TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents (Supplementary Document)

Geon Lee[1], **Wenchao Yu**[2], **Kijung Shin**[1], **Wei Cheng**[2], **Haifeng Chen**[2]

[1]KAIST, [2]NEC Labs

{geonlee0325, kijungs}@kaist.ac.kr, {wyu, weicheng, haifeng}@nec-labs.com

A Designed Prompts for Experiments

We discuss the prompts used in TIMECAP. Each prompt consists of a system prompt and a user prompt. The major prompts used in TIMECAP are described as follows:

- **Contextualization.** In Figure 1, we show the prompt used by $\mathcal{A}_{\text{C}}$ to generate a textual summary about the context of the time series data.
- **Prediction (using time series data).** In Figure 2, we present the template for the prompt that instructs $\mathcal{A}_{\text{P}}$ to predict time series based on the time series data.
- **Prediction (using text summary data).** In Figure 3, we present the template for the prompt that instructs $\mathcal{A}_{\text{P}}$ to predict time series based on the text summary data.
- **Prediction (using in-context examples).** In Figure 4, we present the template for the prompt that instructs $\mathcal{A}_{\text{P}}$ to predict time series based on the text summary data by referring to the in-context examples. The k selected in-context examples are provided prior to the text summary of interest.

B Time Series Contextualization

In this section, we provide examples of contextualized text summaries generated by $\mathcal{A}_{\text{C}}$. Specifically, we present the time series and contextualized summaries for three datasets: **New York** (Figure 5), **S&P 500** (Figure 6), and **Mortality** (Figure 7), one from each domain.

We can observe that the LLM (specifically, $\mathcal{A}_{\text{C}}$) has the capability to understand time series data and generate text summaries that provide insights beyond the raw data. It utilizes domain knowledge and reasoning capabilities to generate these summaries.

C Experimental Settings

In this section, we discuss further details on the experiments.

C.1 Datasets

We provide more details of the datasets we collected for the experiments. The description of each dataset is as follows:

- **Weather**: Datasets in this domain consist of hourly time series data for temperature, humidity, air pressure, wind speed, and wind direction from New York (**NY**), San Francisco (**SF**), and Houston (**HS**). [1] Given the last 24 hours of data, the task is to predict the next-day rain.
- **Finance**: Datasets in this domain contain daily time series data for nine indicators: S&P 500, Nikkei 225, FTSE 100, VIX, Gold Futures, Crude Oil Futures, and exchange rates of EUR/USD, USD/JPY, and USD/CNY. The task is to predict the movement of the S&P 500 (**SP**) or Nikkei 225 (**NK**) prices. Specifically, given the last 20 days of data, the objective is to predict if the price will (1) increase by more than 1%, (2) decrease by more than 1%, or (3) remain otherwise (i.e., between -1% and 1%).
- **Healthcare**: We consider two distinct weekly time series datasets as follows: [2]
  - The mortality (**MT**) dataset includes data on influenza deaths, pneumonia deaths, total deaths, and the mortality ratio from influenza or pneumonia. The task is to predict if the mortality ratio will exceed the average threshold in the coming week.
  - The test-positive (**TP**) dataset includes the total number of specimens tested, positive specimens for Influenza A and B, and their respective ratios. The task is to predict if the ratio of respiratory specimens testing positive for influenza will exceed the average threshold in the coming week.

C.2 Baselines

We discuss more details about the baselines, specifically the specialized models for time series prediction.

- **DLinear** (Zeng et al. 2023) challenges the effectiveness of Transformer-based time series forecasting models. It utilizes two distinct linear layers, independently applied to the trend and seasonal components of the input time series. The outputs from these layers are summed to generate the final prediction.
- **Autoformer** (Wu et al. 2021) incorporates time series decomposition, inspired by classical time series models. It

[1]https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data

[2]https://www.cdc.gov/flu/weekly/overview.htm

features an auto-correlation mechanism that replaces the conventional self-attention layer.

- **Crossformer** (Zhang and Yan 2022) aims to capture cross-channel dependencies for time series forecasting. It segments the input time series into patches and trains a module to capture cross-time and cross-dimension dependencies among these patches.
- **TimesNet** (Wu et al. 2022) employs a 2D CNN to capture temporal dependencies and periodic frequencies of the time series by converting it into a set of 2D tensors.
- **PatchTST** (Nie et al. 2023) segments time series into patches and applies them to the vanilla Transformer encoder. It is channel-independent, where each time series channel is applied individually.

C.3 Hyperprameters

The hyperparameters used for TIMECAP and its competitors are discussed below. For the baselines, we used the default hyperparameters from the Time Series Library [3] unless otherwise specified. For GPT4TS (Zhou et al. 2024) and TimeLLM (Jin et al. 2024), we use their respective open source codes. [4]

- **DLinear:** dropout $\in \{0.0, 0.1, 0.2\}$, moving average $\in \{3, 5\}$, learning rate $\in \{0.0001, 0.001\}$
- **Autoformer:** number of attention layers $\in \{1, 2\}$, number of attention heads $\in \{4, 8, 16\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$
- **Crossformer:** number of attention layers $\in \{1, 2\}$, number of attention heads $\in \{4, 8, 16\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$
- **PatchTST:** number of attention layers $\in \{1, 2\}$, number of attention heads $\in \{4, 8, 16\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$
- **TimesNet:** number of layers $\in \{1, 2\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$
- **PatchTST:** number of attention layers $\in \{1, 2\}$, number of attention heads $\in \{4, 8, 16\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$
- **GPT4TS:** number of attention layers $\in \{1, 2\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$
- **Time-LLM:** number of attention heads $\in \{4, 8, 16\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$

For the multi-modal encoder in TIMECAP, we search the hyperparameters from: number of attention layers $\in \{1, 2\}$, number of attention heads $\in \{4, 8, 16\}$, dropout $\in \{0.0, 0.1, 0.2\}$, learning rate $\in \{0.0001, 0.001\}$ and BERT (`bert-base-uncased`) is used for the Language Model.

[3] https://github.com/thuml/Time-Series-Library

[4] GPT4TS: https://github.com/DAMO-DI-ML/NeurIPS2023-One-Fits-All and TimeLLM: https://github.com/KimMeen/Time-LLM

Table 1: Average F1 scores for different combinations of GPT-3.5 and GPT-4 in the weather domain.

Contextualize ($\mathcal{A}_C$)	Predict ($\mathcal{A}_P$)	F1 Score
GPT-3.5	GPT-3.5	0.370
GPT-3.5	GPT-4	0.423
GPT-4	GPT-3.5	0.533
GPT-4	GPT-4	**0.591**

C.4 Large Language Model

For the OpenAI API, we use the parameters `temperature`=0.7, `max_tokens`=2048, and `top_p`=1. We employed the `gpt-4-1106-preview` version for GPT-4.

This configuration consistently produces robust responses from the LLM, with only minor variations across different trials. As illustrated in Figure **??**, our main experimental results show significant improvements. We conducted a one-tailed t-test comparing (a) predict, (b) contextualize and predict (TIMECP), and (c) contextualize, augment, and predict (TIMECAP). Each enhancement demonstrated a statistically significant improvement with $p < 0.05$.

C.5 Environment

We conducted all the experiments on a machine with RTX 8000 D6 (48GB) GPUs.

D Extra Experiments

We evaluate GPT-3.5 and GPT-4 in two aspects: contextualizing time series data ($\mathcal{A}_C$) and predicting based on the given context ($\mathcal{A}_P$). We measure the performance using different model combinations, and as shown in Table 1, GPT-4 is consistently superior to GPT-3.5 in both aspects. This highlights the importance of the LLM's ability to comprehensively contextualize and accurately predict time series data.

References

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *ICLR*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *ICLR*.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *NeurIPS*.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *AAAI*.

Zhang, Y.; and Yan, J. 2022. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *ICLR*.

System Prompt

Your job is to act as [specific role/job]. You will write a high-quality report that is informative and helps in understanding the current [domain] situation.

User Prompt

Your task is to analyze [description of the time series data]. Review the time-series data provided for the [input length]. Each time-series consists of values separated by a '|' token for the following indicators:

[Time Series Data]

Based on this time-series data, write a concise report that provides insights crucial for understanding the current [domain] situation. Your report should be limited to five sentences, yet comprehensive, highlighting key trends and considering their potential impact on [background]. Do not write numerical values while writing the report.

Figure 1: Prompt for contextualizing time series.

System Prompt

Your job is to act as [specific role/job]. You will be given a time-series data of [data description]. Based on this information, your task is to [task description].

User Prompt

Your task is to [task description]. Review the time-series data provided for the [input length]. Each time-series consists of [resolution] values separated by a '|' token for the following indicators:

[Time Series Data]

Based on this information, respond with either [options]. Do not provide any other details.

Figure 2: Prompt for predicting time series based on the given time series data.

Zhou, T.; Niu, P.; Sun, L.; Jin, R.; et al. 2024. One fits all: Power general time series analysis by pretrained lm.

System Prompt

Your job is to act as [specific role/job]. You will be given a summary of [data description]. Based on this information, your task is to [task description].

User Prompt

Your task is to [task description]. Review the summary provided for the [input length]:

[Text Summary Data]

Based on this information, respond with either [options]. Do not provide any other details.

Figure 3: Prompt for predicting time series based on the given text summary data.

System Prompt

Your job is to act as [specific role/job]. You will be given a summary of [data description]. Based on this information, your task is to [task description].

User Prompt

Your task is to [task description]. First, review the following [number of samples] examples of [domain] summaries and outcomes so that you can refer to when making predictions.

[In-Context Example 1]

⋮

[In-Context Example k]

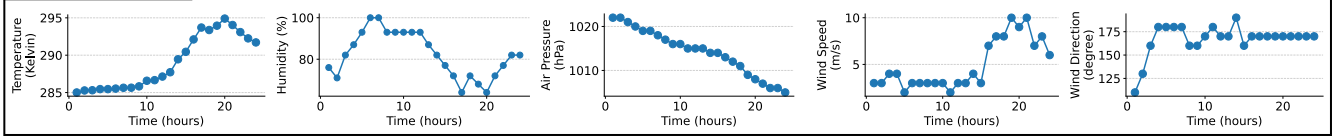
The [situation] is summarized as follows:

[Text Summary Data]

Based on the understanding of the provided examples, predict the outcome of the current [domain] summary. Respond your prediction with either [options]. Response should not include other terms.

Figure 4: Prompt for predicting time series based on the given in-context examples.

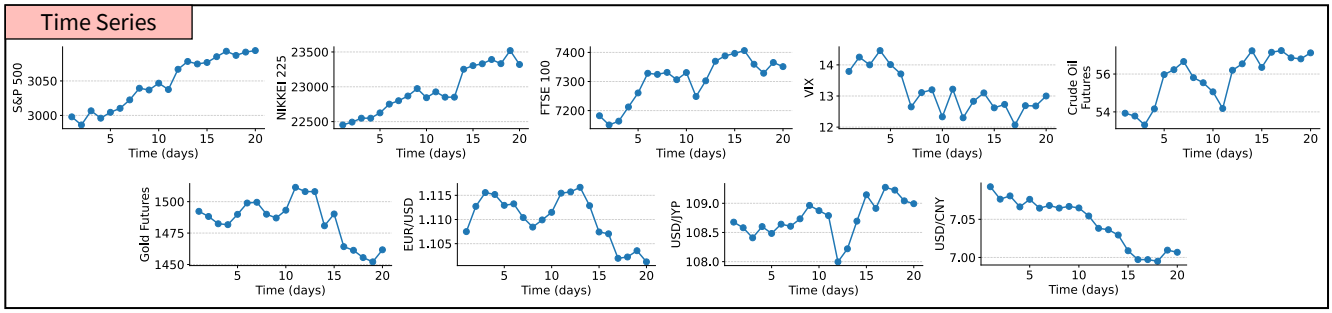
Time Series



Text Summary

Over the past 24 hours, New York City has experienced a gradual increase in temperature, indicating a possible warming trend. Humidity levels have seen fluctuations, reaching full saturation at certain points, which could have led to foggy conditions or precipitation. Air pressure has been consistently decreasing, suggesting an increased likelihood of stormy weather or changes in weather patterns. Wind speeds have generally been mild to moderate, but showed a significant increase in the latter half of the period, which can contribute to cooler perceived temperatures and may signal the approach of a weather system. The wind direction has been predominantly southerly, which typically brings warmer air masses to the region.

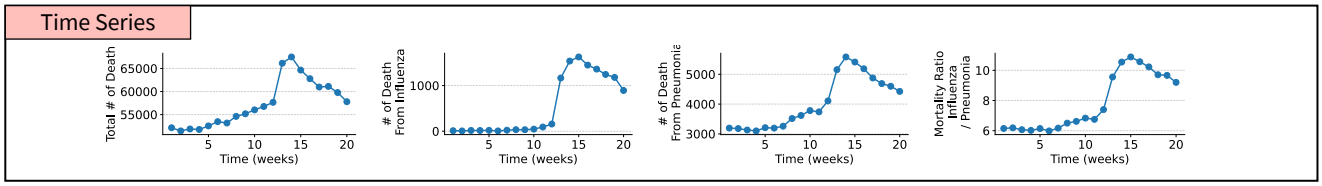
Figure 5: Contextualization of a time series in **New York** in the weather domain.



Text Summary

Over the last 20 market days, the S&P 500 has shown a steady upward trend, increasing from 2997.95 to 3094.04, indicating a positive sentiment among investors and a potential growth period for the US markets. The VIX, commonly known as the "fear index," has hovered around the low-to-mid teens, with a slight increase from 13.79 to 13.00, suggesting that market volatility has remained relatively stable and investor confidence is steady. International markets have seen mixed movements, with Japan's Nikkei 225 experiencing a moderate increase, closing higher at 23319.87 from an opening of 22451.86, while the UK's FTSE 100 displayed slight fluctuations but ultimately dipped from 7182.30 to 7351.20, which could reflect ongoing uncertainties in the region, possibly related to Brexit developments. Commodity prices have seen varying directions, with Gold Futures retracting slightly from 1492.30 to 1461.70, which may indicate a shift in safe-haven demand, while Crude Oil Futures showed a modest increase from 53.93 to 57.12, potentially signaling a cautious optimism in global energy demand. Exchange rates have seen the Euro and the Japanese Yen weaken against the US Dollar, moving from 1.11 to 1.10 for EUR/USD and from 108.68 to 108.99 for USD/JPY respectively, whereas the USD/CNY pair remained relatively stable, signaling a current balance in trade expectations between the US and China amidst ongoing trade discussions.

Figure 6: Contextualization of a time series in **S&P 500** in the finance domain.



Text Summary

Over the last 20 weeks, there has been a significant fluctuation in mortality due to Influenza or Pneumonia in relation to total deaths. Initially, the ratio of mortality from these causes was below the 20-week average of 7.84%, with a gentle increase from 6.15% to 6.83% by the 10th week. However, a notable spike occurred from the 11th week onwards, reaching a peak at the 14th week with a ratio of 10.55%, which correlates with an absolute increase in deaths attributed to Influenza or Pneumonia, particularly Influenza with a dramatic rise from 157 to 1536 deaths between the 11th and 14th weeks. Following this peak, there has been a gradual decline, yet the ratio remains above the average, settling at 9.20% in the 20th week. This trend suggests a potential strain on healthcare resources due to a surge in severe respiratory infections, necessitating increased vigilance and potentially enhanced public health measures.

Figure 7: Contextualization of a time series in **Mortality** in the healthcare domain.