CAUSALPROMPT: ENHANCING LLMS WITH WEAKLY SUPERVISED CAUSAL REASONING FOR ROBUST PERFORMANCE IN NON-LANGUAGE TASKS

Tung-Wei Lin *
UC Berkeley, USA
twlin@berkeley.edu

Vanshaj Khattar *
Virginia Tech, USA
vanshajk@vt.edu

Yuxuan Huang *
University College London, UK
yuxuan.huang.23@ucl.ac.uk

Junho Hong
University of Michigan, USA
jhwr@umich.edu

Ruoxi Jia
Virginia Tech, USA
ruoxijia@vt.edu

Chen-Ching Liu
Virginia Tech, USA
ccliu@vt.edu

Alberto Sangiovanni-Vincentelli
UC Berkeley, USA
alberto@eecs.berkeley.edu

Ming Jin
Virginia Tech, USA
jinming@vt.edu

ABSTRACT

In confronting the pressing issue of climate change, we introduce "CausalPrompt", an innovative prompting strategy that adapts large language models (LLMs) for classification and regression tasks through the application of weakly supervised causal reasoning. We delve into the complexities of data shifts within energy systems, often resulting from the dynamic evolution of sensor networks, leading to discrepancies between training and test data distributions or feature inconsistencies. By embedding domain-specific reasoning in the finetuning process, CausalPrompt significantly bolsters the adaptability and resilience of energy systems to these shifts. We show that CausalPrompt significantly enhances predictions in scenarios characterized by feature shifts, including electricity demand, solar power generation, and cybersecurity within energy infrastructures. This approach underlines the crucial role of CausalPrompt in enhancing the reliability and precision of predictions in energy systems amid feature shifts, highlighting its significance and potential for real-world applications in energy management and cybersecurity, contributing effectively to climate change mitigation efforts.

1 INTRODUCTION

The fight against climate change necessitates innovative solutions for optimizing energy systems. At the heart of these solutions are advanced predictive models and sensor networks within energy grids, which are pivotal for ensuring the efficient operation of energy grids. These models facilitate real-time decisions Pinheiro et al. (2023), from predicting energy demand and generation to adjusting for variable renewable outputs Chen et al. (2023), thereby enhancing grid reliability and reducing carbon emissions. However, the effectiveness of these systems is continually challenged by the dynamic nature of energy grids, such as sensor network upgrades, malfunctions Barrabés et al. (2023), and vulnerabilities to cyberattacks Kulinski et al. (2020); Zaboli et al. (2023). Such complexities necessitate models that are not only accurate but also robust against feature and distribution shifts.

Large Language Models (LLMs) have extended their utility beyond the realm of Natural Language Processing (NLP) to address challenges in diverse domains. RT-2 exemplifies this versatility through its employment of semantic reasoning derived from extensive pretraining corpora for robotic control Brohan et al. (2023). Similarly, HeLM harnesses the medical knowledge embedded within LLMs, combined with multimodal clinical data, to tackle bespoke healthcare issues Belyaeva et al. (2023). Another notable innovation, LIFT, employs a novel framework that transforms tabular data into natural language descriptions Dinh et al. (2022). This approach enables fine-tuning of LLMs

*These authors contributed equally to this work.

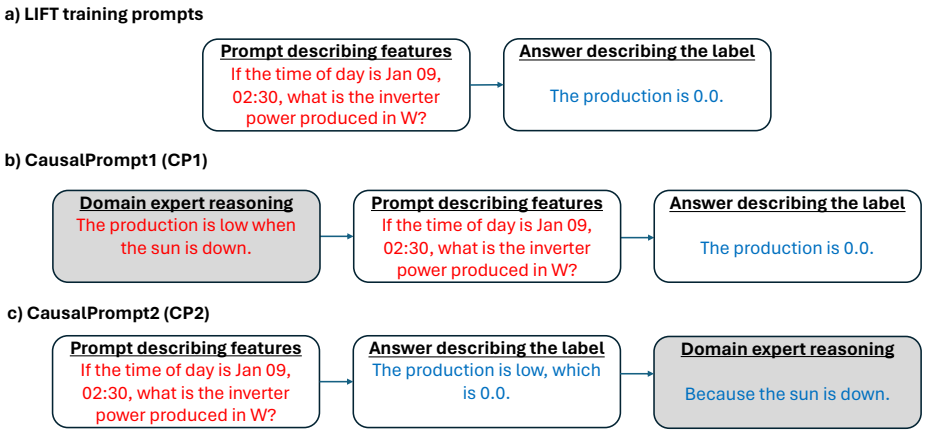

Figure 1: **Comparison between CausalPrompt and LIFT.** Prompts are in red while answers are in blue. Domain expert reasonings are shaded.

for regression and classification tasks, showcasing remarkable efficacy. The core mechanism of LIFT involves a manually designed translation template that converts numerical features and their corresponding labels into natural language sentences, on which an LLM is fine-tuned. Subsequently, the fine-tuned LLM is prompted with feature descriptions of test data generated using the same template to predict the corresponding labels.

Despite its effectiveness, LIFT faces challenges with feature and distribution shifts, which are prevalent in energy systems due to the increasing reliance on renewable sources, climate change, and the evolution of sensor networks. The ability of predictive models to remain robust and deliver accurate predictions despite perturbations is crucial. This resilience ensures continuous operation and optimization of energy sources, contributing significantly to reducing carbon emissions and combating climate change. This paper focuses on the resilience of predictive models like LIFT in the face of such shifts, especially regarding energy production, consumption, and cyberattack detection. We found that LIFT's performance significantly deteriorates under feature shifts (See Table 1 and 2). This weakness highlights the critical gap in leveraging the knowledge and semantic reasoning capabilities derived from large pretraining corpora for LLMs fine-tuned by LIFT.

To address this limitation, we introduce a novel enhancement to LIFT through the integration of causal reasoning during finetuning, facilitated by the incorporation of domain expert reasonings, akin to the idea of weakly supervised learning Zhou (2018). We call our method **CausalPrompt**. Our results demonstrate that this enhancement leads to a substantial improvement in LIFT's performance, not only in standard operating scenarios but also under feature shifts. By enhancing the robustness of predictive models against these challenges, we hope that CausalPrompt can be used to improve the reliability of energy grids and support broader climate change mitigation strategies.

2 Methodology and Experimental Setup

LIFT converts each training data sample into a single sentence with a fixed template and then fine-tunes the LLM with the sentences to predict labels. To enhance LIFT for feature shifts prevalent in energy systems, our idea is to induce causal reasoning in the LLMs Wei et al. (2022); Yao et al. (2023); Sel et al. (2023). We introduce our novel prompting method: **CausalPrompt**, where we induce domain expert reasoning in the prompts during fine-tuning. The main idea is to expose the model to causal reasoning for the observed labels. See Figure 1 for a direct comparison of CausalPrompt and LIFT. We use two different styles of prompts to fine-tune:

1. **CausalPrompt1 (CP1)**: CP1 incorporates domain expert reasoning at the beginning of the training prompt. The LLM is fine-tuned to utilize both the query and the domain expert reasoning to generate the corresponding label.

Table 1: **RMSEs with and without (w/o) feature shifts on datasets 1 and 2.** (Lower is better.) The best results are in bold, and the second best are underlined.

Dataset	CP1 (ours)	CP2 (ours)	LIFT	GPR	LR	MLP	KNN	DTR
Electric demand dataset (w/o shift)	0.25	0.40	0.41	**0.15**	0.50	<u>0.23</u>	0.17	**0.15**
Electric demand dataset (with shift)	**0.41**	<u>0.60</u>	2.63	0.96	0.92	3.16	0.71	0.72
Solar power prediction (w/o shift)	**83.90**	178.81	182.29	<u>121.48</u>	215.92	147.20	109.07	122.07
Solar power prediction (with shift)	**121.01**	304.57	359.36	280.42	<u>216.52</u>	333.60	287.42	299.36

Table 2: **Accuracy (%) on the cybersecurity dataset with and without (w/o) feature shifts.** (Higher is better.)

Dataset	CP1 (ours)	CP2 (ours)	LIFT	LSTM	GRU	RNN
Substation cyber-attack (w/o shift)	**64.28**	35.72	<u>21.42</u>	35.71	35.71	35.71
Substation cyber-attack (with shift)	**57.14**	<u>28.57</u>	7.15	14.29	<u>28.57</u>	21.43

2. **CausalPrompt2 (CP2)**: In CP2, domain expert reasoning is appended after the label. The LLM is fine-tuned to first produce the label followed by the domain expert reasoning.

Note that the domain expert reasoning used in both CP1 and CP2 might not always be precisely accurate or fully representative of the underlying causal process. Instead, these reasoning serve as augmented, albeit weak, indicators. This concept aligns with the principles of weakly supervised learning, where the provided labels are not necessarily exact but are still utilized for training purposes. We validate the effectiveness of the proposed CausalPrompt method on three datasets.

1. **Dataset 1: Electric demand dataset.** We create a synthetic dataset for hourly electricity demands over 2 months. We assume the dual peak model Zanocco et al. (2022), where every day has a morning and evening electric demand peak with prominent peaks on weekdays. The goal is to predict the electricity demand given the time. One example of the CP2 prompt with labels and reasoning is: If time of day is 15:00, what is the electricity demand in KW? The demand is 1.588980326396931. The demand is medium because it's afternoon and the regular time.
2. **Dataset 2: Solar power prediction.** We consider a solar power generation dataset from the CityLearn challenge Vazquez-Canteli et al. (2020). We take the first three months of hourly solar power generation data from climate zone 1. The goal is to predict solar power generation from a solar cell given the time. One example of the CP2 prompt with labels and reasoning is: If the time of day is Jan 01, 06:30, what is the inverter power produced in W? the production is low, which is 0.0. Probably because there is an overcast or it is not close to noon.
3. **Dataset 3: Substation cyber-attack.** The third dataset is the cyber-attack data on the IEC–61850 communication protocol, which is commonly used in digital substations for communication between the merging units Mackiewicz (2006). We specifically consider the data injection (DI) and the network error (NE) attacks in the IEC-61850 protocol for the Sampled Value signals. The goal is to predict whether the substation is under DI attack, NE attack, or is normal given the stream of sampled value counts. The dataset and the domain expert reasoning are from Zaboli et al. (2023). The following two examples are of CP2 training prompts for both attacks:
   (a) The following is the stream of sampled value counts: 2141, 2142, 2143, 2144, 2145, 2146, 2145, 2148, 2149, 2150, 2151, It is a data network error attack because the stream doesn't increment by 1 every time.
   (b) The following is the stream of sampled value counts: 4789, 4790, 4791, 4792, 4793, 4794, 4795, 4796, 4797, 4798, 4799, 4800, It is a data injection attack because there are values out of bound (0, 4799).

Training. After the data are converted into sentences, we randomly shuffle the sentences (data points) and sample 50% of them for training, 25% for testing, and 25% for validation. In this study, we fine-tune the OpenAI GPT-3.5 Turbo model using OpenAI's API.

Testing under feature shifts. The prediction of a test data is parsed from the fine-tuned LLM's response to the corresponding prompt. We test the fine-tuned model under two settings: *1)* without (w/o) feature shifts and *2)* with feature shifts. For without feature shifts, the model is fine-tuned and tested using the same template. For instance, for dataset 2 in CP2 style without feature shift, the model is prompted: If the time of day is Jan 01, 06:30, what is the inverter power produced in W? The LLM then responds with a label and the corresponding causal reasoning. On the other

Table 3: **Percentage (%) drop in the performance after feature shift in the test set.** (* - not applicable). (Lower is better.)

Dataset	CP1	CP2	LIFT	GPR	LR	MLP	KNN	DTR	LSTM	GRU	RNN
Electric demand dataset	64.00	**50.00**	541.46	540.00	80.39	1216.67	343.75	380.00	*	*	*
Solar power prediction	44.23	70.33	97.13	130.83	**0.27**	126.63	163.51	145.23	*	*	*
Substation cyber-attack	**11.10**	20.01	66.61	*	*	*	*	*	59.98	20.01	40.00

hand, when there is feature shift, the template is modified. We provide samples of how the LLM is prompted in CP2 style under feature shift for testing, **Dataset 1.** If time of day is 1 hours and 30 minutes after midday, what is the electricity production? **Dataset 2.** if time of day is 2 hours and 30 minutes before midnight, what is the electricity production? **Dataset 3.** The following is the stream of sampled value counts subtracted from 4799: [-90, -91, -92, -93, -94, -95, -96, -97, -98, -99, -100, -101]. In all the above cases, the feature descriptions and values are different from what the LLM saw during fine-tuning, simulating the feature shift.

Evaluation metrics and baselines. For the evaluation metrics, we use root mean square error (RMSE) for the regression tasks (i.e., datasets 1 and 2) to compare CausalPrompt with LIFT, Gaussian Process regression (GPR), linear regression (LR), multi-layer perceptron (MLP), K-nearest-neighbors (KNN) regression, and decision tree regression (DTR); we use accuracy for the classification task (dataset 3) to compare with LIFT, long short term memory (LSTM), recurrent neural network (RNN), and gated recurrent unit (GRU) since packet data are of varying lengths. For baselines other than LIFT, data is made of extracted numerics from the converted sentences. If the test data with feature shift has fewer features than the training data, the missing features are dummy-filled with zero. For example, the month and date are filled with zero for dataset 2 when tested under feature shift because they were present in the training prompt but absent in the testing prompt.

3 Experimental Results and Discussion

Tables 1, 2, and 3 show the effectiveness of CausalPrompt and its robustness to feature shifts compared with baselines. For the electric demand dataset, we can see from Table 1 that CP1 and CP2 are comparable with all the other baselines when there is no feature shift in the test set. However, when tested with feature shift, the drop in performance is significant for all baselines (Table 3). Notably, the performance drops are lowest for CP1 and CP2, highlighting the robustness of our training prompts to feature shifts.

For solar power prediction, we see similar trends in Table 1 with comparable or better performance compared to other baselines. The RMSE is lowest for CP1 for both with and without feature shifts. Moreover, the performance drop for both CP1 and CP2 is the lowest, as seen in Table 3. The exception is the LR baseline, which did equally bad before and after the feature shift, making the performance drop negligible. We also see similar trends on the classification task on the cyber-attack dataset in Table 2.

Discussion. Incorporating causal reasoning into training prompts enhances the LIFT framework, mitigating degradation from feature shift. This improvement stems from exposing the model to causal relationships between features and labels via domain expert reasoning during fine-tuning. While baselines have comparable or superior performance to CausalPrompt (CP1 and CP2) without feature shifts, their performance significantly declines under feature shifts, underscoring the need for semantic feature understanding to maintain accuracy despite changes in feature descriptions.

Limitations. Although CausalPrompt reduces the impact of feature shifts, its performance also degrades. Even with smaller degradation compared to other baselines, its application to safety-critical settings will be limited. Moreover, CausalPrompt relies on domain expert reasoning, which is not always available. Another limitation is the cost of fine-tuning on the current OpenAI API. Our fine-tuning experiments cost a total of around 50 USD. We anticipate that as open-source models advance, our method could be applied to these models without high fine-tuning costs.

4 Conclusion and Future Work

In this paper, we highlight the innovative approach of CausalPrompt in addressing the challenges of climate change. By embedding weakly supervised causal reasoning into large language models (LLMs), our method significantly improves the adaptability and accuracy of energy systems against data shifts, a common issue in dynamic environments like energy grids. This capability is critical for enhancing reliability and efficiency. We demonstrate not only the potential of CausalPrompt to fortify energy systems against variability but also its broader applicability in climate change mitigation efforts. The success of CausalPrompt in three datasets underscores the importance of integrating domain-specific knowledge and causal reasoning into LLMs, opening new avenues for future exploration in few-shot learning, the impact of different feature shifts, and further refinement of causal reasoning methods. This contribution is a step towards leveraging AI in the fight against climate change, enhancing energy system resilience and sustainability.

References

Míriam Barrabés, Daniel Mas Montserrat, Margarita Geleta, Xavier Giró-i Nieto, and Alexander G Ioannidis. Adversarial learning for feature shift detection and correction. *arXiv preprint arXiv:2312.04546*, 2023.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pp. 86–102. Springer, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Yu-Quan Chen, Iris Hui-Ru Jiang, and Katherine A Kim. Pv inverter control algorithm using reinforcement learning to mitigate the duck curve problem. In *2023 IEEE Applied Power Electronics Conference and Exposition (APEC)*, pp. 2293–2298. IEEE, 2023.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784, 2022.

Sean Kulinski, Saurabh Bagchi, and David I Inouye. Feature shift detection: Localizing which features have shifted via conditional distribution tests. *Advances in neural information processing systems*, 33:19523–19533, 2020.

Ralph E Mackiewicz. Overview of iec 61850 and benefits. In *2006 IEEE Power Engineering Society General Meeting*, pp. 8–pp. IEEE, 2006.

Marco G Pinheiro, Sara C Madeira, and Alexandre P Francisco. Short-term electricity load forecasting—a systematic approach from system level to secondary substations. *Applied Energy*, 332: 120493, 2023.

Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Lu Wang, Ruoxi Jia, and Ming Jin. Algorithm of thoughts: Enhancing exploration of ideas in large language models. *arXiv preprint arXiv:2308.10379*, 2023.

Jose R Vazquez-Canteli, Sourav Dey, Gregor Henze, and Zoltan Nagy. Citylearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management. *arXiv preprint arXiv:2012.10504*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Aydin Zaboli, Seong Lok Choi, Tai-Jin Song, and Junho Hong. Chatgpt and other large language models for cybersecurity of smart grid applications. *arXiv preprint arXiv:2311.05462*, 2023.

Chad Zanocco, Tao Sun, Gregory Stelmach, June Flora, Ram Rajagopal, and Hilary Boudet. Assessing californians' awareness of their daily electricity use patterns. *Nature Energy*, 7(12):1191–1199, 2022.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1): 44–53, 2018.