incorporated various metrics, including the number of COVID-19 hospitalizations per 100k individuals, reported cases per 100k, and the percentage of the population reported infected in the past 12 weeks. Additionally, they accounted for vaccination rates, including the percentages of the population that have received one dose, completed the full vaccine series, and received a booster shot.

For our LSTM, BiLSTM, and GRU models, we transformed the training data into sequences to predict hospitalization rates per 100k for both 1-week and 3-week for each state. To facilitate probabilistic output, we incorporated a dropout layer prior to the final output layer to introduce prediction variability. This process involved generating predictions 100 times and converting continuous hospitalizations into hospitalization trend categories (HTC), as detailed in Method Section 7.1. We then computed the probability of each category to achieve a probabilistic forecast similar to PandemicLLMs. Consistent with the PandemicLLMs' methodology, we utilized data up until September 2022 for training and validation purposes, while data after September 2022 became our test dataset. We employed the Smoothed L1 loss function, defined as

$$l(x,y) = \begin{cases} \frac{0.5(x-y)^2}{\beta}, & \text{if } |x-y| < \beta \\ |x-y| + 0.5\beta, & \text{otherwise} \end{cases} \quad (2)$$

The experiments were implemented using Python 3·8 with PyTorch. We applied a grid search technique to iterate through a predefined set of hyperparameters, which included learning rate, sequence length, number of layers, week range, and hidden layer size. The period before testing was divided into an 80% training and 20% validation split. An early stopping method was activated if there was no further improvement in validation error. Our final model selections were based on MSE and hyperparameters are documented in the subsequent table:

Supplementary table 2: Selected hyperparameters for machine learning baselines

Model	Learning Rate	Sequence Length	Hidden Layer Size	Number of Layers
LSTM	0.00025	4	128	1
BiLSTM	0.0001	8	64	1
GRU	0.0001	8	32	1

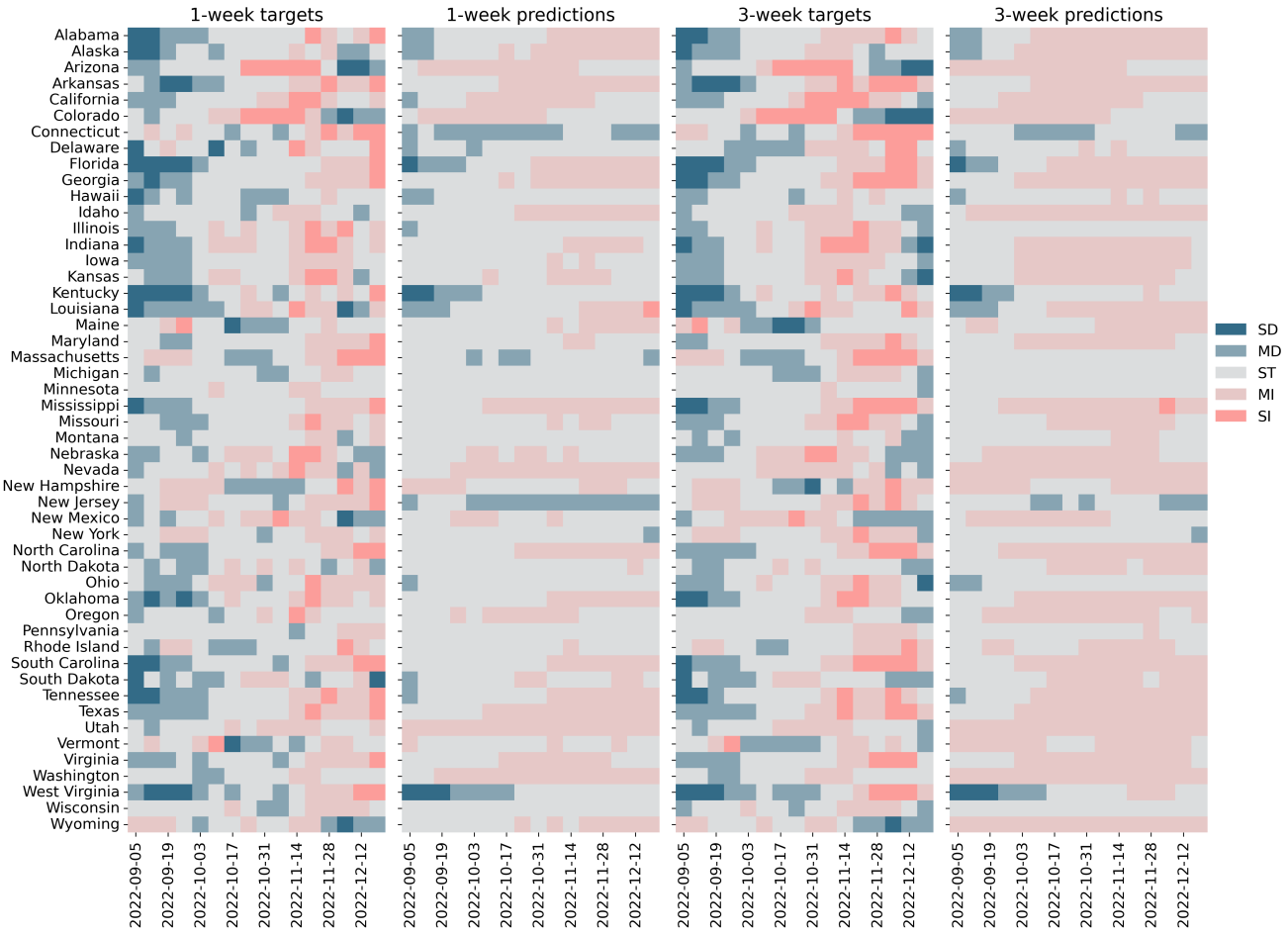
5 Experimental settings

We employ LLaMA2[4] as the backbones for PandemicLLMs. For optimization, AdamW[5] is utilized, accompanied by a warmup-decay learning rate schedule (the warm-up rate is set at 0.1). The batch size is configured to 4, and we fine-tune the model over 1500 steps. To identify the model configuration that yields the lowest Mean Squared Error (MSE) on the validation data, we conduct a grid search across various hyperparameters. We use ChatGPT-3.5 to rewrite the static or dynamic prompts for data augmentation. Details regarding our learning rate, random seed, as well as static and dynamic information augmentation settings are presented in Supplementary Table 3.

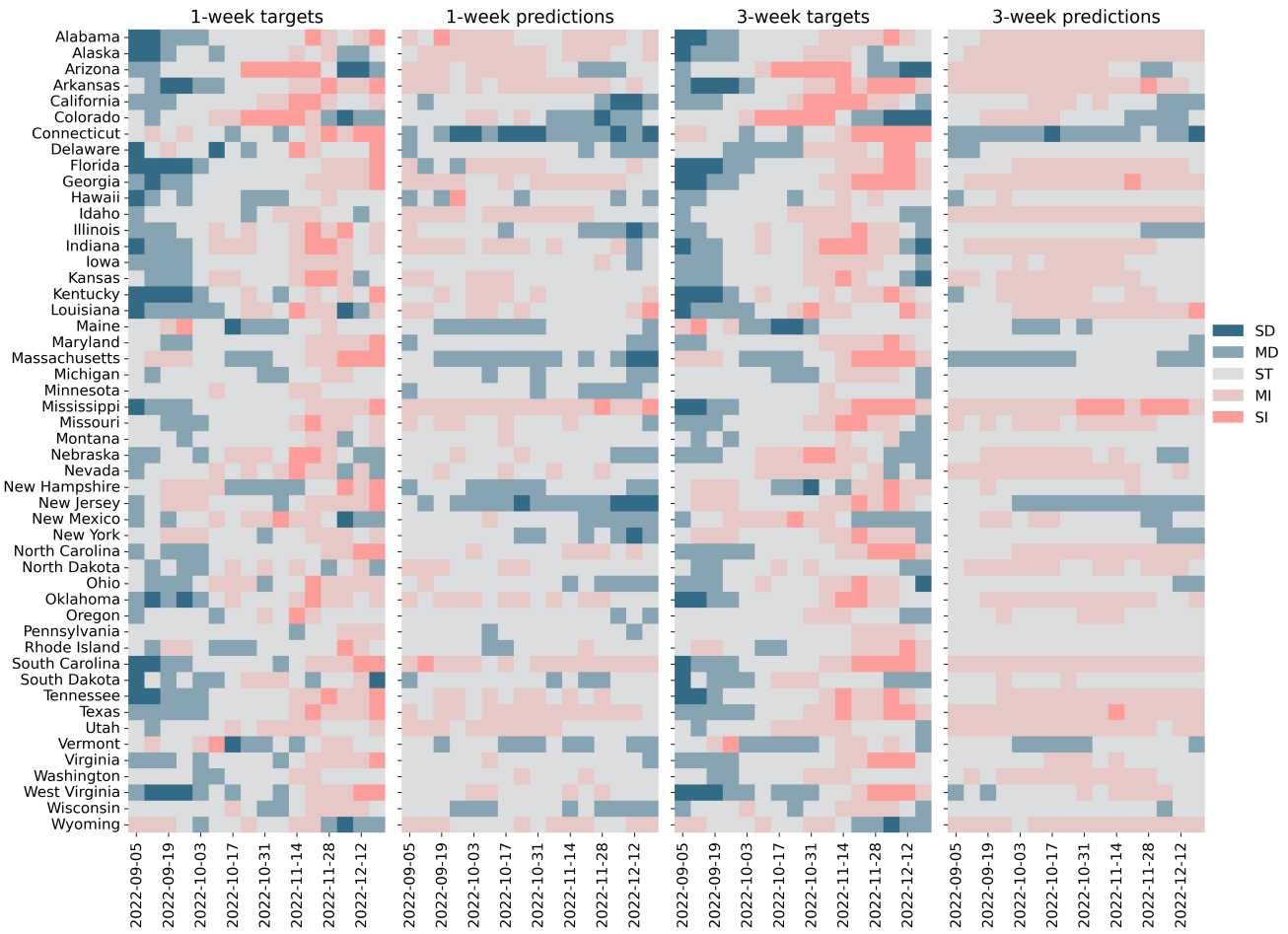
Supplementary Table 3: Hyperparameters settings for PandemicLLMs.

Prediction Target	**Model**	**Hyperparameters**			
		Learning Rate	**Random Seed**	**Static Information Augmentation**	**Dynamic Information Augmentation**
1-week	PandemicLLM-7B	2e-5	2024		
	PandemicLLM-13B	1e-5	2023	✓	
	PandemicLLM-70B	1e-5	2023	✓	✓
3-week	PandemicLLM-7B	2e-5	2024	✓	
	PandemicLLM-13B	2e-5	2023		
	PandemicLLM-70B	2e-5	2023	✓	✓

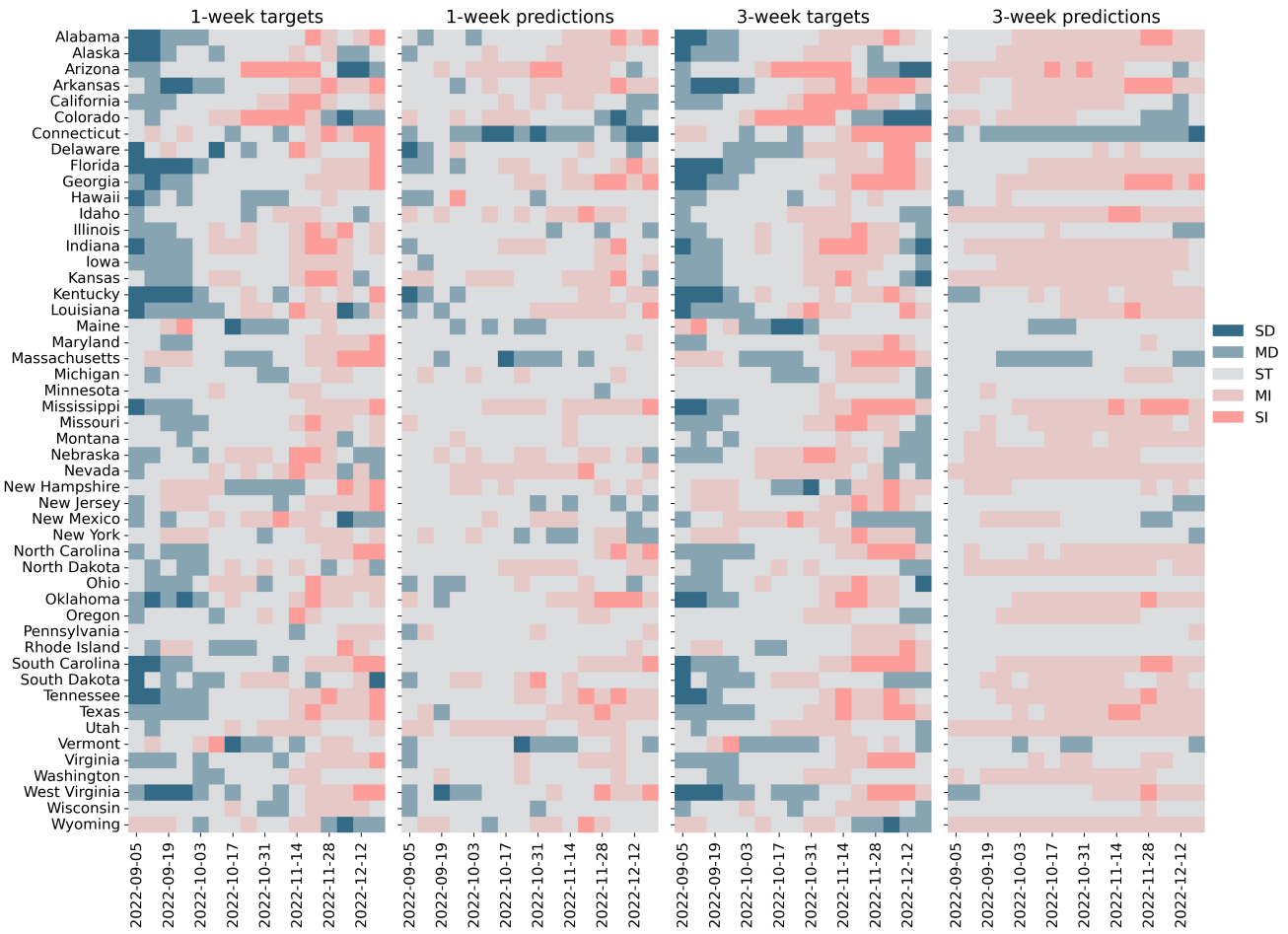
6 Baseline models' predictions



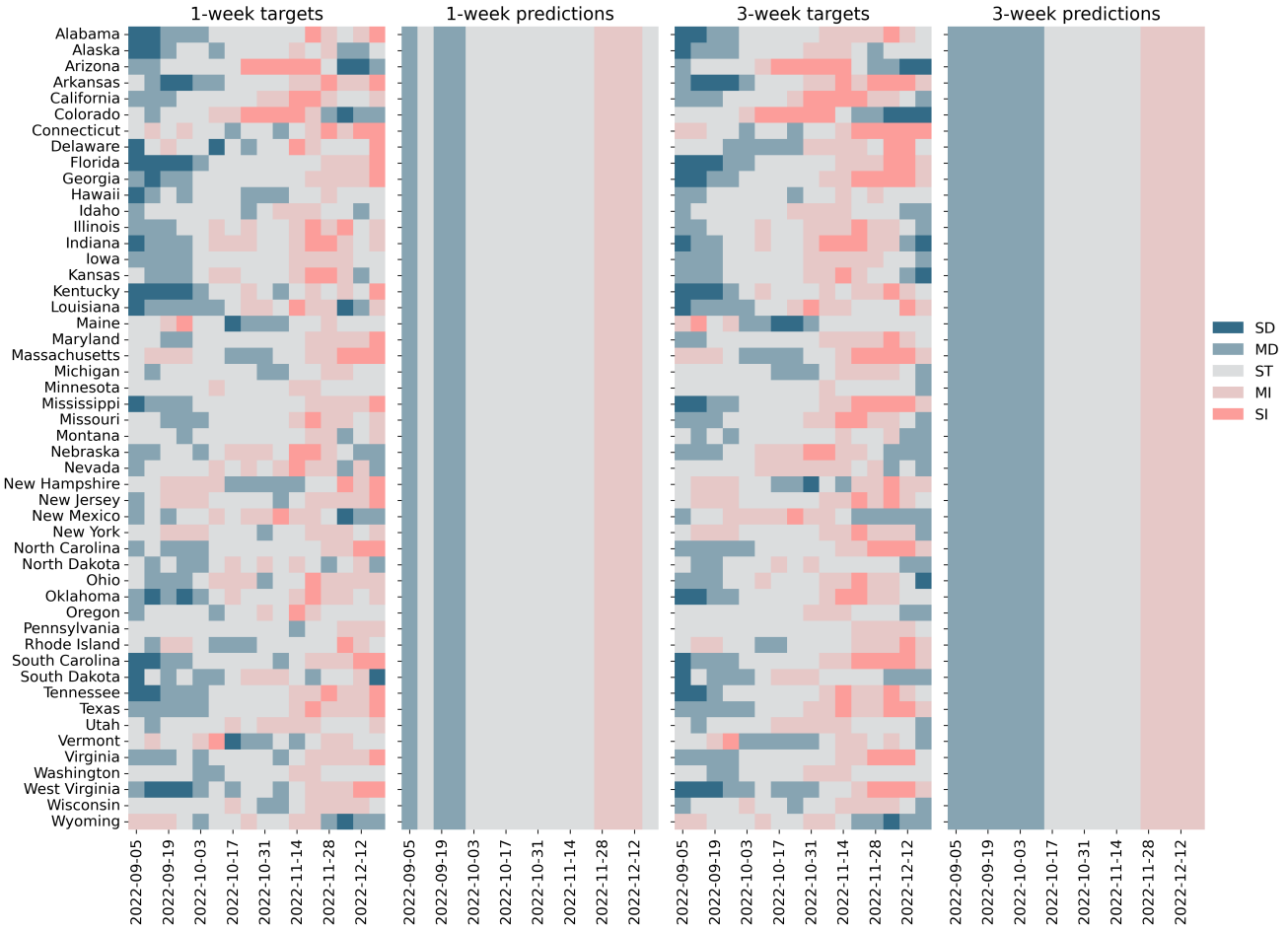
Supplementary figure 3. **LSTM's predictions visualization:** The figure presents the LSTM's predictions stratified by location and time versus the ground truth targets. The color scheme indicates different target categories. SD: Substantial Decrease, MD: Moderate Decrease, ST: Stable, MI: Moderate Increase, SI: Substantial Increase.



Supplementary figure 4. **Bi-LSTM's predictions visualization:** The figure presents the Bi-LSTM's predictions stratified by location and time versus the ground truth targets. The color scheme indicates different target categories. SD: Substantial Decrease, MD: Moderate Decrease, ST: Stable, MI: Moderate Increase, SI: Substantial Increase.

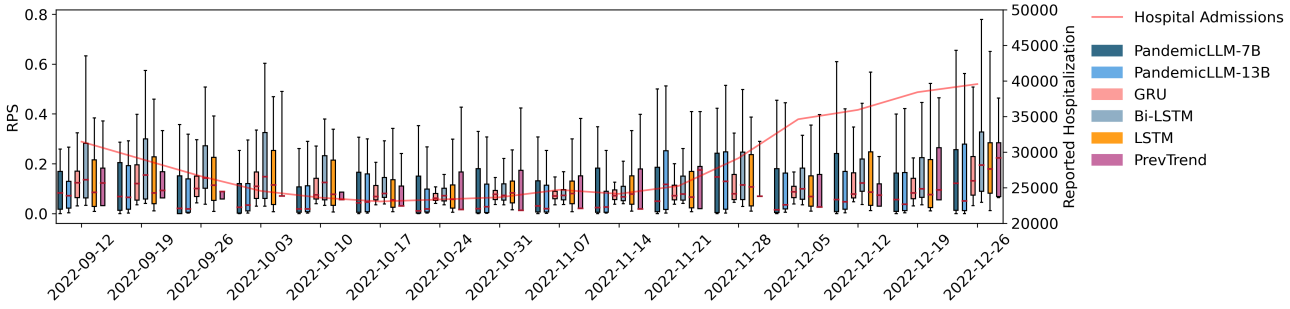


Supplementary figure 5. **GRU's predictions visualization:** The figure presents the GRU's predictions stratified by location and time versus the ground truth targets. The color scheme indicates different target categories.SD: Substantial Decrease, MD: Moderate Decrease, ST: Stable, MI: Moderate Increase, SI: Substantial Increase.

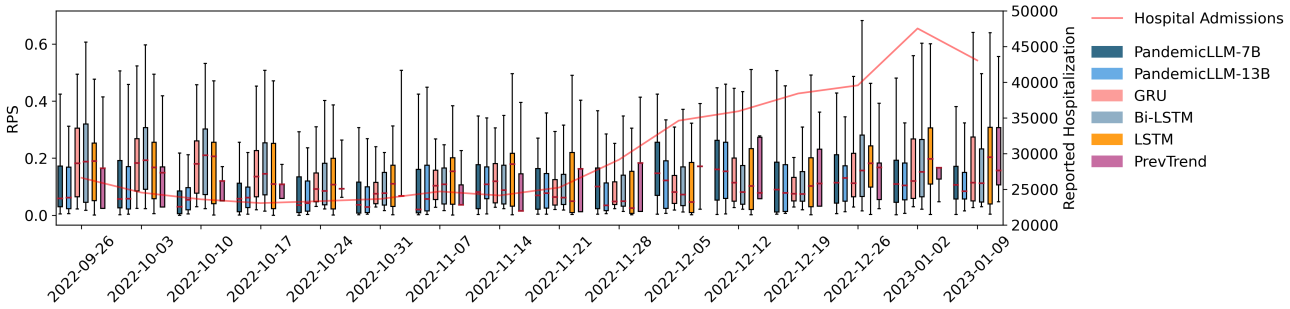


Supplementary figure 6. **PrevTrend's predictions visualization:** The figure presents the PrevTrend's predictions stratified by location and time versus the ground truth targets. The color scheme indicates different target categories. SD: Substantial Decrease, MD: Moderate Decrease, ST: Stable, MI: Moderate Increase, SI: Substantial Increase.

7 Temporal model performance evaluation with alternative error metrics

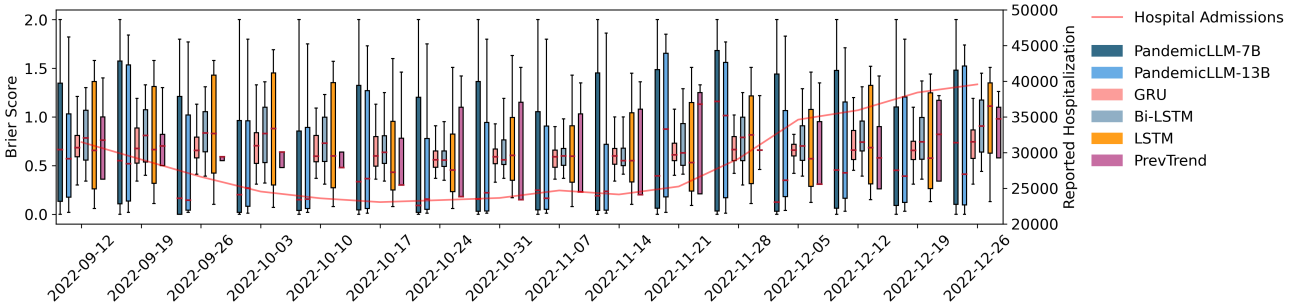


(a) 1-week predictions

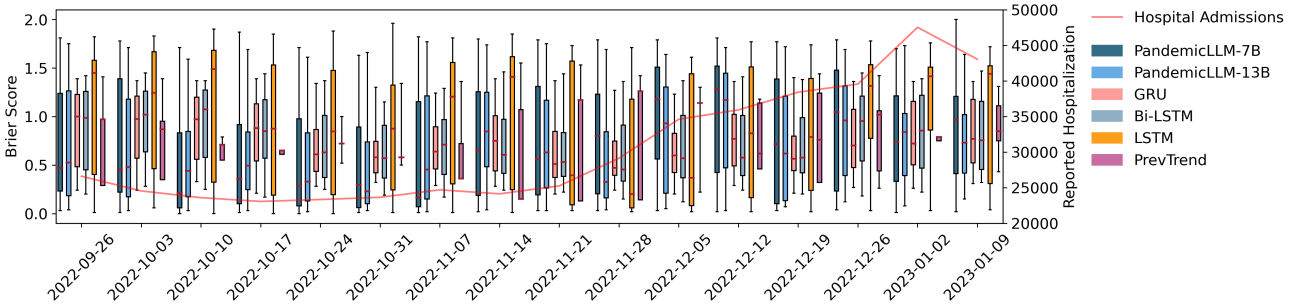


(b) 3-week predictions

Supplementary figure 9: **Performances comparison of PandemicLLMs with baseline and other machine learning models across time.** The red curve on the back represents the weekly reported COVID-19 hospital admission at the national level. The left y-axis represents the scale of RPS, and the right y-axis represents the scale of hospital admission. Each set of bar graphs in the figure represents the distribution of RPS for all states during a specific week. The color bars represent the error distributions for different models. **(a)** 1-week forecasting performance. **(b)** 1-week forecasting performance.



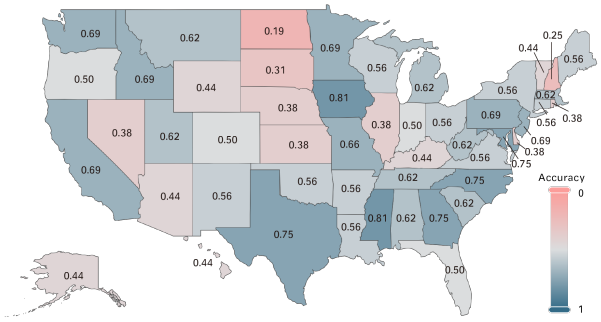
(a) 1-week predictions



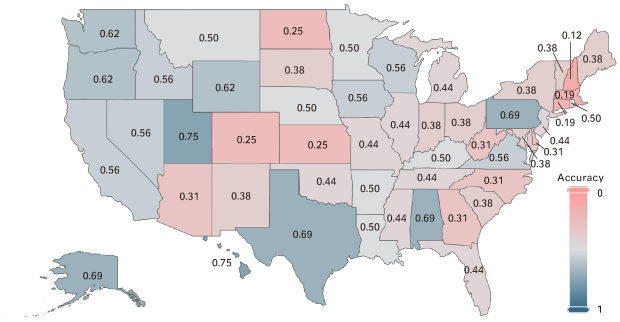
(b) 3-week predictions

Supplementary figure 10: **Performances comparison of PandemicLLMs with baseline and other machine learning models across time.** The red curve on the back represents the weekly reported COVID-19 hospital admission at the national level. The left y-axis represents the scale of the Brier Score, and the right y-axis represents the scale of hospital admission. Each set of bar graphs in the figure represents the distribution of the Brier Score for all states during a specific week. The color bars represent the error distributions for different models. **(a)** 1-week forecasting performance. **(b)** 1-week forecasting performance.

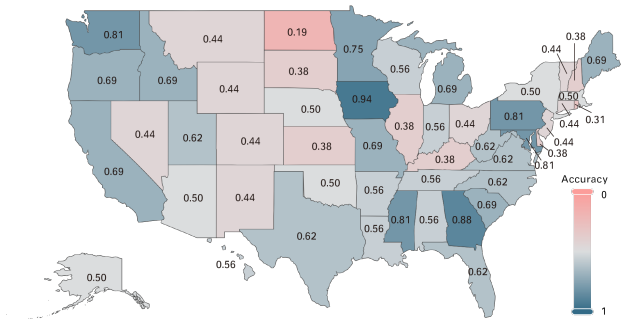
8 Spatial model performance evaluation with alternative error metrics



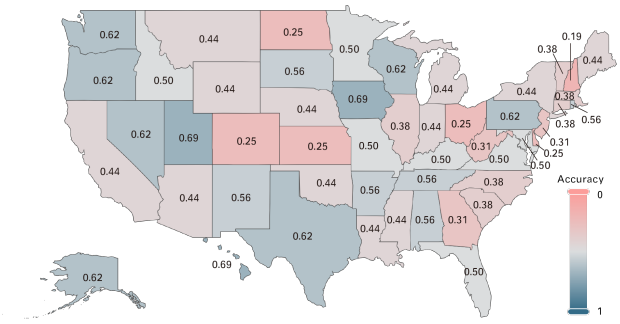
(a) PandemicLLM-7B performance by state (1-week)



(b) PandemicLLM-7B performance by state (3-week)

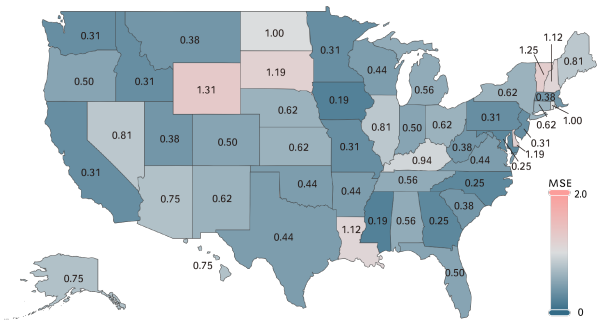


(c) PandemicLLM-13B performance by state (1-week)

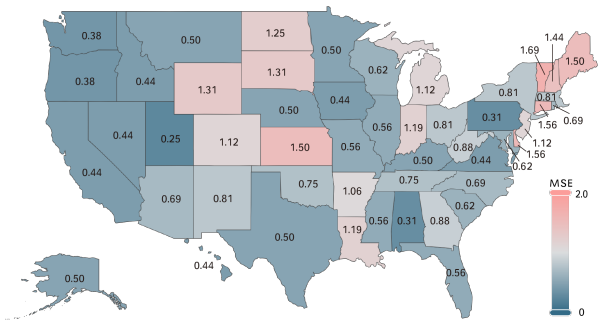


(d) PandemicLLM-13B performance by state (3-week)

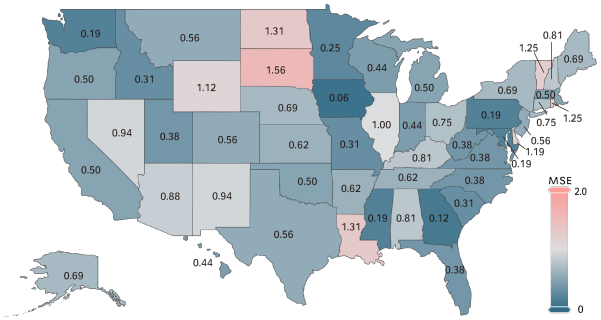
Supplementary figure 11: **Average state performance based on Accuracy by states. (a, b)** 1 and 3-week performance for PandemicLLM-7B. **(c, d)** 1 and 3-week performance for PandemicLLM-13B



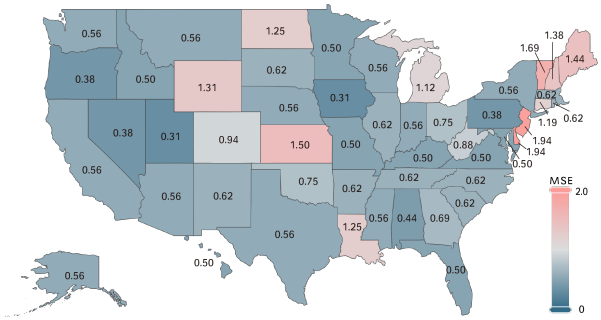
(a) PandemicLLM-7B performance by state (1-week)



(b) PandemicLLM-7B performance by state (3-week)

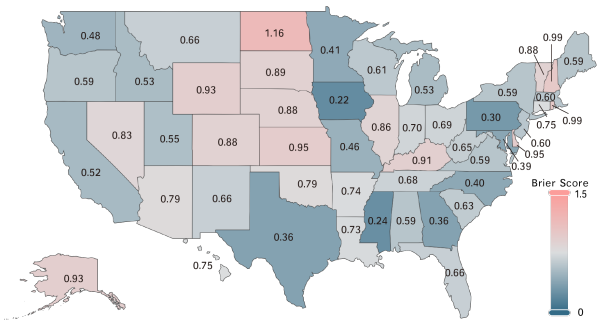


(c) PandemicLLM-13B performance by state (1-week)

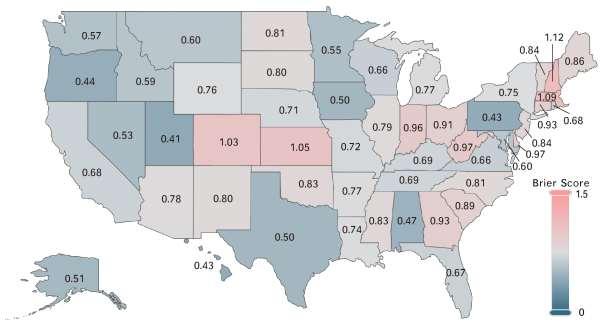


(d) PandemicLLM-13B performance by state (3-week)

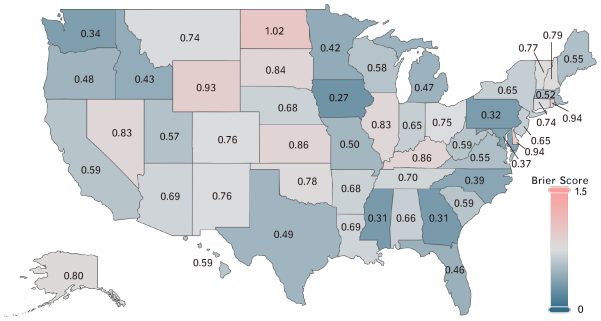
Supplementary figure 12: **Average state performance based on MSE by states.** **(a, b)** 1 and 3-week performance for PandemicLLM-7B. **(c, d)** 1 and 3-week performance for PandemicLLM-13B



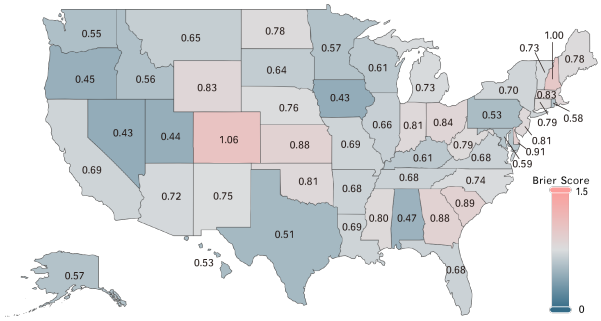
(a) PandemicLLM-7B performance by state (1-week)



(b) PandemicLLM-7B performance by state (3-week)

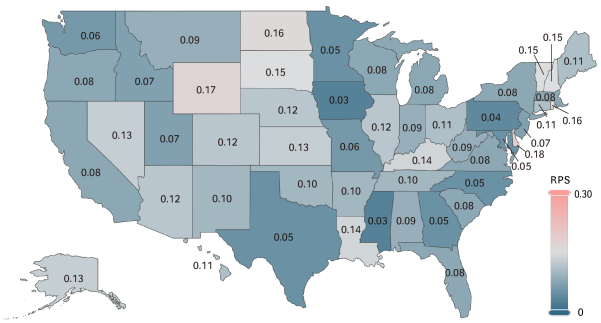


(c) PandemicLLM-13B performance by state (1-week)

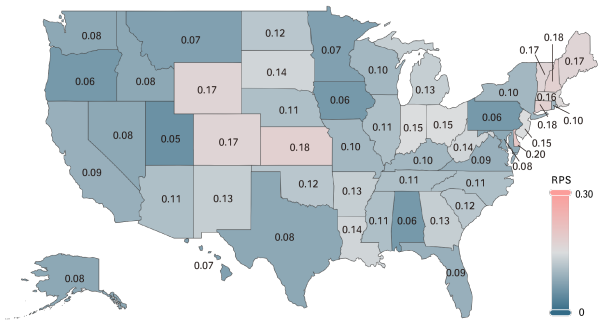


(d) PandemicLLM-13B performance by state (3-week)

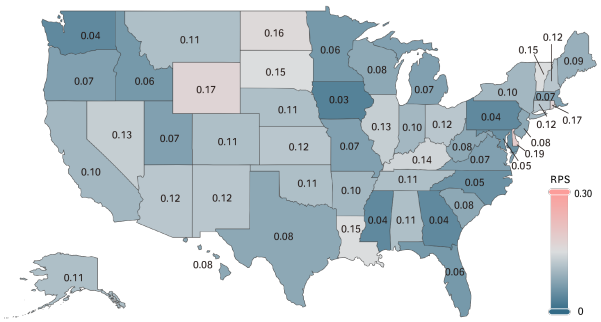
Supplementary figure 13: **Average state performance based on Brier Score by states. (a, b)** 1 and 3-week performance for PandemicLLM-7B. **(c, d)** 1 and 3-week performance for PandemicLLM-13B



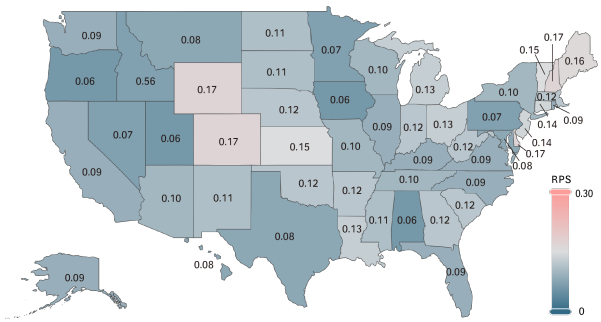
(a) PandemicLLM-7B performance by state (1-week)



(b) PandemicLLM-7B performance by state (3-week)



(c) PandemicLLM-13B performance by state (1-week)



(d) PandemicLLM-13B performance by state (3-week)

Supplementary figure 14: **Average state performance based on RPS by states. (a, b)** 1 and 3-week performance for PandemicLLM-7B. **(c, d)** 1 and 3-week performance for PandemicLLM-13B

References

1. Statement on omicron sublineage BA.2. https://www.who.int/news/item/22-02-2022-statement-on-omicron-sublineage-ba.2. Accessed: 2024-01-01.
2. Epidemiological update: SARS-CoV-2 omicron sub-lineages BA.4 and BA.5. https://www.ecdc.europa.eu/en/news-events/epidemiological-update-sars-cov-2-omicron-sub-lineages-ba4-and-ba5. Accessed: 2024-01-01.
3. TAG-VE statement on Omicron sublineages BQ.1 and XBB. https://www.who.int/news/item/27-10-2022-tag-ve-statement-on-omicron-sublineages-bq.1-and-xbb. Accessed: 2024-01-01.
4. Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
5. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (OpenReview.net, 2019).