Generative AI for Energy: Multi-Horizon Power Consumption Forecasting using Large Language Models

Kevin Roitero
kevin.roitero@uniud.it
University of Udine
Udine, Italy

Gianluca D'Abrosca
dabrosca.gianluca@spes.uniud.it
University of Udine
Udine, Italy

Andrea Zancola
azancola@acegasapsamga.it
AcegasApsAmga SpA, Hera Group
Italy

Vincenzo Della Mea
vincenzo.dellamea@uniud.it
University of Udine
Udine, Italy

Stefano Mizzaro
mizzaro@uniud.it
University of Udine
Udine, Italy

Abstract

We leverage generative NLP-based models, specifically Transformer-Based models, for multi-horizon univariate and multivariate power consumption forecasting. We apply our approach to various datasets, focusing on short-term (1 day) and long-term (1 week) forecasts. We test several lag configurations with and without additional contextual information and achieve promising results. We evaluate the forecasts' effectiveness using a range of metrics, and aggregate the results on a monthly basis for a comprehensive understanding of the performance throughout the year.

CCS Concepts

• **Computing methodologies** → **Natural language generation**.

Keywords

Transformers, Power Consumption Forecasting, Time Series.

ACM Reference Format:
Kevin Roitero, Gianluca D'Abrosca, Andrea Zancola, Vincenzo Della Mea, and Stefano Mizzaro. 2024. Generative AI for Energy: Multi-Horizon Power Consumption Forecasting using Large Language Models . In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679933

1 Introduction

The growing need for efficient energy management and sustainable resource use has intensified the research looking into accurate forecasting of power consumption. Traditional methods like ARIMA and SARIMA have long been used for prediction [9, 16, 24]. However, these methods struggle when dealing with large datasets, especially for medium- and long-term forecasts, due to issues like non-stationarity and non-linearity in the data [11, 25]. Consequently, advanced Deep Learning (DL) models are becoming popular in this field. Among all models and architectures, Transformer-based models, originally introduced for natural language processing (NLP) tasks, have been recently recognized for their potential in temporal sequence modeling and forecasting [20, 22, 27, 28].

This paper investigates the use of Transformer-based models, particularly the T5 model, for multi-horizon univariate power consumption forecasting, covering short and long-term forecasts and evaluating different model's configurations and the impact of contextual information. By customizing modeling and training techniques, the study improves the T5 model's performance on time series forecasting tasks. To the best of our knowledge, this work is the first to use generative NLP for time series energy power consumption forecasting.

The code needed to replicate our experiments is made available to the research community.[1]

2 Background and Related Work

The study of time series forecasting, particularly within the context of energy consumption, has evolved significantly with advancements in machine and deep learning technologies. Initially dominated by statistical methods such as ARIMA and Exponential Smoothing [9, 16, 24], the focus has shifted towards more complex models capable of handling complex patterns and external features present in large datasets [13, 15, 23].

Traditional approaches in this field often rely on Recurrent Networks (RNNs) and their advanced forms, such as Long Short-Term Memory (LSTM) networks [10] or Gated Recurrent Units (GRU) [5]. these models are especially well-suited for modeling scenarios where the prediction of a current value depends critically on previous observations. Recent developments have seen Transformer architectures being adapted for time series forecasting [13]. These architectures offer a complementary approach to classical RNN-based methods, particularly benefiting from the Transformer's ability to handle dependencies within sequences. The flexibility of Transformers has led to their successful application in areas with complex contextual patterns, such as human mobility forecasting [27, 28] and time series classification [22]. However, the direct application of vanilla Transformer models to time series forecasting, particularly in settings with contextual features, remains challenging due to the absence of large-scale, domain-specific training datasets and the need for specialized model architectures [28].

[1] https://osf.io/vmkeu/?view_only=4b3a7fc3ed9045eeaa13e26803907d7c

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679933

Table 1: Summary of datasets used in this study.

Dataset Acronym	Temporal Range	Sampling Periodicity	Training Set Size	Test Set Size
AAA	2014-2018	15 min	140,160	35,040
Spain	2015-2017	1 hour	17,520	8,760
ELD	2011-2013	15 min	70,080	35,040

Our approach builds on the concept of causal language models, specifically leveraging the T5 model [19]. Unlike masked language models like BERT [7], which learn bidirectional contexts, causal models operate under a unidirectional framework. This makes them particularly suitable for time series forecasting, where each prediction is contingent solely on preceding data points and possibly on external features.

3 Experimental Setting

3.1 Aims

We aim to predict future energy consumption using past observations, known as *lags*. Each Transformers-based model is fed with the maximum permissible past timestamps constrained by the model's context size limit, whereas for each baseline model we perform a heuristic search over the lags, epochs, and model parameters and report the scores for the model with higher effectiveness. We train the models under two settings: using only previous lags and incorporating external information including time and weather features. Our goal is to assess the models' effectiveness across two different future horizons. The first forecast, *t+1 day* is achieved considering t+96 for datasets with 15-minute intervals and t+24 for those with hourly intervals; the second long-range forecast, *t+1 week*, corresponds to t+672 for 15-minute intervals and t+168 for hourly intervals datasets.

3.2 Data

We rely on three distinct datasets, each with different characteristics and temporal ranges. Two of these datasets are publicly available, while the third one is proprietary. Table 1 provides the statistics for the datasets used. The first dataset, referred to as the *AAA* dataset, originates from AcegasApsAmga S.p.A., an Italian company that is part of the Hera Group. The company provides distribution services for gas, electricity, water, environmental, and energetic services to approx. 3.4 million citizens and businesses across more than 300 municipalities. This proprietary dataset includes data collected from 2014 to 2018 for energy pods, detailing power consumption on a 15-minute interval and includes the weather information. We use data from 2014 to 2017 for training and 2018 for test.

The second dataset, publicly available and referred to as the *Spain* dataset,[2] is composed of hourly power consumption data from Spain spanning from 2015 to 2017 for training and data from 2018 for testing. It includes data about electrical consumption, generation, pricing, and weather. The third dataset, also publicly available, is referred to as the *ELD* dataset.[3] It includes electricity consumption data from 370 clients from 2011 to 2013 for training and data from 2014 for testing. This dataset contains consumption values for Portuguese clients given in kWh for each 15-minute interval. We specifically focus on the "MT_124" client, the one containing the most data points.

3.3 Measures

We evaluate our models' effectiveness using several standard error metrics, which are assessed monthly and then aggregated by computing the mean and standard deviation over the test year, providing a detailed view of the models' annual performance. The metrics used include Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Maximum Error (ME), and Root Mean Squared Error (RMSE). These metrics assess different aspects of models' effectiveness. MAE provides a straightforward measure of overall error magnitude, MAPE expresses these errors as a percentage offering a scale-relative assessment which is useful for comparing different dataset and scales, ME identifies the largest single forecast error highlighting the worst-case, and RMSE penalizes larger errors more than smaller ones.

3.4 Deep Learning Baselines

We rely on several DL baselines for time series prediction. While traditional methods like ARIMA and SARIMA have been effective for simpler, smaller datasets, they struggle with the non-linearities and non-stationarities typical in large-scale time series data. DL models, on the other hand, are able to model complex patterns and interactions at scale and can integrate multiple types of input features which allows to handle external data and handle missing values [3, 12]. The set of baselines are implemented relying on the PyTorch-Forecasting[4] framework.

We use Long Short-Term Memory (LSTM) [29] and Gated Recurrent Unit (GRU) [8] networks. We also consider the Neural Basis Expansion Analysis for Interpretable Time Series Forecasting (NBeats) [15], which relies on a deep stack of fully connected layers to predict future time windows based on past data, the Temporal Fusion Transformer (TFT) [13] model which employs attention mechanisms to capture long- and short-range dependencies within time series data and, finally, the Deep Autoregressive model (DeepAR) model [23] which leverages recurrent networks to approximate and forecast time series evolution.

3.5 Proposed Models

Our set of models were developed on the PyTorch and HuggingFace platforms, starting from the T5-base model[5], which features a 12-block transformer architecture for a total of 220M parameters [19]. We model the whole problem of time series forecasting as a sequence-to-sequence problem [14]. Thus, we convert the time-series data into textual format, a method which has proven to be effective in different domains such as diagnostic texts [6, 17, 21], human mobility forecasting [26–28], and time series classification [22]. Specifically, we transform past observations and contextual

[2] https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather

[3] https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014

[4] https://pytorch-forecasting.readthedocs.io/en/stable/.

[5] https://huggingface.co/t5-base.

features (i.e., date, time, and weather data) into strings, which we then input to the model. We can represent our general prompting as (braces indicate values while square brackets optional parts):

> context: {contextual features}.
> previous observations:{value} [at {time}], . . ., {value} [at {time}].

E.g., if the contextual features are temperature and humidity and if we rely on 50 lags, a possible instance might be:

> context: temperature 27C, humidity 75%.
> previous observations: 0.85 at time t-1, . . ., 0.10 at time t-50.

We develop the following model variations. T5, the base model as detailed above. We use as the contextual features (i.e., time, date, and weather) and we fill the rest of the context with lags (i.e., previous values of the series). We did try with both the base and large version of T5 but we observed similar results. Then, we develop $T5_{Lags+Quart}$ model, a model variation that inputs lags labeled with their respective quarters. This approach allows the model to integrate seasonal trends and periodic patterns. The input is formatted as "The value was 0.29 at Q1, 0.21 at Q2, . . .", which helps the model to handle recurring temporal patterns. We also use the $T5_{Lags}$ model, which focuses solely on the sequential order of values without including contextual data. This model should gain an advantage when the focus is solely on the sequence of temporal values rather than their specific time occurrence, and treats timestamps merely as subsequent data points. The input is formatted as "0.29, 0.76, . . .". This approach allows for maximizing the number of lags input into the model by minimizing the contextual information. We also develop the $T5_{Lags+MLM}$ model, a version of T5 that performs the masked language modeling task[6], and thus is trained to predict masked portions of the input sequence. We structure the task using curriculum learning [1], thus presenting instances to the model in a increasing order of difficulty. We achieve this by masking 5% of the input sequence and progressively increasing to 25%, then, at the end of the training, masking only the end of the sequence [18].

We also introduce the $T5_{Cont}$ model, which maximizes the number of contextual features included while minimizing the number of lags. This approach prioritizes comprehensive contextual data over sequential lag inputs. To further optimize this, we develop the $T5_{SelCont}$ model, which selectively incorporates only the most influential features identified through a correlation and clustering analysis, allowing to double the number of input lags compared to $T5_{Cont}$. Finally, we introduce the $T5_{Dual}$ model, which incorporates two separate T5 encoders to generate distinct representations: one for the lags and another for contextual features, which are then combined using an attention layer. This models allows to maximize both the number of features and lags in input.

The experiments are conducted on a machine with 2 Nvidia 3090 GPUs. Each model is trained for three epochs using the multi-class cross entropy loss function over the language model vocabulary size [2]. For inference, we use a beam search technique to generate the output sequence auto-regressively. Notably, we found the fine-tuned models capable of generating floating point numbers directly, without the need for explicit constraints.

[6]see https://huggingface.co/docs/transformers/model_doc/t5.

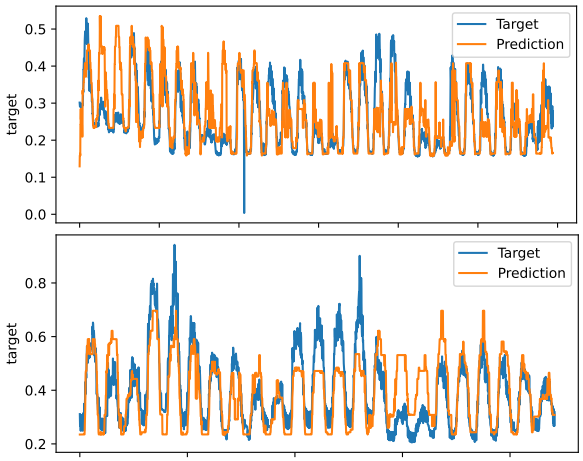

Figure 1: Prediction at 1 day (top) and 1 week (bottom) for $T5_{SelCont}$ and AAA dataset.

4 Experimental Results

Table 2 shows the effectiveness of the different models, with scores computed by taking the average metric on a per-month bases over the year used as test set. The ELD dataset does not come with weather information, so not all the models are used on it. An example of target and predicted scores is available in Figure 1.

We start by detailing the predictions for the subsequent day (i.e., t+1d), shown in the upper part of the table. The results highlight the effectiveness of transformer-based models, particularly $T5_{Lags+MLM}$, $T5_{Cont}$, and $T5_{SelCont}$. Overall, across all datasets, transformers-based models shows lower errors compared to baseline models, with the only exception of ME for the Spain and ELD dataset, where they are still among the most effective models. This suggests the ability of LLMs to handle time series forecasting problems. If we inspect the $T5_{Cont}$ and $T5_{SelCont}$ models, which focus on integrating extensive and selected contextual information respectively, also perform well across datasets, indicating the benefits of incorporating context in input to improve the effectiveness of the forecast. In the ELD dataset, where weather information is not present, transformers models and particularly $T5_{Lags}$ and $T5_{Lags+MLM}$, which focus solely on the sequence of values, continue to show higher effectiveness when compared to baselines. This result suggest that LLMs are effective in handling temporal forecasting problems even in the absence of contextual features, thus when dealing with univariate series [4].

We now turn to inspect the error scores for the prediction over the subsequent week (i.e., t+1w), shown in the lower part of Table 2. Transformer-based models such as $T5_{Lags+MLM}$, $T5_{Lags+Quart}$, and $T5_{Dual}$ again show higher effectiveness when compared to baselines, again with a single exception on the ME metric for the Spain dataset, where they still are among the best models. By inspecting the error metrics we see that $T5_{SelCont}$, with its emphasis on high quality contextual features, performs well in the AAA and Spain datasets, remarking the importance of incorporating contextual features. $T5_{Dual}$, which relies on the dual encoders to handle both lags and contextual data, shows robust results across datasets, again

Table 2: Effectiveness of the forecasting. Lower errors are highlighted in bold.

			AAA				Spain				ELD			
Model	**Lags**	**Pred**	**MAE**	**MAPE**	**ME**	**RMSE**	**MAE**	**MAPE**	**ME**	**RMSE**	**MAE**	**MAPE**	**ME**	**RMSE**
LSTM	–	t+1d	0.066	0.180	0.357	0.089	0.128	0.259	0.503	0.166	0.098	0.245	**0.405**	0.117
GRU	–	t+1d	0.080	0.255	0.332	0.099	0.135	0.341	0.443	0.169	0.119	0.298	0.522	0.149
DeepAR	–	t+1d	0.066	0.187	0.364	0.090	0.121	0.273	**0.382**	0.147	0.106	0.265	0.433	0.127
NBeats	–	t+1d	0.061	0.171	0.347	0.085	0.116	0.254	0.413	0.145	0.106	0.265	0.520	0.129
TFT	–	t+1d	0.062	0.185	0.326	0.084	0.101	0.250	0.417	0.130	0.124	0.310	0.471	0.147
T5	45	t+1d	0.051	0.163	0.340	0.075	0.098	0.241	0.472	0.122	0.048	0.124	0.462	0.069
$T5_{Lags+Quart}$	90	t+1d	0.050	0.160	0.358	0.073	0.086	0.235	0.471	0.112	0.038	0.093	0.451	0.064
$T5_{Lags}$	135	t+1d	0.050	0.160	0.353	0.072	0.068	0.230	0.474	0.102	0.033	0.083	0.431	0.054
$T5_{Lags+MLM}$	135	t+1d	0.046	0.150	0.347	0.067	0.060	**0.223**	0.470	0.093	**0.026**	**0.065**	0.460	**0.040**
$T5_{Cont}$	50	t+1d	0.037	0.116	0.265	0.051	**0.055**	0.293	0.422	**0.090**	–	–	–	–
$T5_{SelCont}$	100	t+1d	**0.036**	**0.111**	0.254	**0.049**	0.059	0.319	0.477	0.095	–	–	–	–
$T5_{Dual}$	120	t+1d	0.038	0.124	**0.247**	0.052	0.069	0.307	0.447	0.101	–	–	–	–
LSTM	–	t+1w	0.063	0.203	0.292	0.077	0.157	0.337	**0.390**	0.182	0.094	0.295	0.511	0.124
GRU	–	t+1w	0.113	0.299	0.479	0.152	0.215	0.455	0.532	0.248	0.122	0.305	0.531	0.151
DeepAR	–	t+1w	0.088	0.288	0.344	0.105	0.142	0.361	0.423	0.169	0.111	0.277	0.524	0.139
NBeats	–	t+1w	0.142	0.260	0.401	0.161	0.140	0.365	0.469	0.171	0.109	0.272	0.512	0.136
TFT	–	t+1w	0.100	0.339	0.349	0.122	0.151	0.388	0.485	0.188	0.127	0.317	0.550	0.158
T5	45	t+1w	0.064	0.206	0.331	0.085	0.121	0.317	0.556	0.167	0.052	0.134	0.483	0.075
$T5_{Lags+Quart}$	90	t+1w	0.062	0.196	0.367	0.085	0.101	0.289	0.558	0.141	0.044	0.102	0.472	0.067
$T5_{Lags}$	135	t+1w	0.061	0.201	0.371	0.083	0.091	0.297	0.566	0.131	0.035	0.088	**0.433**	0.057
$T5_{Lags+MLM}$	135	t+1w	0.058	0.196	0.363	0.078	0.086	**0.257**	0.556	0.123	**0.028**	**0.070**	0.465	**0.041**
$T5_{Cont}$	50	t+1w	0.058	0.186	0.308	0.075	**0.071**	0.264	0.467	**0.106**	–	–	–	–
$T5_{SelCont}$	100	t+1w	**0.054**	**0.178**	**0.294**	0.072	0.073	0.294	0.472	0.108	–	–	–	–
$T5_{Dual}$	120	t+1w	**0.054**	**0.178**	**0.294**	**0.070**	0.089	0.287	0.503	0.123	–	–	–	–

confirming the importance of leveraging a composite approach to maximize both lags and contextual information in input data, which can improve the models' forecasting capabilities. By inspecting the ELD dataset we also see a confirmation that that when contextual features are not available, transformers based models and in particular $T5_{Lags+MLM}$ still outperform baselines.

Overall, the findings from Table 2 suggest the higher capability of transformer-based models to deliver more accurate and reliable energy consumption forecasts both for the next day and the subsequent week across various metrics. This is also confirmed visually when inspecting Table 2, which shows that both for short- and long-term forecast the model is able to accurately follow, though with some errors, the general pattern of a target time-series. The predictions for the other datasets are not shown for space issue but exhibit a similar trend. The higher effectiveness of the NLP-based models' variations offers insights for some guidelines on selecting the most suitable model based on specific scenarios. Notably, when contextual data, especially weather information, is available, it is convenient to employ models that maximize the use of high-quality context and weather features. For instance, the $T5_{SelCont}$ model is ideal when an analysis of the most representative features is available, whereas the $T5_{Dual}$ model can be used to bypass the feature selection process and input both lags and contextual features to the model, delegating this step to the attention layer of the model which is employed after the dual encoding phase. Conversely, in scenarios lacking rich contextual data, NLP-based models still outperform traditional baselines. Among these, models trained with sophisticated strategies like $T5_{Lags+MLM}$ are preferred, although simpler models like $T5_{Lags}$ also demonstrate robust performance across datasets. In summary, our analysis confirms the comprehensive effectiveness of NLP-based models, and particularly of modified large language models, in handling the complexities of time series forecasting.

5 Conclusions and Future Work

This study demonstrates the potential of generative NLP-based models for multi-horizon power consumption forecasting across different datasets, both on their ability to predict energy consumption in short- (next day), and long-term (next week). The experimental results suggest the higher effectiveness of the developed models, especially those relying on external features in capturing the complexities of power consumption patterns more effectively than traditional forecasting methods. The use of contextual features such as time, date, and weather conditions, together with the innovative modeling of input data in a natural language format, has shown to significantly improve the forecasting effectiveness.

Future research will focus on improving predictive accuracy by integrating more diverse contextual data and explore model interpretability to provide clearer insights into energy consumption patterns.

Acknowledgments. This work was partially supported by the REACT-EU project 'Data-Driven Multiutility Grid: Supporto alle Decisioni per Garantire la Sostenibilità dal Real Time al Lungo Termine', PON 2014-2020 AZIONE IV.6 GREEN.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, Quebec, Canada) *(ICML 09)*. Association for Computing Machinery, New York, NY, USA, 41–48. https://doi.org/10.1145/1553374.1553380

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] Shobhit Chaturvedi, Elangovan Rajasekar, Sukumar Natarajan, and Nick McCullen. 2022. A comparative assessment of SARIMA, LSTM RNN and Fb Prophet models to forecast total and peak monthly energy demand for India. *Energy Policy* 168 (2022), 113097. https://doi.org/10.1016/j.enpol.2022.113097

[4] Alice Chuang. 1991. Time Series Analysis: Univariate and Multivariate Methods. *Technometrics* 33, 1 (1991), 108–109. https://doi.org/10.1080/00401706.1991.10484777

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[6] Vincenzo Della Mea, Mihai Horia Popescu, and Kevin Roitero. 2020. Underlying cause of death identification from death certificates using reverse coding to text and a NLP based deep learning approach. *Informatics in Medicine Unlocked* 21 (2020), 100456.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 1597–1600.

[9] Everette S Gardner Jr. 1985. Exponential smoothing: The state of the art. *Journal of forecasting* 4, 1 (1985), 1–28.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[11] Mehdi Khashei, Mehdi Bijari, and Seyed Reza Hejazi. 2012. Combining seasonal ARIMA models with computational intelligence techniques for time series forecasting. *Soft computing* 16 (2012), 1091–1105.

[12] Ashutosh Kumar Dubey, Abhishek Kumar, Vicente García-Díaz, Arpit Kumar Sharma, and Kishan Kanhaiya. 2021. Study and analysis of SARIMA and LSTM in forecasting time series data. *Sustainable Energy Technologies and Assessments* 47 (2021), 101474. https://doi.org/10.1016/j.seta.2021.101474

[13] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764.

[14] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[15] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437* (2019).

[16] Adhistya Erna Permanasari, Indriana Hidayah, and Isna Alfi Bustoni. 2013. SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence. In *2013 International Conference on Information Technology and Electrical Engineering (ICITEE)*. 203–207. https://doi.org/10.1109/ICITEED.2013.6676239

[17] Mihai Horia Popescu, Kevin Roitero, Stefano Travasci, and Vincenzo Della Mea. 2021. Automatic Assignment of ICD-10 Codes to Diagnostic Texts using Transformers Based Techniques. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. IEEE, 188–192.

[18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[20] Kevin Roitero, Cristina Gattazzo, Andrea Zancola, Vincenzo Della Mea, Stefano Mizzaro, et al. 2022. Causal Text-to-Text Transformers for Water Pollution Forecasting. In *CEUR WORKSHOP PROCEEDINGS*, Vol. 3463. CEUR-WS.

[21] Kevin Roitero, Beatrice Portelli, Mihai Horia Popescu, and Vincenzo Della Mea. 2021. DiLBERT: Cheap Embeddings for Disease Related Medical NLP. *IEEE Access* 9 (2021), 159714–159723.

[22] Kevin Roitero, Beatrice Portelli, Giuseppe Serra, Vincenzo Della Mea, Stefano Mizzaro, Gianni Cerro, Michele Vitelli, and Mario Molinara. 2023. Detection of Wastewater Pollution Through Natural Language Generation With a Low-Cost Sensing Platform. *IEEE Access* 11 (2023), 50272–50284. https://doi.org/10.1109/ACCESS.2023.3277535

[23] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36, 3 (2020), 1181–1191.

[24] Robert H. Shumway and David S. Stoffer. 2017. *ARIMA Models*. Springer International Publishing, Cham, 75–163. https://doi.org/10.1007/978-3-319-52452-8_3

[25] Shixiong Wang, Chongshou Li, and Andrew Lim. 2019. Why are the ARIMA and SARIMA not sufficient. *arXiv preprint arXiv:1904.07632* (2019).

[26] Hao Xue, Flora D Salim, Yongli Ren, and Charles LA Clarke. 2022. Translating Human Mobility Forecasting through Natural Language Generation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1224–1233.

[27] Hao Xue, Tianye Tang, Ali Payani, and Flora D Salim. 2024. Prompt mining for language-based human mobility forecasting. *arXiv preprint arXiv:2403.03544* (2024).

[28] Hao Xue, Bhanu Prakash Voutharoj, and Flora D Salim. 2022. Leveraging Language Foundation Models for Human Mobility Forecasting. *arXiv preprint arXiv:2209.05479* (2022).

[29] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.