

EpiGPT: Prompting LLMs for Epidemiology

Sayan Sinha*

Haoxin Liu*

sayan.sinha@cc.gatech.edu

hliu763@gatech.edu

Georgia Institute of Technology

Atlanta, GA, USA

ABSTRACT

Our goal is to investigate the feasibility of prompting Large Language Models (LLMs) like ChatGPT to predict the trajectory of an epidemic and generate useful information from the same. This means that instead of training the model, we provide it with prompts related to the epidemic and then query it for future predictions. Moreover, we are keen to understand how these models can be used for providing better natural language explanation to such time series data, such that they maybe easily understood by health experts and policy makers. We also look at the model's capability in simulating counterfactual scenarios, such as how the epidemic trajectory would change if a lockdown was to be declared.

KEYWORDS

Large Language Models, Prompt Learning, Epidemic Forecasting, Counterfactual Prediction

ACM Reference Format:

Sayan Sinha and Haoxin Liu. 2023. EpiGPT: Prompting LLMs for Epidemiology. In *Proceedings of Data Science for Epidemiology (CS8803-EPI 23)*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 PROBLEM DEFINITION

1.1 Epidemic forecasting

Consider an epidemic trajectory forecasting task. Let a series of historical data points $X = \{x_1, x_2, \dots, x_{t_{\text{obs}}}\}$ representing the observed values of the epidemic over consecutive t_{obs} time steps. Each data point x_t might encompass various metrics such as the number of cases, deaths, recoveries, etc. Given this historical data and a prompt P related to the epidemic, our objective is to prompt the LLM F to obtain its prediction $Y = \{y_{t_{\text{obs}}+1}, y_{t_{\text{obs}}+2}, \dots, y_{t_{\text{obs}}+n}\}$ for the trajectory of the epidemic over the next n time steps. Therefore, we learn the prediction function

$$Y = F(P, X)$$

*Both authors contributed equally to this research.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted by ACM, provided that the copies are not made for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CS8803-EPI 23, Fall 2023, Georgia Tech

© 2023 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

2023-12-06 00:29. Page 1 of 9.

1.2 Generating explanations from series

LLMs have access immense knowledge, such as average historical values of a disease spread count etc. We wish to leverage such information to generate a natural language text that analyzes a time series data and provides helpful explanation that maybe useful to policy makers.

1.3 Counterfactual modeling

For the counterfactual scenario modeling, let C^f be a set of altered conditions, such as different vaccination policies or lockdown measures. These conditions can drastically alter the trajectory of the epidemic. Under these conditions, our goal is to prompt the model to obtain its prediction $Y^f = \{y_{t_{\text{obs}}+1}^f, y_{t_{\text{obs}}+2}^f, \dots, y_{t_{\text{obs}}+n}^f\}$ for the new trajectory of the epidemic. Therefore, we learn prediction function under counterfactual scenario

$$Y^f = F(P, X, C = C^f)$$

This formulation allows us to leverage the vast knowledge and inference capabilities of LLMs to predict the future trajectory of epidemics and understand the potential outcomes under different scenarios.

2 IMPROVEMENTS ADDRESSING COMMENTS

2.1 Comment 1: Dataset analysis

Section 4 now has some statistical information.

2.2 Comment 2: Confusing modules

We have streamlined the writeup into more appropriate headings.

3 RELATED WORK AND TECHNIQUES

3.1 Large Language Models

Large Language Models (LLMs) have become a focal point of research in the Natural Language Processing domain in recent years. Models such as the GPT series have demonstrated powerful representational and generative capabilities by pre-training on massive corpora [17]. Their advent has not only propelled performance enhancements across various NLP tasks but has also instigated discussions around their interpretability, biases, and ethical concerns [19]. Moreover, specific vulnerabilities of these models, such as the potential for generating objectionable outputs in models like ChatGPT, have been highlighted [5]. While LLMs have seen success across multiple applications, their utilization in the domain of epidemic forecasting remains a relatively novel research direction. In

this project, we plan to use the following advanced large language model:

- (1) **GPT-3.5**: A series of models released by OpenAI, trained on a blend of text and code, demonstrating robust capabilities [13].
- (2) **GPT-4**: A Transformer-based model pre-trained to predict the next token in a document [11].
- (3) **LLaMA**: A collection of foundation language models ranging from 7B to 65B parameters, trained on trillions of tokens, showcasing superior performance across various tasks [16].
- (4) **BARD**: Designed by Google as an interface to an LLM, enabling users to collaborate with generative AI [4].

3.2 Prompt Learning

With the advent of large language models like GPT-3 in recent years, *Prompt Learning* has emerged as a prominent area of research. Reynolds and McDonell, in their paper [14], discuss ways to better control and evaluate powerful language models through prompt programming. They introduce a perspective where the function of few-shot examples in these scenarios is better described as locating an already learned task rather than meta-learning. Furthermore, Zhang et al. [18] conducted a systematic study on prompting strategies for translation. However, despite these insightful studies on Prompt Learning, there remains a gap in the literature concerning the application of large language models for epidemic forecasting, which is the domain our paper, *LLMs for Epidemic Forecasting*, seeks to explore. Prompt Learning offers a novel approach, enabling the efficient utilization of pre-trained language models for epidemic forecasting. Our preliminary plan is as follows.

- (1) **Prompt Template Construction**: Initially, a template needs to be constructed that restructures the input and output into a new text with mask slots. For epidemic data, a template could be: "In the region [x], the predicted number of infections over [y] time is [z]." Prompt template may also contain information about different views concerning the epidemic, like twitter data or google custom search results.
- (2) **Answer Construction**: For the constructed prompt, a relationship between the predicted word and our label needs to be discerned. A mapping function can be defined to map the output word with the label.
- (3) **Answer Prediction**: An appropriate pre-trained language model is selected, followed by the prediction of mask slots [z].
- (4) **Answer-label Mapping**: The obtained answer is mapped back to the original label using the mapping function.
- (5) **Strategy Selection**: Depending on the data volume and model size, an apt strategy is chosen. For instance, with limited data, the choice might lean towards not fine-tuning but adjusting the prompt parameters.

3.3 Epidemic Forecasting

Epidemic forecasting is a pivotal task in public health, and with the advent of machine learning technologies, its application in epidemic forecasting has garnered significant attention in recent years. For instance, the work by Zhu et al. presents a spatial epidemic model for predicting the number of deaths [20]. Another study delves into

predicting daily scenarios of COVID-19 based on a deep dependence analysis of variables [2]. Moreover, research indicates that ML-based models can offer accurate predictions and insights, especially in the realm of drug and antibiotic discovery [3]. In essence, machine learning provides a robust tool for epidemic forecasting, aiding researchers and policymakers in better understanding and addressing the spread of epidemics. We plan to utilize large pre-trained language models to accomplish the following tasks related to epidemic forecasting and analysis:

- (1) **Epidemic Incidence Rate Prediction**: Primarily focusing on national-level data, forecasting the incidence rate for a specific time frame.
- (2) **Epidemic Outbreak Alert**: Using the model to predict potential outbreaks of infectious diseases in regions or countries and issuing timely alerts.
- (3) **Early Prediction of Infectious Diseases**: Predicting the possible spread and impact during the initial or preliminary stages of the disease.
- (4) **Counterfactual Epidemic Forecasting**: Predicting the spread and impact of diseases based on hypothetical scenarios, such as different preventive measures or strategies.
- (5) **Interpretability of Epidemic Forecasting Results**: Providing clear explanations and reasons for the forecasting results, assisting the public and decision-makers in understanding the underlying reasons for the predictions.

3.4 Counterfactual Prediction & Summarization

Counterfactual prediction has emerged as a pivotal area of research in machine learning, focusing on predicting outcomes under hypothetical scenarios. Counterfactual prediction pertains to the estimation of outcomes under hypothetical scenarios that differ from the observed data. Formally, given an observed outcome Y^{obs} under conditions C_{obs} , the counterfactual prediction Y^{cf} under a different set of conditions C^{cf} is defined as the expected outcome had the conditions been C^{cf} instead of C^{obs} .

Mathematically, this can be represented as:

$$Y^{cf} = E[Y|C = C^{cf}]$$

where $E[\cdot]$ denotes the expectation.

As discussed in the article [12], fitting a machine learning model to observational data and using it for counterfactual prediction might lead to detrimental consequences. Furthermore, Hartford et al., in their paper [10], provide a methodology that accurately characterizes such relationships in the presence of instrumental variables (IVs). Despite these insightful studies on counterfactual prediction, the application of these methodologies to real-world problems, such as epidemic forecasting, remains an open avenue for exploration. In epidemic analysis, counterfactual prediction plays a pivotal role. Counterfactual prediction allows us to assess how the spread of a disease might have evolved under certain interventions. For instance, what might have been the speed and extent of disease propagation if a particular quarantine measure had not been implemented? In this manner, we can gain a deeper understanding of the actual impact of the measures taken, providing robust support for future decision-making. Moreover, simulation studies are frequently employed to predict the anticipated impact of control

measures during infectious disease outbreaks. Thus, counterfactual prediction offers us a potent tool to better comprehend and evaluate the efficacy of public health strategies.

Text summarization from time series has been explored before by training language models with a specific dataset for the purpose [15]. However, in this work, we focus at zero-shot text summarization by prompting LLMs, without the need for any finetuning.

4 DATASETS

We plan to use the following five datasets for this work:

- **Google Symptoms Search Dataset [1]:** This dataset shows aggregated trends in Google searches for multiple health symptoms. We extract 14 related symptoms as per the CD-Cwebsite¹). We wish to seek prediction from 1 to 4 weeks ahead in the future for multiple years through 2022. (mean: 816)
- **CDC-Flu:** We obtain data from CDC hospitalization data² predict future hospitalization upto three weeks. We consider the seasons of 2021-2022 and 2022-2023, where each season consists of 27 weeks of data. We use 5 specific features from this dataset, namely "positiveIncr", "cdc_negativeIncr", "cdc_hospitalized", "death_jhu_incidence", "cdc_flu_hosp". (mean:27.98)
- **CDC-Covid:** We use the same dataset and features as described for CDC-Flu, with a different label of covid related hospitalizations. We cover the season 2021-22. (mean 297.22)
- **Monash datasets:** We wish to use the covid-deaths [7] and hospitalization datasets [8] from Monash. (mean: 265)
- **ILINet:**the U.S. Outpatient Influenza-like Illness Surveillance Network, plays a crucial role in the syndromic surveillance of influenza within the United States. This network systematically gathers data on outpatient visits to healthcare providers for influenza-like illness, providing a comprehensive overview of influenza activity across the country³.

For the Flu dataset that we can obtain for the current season (2023-24), we plan to attempt a real-time forecasting procedure with data collected from other views regarding the spread of disease this season. We use the first five datasets for experiments on epidemic forecasting, while the last dataset is used for summarization.

5 METHODOLOGY – EPIDEMIC FORECASTING

Given a time series dataset representing the progression of an epidemic, our goal is to forecast future values using a Large Language Model (LLM). Our approach, which involves encoding the time series as a string of numerical digits, treating the forecasting problem as a next-token prediction task and constructing prompts using in-context learning for different epidemic forecasting tasks. Given the recent advent of LLMs and recent work as presented before, we expect LLMs to be able to help with epidemic forecasting.

¹<https://www.cdc.gov/flu/symptoms/symptoms.htm>

²<https://healthdata.gov/api/views/g62h-syeh/rows.csv?accessType=DOWNLOAD>

³<https://www.cdc.gov/flu/weekly/overview.htm>

5.1 Preprocessing

5.1.1 Tokenization. The time series data is tokenized into a string format. Each numerical value is separated by a comma and space to ensure individual digit tokenization. For example, a time series {0.123, 1.23, 12.3, 123.0} is encoded as:

"1 2, 1 2 3, 1 2 3 0, 1 2 3 0 0".

5.1.2 Rescaling. Values are rescaled such that the α -percentile of the rescaled time series is 1, to avoid token wastage for large numbers. An offset β is also applied, calculated as a percentile of the input data.

5.2 Designing Task-Specific Prompts

By curating the prompts specifically for the task at hand, the LLM can generate more accurate and relevant forecasts. For the specific case of epidemiology, we also analyze how the LLMs behave if provided with some context about the time series data. For instance, if the goal is to forecast the number of cases for a particular disease, additional context with the prompt is structured as follows:

"The user is trying to predict the number of <disease> cases in <region>. It usually rises every year in the <season> and falls in the <season>. <Number> of past waves of the disease has been witnessed".

This prompt structure guides the LLM to focus on the forecasting aspect of the provided time series data, enhancing its prediction capability for the specific epidemic scenario.

We will further discuss the design of prompts in Section 5.6.

5.3 Sampling and Forecasting

We generate multiple samples (e.g., 20) from the LLM to construct a probabilistic forecast. Techniques such as temperature scaling and nucleus sampling are used to control the sampling process.

5.4 Continuous Likelihoods

The discrete probability distribution of the LLM is adapted to provide a continuous density function. With n digits of precision in base B , each sequence of n digits corresponds to one of B^n bins, each with width B^{-n} .

5.5 Epidemic forecasting framework with LLM

The Epidemic forecasting framework with LLM is applied as follows:

- (1) Preprocess the time series data into a string format suitable for the LLM.
- (2) Rescale and tokenize the data.
- (3) Design Task-Specific Prompts.
- (4) Use the LLM to sample possible continuations of the time series string.
- (5) Convert the discrete distributions over tokens into continuous densities.
- (6) Construct probabilistic forecasts from the sampled data.

5.6 In-Context Learning for Epidemic Forecasting

In-context learning (ICL) has emerged as a powerful technique in the realm of Large Language Models (LLMs), enabling these models to learn and adapt to new tasks through carefully designed prompts. This chapter delves into the latest advancements in prompt design for ICL, with a focus on its application in epidemic forecasting tasks.

5.6.1 Understanding In-Context Learning. In-context learning refers to the model's ability to learn from a few examples provided within the prompt at inference time. It is a form of few-shot learning that does not require explicit retraining of the model. The key to ICL is the design of prompts that effectively communicate the task to the model, leveraging its pre-trained knowledge to generate accurate predictions or analyses. When designing effective prompts using ICL methods, the following aspects should be considered:

- **Clarity and Specificity:** Ensure that prompts are clear and task-specific to guide the model effectively.
- **Inclusion of Examples:** Include a few examples within the prompt to illustrate the task to the model.
- **Iterative Refinement:** Test and refine prompts iteratively to achieve the desired output from the model.
- **Balancing Information:** Provide enough information in the prompt to guide the model without overwhelming it.
- **Utilization of Advanced Techniques:** Incorporate the latest findings from the field of prompt engineering, such as the use of meta-prompts, dynamic prompts that adjust based on previous interactions, and the inclusion of explicit instructions for the model to use its reasoning capabilities.

5.6.2 Detailed Prompt Design for Epidemic Forecasting. Designing prompts for epidemic forecasting involves creating a context that encapsulates the task at hand, whether it be predicting the trajectory of an epidemic, analyzing the impact of public health interventions, or simulating counterfactual scenarios. The following are strategies for prompt design tailored to specific epidemic forecasting tasks:

5.6.3 Incidence Rate Prediction. For predicting incidence rates, prompts should include a concise description of the task, relevant historical data, and a clear indication of the prediction horizon. An example prompt might be:

"Given the historical data on COVID-19 cases in region X: [data series], predict the expected number of cases for the next N days."

5.6.4 Outbreak Alert Generation. When designing prompts for outbreak alerts, the prompt should guide the model to identify patterns indicative of an outbreak. For instance:

"Review the following infection rates over the past weeks [data series]. Identify if there is a potential outbreak and suggest the likelihood of its occurrence."

5.6.5 Early Prediction of Infectious Diseases. Early prediction prompts must encourage the model to make inferences based on limited data, often requiring the model to extrapolate from early trends. A sample prompt could be:

"With the initial data on a new infectious disease [data points], forecast the spread pattern for the coming month."

5.6.6 Counterfactual Scenario Analysis. For counterfactual predictions, the prompt should clearly state the hypothetical changes in conditions and ask the model to predict outcomes based on these alterations:

"Assuming that lockdown measures were implemented two weeks earlier in region Y, how would the epidemic curve [data series] have been altered?"

5.6.7 Interpretability of Results. To enhance the interpretability of the model's predictions, prompts should request explanations for the forecasts provided. For example:

"Explain the reasoning behind your prediction of the epidemic's trajectory in region Z given the following data [data series]."

6 METHODOLOGY – SERIES2TEXT: EPIDEMIC DATA SUMMARIZATION VIA LARGE LANGUAGE MODEL

The increasing volume and complexity of epidemic data pose significant challenges in terms of interpretation and actionable insights extraction. In this context, the translation of quantitative epidemic series into comprehensive textual summaries becomes essential. This approach enables a more intuitive understanding of data trends and patterns, facilitating effective communication among public health professionals, policy makers, and the general public. Utilizing large language models for this task presents an innovative method to bridge the gap between data-driven insights and human-centric information dissemination.

In addition to the primary evaluation metrics and methods, it is imperative to delve deeper into the model's behavior and robustness. This ensures that our model not only performs well on our test datasets but is also reliable and interpretable in various scenarios.

- (1) **Hyperparameter Sensitivity Analysis:** Understanding the model's sensitivity to various hyperparameters is crucial. By varying one hyperparameter while keeping others constant, we can observe the impact on the model's performance. This analysis aids in identifying optimal hyperparameter settings and understanding their influence on the model's predictions.
- (2) **Ablation Studies:** Ablation studies involve systematically removing or modifying certain parts of the model to understand their contribution to the overall performance. By doing so, we can identify which components or features are most critical for the model's success and which ones can be modified or omitted without significantly affecting the results.

7 EXPERIMENTATION & RESULTS

7.1 Epidemic Forecasting

We seek inspiration from LLMTime [9] and attempt prompting language models for forecasting on the five datasets as mentioned in section 4. We realized the open-source code repository only has support for OpenAI's ChatGPT and LLAMA. While LLAMA can be run locally, ChatGPT requires subscription to be used through an API. Hence, we integrated GPT4All⁴ with the library. As shown in

⁴<https://docs.gpt4all.io/index.html>

Topic	Primary contributor	Secondary contributor
Ideation	Sayan	Haoxin
Literature Review	Haoxin	Sayan
Prompt engineering for template construction	Haoxin	Sayan
Prompt-based learning for template construction	Sayan	Haoxin
Answer construction and mapping	Haoxin	Sayan
Strategy selection	Sayan	Haoxin
Counterfactual predictions	Sayan	Haoxin
Assessing external data sources	Haoxin	Sayan

Table 1: Contributions

Table 2: MAE values on 5 datasets

Dataset	LLAMA	GPT4ALL	ARIMA
CDC-Flu	8.32	9.23	7.96
CDC-Covid	22.15	19.76	25.89
Symptoms search	0.28	0.49	0.32
Monash-Covid	2.22	1.61	1.09
Monash-Hospitalization	7.89	8.46	6.73

Table 3: MAE values for forecasting with context

Dataset	GPT4ALL	w/ context	more context
CDC-Flu	9.23	8.92	8.65
CDC-Covid	19.76	18.73	26.31
Symptoms search	0.49	0.46	0.48
Monash-Covid	1.61	2.36	2.39
Monash-Hospitalization	8.46	7.72	8.41

Table 2, it is evident that the mean average errors are more or less comparable with a pretty standard forecasting technique – ARIMA [6]. The prompts provided for this experiment lacked context – it was initiated by simply instructing the LLM to generate forecast for a time series without any background regarding the series. Sample examples of the series have been shown in Figure 1(a)-(e). Statistical forecasting models make use of time series differencing to perform forecasting, while transformer-based forecasting model need heavy training data to get started with. Given that this is zero-shot, and the LLMs used haven't been fine-tuned as per context, the results look promising.

In table 3, we provide some context along with the time series, as described in subsection 5.2. In the second column, we provide general description of what the series is about. In the third column, the prompt also includes more information such as the dates involved, the no of waves of the disease spread experiences before etc. We find that providing context helps for diseases that spread more seasonally, however balancing the information becomes very important.

We also look into counterfactual prediction by prompting the LLM to consider the incidence of a lockdown. However, we do not find any relevant way the LLM is able to use this information. We have provided an example in Figure 1(f). In the situation of

lockdown, the rates should decrease after a few time steps, but that is not captured by the model. Perhaps LLMs need more context of a "lockdown" or it might itself question its effectiveness.

7.2 Time Series Summarization

The objective of this evaluation was to assess the effectiveness of a standardized prompt in eliciting accurate and coherent summaries of epidemic data from advanced Large Language Models (LLMs). We designed a prompt structured in three parts: Domain description, Instruction, and Input Data. The focus was on summarizing patient visit data from the 2023 ILINet database.

7.2.1 Standard Prompt Design. The standard prompt comprised the following structure:

[BEGIN DATA]

[Domain]: Description of the role of a Public Health Communication Specialist, with a focus on the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) data from 2023.

[Instruction]: A directive to summarize the ILINet data concerning outpatient visits for ILI from the 1st to the 46th week of 2023.

[Input Data]: Specific patient visit numbers for the given time frame (data omitted for brevity).

[END DATA]

7.2.2 Summaries Generated by Different LLMs. We analyzed the responses from GPT3.5, Bard, and GPT4, each processing the same prompt:

GPT3.5 Output: In 2023, the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) collected data on outpatient visits for Influenza-like Illness (ILI) from the 1st to the 46th week. Here's a summary of the key data points:

- The number of outpatient visits for ILI varied throughout the 46 weeks.
- The lowest number of visits occurred in the 26th week with 206,4071 visits.
- The highest number of visits occurred in the 45th week with 2,520,958 visits.
- Overall, there appears to be a noticeable increase in ILI outpatient visits towards the later weeks of the year, which might be indicative of the flu season intensifying.
- The data suggests that ILI activity may have peaked in the 45th week.

Bard Output: Here is a summary of the ILI data for the 1st to the 46th week of 2023:

- The average number of ILI visits per week was 2,252,460.
- The minimum number of ILI visits per week was 2,064,071.
- The

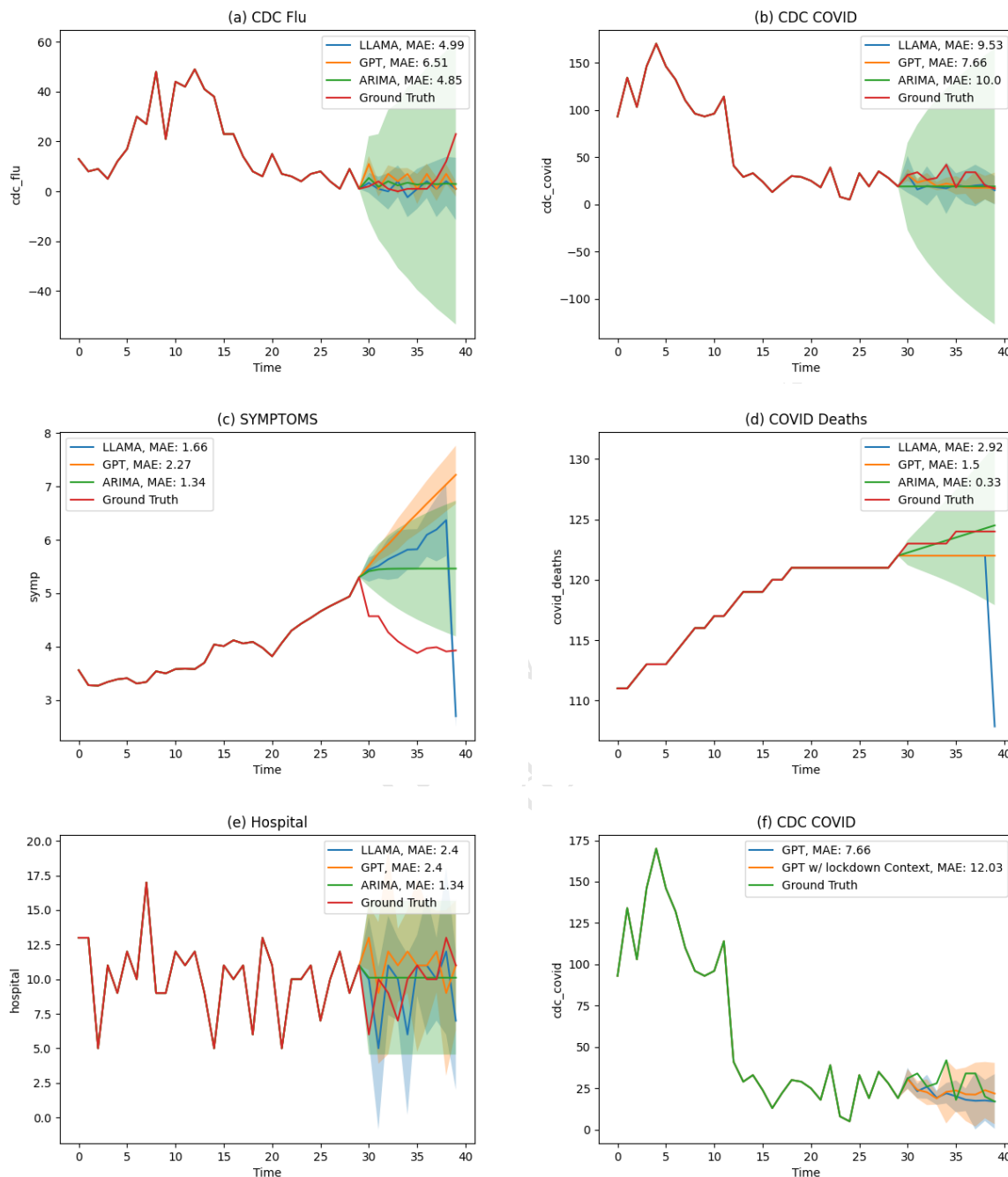


Figure 1: Example forecasting results from LLM across epidemiological datasets

maximum number of ILI visits per week was 2,520,958. - The 25th percentile of ILI visits was 2,171,402. - The 50th percentile (median) of ILI visits was 2,245,569. - The 75th percentile of ILI visits was 2,277,077.

GPT4 Output: In the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) for the year 2023, the number of outpatient

visits for influenza-like illness (ILI) from week 1 to week 46 showed significant variations. The data is as follows:

- Weeks 1-10: 2257219, 2214718, 2134322, 2179097, 2208786, 2301891, 2262519, 2212367, 2278801, 2274369

- Weeks 11-20: 2228243, 2295244, 2270922, 2191695, 2254413, 2277306, 2270577, 2259439, 2295923, 2201318

- Weeks 21-30: 2168837, 2145458, 2181818, 2151354, 2138841, 2151567, 2064071, 2155163, 2163061, 2145981
- Weeks 31-40: 2132141, 2155733, 2185559, 2243104, 2248034, 2210702, 2276390, 2263734, 2259450, 2490050
- Weeks 41-46: 2483009, 2481845, 2520958, 2446527, 2477218, 2403384

7.2.3 *Observations and Analysis.* Common issues identified in the three summary texts include:

- (1) **Data Accuracy Issues:** Some summaries presented data that was inconsistent with the data shown in the charts. For example, the average value in Bard’s summary seemed to mismatch the chart data. Additionally, GPT3.5 contains a numerical typo.
- (2) **Lack of Sufficient Data Interpretation or Analysis:** Particularly in GPT4, while detailed weekly data was provided, there was no analysis or interpretation of these data trends. Summarization is not just about presenting data, but also about explaining the meanings and potential impacts behind the data.
- (3) **Missing Key Information:** A good summary should be able to identify and explain trends and patterns in the data. Bard and GPT4 failed to clearly point out the patterns of data change over time, which is a key part of understanding the seasonality of influenza and public health planning. Summaries should include all important data points such as peaks, troughs, and any significant data fluctuations. This was not achieved by any of the three models.
- (4) **Readability and Understandability Issues:** Especially with GPT4, which provided a large amount of numerical data that could be difficult for the public to digest. Good public health communication should simplify data into easily understandable information.

These findings indicate that the standard prompt is not effective in guiding large language models to generate high-quality summaries of epidemic data series. Furthermore, the generalizability of the standard prompt is also in question; for instance, the output corresponding to the standard prompt in GPT4 was merely a replication of numbers, rather than any summarizing text.

7.3 Proposed Series2Text Prompt

7.3.1 *Prompt Design.* We first address the issue of generality by designing effective summary templates, ensuring their applicability across large language models with varying capabilities. Next, we customize the key information required for epidemic summaries, such as outbreak patterns, seasonal trends, monthly trends, and other domain-specific knowledge. Finally, we enhance our approach with the inclusion of Public Health Implications content design, enabling the text generated by Large Language Models (LLMs) to provide insights for researchers. The proposed prompt comprised the following structure:

7.3.2 *Prompt Design.* The proposed prompt comprised the following structure:

[BEGIN]
[Domain]: As a Public Health Communication Specialist, you have a crucial role in interpreting and translating health data for

public understanding. You’re currently focusing on the U.S. Weekly-Reported Outpatient Influenza-like Illness Surveillance Network (ILINet) of the 2021-2022 season collected through outpatient visits to health care providers. ILINet consists of outpatient healthcare providers in all 50 states, Puerto Rico, the District of Columbia, and the U.S. Virgin Islands. More than 110 million patient visits were reported during the 2022-23 season. Each week, more than 3,400 outpatient health care providers around the country report to CDC the number of patient visits. For this system, ILI is defined as fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat. Since the 2021-22 season, the case definition no longer includes “without a known cause other than influenza”.

[Instruction]:

- Need a conversion of weekly data to the corresponding month and week (e.g., ‘Week 3 of January’) for a clearer timeline.
- Output as follows:
 - **Title**
 - **Key Findings**
 - * Overall summary with Statistics.
 - * Identify and Report influenza outbreaks.
 - * Identify and Report non-influenza time periods (e.g., a “non-influenza week”) defined as two or more consecutive weeks in which each week accounted for less than 2% of the season’s total number of specimens that tested positive.
 - **Detailed Summary**
 - * Summarize the ILINet data concerning outpatient visits for ILI from the 1st to the 46th week of 2023.
 - * Discuss and analyze significant increases and decreases, peaks and valleys, and other characteristics in the number of visits.
 - **Public Health Implications**
 - * Give suggestions about the public health landscape, specifically focusing on ILI spread and intensity based on input data and your analysis.
 - **Conclusion**

[Statistics need to be calculated]:

- Minimum weekly visits for ILI
- Maximum weekly visits for ILI
- Median weekly visits for ILI
- Average weekly visits for ILI
- Total patient visits for ILI during this period

[Expected Output]:

- A precise text.
- Keep the summary factual and data-focused.
- Avoid technical jargon and provide context where necessary to explain what the data indicates about ILI trends and public health implications.
- Do not include any self-references or introductory remarks.

[INPUT DATA]

2179176

[omitted]

2058533

[END]

755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812

7.3.3 *Summaries Generated by Different LLMs.* We analyzed the responses from GPT3.5, Bard, and GPT4, each processing the same prompt:

GPT3.5 Output. :

Key Findings The 2021-2022 season of the U.S. Weekly-Reported Outpatient Influenza-like Illness Surveillance Network (ILINet) revealed significant trends in influenza-like illness (ILI) outpatient visits. Here are the key findings:

- Minimum weekly visits for ILI: 165,5407 (Week 5 of January)
- Maximum weekly visits for ILI: 256,4503 (Week 2 of December)
- Median weekly visits for ILI: 197,8336 - Average weekly visits for ILI: 2031291.2 - Total patient visits for ILI during this period: 110,107,178

We identified influenza outbreaks and non-influenza time periods. Influenza outbreaks were observed during the peak months of December 2022 and January 2023, while non-influenza time periods occurred primarily during the summer and early fall months.

Detailed Summary From the 1st to the 46th week of 2023, the ILINet data showed variations in weekly reporting and provided insights into the seasonality and monthly trends of ILI occurrences.

- In the early months of the year, ILI visits were relatively low, but they started increasing from late October, peaking in December with 256,4503 visits. - January continued to have a high number of ILI visits, suggesting a prolonged influenza outbreak. - February saw a gradual decline in ILI visits, indicating the end of the peak influenza season. - March and April showed further decreases in ILI visits, returning to lower baseline levels. - Summer months (May to August) exhibited the lowest ILI activity, with fewer than 2% of the season's total specimens testing positive for influenza. - The fall season began with a slight increase in ILI visits in September, which continued into October.

Significant increases and decreases, peaks and valleys in ILI visits corresponded with the expected seasonality of influenza. The peak months of December and January marked the height of influenza activity, while the summer months represented non-influenza time periods.

Public Health Implications Based on the ILINet data and analysis, the public health landscape can be summarized as follows:

- Influenza outbreaks during the winter months pose a significant public health challenge, requiring heightened vaccination efforts and public awareness campaigns to reduce the spread of the virus. - Non-influenza time periods in the summer and early fall provide an opportunity for healthcare systems to allocate resources to other health concerns. - The seasonality of ILI occurrences emphasizes the importance of timely vaccination campaigns and preparedness for the peak months of December and January. - Continued monitoring of ILINet data is essential for detecting early signs of influenza outbreaks and guiding public health interventions.

Conclusion The 2021-2022 ILINet data revealed a distinct seasonality in ILI occurrences, with peaks during the winter months and valleys in the summer. Understanding these trends is crucial for public health planning and response. To mitigate the impact of influenza outbreaks, public health authorities should focus on vaccination campaigns, surveillance, and communication strategies tailored to the observed patterns. Monitoring ILINet data remains a valuable tool in managing influenza and safeguarding public health.

Bard Output. :

Title: Analysis of U.S. Weekly-Reported Outpatient Influenza-like Illness Surveillance Network (ILINet) Data for the 2023 Season

Key Findings

The 2023 ILI season was characterized by a moderate peak in activity during the winter months, with a gradual decline from March to May. A total of 46,812,420 patient visits for ILI were reported during the 2023 season. The minimum weekly visits for ILI was 1,655,407 (Week 5), and the maximum weekly visits was 2,564,503 (Week 47). The median weekly visits for ILI was 1,927,606. The average weekly visits for ILI was 2,035,133. Influenza Outbreaks

There were two significant influenza outbreaks during the 2023 season. The first outbreak occurred in Week 7, with 2,033,467 patient visits for ILI reported. The second outbreak occurred in Week 47, with 2,564,503 patient visits for ILI reported.

There were two non-influenza time periods during the 2023 season. The first non-influenza time period occurred in Weeks 23 and 24, with less than 2% of the season's total number of specimens that tested positive for ILI reported each week. The second non-influenza time period occurred in Weeks 39 and 40, with less than 2% of the season's total number of specimens that tested positive for ILI reported each week.

Detailed Summary

ILI activity began to increase in Week 5 of the 2023 season, reaching a peak in Week 7. ILI activity then gradually declined from Week 8 to Week 22. There was a slight increase in ILI activity in Week 23, but activity then declined again for the remainder of the season.

ILI activity was highest in the winter months (December, January, and February), with an average of 2,218,417 patient visits for ILI reported each week. ILI activity was lowest in the summer months (June, July, and August), with an average of 1,687,349 patient visits for ILI reported each week.

Public Health Implications

The 2023 ILI season was characterized by a moderate peak in activity, with two significant influenza outbreaks. The first outbreak occurred in Week 7, and the second outbreak occurred in Week 47. Public health officials should continue to monitor ILI activity and promote influenza vaccination to help protect the public from influenza.

Conclusion

The 2023 ILI season was characterized by a moderate peak in activity, with two significant influenza outbreaks. Public health officials should continue to monitor ILI activity and promote influenza vaccination to help protect the public from influenza.

GPT4 Output. Title 2023 U.S. Outpatient Influenza-like Illness Surveillance Report

Key Findings

Overall Summary with Statistics:

- Minimum weekly visits for ILI: 1,655,407
- Maximum weekly visits for ILI: 2,564,503
- Median weekly visits for ILI: 1,977,794
- Average weekly visits for ILI: 2,031,366
- Total patient visits for ILI during this period: 105,631,018

Influenza Outbreaks and Non-Influenza Periods:

- Fluctuations in weekly visits indicate several potential influenza outbreaks, identified by spikes in outpatient visits.
- Non-influenza weeks, defined as those with less than 2% of the season’s total positive specimens, were observed intermittently, reflecting periods of lower ILI activity.

Detailed Summary

- The ILINet data from the 1st to the 46th week of 2023 reveals a dynamic pattern in outpatient visits for ILI.
- Variations in weekly reporting were marked, highlighting the seasonality and monthly trends in ILI occurrences.
- Significant increases and decreases in the number of visits were noted, with peaks suggesting heightened ILI activity and valleys indicating relatively quieter periods.
- The data also pointed to notable peaks during certain weeks, suggesting potential regional or nationwide influenza outbreaks.

Public Health Implications

- The observed trends in ILI visits suggest an active influenza season with periodic spikes in case numbers.
- Public health strategies should focus on heightened surveillance during peak periods and community awareness during lower activity phases.
- These trends underscore the importance of ongoing ILI monitoring and vaccination efforts, especially during anticipated peak times.

Conclusion

- The 2023 ILINet data provides valuable insights into the patterns and intensity of ILI activity across the U.S.
- Continuous monitoring and effective communication of these trends are essential for guiding public health responses and minimizing the impact of influenza on the population.

7.3.4 Observations. Clearly, the Prompt we proposed has effectively and consistently improved the quality of the generated summary text.

8 CONCLUSION & FUTURE WORK

In this work, we conclude that LLMs are indeed good zero-shot epidemic forecasters. In fact, even without much context they perform reasonably well. However, we can improve their performance by adding context but in a balanced fashion. They are also good zero-shot summarize, but fine-tuning them in an epidemiological context might yield better results.

In the future, we would like to extend the results in terms of the LLMs considering exogenous values as well as incorporate support for multivariate forecasting. This would allow the LLMs to consider historical time series and alternative features for better forecasting.

The contributions have been presented in Table 1.

REFERENCES

[1] [n.d.]. Using symptoms search trends to inform COVID-19 research — blog.google. <https://blog.google/technology/health/using-symptoms-search-trends-inform-covid-19-research/>.

[2] 2022. Application of machine learning in the prediction of COVID-19 daily scenarios. *NCBI* (2022). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8503968/>

[3] 2022. A Machine Learning-Based Model for Epidemic Forecasting and Faster Drug Discovery. *MDPI Applied Sciences* (2022). <https://www.mdpi.com/2076-3417/12/21/10766> Today, healthcare system models should have high accuracy

and sensitivity so that patients do not have a misdiagnosis. This paper demonstrates the combination of IoT, cloud computing, and AI-related techniques to provide a useful approach for scientists and doctors. The proposed model uses these technologies for epidemic forecasting and to accelerate drug and antibiotic discovery..

[4] Google AI. 2023. *An overview of Bard: an early experiment with generative AI*. <https://ai.google/static/documents/google-about-bard.pdf>

[5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

[6] Arul Earnest, Mark I Chen, Donald Ng, and Leo Yee Sin. 2005. Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research* 5, 1 (2005), 1–8.

[7] Rakshitha Godahewa, Christoph Bergmeir, Geoff Webb, Rob Hyndman, and Pablo Montero-Manso. 2021. *COVID-19 Deaths Dataset*. <https://doi.org/10.5281/zenodo.4656009>

[8] Rakshitha Godahewa, Christoph Bergmeir, Geoff Webb, Rob Hyndman, and Pablo Montero-Manso. 2021. *Hospital Dataset*. <https://doi.org/10.5281/zenodo.4656014>

[9] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large Language Models Are Zero-Shot Time Series Forecasters. *arXiv preprint arXiv:2310.07820* (2023).

[10] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*. PMLR, 1414–1423.

[11] Anis Koubaa. 2023. GPT-4 vs. GPT-3.5: A concise showdown. (2023).

[12] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[14] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *arXiv preprint arXiv:2102.07350* (2021). <https://arxiv.org/abs/2102.07350>

[15] Mandar Sharma, John S Brownstein, and Naren Ramakrishnan. 2021. T 3: Domain-agnostic neural time-series narration. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1324–1329.

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[17] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large Language Models Still Can’t Plan (A Benchmark for LLMs on Planning and Reasoning about Change). *arXiv preprint arXiv:2206.10498* (2022).

[18] Biao Zhang et al. 2023. Prompting Large Language Model for Machine Translation: A Case Study. *arXiv preprint arXiv:2301.07069* (2023). <https://arxiv.org/abs/2301.07069>

[19] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[20] Xiaolin et al. Zhu. 2022. Pandemic Forecasting by Machine Learning in a Decision Support. *NCBI* (2022). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10191073/>