LLM-based Multivariable Missing Data Restoration for Photovoltaic-Storage Integrated Systems

Zhenghao Zhou, Yiyan Li*, Zelin Guo
Runlong Liu
College of Smart Energy
Shanghai Jiao Tong University
Shanghai, China
{zhenghao.zhou, yiyan.li*, gz11996,
runlong_liu}@sjtu.edu.cn

Zheng Yan
School of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China
yanz@sjtu.edu.cn

Mo-Yuen Chow
University of Michigan - Shanghai Jiao
Tong University Joint Institute
Shanghai Jiao Tong University
Shanghai, China
moyuen.chow@sjtu.edu.cn

***Abstract*—As one of the key renewable energy resources, photovoltaic (PV) has significant volatility and usually collaborates with energy storage system (ESS), formulating PV-ESS systems. Collecting multivariable data from PV-ESS systems, such as power output, weather condition and state of charge (SOC), is critical to understanding and optimizing the system operation. However, the occurrence of missing data impairs the corresponding data-driven studies. This paper introduces an approach using large language model (LLM) for multivariable missing data restoration of PV-ESS systems. A method for constructing an instruction-based fine-tuning dataset that aligns different data modalities is proposed, which can be used for supervised fine-tuning. Evaluations on three LLMs with varying parameter sizes show that larger models and the use of instructional prompts can improve the performance in missing data restoration. It is also observed that larger models have higher computational costs, highlighting the need for a trade-off between model performance and efficiency.**

Keywords—Multivariable missing data restoration, large language model, photovoltaic, energy storage system.

I. INTRODUCTION

Photovoltaic (PV) energy is a promising alternative to fossil fuels. Due to the inherent volatility of PV power generation, utility-scale solar farms are frequently co-located with energy storage system (ESS), forming integrated PV-ESS plants [1]. For operational PV-ESS plants, the presence of missing or anomalous values—such as null entries or flat-lined data resulting from equipment malfunctions or communication interruptions—is an unavoidable problem[2]. The existence of such data anomalies severely degrades data quality, thereby impeding various downstream applications, including but not limited to, power generation forecasting and operational anomaly detection.

Therefore, the imputation of missing data is of paramount importance in the analysis of PV-ESS. Properly imputed data, which should closely mirror the statistical properties of the authentic data, are often considered a credible substitute for the true values. Existing techniques for missing data recovery are broadly categorized into two main classes: model-based methods and data-driven methods. Model-based approaches leverage the physical principles of the system to simulate its response to external disturbances, aiming to reconstruct the missing data segments (MDS). In contrast, data-driven methods offer an end-to-end solution by learning directly from the historical data.

Model-based methods are fundamentally grounded in the mathematical representation of the physical behavior of system components, such as PV arrays and battery storage units. These methods can be effective when an accurate system model is available. However, their performance is often constrained by the fidelity of the model and the difficulty in precisely capturing all dynamic operational characteristics and external environmental influences, leading to potential inaccuracies in the imputed values.

Data-driven approaches do not rely on explicit physical models. Instead, they identify and learn the complex, non-linear relationships and temporal dependencies directly from historical operational data [3]. This category encompasses a wide range of techniques, from classical methods like linear regression to more advanced machine learning models. The inherent strength of these methods lies in their ability to adapt to complex data patterns without prior assumptions about the system's physical laws. Within the data-driven paradigm, there is a strong correlation between variables in a PV-ESS plant; for instance, the power output of the PV array and the charging/discharging behavior of the energy storage are intrinsically linked [4].

Recently, a new frontier of data-driven techniques, generative Artificial Intelligence (AI), has shown exceptional promise for time series imputation. Methods based on generative adversarial network [5] and transformer, such as Bidirectional Encoder Representations from Transformers (BERT) [6], have been successfully applied to generate synthetic data that captures the complex distributions of real-world time series. Following this trajectory, Large Language Model (LLM), which have achieved transformative success in natural language processing and other domains, are emerging as a powerful new tool. Due to the powerful functionality, LLMs attract some attention in power system, such as document summarization [7], energy management [8] and interactive load forecasting [9]. Nevertheless, the application of LLM to the specific challenge of data imputation in power systems remains in a nascent stage [10].

This paper presents a novel approach that leverages the strong capabilities of LLMs in processing diverse types of data. The method first converts textual instructions into token

This work was supported in part by National Natural Science Foundation of China under Grant 52307121, and in part by Shanghai Sailing Program under Grant 23YF1419000. (Corresponding author: Yiyan Li.)

embeddings using the model's tokenizer, while temporal data is transformed into time-series embeddings through an dedicated input layer. These two embeddings are then concatenated and fed into the LLM, enabling it to clearly understand the context and objectives of the task. By doing so, the approach exploits the LLM's powerful generalization ability to achieve a fine-grained understanding of the interdependencies among variables. Our method is capable of imputing multiple variables simultaneously, offering a more flexible and effective solution to the critical problem of missing data in power systems.

II. Methodology

A. Problem Formulation of Multivariable Missing PV-ESS Data Restoration

Denote a historical sensors time series matrix of PV-ESS $\mathbf{X}$ as

$$\mathbf{X}=\begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_L^1 \\ x_1^2 & x_2^2 & \cdots & x_L^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^E & x_2^E & \cdots & x_L^E \end{bmatrix} \tag{1}$$

where E is the number of variables and L is the length of the time series.

For every variable, it can be denoted as:

$$\mathbf{X}^N=\begin{bmatrix} x_1^N & x_2^N & \cdots & x_L^N \end{bmatrix} \tag{2}$$

We define a missing data segment in a special variable as $\mathbf{X}_{mds}^N$. We can divide the total segment into three periods: $[\mathbf{X}_{mds}^N, \mathbf{X}_{mds}^{other}]$ as the restoring data period, $[\mathbf{X}_{pre}^N, \mathbf{X}_{pre}^{other}]$ as the pre-missing period, $[\mathbf{X}_{post}^N, \mathbf{X}_{pre}^{other}]$ as the post-missing period. Thus, one single missing data restoration problem can be described as:

$$\hat{\mathbf{X}}_{mds}^N = f_\theta(\mathbf{X}_{pre}^N, \mathbf{X}_{pre}^{other}, \mathbf{X}_{mds}^{other}, \mathbf{X}_{poster}^N, \mathbf{X}_{post}^{other}) \tag{3}$$

where f_θ is the mapping function.

B. Supervised Fine-tuning Strategy

1) Data Preprocessing and Prompt Engineering

The supervised fine-tuning (SFT) dataset for LLM consists of three parts: input, output and instruction. After the data preprocessing, the raw data $\mathbf{X}$ is converted into standardized input and output data. The data preprocessing steps are as follows:

Firstly, each temporal variable is individually normalized using min-max scaling, transforming the data to lie within the interval [0,1] to obtain the standardized dataset $\tilde{\mathbf{X}}$:

$$\tilde{x}=\frac{x-x_{min}}{x_{max}-x_{min}} \tag{4}$$

Secondly, we define a mask matrix $\mathbf{M}$, which consists of binary values, 0 and 1. The size of $\mathbf{M}$ is similar to $\mathbf{X}$.

$$\mathbf{M}=\begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_L^1 \\ m_1^2 & m_2^2 & \cdots & m_L^2 \\ \vdots & \vdots & \ddots & \vdots \\ m_1^E & m_2^E & \cdots & m_L^E \end{bmatrix} \tag{5}$$

As equation (6)-(8) and Fig. 1, in the mask matrix, we randomly designate contiguous segments to be marked as 1, with segment lengths drawn from a normal distribution. The parameters of the distribution—its mean and variance—are tunable and can be adapted according to the specific characteristics of the dataset in use. The process can be formulated:

$$\ell=\max\left(0,\left\lfloor \mathcal{N}(\mu,\sigma^2)\right\rfloor\right) \tag{6}$$

$$s\sim\mathcal{U}(1,L-\ell+1) \tag{7}$$

$$M[s:s+\ell-1]\leftarrow 1 \tag{8}$$

where ℓ is the length of the MDS, $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with a mean of μ and a variance of σ^2, s is the started point of MDS, $\mathcal{U}$ is a uniform distribution and L is the length of total segments.

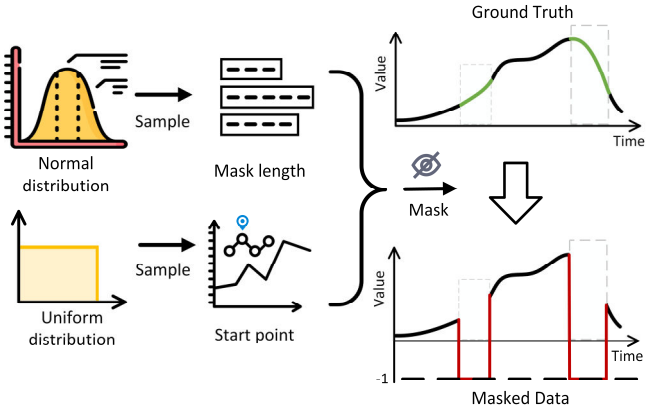

Fig. 1. The data preprocessing for supervised fine-tuning.

The effects of mask matrix are two-fold: (1) It is used to mask true values in the input data, thereby facilitating the construction of SFT training pairs; (2) It serves to identify the masked positions and facilitates the calculation of the loss function by focusing on the relevant regions of the input.

Thirdly, based on the mask matrix, all elements in the dataset $\tilde{\mathbf{X}}$ that correspond to positions marked as 1 in the mask are replaced with -1:

$$\bar{\mathbf{X}}=(\mathbf{1}-\mathbf{M})\odot\tilde{\mathbf{X}}+(-1)\cdot\mathbf{M} \tag{9}$$

Fourthly, prompt engineering is a methodology in natural language processing that focuses on designing effective prompts for language models. It aims to guide models to generate desired outputs by formulating specific instructions or queries. Well-crafted prompts can significantly enhance model performance across a variety of tasks, such as data restoration. As shown in Fig. 2, the instruction $\mathbf{I}$ can be seen as a prompt in SFT. Finally, the SFT dataset consist of input $\bar{\mathbf{X}}$, output $\tilde{\mathbf{X}}$, and instruction $\mathbf{I}$. The process can be formulated as:

$$\hat{\mathbf{X}}=LLM(\mathbf{I},\bar{\mathbf{X}}) \text{ with } Loss(\hat{\mathbf{X}},\tilde{\mathbf{X}},\mathbf{M}) \tag{10}$$

2) Modal alignment embedding

General-purpose LLMs are primarily designed for natural language processing and are not inherently suitable for directly handling time series data, which belongs to a different modality. In addition to the time series inputs, we also provide textual instructions as prompts to guide model behavior. Therefore, before feeding the data into the LLM,

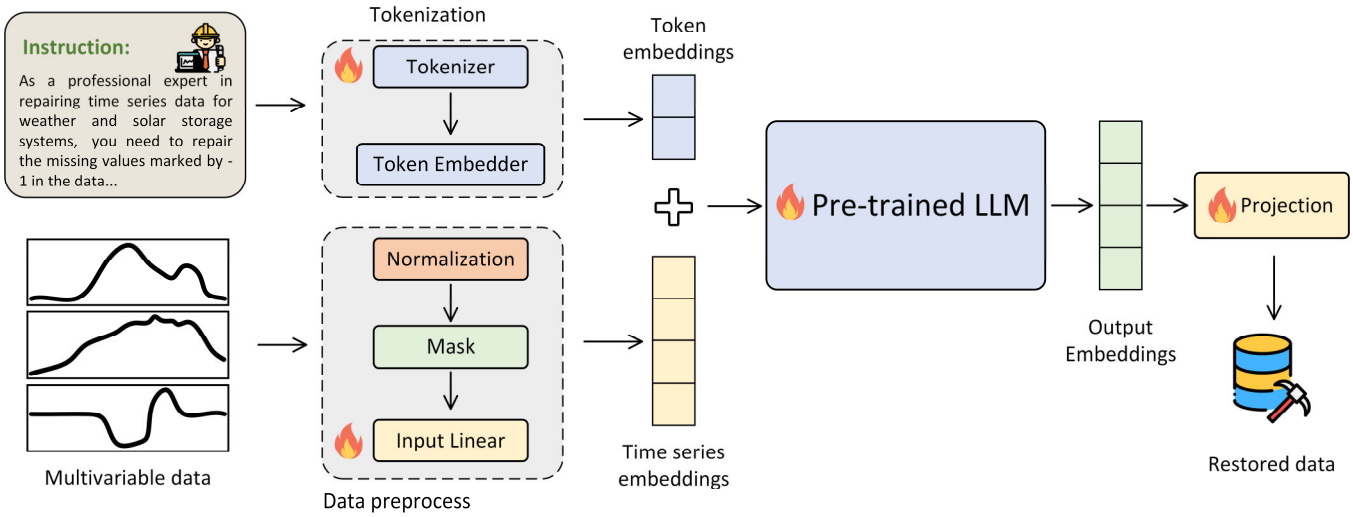

Fig. 2. The Large language model-empowered multivariable restoration framework.

we introduce a modality alignment step to process both the textual and temporal information and form a unified input representation.

Specifically, we utilize the tokenizer and token embedder of the LLM to encode the instruction text into word embeddings, as shown in Fig. 2. Meanwhile, the time series data is embedded into a temporal vector using a fully connected layer. Finally, the resulting word embeddings and temporal vector are concatenated to form a joint representation that aligns the two modalities, which is then fed into the large language model.

3) Physical information embedding

PV-ESS is a system governed by clear physical laws, where multiple variables are connected through well-defined dependencies. For example, irradiance has a direct impact on the output of the PV unit; the PV and storage systems work together to regulate the total output; and there is a direct mapping between the energy storage output and its SOC. These physical relationships provide valuable prior knowledge that can effectively support data imputation across multiple variables.

In practice, when data is missing, it is rare for multiple variables to be simultaneously missing values. In most cases, only a few variables have missing entries, while the rest remain intact. In such situations, the available complete variables can offer reliable information to assist the model in reconstructing the missing parts.

To enable the model to understand these physical relationships in a simple yet effective way, we incorporate the physical knowledge into the model via prompts, which is instructions in SFT. This allows the model to recognize the task type, the physical meaning of each variable, and the underlying physical relationships between strongly correlated variables.

4) Training loss

In SFT, LLM is trained with specially designed loss functions. The loss function includes 2 terms: the accuracy loss (L_{acc}) and the mask loss (L_{mask}) as shown in (11)–(13). λ_1 is the weight to balance.

$$\min\{L_{acc} + \lambda_1 L_{mask}\} \tag{11}$$

$$L_{acc} = \left\|\hat{\mathbf{X}} - \bar{\mathbf{X}}\right\|_2^2 \tag{12}$$

$$L_{mask} = \left\|\hat{\mathbf{X}} \odot \mathbf{M} - \bar{\mathbf{X}} \odot \mathbf{M}\right\|_2^2 \tag{13}$$

L_{acc} is a comprehensive loss to understand the changing pattern of PV-ESS time series data. L_{mask} employs mean squared error to minimize the point-to-point discrepancies in the mask segments.

III. Case Study

In this section, we evaluate the abilities of different size the fine-tuned large language models to recover PV-ESS missing data segments. As part of our experimental setup, the Qwen model—known for its stability and widespread adoption—is selected as the baseline. We further assess models of different scales, specifically those with approximately 0.6B, 8B, and 32B parameters. Both prompt-based and prompt-free settings are considered to comprehensively evaluate model behavior. In this case, the models undergo 60 epochs of fine-tuning on NVIDIA A100 GPUs.

A. Data Preparation

The PV-ESS dataset used in this case consists of 15-min resolution smart meter data, which was from a real PV-ESS plant in China. We select 365 daily profiles to set up a SFT training dataset, which covers 6 kinds of features, including irradiance, temperature, PV output, total output, store output and store stage of charge (SOC).

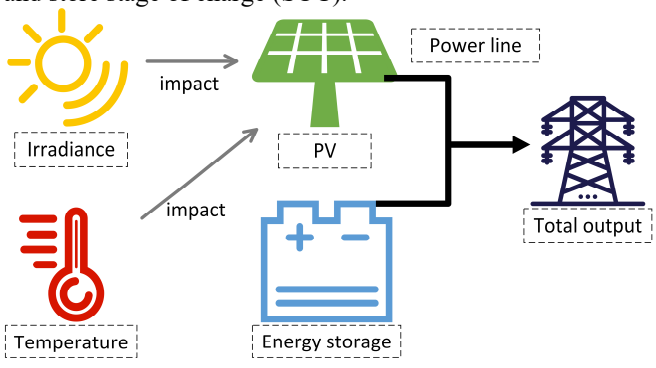

Fig. 3. The influence and correlation relationships of PV-ESS.

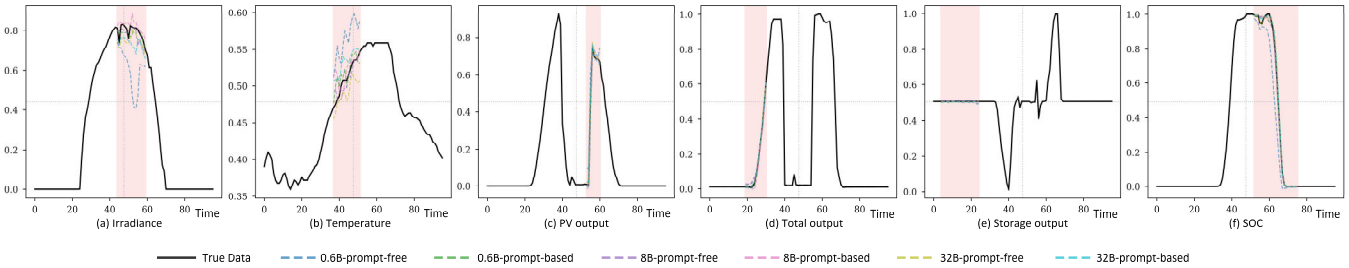

Fig. 4. Examples of missing data restoration with a mask using different size models.

TABLE I
MODEL PERFORMANCES ON ACCURACY

	Model	0.6B-prompt-free		0.6B-prompt-based		8B-prompt-free		8B-prompt-based		32B-prompt-free		32B-prompt-based	
	Metric	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Mask Positions	Irradiance	0.0342	0.0636	0.0273	0.0546	0.0274	0.0529	0.0283	0.0531	0.0218	0.0428	0.0212	0.0426
	Temperature	0.0206	0.0323	0.0176	0.0277	0.0170	0.0281	0.0196	0.0301	0.0162	0.0265	0.0168	0.0302
	PV output	0.0180	0.0331	0.0157	0.0302	0.0182	0.0318	0.0146	0.0308	0.0120	0.0256	0.0124	0.0289
	Total output	0.0241	0.0474	0.0184	0.0399	0.0216	0.0428	0.0196	0.0429	0.0160	0.0365	0.0148	0.0386
	Storage output	0.0176	0.0320	0.0139	0.0251	0.0318	0.0227	0.0144	0.0246	0.0106	0.0185	0.0091	0.0176
	SOC	0.0195	0.0529	0.0148	0.0403	0.0158	0.0334	0.0165	0.0356	0.0110	0.0248	0.0096	0.0236
	Overall	0.0223	0.0452	0.0176	0.0376	0.0187	0.0367	0.0173	0.0323	0.0146	0.0302	0.0143	0.0301
All Positions	Irradiance	0.0184	0.0340	0.0126	0.0268	0.0129	0.0259	0.0129	0.0260	0.0137	0.0261	0.0130	0.0246
	Temperature	0.0176	0.0244	0.0129	0.0184	0.0131	0.0187	0.0203	0.0145	0.0146	0.0205	0.0168	0.0229
	PV output	0.0124	0.0218	0.0107	0.0181	0.0134	0.0189	0.0087	0.0168	0.0083	0.0157	0.0088	0.0151
	Total output	0.0151	0.0274	0.0124	0.0221	0.0151	0.0237	0.0112	0.0222	0.0107	0.0209	0.0096	0.0195
	Storage output	0.0123	0.0225	0.0101	0.0166	0.0079	0.0150	0.0104	0.0160	0.0091	0.0154	0.0069	0.0117
	SOC	0.0126	0.0290	0.0088	0.0208	0.0096	0.0289	0.0094	0.0192	0.0096	0.0198	0.0066	0.0146
	Overall	0.0147	0.0268	0.0113	0.0207	0.0120	0.0205	0.0112	0.0204	0.0110	0.0201	0.0103	0.0187

TABLE II
MODEL PERFORMANCES ON CORRELATION

Model	0.6B-prompt-free	0.6B-prompt-based	8B-prompt-free	8B-prompt-based	32B-prompt-free	32B-prompt-based
PV Correlation Difference	0.0177	0.0050	0.0004	0.0003	0.0021	0.0080
Overall Correlation Discrepancy	-0.0376	-0.0208	-0.0192	-0.0188	-0.0192	-0.0131
PV-ESS Output Deviation	0.0136	0.0106	0.0093	0.0090	0.0094	0.0081

As shown in Fig. 3, irradiance and temperature have a direct influence on the output of the PV system. The PV and ESS work collaboratively, with the PV unit generating power unidirectionally, while the ESS enables bidirectional regulation. When the PV output exceeds the actual demand, the excess energy is stored by charging the battery, resulting in an increase in SOC. Conversely, when the PV output is insufficient to meet the demand, the ESS actively discharges to compensate for the deficit, leading to a decrease in SOC.

B. Accuracy Evaluation

We calculate the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) between the real data and the restored data:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

where y_i is the ground truth, $\hat{y_i}$ is the restored data and n is the number of data points. Examples of missing data restoration are shown in Fig. 4, and the results are calculated in Table I. The following observations are made:

1. As model size increases, significant improvements are observed in both reconstructing masked positions and capturing temporal rules. From 0.6B to 32B parameters, MAE and RMSE consistently decrease across multiple forecasting tasks, indicating stronger capability in imputing missing values and modeling temporal dynamics. Notably, the gains become more pronounced when scaling beyond 8B parameters, suggesting that larger models better capture complex temporal patterns. As shown in Fig. 4(a)(b)(f), 0.6B-prompt-free model easily occurs unstable fluctuations, but larger ones almost never significantly deviating from the ground truth.

2. The integration of prompts further enhances model performance across all model sizes, particularly for larger variants. At smaller scales (e.g., 0.6B), prompts yield moderate improvements, while at larger scales (e.g., 32B), they enable finer-grained control and accuracy, as shown by additional reductions in MAE and RMSE. This indicates that

prompts effectively complement increased model capacity, enabling more precise and robust restorations in time-series modeling.

C. Correlation Evaluation

Considering the interdependencies among multiple variables, we assess the effectiveness using three metrics: 1) Given the strong correlation between irradiance and PV output, we compute the difference between the correlation coefficient of irradiance and PV output in the restored data and that in the true data. 2) We calculate the overall discrepancy by summing the differences between the correlation coefficients of all feature pairs in the restored data and those in the true data. 3) Based on the coupling relationship between PV units and storage systems, the total output should theoretically equal the sum of PV output and storage output. Therefore, we compute the deviation between the measured total output and the expected sum.

From the results in Table II, we made the following observations:

1. As model size increases, the ability to capture multivariate correlations improves significantly, as indicated by decreasing PV correlation difference and overall correlation discrepancy. Larger models better align with true correlation structures, especially between PV output and other variables, leading to more accurate representation of complex system behaviors. This enhanced correlation modeling contributes directly to improved understanding of underlying physical relationships, such as those between solar generation and energy storage dynamics, as reflected in reduced PV-ESS output deviation.

2. The integration of prompts further strengthens the model's capacity to capture variable interactions, particularly for larger architectures. Prompt-based models achieve lower correlation discrepancies and output deviations, indicating more precise characterization of both statistical dependencies and physical constraints. These improvements enable finer-grained understanding of energy system dynamics, where prompts help align model inference with domain-specific physical laws, resulting in more physically consistent and interpretable predictions.

D. Cost Analysis

There is a trade-off between model parameter scale and training cost. Larger models generally require more time and computational resources to train. Although lower error rates are desirable, they come at the expense of increased computational time and financial costs. In this study, since the models are trained on our own hardware rather than accessed via an external Application Programming Interface (API), there is no direct monetary cost. As model developers, we aim to strike a balance between performance and efficiency. To improve training efficiency, we employ Low-Rank Adaptation (LoRA) during fine-tuning. Training time and parameter counts are summarized in Table III. As model size and parameter count increase, so does training time. The use of prompts further extends training duration, mainly due to higher GPU memory consumption, which requires reducing the batch size to avoid out-of-memory errors. Considering both accuracy and computational cost, we recommend the 8B model. Using a larger model leads to unnecessary resource consumption with only marginal improvements in accuracy.

TABLE III
MODEL TRAINING COST

Model	Training time (h)	Training parameter	Training parameter percentage (%)
0.6B-prompt-free	0.19	10237958	1.6857
0.6B-prompt-based	0.46		
8B-prompt-free	0.71	34900998	0.4241
8B-prompt-based	1.92		
32B-prompt-free	2.51	84977670	0.2587
32B-prompt-based	10.12		

IV. CONCLUSION

In this paper, we introduce LLMs for missing data restoration in PV-ESSs, where they demonstrate superior performance in terms of both restoration accuracy and correlation preservation. Results show that even the 0.6B model without prompts can achieve restoration error below 5% and correlation error below 3%. As the model size increases, both accuracy and correlation improves. However, the marginal gains diminish as the parameter scale grows larger. Results also show that incorporating prompts enhances the model's understanding of the specific task, leading to further performance improvements.

REFERENCES

[1] C. S. Lai, Y. Jia, L. L. Lai, Z. Xu, M. D. McCulloch, and K. P. Wong, "A comprehensive review on large-scale photovoltaic system with applications of electrical energy storage," Renew. Sustain. Energy Rev., vol. 78, pp. 439–451, Oct. 2017.

[2] W. Zhang, Y. Luo, Y. Zhang, and D. Srinivasan, "SolarGAN: Multivariate Solar Data Imputation Using Generative Adversarial Network," IEEE Trans. Sustain. Energy, vol. 12, no. 1, pp. 743–746, Jan. 2021.

[3] C. Bülte, M. Kleinebrahm, H. Ü. Yilmaz, and J. Gómez-Romero, "Multivariate time series imputation for energy data using neural networks," Energy AI, vol. 13, p. 100239, Jul. 2023.

[4] H. Demirhan and Z. Renwick, "Missing value imputation for short to mid-term horizontal solar irradiance data," Appl. Energy, vol. 225, pp. 998–1012, Sep. 2018.

[5] Y. Li et al., "Load Profile Inpainting for Missing Load Data Restoration and Baseline Estimation," IEEE Trans. Smart Grid, vol. 15, no. 2, pp. 2251–2260, Mar. 2024.

[6] Y. Hu, K. Ye, H. Kim, and N. Lu, "BERT-PIN: A BERT-Based Framework for Recovering Missing Data Segments in Time-Series Load Profiles," IEEE Trans. Ind. Inform., vol. 20, no. 10, pp. 12241–12251, Oct. 2024.

[7] S. Majumder et al., "Exploring the capabilities and limitations of large language models in the electric energy sector," Joule, vol. 8, no. 6, pp. 1544–1549, Jun. 2024.

[8] X. Yang, C. Lin, H. Liu, and W. Wu, "RL2: Reinforce Large Language Model to Assist Safe Reinforcement Learning for Energy Management of Active Distribution Networks," IEEE Trans. Smart Grid, pp. 1–1, 2025.

[9] Y. Zuo, D. Qin, and Y. Wang, "Large Language Model-Empowered Interactive Load Forecasting," May 23, 2025, arXiv: arXiv:2505.16577.

[10] Y. Hu, H. Kim, K. Ye, and N. Lu, "Applying fine-tuned LLMs for reducing data needs in load profile analysis," Appl. Energy, vol. 377, p. 124666, Jan. 2025.