

SUPPLEMENTARY MATERIALS FOR *PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting*

This Supplementary Materials document is organized as follows:

- Section A introduces the hosting and licensing of our PISA dataset.
- Section B presents the *Datasheet for Datasets* for our PISA dataset.
- Section C provides additional implementation details of the methods included in our benchmark.
- Section D conducts an ablation study to investigate the PromptCast performance of two simplified types of prompts.
- Section E investigates the PromptCast performance when the observation and prediction lengths are varied.
- Section F carries out an exploration of applying PromptCast for multivariate time series forecasting.
- Section G conducts case studies to further investigate language models for time series forecasting under the proposed PromptCast.
- Section H further discusses the future of PISA.

A PISA HOSTING AND LICENSING

Source Data.

- CT: The source data for the CT sub-set can be accessed through the Average Daily Temperature Archive¹². Based on its description, this source data is available for research and non-commercial purposes only.
- ECL: The original data is available at UCI Machine Learning Repository¹³. We used a processed version of the original data provided by Informer repository which is licensed under Apache License 2.0¹⁴.
- SG: We access the raw SafeGraph Weekly Patterns data through SafeGraph Data for Academics¹⁵. “SafeGraph¹⁶, a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places, via the SafeGraph Community. To enhance privacy, SafeGraph excludes census block group information if fewer than five devices visited an establishment in a month from a given census block group.” According to the policy, it is against SafeGraph’s terms of service to directly re-share raw SafeGraph data. However, it is acceptable under SafeGraph’s terms of service to share aggregated and derived data, and to include data in chart and visual forms. We strongly recommend users to register SafeGraph Data for Academics before accessing our PISA dataset.

PISA. The PISA dataset and codes for models reported in the benchmark are available at <https://anonymous.4open.science/r/PISA-ICLR23-1474>. In this repository, we also show some generated examples of language models for the PromptCast task. Please note that only validation sets are provided as PISA examples in the above repository during the submission period. After the acceptance decision notification, the full PISA dataset (including train/val/testing sets) will be uploaded to the same repository and publicly available. The full dataset will also be available through HuggingFace Dataset¹⁷, which will make it easier to use our dataset with HuggingFace models.

¹²<https://academic.udayton.edu/kissock/http/Weather/default.htm>

¹³UCI Machine Learning Repository

¹⁴<https://github.com/zhouhaoyi/Informer2020/blob/main/LICENSE>

¹⁵<https://www.safegraph.com/academics>

¹⁶<https://www.safegraph.com/>

¹⁷<https://huggingface.co/datasets>

The PISA dataset will be distributed under Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)¹⁸.

B PISA DATASHEET

B.1 MOTIVATION

For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)

The PISA dataset is created to support the research of novel PromptCast task proposed in this paper.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The authors of this paper created this PISA dataset.

Who funded the creation of the dataset? (If there is an associated grant, please provide the name of the grant or and the grant name and number.)

The creator of the dataset was supported by XXX.

Any other comments?

None.

B.2 COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)

For the CT sub-set, the instances represent the daily temperature of a city. For the ECL sub-set, the instances represent the daily electricity consumption of a user. For the SG sub-set, the instances represent the daily visitor counts of a POI.

How many instances are there in total (of each type, if appropriate)?

The proposed PISA dataset includes 311,932 instances in total. The details of the partitions are presented in Table 1.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).)

For the CT sub-set, 110 international cities are randomly selected to form the dataset. For the ECL sub-set, We filtered users with missing values and randomly selected 50 users (from 321 users) with full records of the entire data collection period. For the SG sub-set, we randomly selected 324 POIs with full records. We list the data value range of each sub-set in Table 1 and show the data distributions in Figure 2.

What data does each instance consist of? (“Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.)

Each instance consists of the input prompt and the output prompt. The input prompt is the input of a model and the output prompt is the desired output of the model. Our PISA provides the input prompts and the corresponding output prompts in separate text files (e.g., val_x_prompt.txt and val_y_prompt.txt).

Is there a label or target associated with each instance? If so, please provide a description.

Yes. As described in the above response, the output prompt is considered as the label.

¹⁸<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Is any information missing from individual instances? (If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.)

There is no missing data (beyond what was intentionally omitted, *e.g.*, the POI geo-location of the SG sub-set).

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? (If so, please describe how these relationships are made explicit.)

Since the sliding window approach is applied (see Section 3.1), the observation data of different instances might belong to the same object-of-interest. However, each instance should be considered and treated as an independent instance with no relationship to other instances in PISA.

Are there recommended data splits (e.g., training, development/validation, testing)? (If so, please provide a description of these splits, explaining the rationale behind them.)

Yes. Each sub-set in our PISA is divided into train/val/test at the ratio of 7:1:2 by the chronological order. Please refer to Section 3.1 and Table 1 for more details about the data splits.

Are there any errors, sources of noise, or redundancies in the dataset? (If so, please provide a description.)

No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)

The PISA dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? (If so, please provide a description.)

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? (If so, please describe why.)

No.

Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? (If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)

N/A.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? (If so, please describe how.)

N/A.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? (If so, please provide a description.)

N/A.

Any other comments?

None.

B.3 COLLECTION PROCESS

How was the data associated with each instance acquired? (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)

The data associated with each instance is acquired and derived from three data sources: CT, ECL, and SG. We have examined and verified the data. These data sources have also been used in the literature forecasting tasks.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? (How were these mechanisms or procedures validated?)

N/A. Our PISA dataset is based on existing data sources. We did not collect the data by ourselves.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

For the CT sub-set, 110 international cities are randomly selected to form the dataset. For the ECL sub-set, We filtered users with missing values and randomly selected 50 users (from 321 users) with full records of the entire data collection period. For the SG sub-set, we randomly selected 324 POIs with full records.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

All collection and annotation was done by the first author.

Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)

The data collection period of the data sources are reported in Table 1. The collection period of CT is 2017/01/01 - 2020/04/30. The collection period of ECL is 2012/01/01 - 2014/12/31. The collection period of SG is 2020/06/15 - 2021/09/05.

Were any ethical review processes conducted (e.g., by an institutional review board)? (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)

N/A. The original data sources are publicly available and have been widely used in the literature.

Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A.

Were the individuals in question notified about the data collection? (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)

N/A.

Did the individuals in question consent to the collection and use of their data? (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)

N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)

N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)

N/A.

Any other comments?

None.

B.4 PREPROCESSING/CLEANING/LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)

Yes. The details of the preprocessing/cleaning/labeling of the raw data sources are described in Section 3.1. The prompting process of our PISA dataset is provided in Section 3.2.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? (If so, please provide a link or other access point to the "raw" data.)

For the CT and ECL sub-sets, the raw data sources are publicly available. For the SG sub-set, due to SafeGraph policy, the raw data will not be provided. Please refer to Section A for more details.

Is the software used to preprocess/clean/label the instances available? (If so, please provide a link or other access point.)

Yes. In the anonymized GitHub repository (<https://anonymous.4open.science/r/PISA-ICLR23-1474>), we provide the codes of transferring the source numerical data to prompts.

Any other comments?

None.

B.5 USES

Has the dataset been used for any tasks already? (If so, please provide a description.)

No. This PISA dataset is designed and introduced for the novel PromptCast task proposed in the paper.

Is there a repository that links to any or all papers or systems that use the dataset? (If so, please provide a link or other access point.)

Considering that the proposed dataset PISA is associated with the novel task PromptCast formulated in this paper, there is no repository for the time being. In the future, we do have a plan to create such a repository to summarize the papers related to this dataset or this task.

What (other) tasks could the dataset be used for?

The PISA dataset could also be used for the conventional numerical-based time series forecasting task.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal

risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)

No.

Are there tasks for which the dataset should not be used? (If so, please provide a description.)

No. We encourage other researchers to try our dataset on other related tasks.

Any other comments?

None.

B.6 DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. During the reviewing period, the dataset (*i.e.*, the validation sets as examples) is available at the anonymized GitHub repository <https://anonymous.4open.science/r/PISA-ICLR23-1474>. The full dataset will be publicly available on the GitHub page for download by all interested third parties for research purpose after the acceptance decision notification.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed through the GitHub page and HuggingFace dataset page. Please refer to Section A for more details.

When will the dataset be distributed?

The full dataset will be made publicly available after the acceptance decision notification.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Our PISA dataset will be distributed under the CreativeCommons Attribution-NonCommercial-ShareAlike license (CC-BY-NC-SA). The terms of this license may be found at <https://creativecommons.org/licenses/by-ncsa/2.0/legalcode>.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No. There are no third party restrictions on the dataset.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No export controls or other regulatory restrictions apply to the dataset. However, we strongly recommend users to register SafeGraph Data for Academics before considering our PISA dataset. Please refer to Section A for more details.

Any other comments?

None.

B.7 MAINTENANCE

Who will be supporting/hosting/maintaining the dataset?

The authors of this paper will support/host/maintain the proposed PISA dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

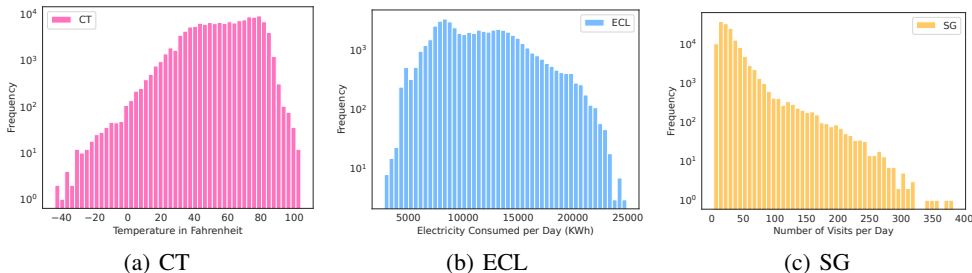


Figure 2: The distribution plots of three sub-sets.

The owner/curator/manager of the dataset can be contacted via: XXX@XXX.

Is there an erratum? If so, please provide a link or other access point.

No erratum for the current version of PISA dataset.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset will be updated in the future by the authors of this paper. The updates will focus on adding new forecasting scenarios, introducing more prompt templates, or expanding PISA to multivariate time series setting. The updated will be communicated to users through the GitHub page and the HuggingFace page.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, all the versions of the dataset will be supported/hosted/maintained by the authors of this paper. The versioning information will be communicated to users through the GitHub page and the HuggingFace dataset page.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

If others want to extend/augment/build on/contribute to the dataset, the authors of this paper should be first contacted. Other researchers are encouraged to pull requests on the GitHub page. We will validate and verify the contributed data before merge the contributed data.

Any other comments?

None.

C DETAILS OF EXPERIMENTS

Data Distribution. As a supplement to the statistics reported in Table 1 (of the main paper), Figure 2 shows the distribution plots of three sub-sets in our PISA dataset. It can be observed that the distributions of the three sub-sets are different. As shown in the figure and according to the value ranges reported in Table 1, these three selected data sources ensure the data diversity of our PISA datasets. Specifically, the proposed PISA covers negative numerical values (the CT sub-set), large values (the ECL sub-set), and small/regular values (the SG sub-set).

Table 7: Details of HuggingFace Pre-trained Models

Model	HuggingFace Key	Pre-trained Model Size
T5	t5-base	891.7 MB
Bart	facebook/bart-base	557.8 MB
Blenderbot	facebook/blenderbot_small-90M	350.4 MB
LED	allenai/led-base-16384	647.7 MB
Pegasus	google/pegasus-xsum	2.3 GB
ProphetNet	microsoft/prophetnet-large-uncased	1.6 GB
Bigbird	google/bigbird-pegasus-large-arxiv	2.3 GB
Electra	google/electra-base-generator	135.0 MB
BERT	bert-base-uncased	440.5 MB
RoBERTa	roberta-base	501.2 MB

HuggingFace Pre-trained Models. In the benchmark, we evaluate 10 language models for PromptCast. The implementations of these 10 models are based on HuggingFace Transformers Library¹⁹. The configuration details are listed in Table 7. By searching <https://huggingface.co/models> with the model key given in the table, the corresponding pre-trained model can be accessed and downloaded.

Implementations. For the evaluated numerical forecasting methods, the implementations are based on the official Autoformer²⁰ repository which also includes the implementation of Transformer and Informer. We follow the default hyperparameter settings except for the *pred_len* parameter (used in Informer and Autoformer). For numerical methods, the data normalization process is also included when we processed the numerical data. In our experiments, the *pred_len* parameter is set to 7 (around half of the observation length, which is 15 in PISA). As for the hyperparameter searching, factor *c* is the one that could affect the forecasting performance of numerical forecasting methods (Xu et al., 2021)). The default factor given by the official implementation is 3 and this default factor has also been used for different datasets according to the official implementation²¹. This justifies that it is a reasonable choice when we evaluate Autoformer performance on our PISA dataset.

For the three different temporal embeddings used in numerical methods benchmarking, we also follow the Autoformer implementations²². Basically, the *timeF* embedding is accomplished via *nn.Linear()* function and the *fixed* and *learned* are based on *nn.Embedding()* function. The *fixed* embedding has fixed non-trainable parameters (similar to the original sin/cos position embedding weight calculation in the vanilla Transformer) for the *nn.Embedding()* layer, whereas the parameters for the *learned* embedding are trainable.

For the language models in the benchmark, their implementations can be grouped into two categories. The first category follows the *EncoderDecoderModel*²³ framework. Three language models are implemented under this category: BERT, RoBERTa, Electra. The rest 7 language models are accomplished through the second category, which is the *ConditionalGeneration* in HuggingFace (e.g., *BartForConditionalGeneration* class²⁴). For the fine-tuning process in PromptCast, it is based on the standard Trainer provided by HuggingFace. Specifically, the sequence-to-sequence trainer is applied as our downstream time series forecasting task in PromptCast is a sequence-to-sequence task. Additionally, no modifications are introduced to the loss function during fine-tuning. For decoding the generation from the language models, the standard tokenizers provided by HuggingFace are used to detokenize the direct output tokens to yield sentences and there are no extra regularization steps on the yielded sentences to acquire numerical predicted outputs. We simply decode the numerical values through string parsing.

¹⁹<https://huggingface.co/docs/transformers/index>

²⁰<https://github.com/thuml/Autoformer>

²¹<https://github.com/thuml/Autoformer/tree/main/scripts>

²²<https://github.com/thuml/Autoformer/blob/main/layers/Embed.py>

²³https://huggingface.co/docs/transformers/model_doc/encoder-decoder

²⁴https://huggingface.co/docs/transformers/v4.19.2/en/model_doc/bart#transformers.BartForConditionalGeneration

Table 8: Basic prompting templates.

		Template
CT	Input Prompt (Source)	The average temperature of was $\{x_{t_1:t_{\text{obs}}}^m\}$ degree on each day. What is the temperature going to be on tomorrow?
	Output Prompt (Target)	The temperature will be $\{x_{t_{\text{obs}}+1}^m\}$ degree.
ECL	Input Prompt (Source)	The client consumed $\{x_{t_1:t_{\text{obs}}}^m\}$ kWh of electricity on each day. What is the consumption going to be on tomorrow?
	Output Prompt (Target)	This client will consume $\{x_{t_{\text{obs}}+1}^m\}$ kWh of electricity.
SG	Input Prompt (Source)	There were $\{x_{t_1:t_{\text{obs}}}^m\}$ people visiting the POI on each day. How many people will visit the POI on tomorrow?
	Output Prompt (Target)	There will be $\{x_{t_{\text{obs}}+1}^m\}$ visitors.

We would like to emphasize that the language models and numerical forecasting methods are treated and processed equally and fairly in our benchmark. There is also no specific hyperparameter tuning for language models under PromptCast. For both language models and numerical models, we all use the default settings provided/recommended by the official implementations. Thus, we believe that the comparison is fair and convincing. This hyperparameter searching-free characteristic could also reflect another benefit of PromptCast in real-world applications, that is, no need to conduct complicated and time-consuming hyperparameter tuning processes. The forecasting models can then be deployed more quickly for new forecasting scenarios.

All the above scripts are provided in the anonymized GitHub repository²⁵. The experiments were performed with PyTorch on a Linux server equipped with Nvidia V100 GPUs. Note that in our experiments, only 1 GPU is enabled per run for each method.

D PROMPTS ABLATION STUDY

To further investigate the prompts in the PromptCast task, we conduct an ablation study on our prompting templates and two simplified prompts are developed as follows.

- **Basic Prompt:** as shown in Table 8, we remove auxiliary information (*e.g.*, date information) and only keep the core information (*i.e.*, the historical observation values $x_{t_1:t_{\text{obs}}}^m$) in the input prompts. The output prompts remain the same.
- **Minimum Prompt:** This is the simplest and the most straightforward version of prompts. We use commas to convert the sequential numerical values $x_{t_1:t_{\text{obs}}}^m$ into comma-delimited strings as the input prompts (*e.g.*, ``78, 81, 83, 84, 84, 82, 83, 78, 77, 77, 74, 77, 78, 73, 76''). For the output prompts, the prediction targets $x_{t_{\text{obs}}+1}^m$ are directly used as single-word strings (*e.g.*, ``78'').

The prediction results of using these two types of prompts on our PISA dataset are presented in Table 9. We can notice that both two simplified prompts lead to worse performance compared to the default template reported in the main paper (Table 2). Although the minimum prompts can yield reasonable results, for the same language model, the performance of the basic prompt outperforms the performance of the minimum prompt. These comparison results demonstrate that: (1) including proper contexts (even if these contexts introduce no extra data, *e.g.*, the basic prompt) in the prompt is beneficial; and (2) the auxiliary information is a significant component in the prompt for better prediction performance.

E VARYING OBSERVATION AND PREDICTION LENGTHS

The proposed language foundation models-based PromptCast forecasting paradigm is also suitable with varying observation lengths (*i.e.*, input lengths) and prediction horizons (*i.e.*, output lengths). For the multi-step forecasting (*i.e.*, larger prediction horizons), the question part of the input prompt

²⁵<https://anonymous.4open.science/r/PISA-ICLR23-1474>

Table 9: Results of language models with different prompt types.

Prompt	Model	CT				ECL				SG			
		RMSE		MAE		RMSE		MAE		RMSE		MAE	
		mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Basic	Bart	6.512	0.016	4.790	0.012	603.571	2.461	380.583	1.695	8.677	0.055	5.912	0.013
	Pegasus	6.496	0.012	4.767	0.007	595.942	2.246	366.677	1.373	8.530	0.032	5.879	0.010
	Bigbird	6.478	0.033	4.752	0.019	609.315	3.071	378.100	2.112	8.576	0.035	5.922	0.009
Minimum	Bart	6.564	0.082	4.830	0.083	612.487	3.660	382.931	2.509	8.686	0.117	5.937	0.021
	Pegasus	6.560	0.049	4.818	0.027	597.459	3.763	368.174	1.178	8.632	0.046	5.915	0.016
	Bigbird	6.496	0.042	4.766	0.026	618.907	4.422	384.327	0.861	8.774	0.065	5.950	0.015

template can be updated to “how many people will visit POI in the next Y days?” and the output prompt is also needed to be revised to reflect the multi-step future data values such as “For day X to day Y, there will be A, B, C, ... visitors”. Here is an example of predicting the next 7-time steps with the SG sub-set. **Input Prompt:** From April 24, 2021, Saturday to May 08, 2021, Saturday, there were 6, 8, 6, 15, 12, 4, 13, 7, 8, 12, 16, 9, 11, 18, 10 people visiting POI 1 on each day. How many people will visit POI 1 in the next 7 days? **Output Prompt:** From May 09, 2021, Sunday to May 15, 2021, Saturday, there will be 11, 11, 8, 13, 10, 18, 19 visitors. Based on these modified prompts, we examine the three language models (Bart, Bigbird, Pegasus) with different observation length settings (7 days and 15 days) and prediction horizons (1 day, 4 days, and 7 days), resulting in a total of 6 length configuration combinations. The performance of these three language models are listed in Table 10.

From the table, we can see that PromptCast can be easily adapted to multi-step forecasting while using the exact same forecasting model architectures (*i.e.*, the language foundation models) for different settings. Generally, compared to the single-step forecasting performance, increasing the prediction horizon leads to larger missing rates. Specifically, Bart shows a worse performance as evidenced by the large missing rates on different sub-sets. However, Bigbird and Pegasus can still yield reasonably good predictions with extreme small missing rates, which demonstrates language models are still powerful in predicting multiple steps ahead. We also notice that when predicting the same future length, using 15 days as the observation often leads to better performance. As a 15-day period could cover weekly patterns in the observation, it is beneficial for yielding more accurate forecasting.

To further provide a glimpse of the multi-step forecasting performance, more examples of three sub-sets of predicting the next 7 days setting are also available in the anonymized GitHub repository²⁶. This exploration of PromptCast on different length settings illustrates that our proposed PromptCast is robust to dynamic observation/prediction lengths and presents a promising research direction for time series forecasting.

F MULTIVARIATE TIME SERIES FORECASTING

As the first attempt to leverage language models for time series forecasting, we take the basic forecasting setting to demonstrate the concept of PromptCast. However, we would like to emphasize that the proposed new paradigm is flexible and suitable for other forecasting settings such as the multivariate time series forecasting setting. We only need to update the prompt templates accordingly when PromptCast is adapted to multivariate time series forecasting setting. Under the new setting, although input/output prompts might be updated, the core prediction model (*i.e.*, language foundation models) in PromptCast can remain the exact same (prompt-agnostic) and same pre-trained weights can also be used. For example, in the experiments under Section D, different prompts are applied. But we can also directly use the same models (without any changes to the model structures) to yield predictions. For conventional numerical forecasting methods, however, necessary updates on the deep learning model (*e.g.*, update the encoder part to fit multivariate features) must be introduced when the forecasting setting is changed. Additionally, processes such as extra hyperparameter search are often

²⁶https://anonymous.4open.science/r/PISA-ICLR23-1474/Dataset/PISA_Plus_Multistep_examples/README.md

Table 10: Results of language models with different length settings.

Observation	Prediction		RMSE	CT MAE	Missing Rate
7 Days	1 Day	Bart	6.762±0.090	5.017±0.083	0
		BigBird	6.425±0.024	4.716±0.013	0
		Pegasus	6.820±0.044	5.088±0.031	0
7 Days	4 Days	Bart	10.087±0.064	7.478±0.058	1.622%±0.721%
		BigBird	9.213±0.117	6.880±0.073	0
		Pegasus	9.422±0.080	7.003±0.047	0
7 Days	7 Days	Bart	11.681±0.183	8.620±0.058	22.745%±1.926%
		BigBird	10.200±0.120	7.601±0.083	0.036%±0.018%
		Pegasus	10.144±0.188	7.526±0.131	0.022%±0.014%
15 Days	1 Day	Bart	6.432±0.040	4.759±0.027	0
		Bigbird	6.351±0.016	4.707±0.019	0
		Pegasus	6.379±0.023	4.727±0.014	0
15 Days	4 Days	Bart	10.135±0.043	7.631±0.033	1.799%±1.651%
		BigBird	9.432±0.136	7.075±0.096	0
		Pegasus	9.582±0.118	7.166±0.085	0
15 Days	7 Days	Bart	12.033±0.605	8.773±0.109	27.107%±1.159%
		Bigbird	10.329±0.162	7.745±0.120	0.023%±0.031%
		Pegasus	10.522±0.187	7.899±0.142	0.179%±0.265%

Observation	Prediction		RMSE	ECL MAE	Missing Rate
7 Days	1 Day	Bart	527.772±7.282	360.066±3.688	0
		BigBird	530.929±4.792	365.122±2.184	0
		Pegasus	527.735±5.921	360.702±2.352	0
7 Days	4 Days	Bart	1075.721±218.227	542.006±16.379	6.791%±2.949%
		BigBird	750.708±28.990	493.466±9.704	0.047%±0.015%
		Pegasus	741.931±9.613	482.426±4.749	0.120%±0.116%
7 Days	7 Days	Bart	1332.287±266.259	612.469±32.104	29.310%±20.472%
		BigBird	946.462±57.607	551.675±6.601	0.537%±0.314%
		Pegasus	851.768±8.422	545.536±2.627	3.669%±1.086%
15 Days	1 Day	Bart	527.350±10.608	355.390±2.751	0
		Bigbird	519.665±3.440	350.699±1.953	0
		Pegasus	537.186±11.296	361.135±4.728	0
15 Days	4 Days	Bart	855.906±60.640	528.250±15.276	25.701%±3.393%
		BigBird	700.280±14.048	481.145±7.026	0.041%±0.032%
		Pegasus	769.008±17.279	476.013±3.494	0.096%±0.112%
15 Days	7 Days	Bart	1190.564±111.116	611.117±55.916	50.259%±28.139%
		Bigbird	1063.650±222.601	541.538±7.397	0.329%±0.190%
		Pegasus	835.386±23.031	525.537±6.067	1.422%±0.465%

Observation	Prediction		RMSE	SG MAE	Missing Rate
7 Days	1 Day	Bart	8.427±0.016	5.909±0.023	0
		BigBird	8.526±0.052	5.962±0.026	0
		Pegasus	8.406±0.063	5.957±0.028	0
7 Days	4 Days	Bart	10.156±0.164	6.982±0.101	2.655%±1.285%
		BigBird	9.619±0.824	6.654±0.495	0
		Pegasus	8.813±0.039	6.142±0.005	0
7 Days	7 Days	Bart	12.453±0.894	8.375±0.169	32.716%±5.301%
		BigBird	10.851±0.612	7.433±0.410	0.034%±0.026%
		Pegasus	8.926±0.017	6.247±0.011	0.013%±0.005%
15 Days	1 Day	Bart	8.279±0.053	5.785±0.023	0
		Bigbird	8.326±0.048	5.841±0.031	0
		Pegasus	8.289±0.016	5.817±0.013	0
15 Days	4 Days	Bart	9.476±0.079	6.559±0.029	1.212%±0.325%
		BigBird	9.392±0.571	6.436±0.313	0
		Pegasus	8.743±0.062	6.082±0.026	0
15 Days	7 Days	Bart	11.048±0.325	7.653±0.129	38.328%±5.362%
		Bigbird	9.731±0.427	6.658±0.200	0.007%±0.011%
		Pegasus	8.802±0.031	6.139±0.016	0.011%±0.006%

Table 11: Examples of two types of prompts under the multivariate time series setting.

Input Prompt A&B	From May 13, 2016, Friday to May 27, 2016, Friday, the PM2.5 was 32, 23, 20, 53, 82, 113, 133, 94, 64, 83, 20, 4, 18, 48, 57; PM10 was 32, 40, 57, 87, 90, 113, 133, 94, 64, 83, 20, 20, 18, 48, 103; and SO2 was 2, 3, 4, 11, 29, 27, 26, 33, 13, 14, 6, 2, 2, 6, 16 on each day. What are the pollutant values going to be on May 28, 2016, Saturday?
Output Prompt A	The pollutant values will be 3, 23, 2.
Output Prompt B	The PM2.5 will be 3. The PM10 will be 23. The SO2 will be 2.
Input Prompt C	From May 13, 2016, Friday to May 27, 2016, Friday, the PM2.5 was 32, 23, 20, 53, 82, 113, 133, 94, 64, 83, 20, 4, 18, 48, 57 on each day. What is the PM2.5 value going to be on May 28, 2016, Saturday? From May 13, 2016, Friday to May 27, 2016, Friday, the PM10 was 32, 40, 57, 87, 90, 113, 133, 94, 64, 83, 20, 20, 18, 48, 103 on each day. What is the PM10 value going to be on May 28, 2016, Saturday? From May 13, 2016, Friday to May 27, 2016, Friday, the SO2 was 2, 3, 4, 11, 29, 27, 26, 33, 13, 14, 6, 2, 2, 6, 16 on each day. What is the SO2 value going to be on May 28, 2016, Saturday?
Output Prompt C	The PM2.5 will be 3. The PM10 will be 23. The SO2 will be 2.

Table 12: Results of language models on multivariate time series forecasting.

Prompt		RMSE	MAE	MissingRate
N/A	CopyYesterday	73.058	43.274	N/A
	HistrocialAverage	63.289	40.772	N/A
	CopyLastWeek	87.250	52.796	N/A
Prompt A	Bart	79.218±4.948	46.553±2.838	0
	BigBird	75.783±1.988	45.590±2.105	2.681%±3.471%
	Pegasus	82.859±0.588	48.168±0.444	0
Prompt B	Bart	75.268±2.150	41.869±0.915	0
	BigBird	81.510±2.524	48.913±0.639	0
	Pegasus	75.706±1.736	45.206±1.156	0
Prompt C	Bart	68.282±2.345	40.690±1.131	0
	BigBird	67.729±1.880	41.547±1.071	0
	Pegasus	67.834±0.859	39.893±0.691	0

required when the model structure is modified. This could also reflect the “code less” benefits and robustness of the proposed PromptCast paradigm.

To investigate the capability of the proposed PromptCast on multivariate forecasting, we further conduct a pilot study of multivariate PromptCast on Beijing Air Quality Data²⁷ in addition to the three sub-sets in the main PISA dataset. The data collection period of this Air Quality (AQ) set is from 2013/03/01 to 2017/02/28. Similar to the main PISA, we split this period into training set (2013/03/01-2015/12/18), validation set (2015/12/19-2016/05/12), and test set (2016/05/13-2017/02/28). After filtering missing values, we focus on three types of air pollutants: PM2.5, PM10, and SO2, which means the feature dimension is 3.

To apply PromptCast for multivariate forecasting, we update the prompt template to include multiple features as Table 11. In total, we design and examine three different prompts. For Prompt A and B, the input prompt consists of the description of all 3 features and the main difference of these two prompts is in the output prompt part. Output Prompt A jointly describes the three features, whereas the three feature values are given separately in three short sentences in Output Prompt B. Prompt C can be seen as an extended version of the prompt template (*e.g.*, Table 2) for the univariate forecasting. It decomposes the multivariate features (*i.e.*, three pollutants in this AQ dataset) into several univariate features.

The results of PromptCast with language foundation models using different prompts are reported in Table 12. Additionally, we list the performance of three naive forecasting methods as baselines. From the table, we can observe that: (1) All three language models can generate plausible predicted future descriptions as evidenced by almost all zero missing rates. (2) The design of prompts could definitely affect the prediction performance. Specifically, Prompt C leads to better performance than Prompt A and B on this multivariate forecasting dataset. (3) The naive historical average forecasting outperforms PromptCast by a small margin. However, using Prompt C still shows a comparable performance. We think the potential reason of why PromptCast is not the best performer in this comparison is because that the current prompts cannot fully characterize the relation of different features. This also explains why Prompt C outperforms the other two prompts.

Open Question and Future Work.

The above analysis points out an open question for PromptCast, that is, how to design prompts for the more challenging multivariate time series. One potential working direction is to learn the internal correlations between different temporal features and then develop prompts to represent the learned correlations. In the future, we will focus on this direction and also explore techniques such as learnable prompts for the multivariate PromptCast. We hope the proposed PISA dataset and the corresponding benchmark could encourage other researchers as well to investigate this interesting PromptCast topic.

G CASE STUDIES

To further investigate the why language models can work well for forecasting time series data under the proposed PromptCast setting, we conduct two case studies and visualize the attentions (between the input sentence and the output sentence) learned by the language models (using Bart model in these case studies as examples).

- *Case Study 1*
 - **Input:** From September 06, 2019, Friday to September 20, 2019, Friday, the average temperature of region 1 was 65, 66, 70, 71, 71, 77, 78, 65, 70, 76, 74, 70, 64, 61, 64 degree on each day. What is the temperature going to be on September 21, 2019, Saturday?
 - **Predicted:** The temperature will be 68 degree.
 - **Ground Truth:** The temperature will be 71 degree.
- *Case Study 2*

²⁷<https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data>

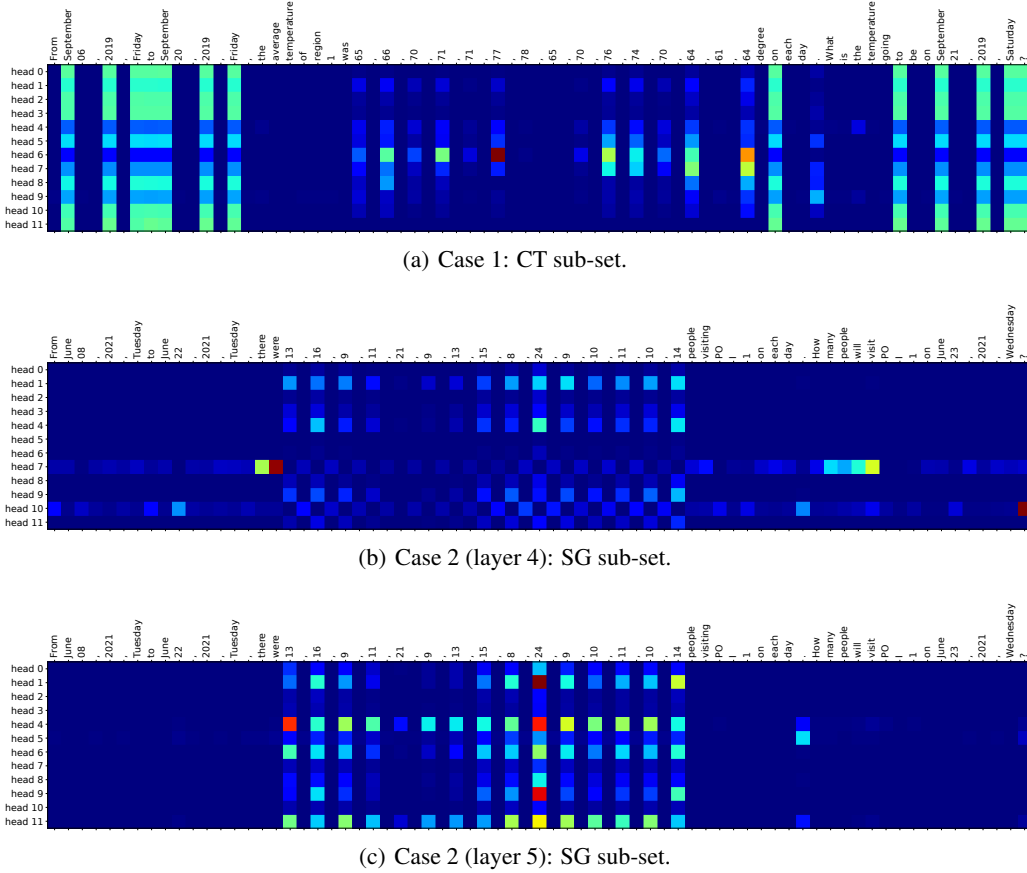


Figure 3: Visualizations of attentions in language models under the proposed PromptCast setting.

- **Input:** From June 08, 2021, Tuesday to June 22, 2021, Tuesday, there were 13, 16, 9, 11, 21, 9, 13, 15, 8, 24, 9, 10, 11, 10, 14 people visiting POI 1 on each day. How many people will visit POI 1 on June 23, 2021, Wednesday?
- **Predicted:** There will be 11 visitors.
- **Ground Truth:** There will be 12 visitors.

Specifically, in Figure 3, we show the attentions between the numerical predicted values in the generated outputs and all the tokens in the input prompts. For each heatmap plot in which a hotter region means a larger attention value, the horizontal axis stands for the input prompt (in the token format) and the vertical axis represents different attention heads.

For Case 1 (Figure 3 (a)), we can clearly observe that head 6 pays more attention to the numerical sequential data in the input prompt while the other heads have higher attention to other semantic auxiliary information such as *September* and *Friday*. Moreover, head 7 shows attentions to the input numerical historical observations (e.g., 76, 74, and 64) and semantic tokens simultaneously. For Case 2, the attention visualization of two layers are displayed in Figure 3 (b) and Figure 3 (c). It can be noticed that the learned attentions between the predicted value and the numerical historical values in the input prompt are mainly in layer 5 (Figure 3 (c)) of the Bart model. Compared to layer 5, layer 4 (Figure 3 (b)) More specifically, head 7 demonstrates large attentions to semantic tokens like *there were* and *How many people will visit* whereas head 10 highlights the auxiliary date token 22 as well as the punctuation token ?.

Based on the analysis of these two studies, we show that the language models can jointly learn the relation of numerical value tokens (historical values) at different time steps and the influence of the semantic tokens under the PromptCast paradigm. It further justifies the rational of leveraging language models for forecasting time series data.

H FURTHER DISCUSSIONS

About Evaluation Metrics.

In addition to the typical RMSE and MAE metrics for evaluating the quality of predicted numerical values, we introduce a new Missing Rate metric in PromptCast. We have also considered other metrics (e.g., BLEU score) that are often used in the NLP domain. However, we noticed that different language models have very close BLEU scores. This is because the generated sentences follow the structure of the output prompt (e.g., There will be X visitors) after fine-tuning. It can be observed in the provided examples in our repository: the only difference between the generated outputs and the ground truth sentences is the numerical values (predicted numbers), which indicates that using RMSE/MAE can well evaluate the prediction performance in PromptCast. Thus, we decided not to use BLEU in the current stage of PromptCast. With the development of PromptCast in the future, we will revisit the evaluation system to investigate the need of introducing other metrics.

Future Source Datasets.

The current version of PISA includes three different real-world forecasting scenarios. In total, the three sub-sets in PISA have 311,932 data instances based on raw data collected from 2,760 days. The future versions of PISA datasets will be designed to include more and larger datasets such as the multivariate Air Quality. Additionally, more challenging heterogeneous datasets that include multiple data sources could also be considered for PromptCast. We also welcome other researchers to contribute to the PISA datasets to cover more diverse forecasting scenarios.