A Unified LLM–Transformer Approach for Probabilistic Electricity Demand and Generation Forecasting

Zehuan Hu[a], Yuan Gao[a,*], Gangwei Cai[b], Mingzhe Liu[c], Wei-An Chen[c], Yan Ke[d,**], Weijun Gao[e]

[a] *The Center for Energy Systems Design (CESD), International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, Japan*
[b] *Department of Architecture, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan*
[c] *Department of Multidisciplinary Engineering, Texas A&M University, College Station, TX, USA*
[d] *Department of Industrial Engineering, University of Roma Tor Vergata, Rome, Italy*
[e] *Faculty of Environmental Engineering, The University of Kitakyushu, 808-0135, Kitakyushu, Japan*

Abstract

Accurate forecasting of electricity demand and generation is essential for secure and efficient power system operations. While deep learning has improved prediction accuracy, most models are limited to deterministic outputs and lack uncertainty quantification. This study proposes LLMformer, a novel probabilistic forecasting framework that integrates a frozen large language model (LLM) encoder, a Transformer decoder, a lightweight probabilistic output module, and an adaptive loss function. Experiments using real-world data from five Japanese regions demonstrate that LLMformer outperforms state-of-the-art models in both deterministic and probabilistic forecasting, achieving up to 31.8% error reduction and strong generalization under domain shifts. The proposed method enables accurate and robust decision-making across diverse energy scenarios.

Keywords: Large language model; Electricity demand and generation forecasting; Probabilistic forecasting; Model generalization ability; Adaptive loss function

*Corresponding author
**Corresponding author2
Email addresses: yuangao1120@gmail.com (Yuan Gao), yanke@mail.zjgsu.edu.cn (Yan Ke)

1. Introduction

1.1. Background

In recent decades, global electricity demand has been steadily increasing, driven by economic growth, population expansion, and the widespread electrification of industries and transportation [1]. Concurrently, there has been a significant shift towards renewable energy sources such as solar, wind, and hydropower, driven primarily by environmental concerns and policy-driven decarbonization targets [2, 3]. This transition poses unique challenges due to the intermittent and fluctuating nature of renewable energy generation, thereby creating new demands for efficient management and accurate forecasting of electricity demand and supply [4, 5].

Ensuring stable and reliable power system performance hinges on accurate forecasting of electricity demand and generation [6]. Inaccurate predictions can lead to severe economic losses, grid instability, and even blackouts due to the mismatch between supply and demand [7, 8]. Conversely, precise forecasting enables optimal operation and planning of generation resources, efficient market participation, improved grid resilience, and facilitates the effective integration of renewable energy sources [9]. Therefore, developing robust forecasting methodologies remains essential for enhancing the operational efficiency and sustainability of modern electricity systems [10].

1.2. Electricity demand and generation forecasting

Currently, most forecasting methodologies for electricity demand and generation are primarily deterministic, aiming to predict single-point values [11]. However, deterministic forecasts inherently carry significant risks, particularly in energy system management and control, due to the inevitable uncertainty in real-world data and operational conditions [12]. Probabilistic forecasting, also known as interval forecasting, has emerged as a promising approach to mitigate such uncertainties by providing prediction intervals that quantify the

range within which actual outcomes are likely to fall with a specified confidence level [13].

Present probabilistic forecasting methods can be broadly classified into two categories. The first category involves models directly outputting specific quantile predictions (such as the 10th, 50th, and 90th percentiles), from which mean values and confidence intervals can be derived [14]. However, a notable limitation of this approach is its inflexibility, as a model trained in this manner generates intervals at fixed quantiles and thus cannot adapt dynamically to varying confidence requirements. The second category assumes that prediction errors conform to a known probability distribution, typically Gaussian, thereby allowing models to output parameters such as mean and standard deviation [15]. This approach provides greater flexibility, effectively catering to a wide range of forecasting tasks and confidence levels.

Recently, there has been growing interest in hybrid forecasting systems that combine deterministic and probabilistic techniques, or integrate multiple model architectures such as CNN-LSTM, attention-based RNNs, and Transformer hybrids [16, 17]. These models often improve accuracy but typically require task-specific designs, high computational cost, and complex data preprocessing. Moreover, most hybrid methods lack flexibility for uncertainty quantification across varying forecast horizons. Despite its potential advantages, research specifically focused on probabilistic forecasting for electricity demand and generation remains relatively limited [18]. Recent representative studies on probabilistic forecasting are summarized in Table 1. Although numerous studies on probabilistic forecasting have emerged recently, most have relied on traditional machine learning methods such as regression, requiring extensive data analysis and preprocessing.

Despite progress in deep learning-based probabilistic forecasting, challenges remain in achieving generalization under domain shifts, integrating both deterministic and probabilistic outputs within a unified framework, and enabling practical deployment in resource-constrained settings. These gaps motivate the development of a lightweight, generalizable, and interpretable forecasting archi-

tecture that can robustly handle diverse tasks with minimal tuning.

Table 1: Literature review for probabilistic forecasting.

Ref.	**Methods**	**Targets**
[19]	Autoregressive moving average model (ARIMA); Gaussian process regression; maximum likelihood estimation	Renewable energy markets
[15]	LSTM with quantile regression	Wind speed
[14]	Quantile regression forest machine learning	Dissolved oxygen concentration
[20]	Conditional normalizing flow	Wind power
[21]	Ensemble probability patch transform and monotonic composite quantile causal temporal convolutional networks	Carbon prices
[22]	Transformer-based model	Charging load of electric vehicle charging stations
[23]	Encoder–Decoder sequence-to-sequence	Renewable energy sources
[24]	Monotone broad learning system and Copula theory	Photovoltaic power
[25]	Quantile regression random forest	Short-term load of power systems

1.3. Large Language Models for time-series forecasting

Large Language Models (LLMs), such as ChatGPT, Geimini and LLaMA, have brought transformative advancements to a wide range of fields, including Natural Language Processing (NLP), machine translation, and text gener-

ation [26]. Distinguished by their large parameter scales and ability to handle massive datasets, these models excel in understanding and producing human-like language [27]. Beyond traditional NLP tasks, LLMs are being increasingly adopted in domains like healthcare, legal analysis, and financial services, where complex and context-aware decision-making is required [28]. The rapid evolution of Transformer-based architectures underpinning these LLMs has further facilitated efficient modeling of complex sequential patterns, opening potential avenues for applications beyond textual data [29].

Recent research has begun investigating the deployment of LLMs in time-series forecasting, capitalizing on their advanced capabilities in capturing sequential dependencies and extensive temporal correlations [30]. Traditional methods, such as ARIMA or LSTM, often struggle to represent full contextual information within longer sequences [31]. LLMs effectively overcome these challenges through parallel processing and improved management of long-range dependencies [32]. Existing studies suggest that forecasting techniques leveraging LLMs frequently achieve better performance compared to conventional forecasting models across various time-series tasks.

Currently, methodologies for applying LLMs to time-series forecasting are broadly divided into two categories. The first approach uses prompt-based reprogramming, embedding specifically crafted prompts within raw time-series inputs to direct pretrained LLMs toward accurate forecasting without altering model parameters [33]. Drawing inspiration from the LLMs' strong conversational and pattern-recognition capabilities, this method effectively utilizes their innate capacity to understand temporal patterns. The second category involves explicitly fine-tuning LLMs on targeted time-series datasets [34, 35]. For example, Jin et al. [36] proposed a novel reprogramming framework that leverages prompt-based inputs to align frozen LLMs with time-series data, significantly enhancing forecasting performance in few-shot and zero-shot scenarios. Chang et al. [37] proposed a two-step fine-tuning strategy, initially aligning the model's core representation with temporal features before fine-tuning a streamlined predictive component. A detailed summary and analysis of existing research on

applying LLMs for time-series forecasting can be found in Table 2. While many studies have demonstrated the feasibility of using LLMs for time-series forecasting, most applications have been restricted to deterministic predictions. Additionally, effective use of LLMs typically requires task-specific fine-tuning or carefully designed prompt engineering.

Table 2: Review of time-series forecasting methods using LLMs.

Ref.	**Forecast target**	**Methods for applying LLMs**
[38]	Energy load	Predefined template prompts
[39]	Personalized glucose	Prompt tuning
[40]	Wind speed	Spatial prompts and temporal prompts
[41]	Carbon price movements	Directly prompts
[42]	Stock return	Fine-tuning
[43]	Wind power	Fine-tuning
[44]	Climate	Predefined template prompts
[45]	Energy load	Fine-tuning
[46]	Wind power	Proposed soft and hard prompts
[47]	Methanol price	Fine-tuning

1.4. Research contribution

Based on the above background, existing research on electricity demand and generation forecasting, as well as the use of LLMs in time-series forecasting tasks, presents several limitations:

1) Although LLMs have demonstrated effectiveness in deterministic forecasting, research on their application to probabilistic forecasting remains scarce, and their full potential in probabilistic tasks is yet to be explored.

Additionally, whether traditional time-series forecasting models can be straightforwardly adapted for probabilistic forecasting remains uncertain.

2) LLMs typically consist of numerous modules and are trained on extensive datasets, complicating decisions regarding which parts should be fine-tuned. Fine-tuning LLMs for specific tasks can enhance accuracy but may simultaneously compromise their generalization capabilities.

3) Prompt engineering is widely used to enhance LLM performance in practical tasks. However, designing effective prompts tailored to different time-series tasks and their unique characteristics remains challenging.

4) Japan features significant regional differences in electricity demand, power generation, and climate conditions due to its extensive north-south geography. Such variability demands strong generalization capabilities from forecasting models. Nonetheless, research specifically targeting Japanese electricity data remains limited. Additionally, increasing extreme weather events and evolving policies could degrade model performance, yet studies evaluating model robustness under such extreme conditions are lacking.

To address these limitations, this study introduces the following innovations and contributions:

1) We propose a novel, simplified probabilistic forecasting module enabling traditional deterministic time-series models to output both the mean and standard deviation for probabilistic forecasts. A new loss function is also introduced to simultaneously ensure accuracy in both probabilistic distributions and deterministic predictions.

2) We propose a novel LLM-based probabilistic forecasting model that significantly enhances prediction accuracy and generalization ability without requiring fine-tuning or prompt engineering. Importantly, this model maintains comparable size and training times to traditional forecasting models.

3) To validate the proposed probabilistic forecasting method and LLM model, an extensive experimental analysis is conducted using electricity demand and generation data from five regions in Japan. Model performance is evaluated in detail, including robustness under extreme weather conditions and data drift scenarios.

2. Methodology

2.1. Problem statement

Traditional time series forecasting models, as illustrated in Fig. 1(a), usually require complete training of the entire model structure for each specific task [48]. These models are typically limited to either deterministic predictions (outputting only point estimates) or probabilistic predictions (outputting distributional results), and thus lack the flexibility to handle diverse forecasting tasks. Additionally, as the complexity of forecasting tasks increases, these models often become bulky and computationally expensive, limiting their practical efficiency and scalability.

In recent years, LLMs have shown promise in time series forecasting tasks. However, existing methods leveraging LLMs still face several challenges [49] (Fig. 1(b)). Typically, these approaches require task-specific fine-tuning of certain model components, which can be costly and time-consuming [50]. Additionally, effective application of LLMs demands carefully designed prompts tailored to specific tasks and datasets, complicating their deployment [51]. In order to solve the above problems, the proposed LLMformer method (Fig. 1(c)) introduces a lightweight and efficient forecasting framework that can simultaneously deliver both deterministic and probabilistic predictions without the need for fine-tuning the entire model or elaborate prompt engineering.

2.2. Proposed LLMformer model

Unlike conventional deep learning (DL) models that rely solely on numerical feature extraction, LLMs inherently learn hierarchical contextual dependencies

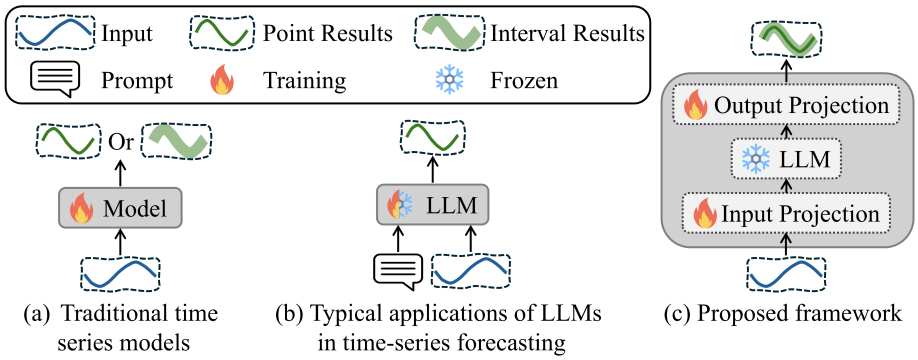


Figure 1: Illustrations of traditional time series forecasting methods, challenges in LLM-based forecasting, and the proposed approach.

from massive sequential data. Such representations can implicitly encode physical correlations (e.g., temperature–demand causality) through learned semantic patterns [52]. By reprogramming numerical sequences into tokenized embeddings, the LLM encoder provides a high-level contextual abstraction of temporal dynamics. This allows the subsequent Transformer decoder to focus on fine-grained temporal alignment, thereby enhancing both deterministic accuracy and probabilistic calibration.

Moreover, freezing the LLM ensures stability and prevents catastrophic forgetting, while maintaining its pre-trained knowledge on sequential dependencies [53]. This design bridges linguistic and temporal representations without requiring costly fine-tuning or handcrafted prompt engineering, enabling more generalizable forecasting across domains.

To address the existing challenges associated with prompt engineering and task-specific fine-tuning of LLMs, we propose a novel model architecture termed the LLMformer, as illustrated in Fig. 2. The LLMformer employs a pre-trained LLM as the encoder and a Transformer-based model as the decoder. Specifically, the encoder re-encodes input sequences into representations comprehensible to the LLM via a multi-head self-attention mechanism, thereby circumventing the need for complex prompt design. Meanwhile, the decoder, utilizing the Transformer structure, effectively translates the outputs of the LLM into desired

forecasting targets. Importantly, our approach avoids fine-tuning the LLM, thus preserving the generalization capability inherent to the pre-trained model.

The computational process of the proposed LLMformer model is detailed as follows. Initially, a word projection layer is introduced to map the original LLM vocabulary—which includes numerous words irrelevant to time-series contexts—onto a smaller, domain-specific set of word embeddings. This targeted vocabulary mapping allows the model to focus explicitly on representations most pertinent to time-series forecasting, thus significantly enhancing computational efficiency and accuracy. This approach mitigates redundancy inherent in standard LLM vocabularies when applied to time-series data. The mathematical formulation of the Word Projection layer is expressed as:

$$\boldsymbol{E} = \boldsymbol{V}\boldsymbol{W}_{\text{wproj}} + \boldsymbol{b}_{\text{wproj}}, \quad \boldsymbol{W}_{\text{wproj}} \in \mathbb{R}^{d_{\text{orig}} \times d_{\text{wproj}}}, \quad \boldsymbol{b}_{\text{wproj}} \in \mathbb{R}^{d_{\text{wproj}}} \tag{1}$$

where $\boldsymbol{E}$ is the output of this layer; $\boldsymbol{B}$ is the original LLM vocabulary embeddings; d_{orig} is the original vocabulary dimension of LLM; d_{wproj} is the dimension after word projection layer, which is 1000 in this study; $\boldsymbol{W}_{\text{wproj}}$ and $\boldsymbol{b}_{\text{wproj}}$ are the learnable weights of word projection layer.

Following the Word Projection layer, a multi-head attention mechanism is employed to integrate the transformed input features. This module adopts a query-key-value (QKV) structure, enabling each feature to interact comprehensively with others, effectively capturing complex interdependencies among variables. Consequently, the model adeptly translates time-series information into natural language representations comprehensible by the LLM. The multi-head attention mechanism is formally defined as follows:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_k, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_v \tag{2}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ indicate the query, key, and value matrices respectively; $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{f \times d}$, represent the learnable weights. In this model, the inputs is used as query, and the word vectors are used as key and value. Then the attention map $\boldsymbol{A}$ and output $\boldsymbol{Z}$ are

$$A(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_k})\boldsymbol{V} \tag{3}$$

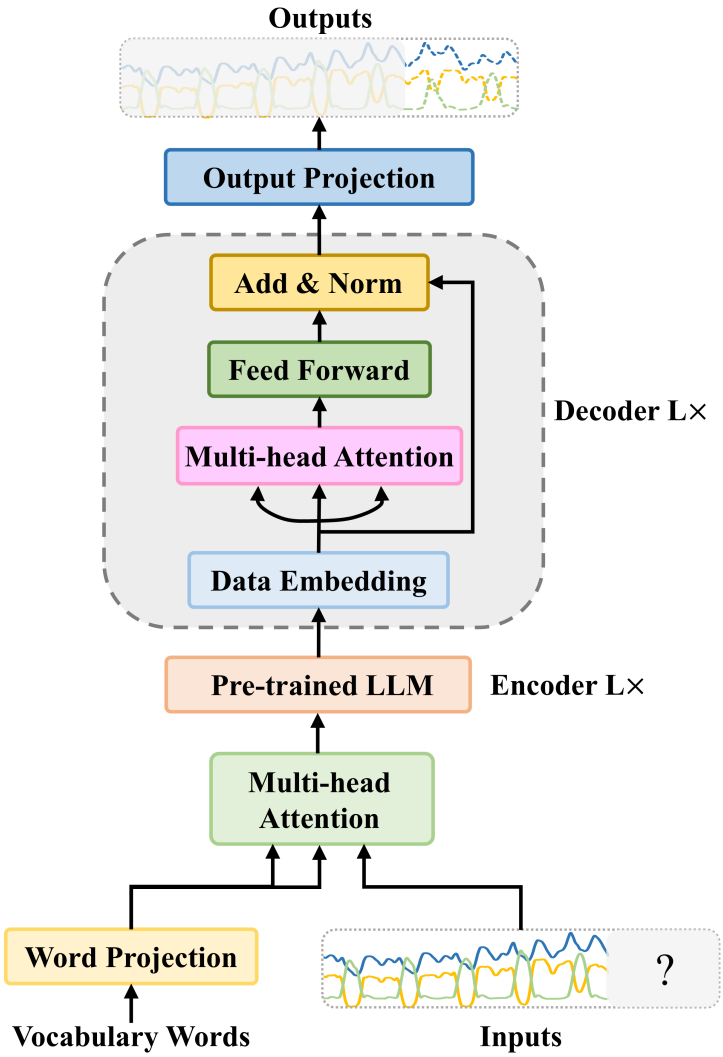


Figure 2: Proposed LLMformer architecture (The model consists of five main components: (1) Word projection layer; (2) Multi-head attention module; (3) Frozen pretrained LLM used as the encoder; (4) Transformer-based decoder; (5) Output projection layer).

$$\boldsymbol{Z} = A(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})\boldsymbol{W}_o, \quad \boldsymbol{W}_o \in \mathbb{R}^{d\times d}, \quad \boldsymbol{Z} \in \mathbb{R}^{T\times d} \tag{4}$$

where d_k is the dimension of queries, keys and values, which used to reduce the impact of the input data dimension on the results.

The subsequent layer involves the pre-trained LLM. By maintaining the integrity of the LLM, we leverage the rich linguistic comprehension capabilities inherent in the pre-trained model. Simultaneously, we focus on effectively reprogramming the inputs to align with the strengths of the LLM. The computational formulation for this layer is represented as follows:

$$\boldsymbol{H} = \mathrm{LLM}_{\mathrm{frozen}}(\boldsymbol{Z} + \mathrm{PE}), \quad \boldsymbol{H} \in \mathbb{R}^{T\times d} \tag{5}$$

Finally, the decoder layer transforms the natural language outputs from the LLM into desired time-series forecasting targets. Specifically, this layer adopts a Transformer decoder structure, processing the outputs from the LLM initially through a data embedding layer, followed by multi-head attention, a feed-forward neural network, and concluding with normalization. The multi-head attention mechanism within this decoder layer follows the same computation as previously defined, using the outputs from the LLM as query, key, and value components in the self-attention operation.

The feed-forward layer within the decoder is mathematically described as:

$$FFN(\boldsymbol{S}) = \max(0, \boldsymbol{S}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2, \quad \boldsymbol{W}_1 \in \mathbb{R}^{d\times d_{ff}}, \quad \boldsymbol{W}_2 \in \mathbb{R}^{d_{ff}\times d} \tag{6}$$

where we used a two-layer feedforward network, $\boldsymbol{W}_1$, $\boldsymbol{b}_1$, $\boldsymbol{W}_2$, and $\boldsymbol{b}_2$ represent learnable weights and biases.

After passing through these layers, the decoder produces the final forecasted time-series outputs, effectively converting linguistic representations back into quantitative predictions, thus completing the translation from the natural language domain to the time-series forecasting task.

2.3. Benchmark

To validate the predictive accuracy and generalization capability of our proposed model, we selected six widely utilized and well-performing DL models from previous time-series forecasting studies and one traditional regression

method. These DL models include basic recurrent neural network such as Long-short term memory network, Transformer-based models such as Transformer, Informer, and iTransformer, and two state-of-the-art models, TimesNet and PatchTST. Quantile regression is selected as traditional method.

2.3.1. Long-short term memory network (LSTM)

Long Short-Term Memory (LSTM) is a classical recurrent neural network specifically designed to address the issues of vanishing and exploding gradients encountered in traditional RNNs [54]. By incorporating gating mechanisms, LSTM effectively captures feature's dependencies in sequential data, making it a foundational model in time-series forecasting.

2.3.2. Transformer

Transformer models leverage attention mechanisms rather than recurrent or convolutional structures to process sequential data [27]. The attention mechanism allows Transformers to efficiently handle long-range dependencies, making them particularly powerful for time-series forecasting tasks that involve extensive temporal contexts.

2.3.3. Informer

Informer is an advanced Transformer-based architecture designed for handling extremely long time-series data [55]. It incorporates ProbSparse self-attention and distillation techniques to reduce computational complexity, thereby enabling efficient forecasting over extended prediction horizons.

2.3.4. iTransformer

The iTransformer model improves upon the Transformer architecture by integrating an interactive attention mechanism [56]. This enhancement allows for a more dynamic representation of temporal dependencies and interactions among variables, significantly improving forecasting accuracy, particularly for multivariate time-series.

2.3.5. TimesNet

TimesNet is a State-Of-The-Art (SOTA) time-series forecasting model that employs adaptive decomposition techniques and advanced attention mechanisms [57]. This approach allows TimesNet to effectively capture complex temporal patterns and variable interactions, achieving excellent forecasting accuracy across diverse datasets and scenarios.

2.3.6. PatchTST

PatchTST is another SOTA model designed to efficiently handle complex time-series data by segmenting the data into smaller patches [58]. It leverages transformer-based architectures and multi-scale patch embedding strategies, significantly enhancing its ability to detect localized patterns and long-range dependencies, thus delivering robust forecasting performance.

2.3.7. Quantile Regression

Quantile Regression (QR) is a widely used classical approach for probabilistic forecasting [59]. It directly estimates specific quantiles (e.g., 10th, 50th, and 90th percentiles) of the target distribution, enabling interval forecasts without assuming any parametric form. Despite its simplicity and interpretability, QR often struggles with high-dimensional multivariate time series and long forecasting horizons, as it lacks temporal modeling capacity.

2.4. Probabilistic forecasting module

Probabilistic forecasting has emerged as a crucial advancement over traditional deterministic forecasting, as it provides not only predictions of expected outcomes but also quantifies uncertainties inherent to future scenarios. Unlike deterministic forecasts, which yield single-point estimates, probabilistic approaches offer distributions or intervals indicating the potential range and likelihood of future values, thereby significantly enhancing decision-making reliability in energy system management.

Traditionally, time-series forecasting models have predominantly focused on deterministic outputs, resulting in fixed predictions for each target. To overcome this limitation, we propose a simple yet effective probabilistic forecasting module capable of generating distributional outputs from models initially designed for deterministic predictions. This approach requires no additional input information, modifying only the output projection layer. Specifically, we replace the original weight-and-bias-based mapping with a Gaussian negative log-likelihood-based mapping, enabling the model to simultaneously output both the mean and standard deviation of the prediction targets, thus obtaining a complete probabilistic distribution.

The Gaussian likelihood function used in this approach is defined as follows:

$$N(x|\mu,\sigma) = -\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{7}$$

Based on the above formulation, the mean and standard deviation are computed through the following expressions:

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{W}_\mu + \boldsymbol{b}_\mu, \quad \boldsymbol{\sigma} = \log(1+\exp(\boldsymbol{X}\boldsymbol{W}_\sigma + \boldsymbol{b}_\sigma)) \tag{8}$$

where $\boldsymbol{X}$ represents the model output prior to the output projection layer. In typical time-series forecasting models, the dimension of $\boldsymbol{X}$ is (batch size, forecast horizon, dimension of model). After the transformations defined above, the resulting mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. have dimensions (batch size, forecast horizon, number of target features). Additionally, $\boldsymbol{W}_\mu$, $\boldsymbol{b}_\mu$, $\boldsymbol{W}_\sigma$ and $\boldsymbol{b}_\sigma$ are learnable parameters optimized during model training.

Employing this simple modification, any deterministic forecasting model can effortlessly output probabilistic forecasts by adding just two linear layers. Consequently, this method substantially enhances forecasting robustness without altering the underlying model architecture, ensuring broad applicability across various forecasting tasks and scenarios.

2.5. Adaptive loss function for deterministic forecasting and probabilistic forecasting

In addition to the proposed probabilistic forecasting module, we design an adaptive loss function that requires only the predicted mean and variance as inputs, without the need for any additional information. Compared with conventional loss functions such as Mean Squared Error (MSE) and Negative Log-Likelihood (NLL), this adaptive loss formulation better balances accuracy (closeness to the true value) and precision (sharpness of the predicted distribution). This enables the model to serve both deterministic and probabilistic forecasting tasks within a unified framework.

Furthermore, unlike fixed-weight strategies that require manual tuning and are often prone to bias, our proposed method introduces trainable weights, thus eliminating the need for manual intervention and ensuring better stability during training. The total loss is computed as:

$$\ell_{\text{adap}} = \frac{1}{2}(w_{\text{MSE}}\ell_{\text{MSE}} + w_{\text{NLL}}\ell_{\text{NLL}} + \log(w_{\text{MSE}}) + \log(w_{\text{NLL}})) \tag{9}$$

where w_{MSE} and w_{NLL} are learnable weights; ℓ_{MSE} and ℓ_{NLL} represent the Mean Squared Error and the Negative Log-Likelihood loss terms, respectively. These components are computed as follows:

$$\ell_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_i)^2 \tag{10}$$

$$\ell_{\text{NLL}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(y_i - \mu_i)^2}{2\sigma_i^2} + \frac{1}{2}\log(2\pi\sigma_i^2)\right] \tag{11}$$

Where, y_i denotes the ground truth value, μ_i and σ_i represent the predicted mean and standard deviation; N is the total number of predictions. The MSE term encourages accurate point forecasts, while the NLL term ensures the model outputs a calibrated probabilistic distribution. By jointly optimizing both objectives with adaptive weighting, the model effectively balances deterministic accuracy and probabilistic reliability.

In addition, in order to verify the effectiveness of our proposed method, we also compared it with the loss function of fixed weights. The loss function of

fixed weights is calculated as follows:

$$\ell_{\text{fix}} = \ell_{\text{MSE}} + \ell_{\text{NLL}} \tag{12}$$

2.6. Model setup

Hyperparameters for all baseline models are configured based on the optimal settings described in their original studies. Each parameter set is chosen to balance complexity and computational practicality, aiming to ensure efficient training and optimal forecasting results within available computational resources. These configurations are further tailored to match the specific attributes of the time-series data, including input dimensionality and sampling intervals, thus aligning closely with forecasting requirements. Table 3 provides an overview of the parameters for each model, detailing layer counts, hidden dimension sizes, attention heads, and other significant settings.

Table 3: Configuration of hyper-parameters for benchmarks and LLMformer (d_{model}: Dimension of model; d_{ff}: Dimension of feed-forward layer; l_{encoder}: Number of encoder layers; l_{decoder}: Number of decoder layers; f_{attn}: Attention factor).

Model	**Model hyper-parameter**				
	d_{model}	d_{ff}	l_{encoder}	l_{decoder}	f_{attn}
LSTM	256		2		
Transformer	256	1024	2	1	3
Informer	512	2048	4	1	5
iTransformer	512	512	3	1	3
TimesNet	32	32	4	1	
PatchTST	256	256	3	1	3
LLMformer	32	64	10	3	3

To guarantee a fair evaluation, a consistent training procedure is adopted across all models. Each model is trained using the Adam optimizer with an initial learning rate of 0.001. Training is conducted for 10 epochs, and the

learning rate is reduced to 95% of its previous epoch value at the end of each epoch to support stable convergence.

Experiments were executed in a uniform computational environment to ensure reproducibility. All models were implemented in PyTorch and trained on a computer equipped with an GPU acceleration which was provided by an NVIDIA GeForce RTX 5090 with 32 GB of memory.

3. Case study

3.1. Introduction of the dataset

The case study is based on a comprehensive dataset comprising hourly electricity data (both demand and generation) and weather data. The electricity dataset spans the years 2019 to 2023 and covers five representative regions across Japan, arranged geographically from north to south: Hokkaido, Tohoku, Tokyo, Kansai, and Kyushu. The dataset includes hourly records of total electricity demand as well as electricity generation broken down by energy source for five regions. All electricity data were collected from publicly available reports provided by regional power companies.

Fig. 3 presents the electricity demand and generation composition by source for the year 2023 in the selected regions. Among them, Kyushu and Kansai exhibit comparable energy patterns, with electricity demand and generation both around 100×10^6MWh. These two regions also display a relatively balanced mix between fossil fuels and renewable energy. In contrast, the remaining regions rely more heavily on fossil fuels for electricity generation. Tohoku's electricity demand is slightly lower than Kyushu and Kansai, while Hokkaido's demand is approximately 30×10^6MWh. Tokyo, which consists of a single metropolitan area, has the lowest electricity demand at approximately 28×10^3MWh, and its energy mix is dominated by fossil fuels, accounting for 83% of its generation.

Weather conditions in five corresponding locations—Sapporo (Hokkaido), Sendai (Tohoku), Tokyo (Tokyo), Osaka (Kansai), and Fukuoka (Kyushu)—play a critical role in influencing both electricity demand and renewable generation.

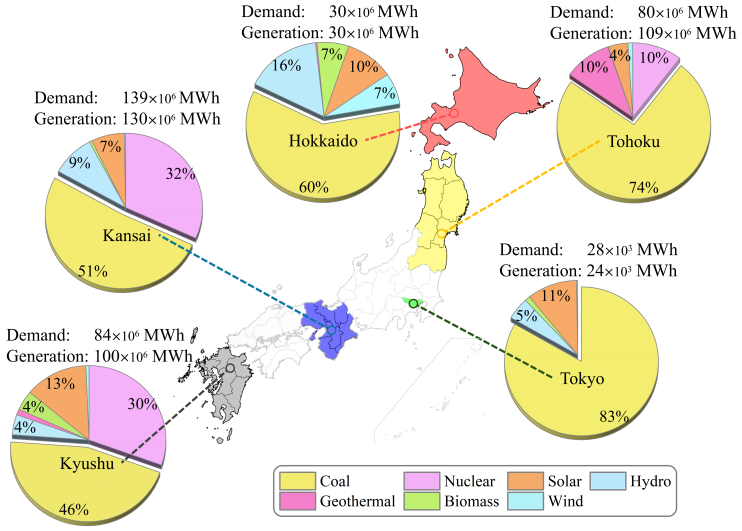


Figure 3: Summary of 2023 electricity demand and generation for five regions in Japan.

The weather dataset, provided by the Japan Meteorological Agency, spans from 2016 to 2023. As shown in Fig. 4, weather variables for 2023 include daily maximum and minimum temperatures and solar radiation, offering key insights into seasonal energy consumption and generation patterns. Each region exhibits distinct climatic characteristics:

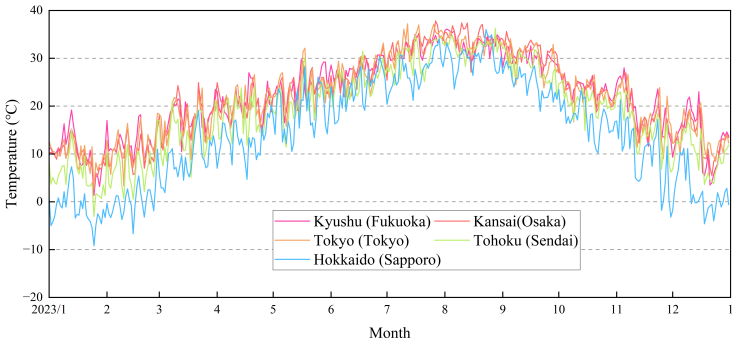
1) Hokkaido (Sapporo) experiences the coldest climate with long, harsh winters and significant seasonal temperature variations.

2) Tohoku (Sendai) has a temperate climate with cold winters and mild summers.

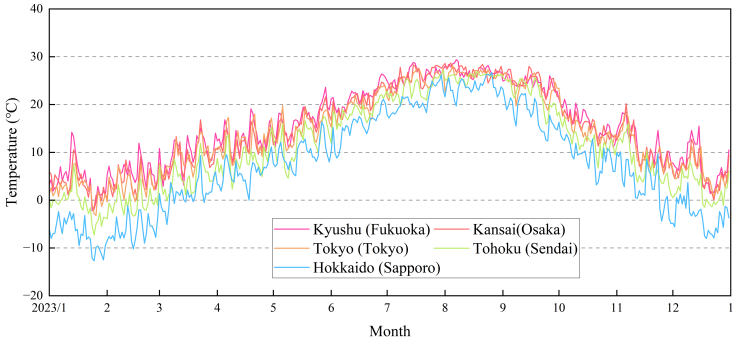
3) Tokyo (Tokyo) is characterized by a humid subtropical climate with hot summers and relatively mild winters.

4) Kansai (Osaka) sees warm and humid conditions year-round with a pronounced summer peak.

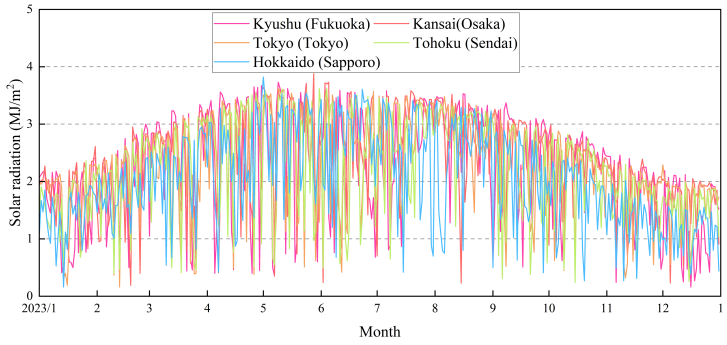
5) Kyushu (Fukuoka) enjoys a subtropical climate, with mild winters and the warmest annual temperatures among the selected regions.



(a) Daily maximum temperature



(b) Daily minimum temperature



(c) Daily maximum solar radiation

Figure 4: Daily meteorological data for representative regions in Japan in 2023.

A detailed summary of all features included in our dataset is provided in Table 4. These five regions were selected to ensure comprehensive representation of Japan's diverse climatic zones, ranging from the cooler northern areas to the warmer southern regions. Moreover, the diversity in energy supply structure and the wide range in electricity demand—spanning over four orders of magnitude—allow for a thorough evaluation of the proposed model's generalization performance under varied climatic and power generation conditions.

Table 4: Overview of Dataset Variables (Note: renewable energy generation is calculated as the total electricity generation minus fossil fuel-based generation; all data is recorded at one-hour intervals).

Category	**Variables**
Demand data	Total electricity demand
Electricity generation data	Fossil
	Nuclear
	Hydro
	Geothermal
	Biomass
	Solar
	Wind
	Renewable (non-fossil) energy
Weather data	Temperature
	Relative humidity
	Precipitation
	Dew point
	Vapor pressure
	Wind speed
	Sunshine duration
	Snowfall
	Global horizontal irradiance

3.2. Generalization to anomalous conditions

With the increasing frequency of extreme weather events driven by global climate change, electricity demand has become more volatile and prone to sudden surges [60]. In addition, gradual changes such as rising average temperatures or shifts in energy policy—often triggered by energy crises—can lead to year-over-year increases in baseline electricity demand [61]. Therefore, it is essential to assess a model's ability to generalize to data distributions that differ from the training conditions.

To simulate two common types of distribution shifts, we design two specific scenarios (Fig. 5):

1) Sudden demand surges from extreme events: To emulate the impact of extreme weather conditions or abrupt increases in electricity load, we select the period from July to September, which typically exhibits the highest annual electricity demand. During this interval, we randomly increase the peak demand values and their surrounding 6-hour window by 20% to 50%. This modification allows us to evaluate each model's ability to respond effectively to sharp and unexpected increases in power consumption. Meanwhile, the temperature will be randomly increased by 10-20%.

2) Gradual shifts due to climate or policy changes: To mimic long-term changes such as gradual warming or demand growth from policy-driven electrification, we increase the entire year's electricity demand by a random value ranging from 10% to 20%. This scenario assesses whether the models can maintain robust forecasting performance under slow, systemic shifts in the input distribution.

Given that renewable energy sources—such as solar, wind, and hydro—are inherently dependent on environmental conditions and cannot be arbitrarily scaled, we assume no increase in renewable generation [62]. To match the increased electricity demand in the above scenarios, we compensate by proportionally increasing fossil-fuel-based generation by the same absolute amount. This setup ensures energy balance while reflecting realistic operational constraints.

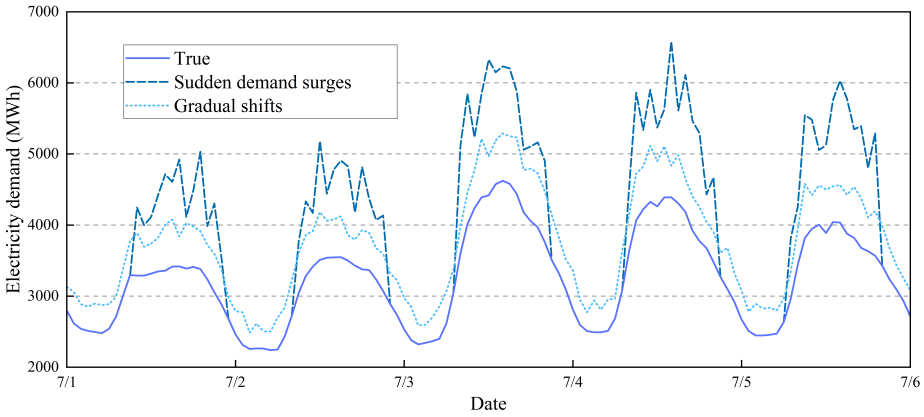


Figure 5: Electricity demand sudden surges and gradual shifts under extreme events and distribution shifts.

3.3. Data partitioning and preprocessing

To accurately evaluate model performance, the four-year dataset is divided into training, validation, and testing subsets. Specifically, the first two years are used for training, the third year for validation, and the final year for testing. During training, the model checkpoint with the best performance on the validation set is saved and used for final evaluation on the test set.

To prevent information leakage and ensure that the model does not have access to future information during prediction, we adopt a fixed-window forecasting strategy. In this setup, each model input consists of a complete historical window of fixed length (historical horizon), and the model outputs the full sequence of future values over a specified forecast horizon. Predictions are made at regular intervals equal to the length of the historical window, ensuring non-overlapping windows during evaluation and simulating realistic operational conditions.

Additionally, due to the differing scales of input variables—including weather features, electricity demand, and generation data—we apply data normalization to accelerate convergence during training. Normalization helps bring all variables to a common scale, which facilitates better optimization and reduces regression errors while preserving the internal structure of the dataset [63]. In this

study, we apply Z-score normalization, which transforms each variable to have zero mean and unit variance [64]. The normalization is computed as follows:

$$x' = \frac{x_{\text{ori}} - \mu_{\text{dataset}}}{\sigma_{\text{dataset}}} \tag{13}$$

where x' and x_{ori} represent the standardized data and original data, respectively; and μ_{dataset} and σ_{dataset} represent the mean and the standard deviation of the original data, respectively.

To interpret the model outputs in their original units, inverse normalization is applied during inference. The denormalization process is defined as:

$$x = x' \times \sigma_{\text{dataset}} + \mu_{\text{dataset}} \tag{14}$$

$$\mu = \mu' \times \sigma_{\text{dataset}} + \mu_{\text{dataset}} \tag{15}$$

$$\sigma = \sigma' \times \sigma_{\text{dataset}} \tag{16}$$

where μ' and σ' are the mean and the standard deviation of model outputs for probabilistic forecasting task; μ and σ are the denormalized mean and the standard deviation values.

This ensures that all evaluation metrics and forecast interpretations reflect realistic and meaningful values in the original data space.

3.4. Evaluation metrics

To comprehensively evaluate both the deterministic and probabilistic forecasting performance of the proposed and baseline models, we employ a set of widely-used quantitative metrics.

3.4.1. Deterministic forecasting metrics

To assess the accuracy of deterministic predictions, we adopt four commonly used evaluation metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE),and Coefficient of Determination (R^2). MAE provides a direct measurement of the average magnitude of prediction errors, regardless of direction. MAPE expresses the prediction error as a percentage of actual values, making it scale-independent. R^2 quantifies how well the predicted values fit the

actual data, indicating the proportion of variance explained by the model. The formulations for the four metrics are as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{17}$$

$$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{18}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{19}$$

where y_i, $\hat{y}_i$, and $\bar{y}$ represent the measured value, predicted value, and mean of measured value, respectively; N is the length of sequence.

3.4.2. Probabilistic forecasting metrics

To evaluate the probabilistic forecasting performance, we utilize three metrics: Continuous Ranked Probability Score (CRPS), Prediction Interval Coverage Probability (PICP), and Prediction Interval Width (PIW). CRPS measures the distance between the predicted cumulative distribution and the actual observation, offering a comprehensive score for distributional accuracy. PICP evaluates the proportion of true values that fall within the predicted confidence interval, assessing the reliability of uncertainty quantification. PIW quantifies the average width of the prediction interval, reflecting the sharpness of the forecast distribution.

As PICP and PIW often present a trade-off—higher coverage typically leads to wider intervals—CRPS serves as an integrated metric that balances calibration and sharpness, offering a more intuitive reflection of overall probabilistic performance. The formulations for these metrics with Gaussian outputs are as follows:

$$\mathrm{CRPS}_i = \sigma_i\left[\frac{y_i - \mu_i}{\sigma_i}\left(2\left(\frac{y_i - \mu_i}{\sigma_i}\right) - 1\right) + 2\left(\frac{y_i - \mu_i}{\sigma_i}\right) - \frac{1}{\sqrt{\pi}}\right] \tag{20}$$

$$\mathrm{PICP}_\alpha = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}, \{y_i \in [\mu_i - z_\alpha \cdot \sigma_i, \mu_i + z_\alpha \cdot \sigma_i]\} \tag{21}$$

$$\mathrm{PIW}_{\alpha} = \frac{1}{N}\sum_{i=1}^{N} 2z_{\alpha} \cdot \sigma_i \tag{22}$$

where z_{α} is the critical value corresponding to the desired confidence level; α represents the width of the confidence interval, which is 90% in this study.

3.4.3. Selection of pre-trained LLMs

Given the wide range of available LLMs with varying sizes and architectures, selecting an appropriate model requires careful consideration of trade-offs between predictive accuracy, training time, and memory usage—especially when the goal is to deploy these models on local machines with limited computational resources.

In this study, we evaluate four representative and relatively lightweight LLMs: BERT (420 MB) [65], ChatGPT2 (522 MB) [66], LLaMA-3.2-1b (2.30 GB) [67], and LLaMA-3.2-3b (5.97 GB) [67]. Fig. 6 illustrates the trade-offs among these models in terms of memory footprint, training time per iteration, and forecasting performance measured by Mean Absolute Error (MAE).

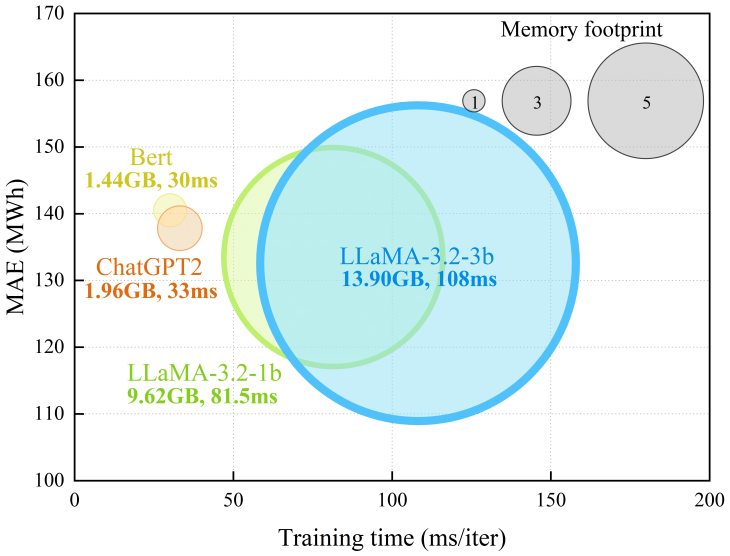


Figure 6: LLM efficiency comparison for proposed model (input horizons: 72; forecast horizons: 24).

The resulting MAEs for the four models are 140, 137, 133, and 132 MWh, respectively. While LLaMA-3.2-3b achieves the best forecasting accuracy, its training time is over three times longer than GPT2, and its memory usage is more than ten times higher. Given these resource constraints, we select GPT2 for the remainder of our experiments, as it provides a reasonable balance between computational cost and predictive performance.

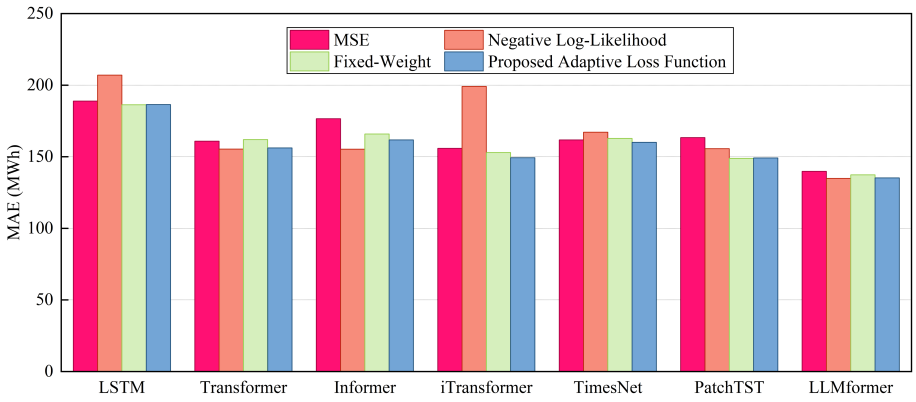
4. Results and discussions

4.1. Performance of the proposed probabilistic forecasting module and adaptive loss function

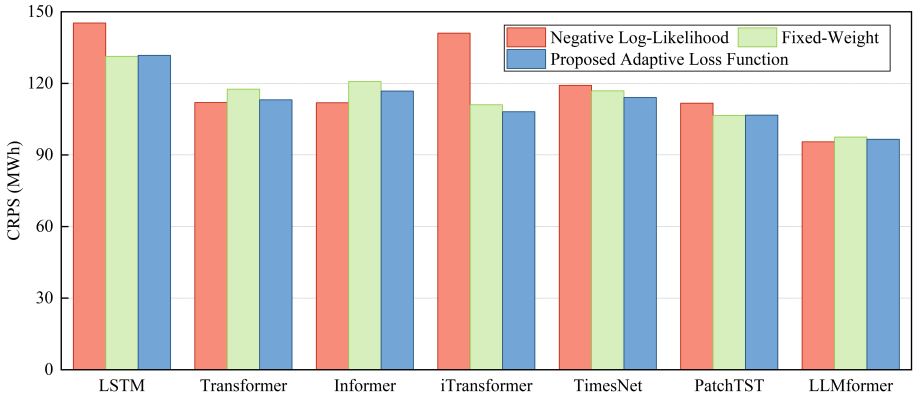
We first evaluate the effectiveness of the proposed probabilistic forecasting module and the adaptive weighted loss function. All baseline models are modified to incorporate our Gaussian-based probabilistic output layer, and their performance is compared under four different loss settings: Mean Squared Error (MSE), Negative Log-Likelihood (NLL), fixed-weight combination of MSE and NLL, and the proposed adaptive-weight loss function. The average performance across three forecasting targets is summarized in Fig. 7.

For deterministic forecasting tasks, we compute metrics using only the predicted mean μ from the models. It is worth noting that when training with MSE, the model does not utilize the predicted standard deviation σ, making our proposed module functionally equivalent to traditional deterministic output mappings in that case. As shown in Fig. 7(a), models trained with NLL loss generally yield the worst deterministic accuracy. The fixed-weight method performs better but is still inferior to MSE and the proposed adaptive loss. Our adaptive loss achieves comparable or even superior results to MSE in many models, demonstrating its ability to retain deterministic accuracy while supporting probabilistic output.

For probabilistic forecasting tasks, models trained with MSE are excluded as they do not output distributions. We therefore compare the NLL loss, the fixed-weight loss, and our adaptive-weight loss. As shown in Fig. 7(b), NLL loss



(a) MAE for deterministic forecast



(b) CRPS for probabilistic forecast

Figure 7: Performance of the proposed adaptive loss function for deterministic forecasting task and probabilistic forecasting task (input horizons: 72; forecast horizons: 24).

tends to achieve the lowest CRPS scores in most models, as expected. However, our adaptive loss achieves nearly identical CRPS values while also preserving deterministic accuracy, making it a balanced and robust solution. The fixed-weight approach provides a compromise between the two objectives but lacks consistency across models. For instance, Transformer and Informer perform best under NLL loss, while iTransformer shows better accuracy when trained with MSE. This highlights that a fixed weighting strategy may not yield optimal results across different architectures.

In contrast, the proposed adaptive loss function automatically learns appropriate weights during training, eliminating the need for manual tuning and enabling simultaneous optimization for both deterministic and probabilistic tasks. These results also validate the effectiveness of the proposed Gaussian-likelihood-based probabilistic output module. Without requiring any changes to the core architecture of existing deterministic time-series models, it enables them to deliver strong probabilistic forecasting performance.

4.2. Comparison between the proposed LLMformer and benchmark models

4.2.1. Forecasting performance across variables

The forecasting results for electricity demand and generation using both the proposed model and benchmark models are summarized in Table 5, which compares the performance across three targets: electricity demand, renewable generation, and fossil-based generation. The results demonstrate that LLMformer consistently achieves the best performance in both deterministic (MAE, MAPE, R^2) and probabilistic (CRPS, PICP, PIW) forecasting tasks. Notably, the LLMformer achieves the lowest MAE and CRPS for all three target variables, reflecting its strong capability in capturing both the central tendency and the uncertainty of time-series dynamics. Among all the benchmark methods, QR has the weakest performance, which shows that traditional regression-based probabilistic prediction methods are completely unsuitable for predicting complex tasks. Among all the DL models, LSTM has the weakest performance, while the results of Transformer, Informer, and iTransformer are relatively close.

Combining the indicators of the two tasks, PatchTST and TimesNet have the best results among the benchmark models.

Table 5: Electricity Demand and Generation Forecasting Performance (Input sequence: 72; Prediction horizon: 24; MAE and CRPS reported in MWh; Best and second-best results are highlighted in bold and underlined, respectively).

Target variable	Metrics	Model							
		QR	LSTM	Transformer	Informer	iTransformer	TimesNet	PatchTST	**LLMformer**
Electricity demand	MAE	301	198	158	175	140	180	162	**135**
	MAPE	0.09	0.06	0.05	0.05	0.04	0.05	0.05	**0.04**
	R^2	0.68	0.85	0.90	0.88	0.92	0.87	0.88	**0.92**
	CRPS	-	140	114	126	102	128	115	**96**
	PICP	0.79	0.87	0.80	0.79	0.93	**0.90**	0.87	0.86
	PIW	996	738	510	537	700	748	620	**489**
Generation (Renewable energy)	MAE	134	145	115	114	120	106	106	**101**
	MAPE	0.25	0.26	**0.19**	0.21	0.29	0.23	0.22	0.20
	R^2	0.71	0.67	0.78	0.79	0.78	0.81	0.81	**0.82**
	CRPS	-	101	82	80	87	76	74	**71**
	PICP	0.70	0.81	0.83	0.84	0.95	**0.93**	0.87	0.88
	PIW	406	400	**316**	380	597	481	387	365
Generation (Fossil energy)	MAE	318	216	196	196	188	194	179	**170**
	MAPE	0.15	0.10	0.09	0.09	0.09	0.09	0.08	**0.08**
	R^2	0.48	0.73	0.77	0.77	0.80	0.79	0.80	**0.82**
	CRPS	-	154	143	144	135	138	130	**122**
	PICP	0.77	0.83	0.74	0.74	0.91	**0.90**	0.87	0.85
	PIW	999	702	612	**534**	838	824	729	540
Average	MAE	251	186	156	162	149	160	149	**135**
	MAPE	0.17	0.14	0.11	0.12	0.14	0.12	0.12	**0.11**
	R^2	0.62	0.75	0.82	0.81	0.83	0.82	0.83	**0.85**
	CRPS	-	132	113	117	108	114	107	**97**
	PICP	0.75	0.83	0.79	0.79	0.93	**0.91**	0.86	0.87
	PIW	800	614	**456**	494	712	684	578	479

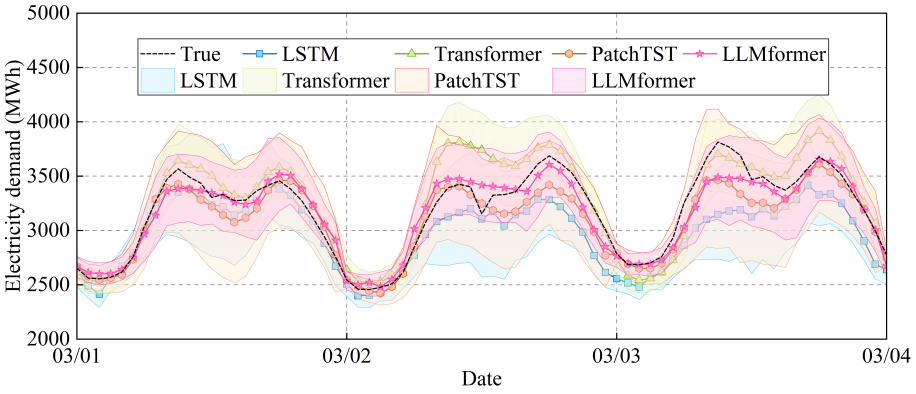
From the MAPE results, we observe that electricity demand forecasts are generally the most accurate across all models, with average percentage errors around 5%. This can be attributed to the relatively stable and predictable nature of electricity consumption throughout the year. In contrast, the prediction of renewable energy generation yields the highest MAPE, ranging from 20% to 30%, due to the inherently volatile and weather-dependent nature of solar, wind, and hydro resources. Fossil fuel generation falls in between, with moderate volatility and predictability.

4.2.2. Forecasting details and uncertainty quantification

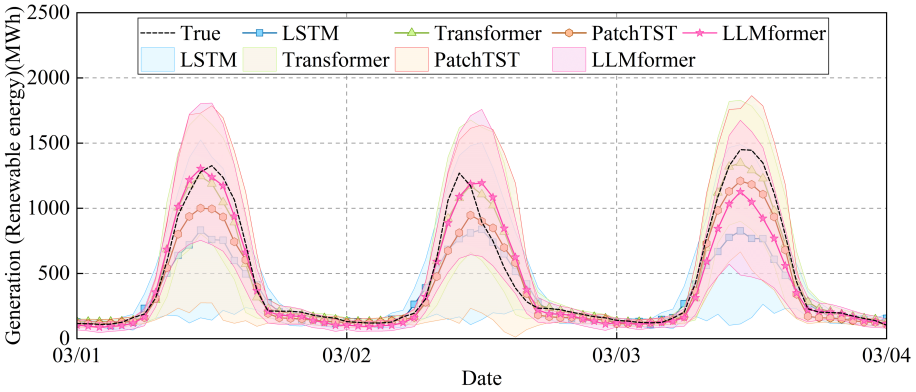
Fig. 8 visualizes detailed prediction results for electricity demand, renewable energy generation, and fossil fuel generation over a three-day period. We compare the proposed LLMformer with LSTM, Transformer, and PatchTST—the best-performing benchmark model. Across all three variables, LLMformer produces forecasts that not only align closely with the true values but also exhibit the narrowest 90% confidence intervals.

This improvement in both accuracy and sharpness of uncertainty bands is particularly noticeable during periods of high fluctuation. For instance, LLMformer maintains tight intervals and accurate predictions even during peak demand hours and rapid renewable generation spikes. Moreover, the prediction intervals for renewable generation are consistently the widest, reflecting the high variability inherent to weather-dependent energy sources. These observations confirm the model's ability to adapt to different uncertainty profiles across prediction tasks.

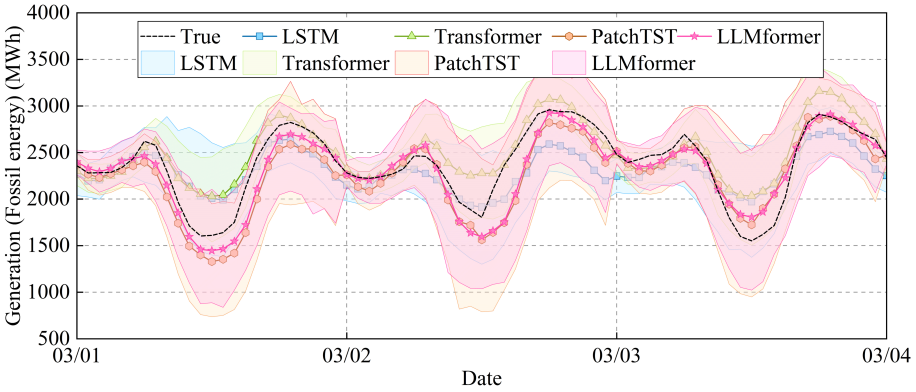
The interpretability of the proposed framework can be understood through the attention mechanisms in both the LLM encoder and Transformer decoder. The multi-head attention in the frozen LLM encoder captures latent associations between semantically similar temporal segments, while the decoder attention focuses on short-term dependencies and local variations. This hierarchical attention interaction allows the model to capture long-term seasonality and short-term weather-driven fluctuations simultaneously. The frozen LLM encoder, through its pre-trained linguistic and sequential knowledge, implicitly learns temporal semantics such as seasonal load variations, weekday–weekend cycles, and periodic demand peaks. Meanwhile, the decoder's attention layers are trained to associate recent input variations—such as sudden temperature increases—with corresponding shifts in demand or renewable generation. This division of labor allows the model to simultaneously capture broad seasonal trends and short-term causal signals like weather–demand interactions, enhancing both interpretability and performance.



(a) Electricity demand



(b) Electricity generation from renewable energy



(c) Electricity generation from fossil energy

Figure 8: Prediction results of median values and 90% confidence interval for electricity demand, renewable energy generation, and fossil fuel generation (input horizons: 72; forecast horizons: 24).

4.2.3. Model efficiency and resource utilization

Fig. 9 illustrates the model efficiency in terms of MAE, CRPS, training time, and memory footprint. Bubble size denotes memory usage during training, the bar height reflects training time per iteration, and the positions of the bubbles and red dashed lines correspond to MAE and CRPS, respectively. The forecasting results indicate that, compared to deterministic forecasting, probabilistic forecasting provides richer and more stable information. This is particularly evident at time points characterized by sudden data shifts or inherent volatility. Under these conditions, probabilistic forecasting consistently offers more robust insights, effectively capturing uncertainties and thus better supporting informed decision-making.

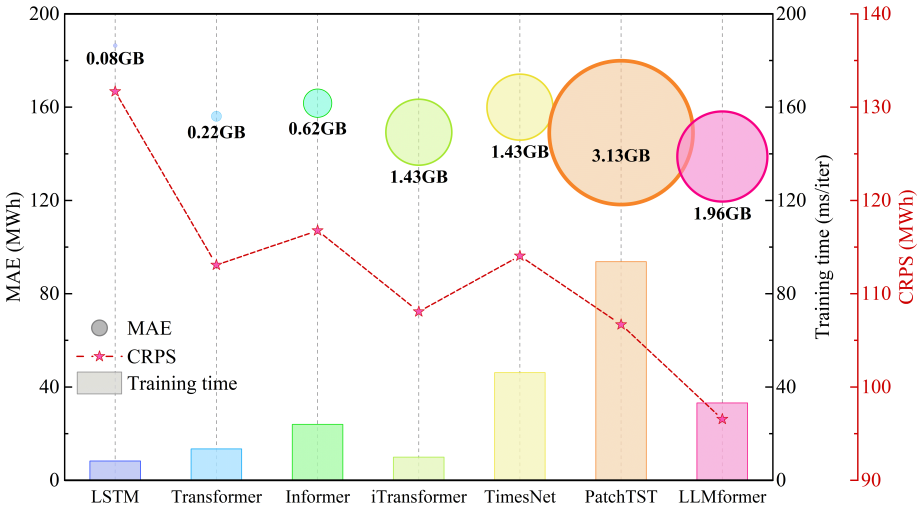


Figure 9: Model efficiency comparison (input horizons: 72; forecast horizons: 24).

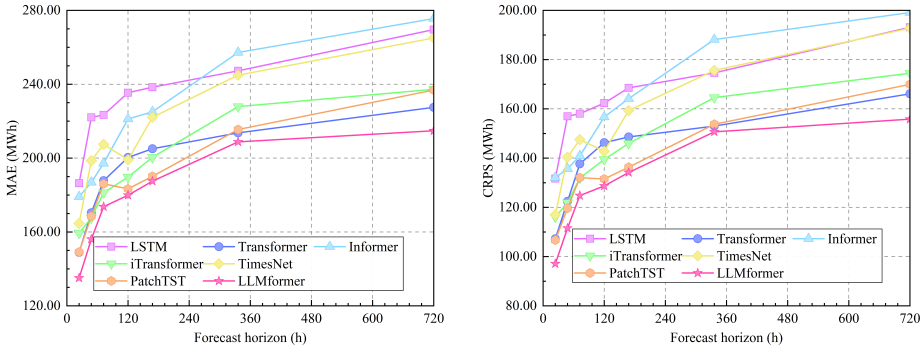
The LLMformer shows comparable training time and memory consumption to state-of-the-art time-series models such as iTransformer, TimesNet, and PatchTST. Specifically, compared to PatchTST, the LLMformer reduces memory usage by 37.38% and training time by 64.62%, while improving MAE by 9.40% and CRPS by 9.35%. These results indicate that the LLMformer achieves superior accuracy and uncertainty estimation without sacrificing efficiency, of-

fering a highly scalable and deployable solution for practical probabilistic forecasting.

In terms of computational efficiency, the proposed framework is compatible with standard GPU hardware and can be fine-tuned or extended with minimal resource overhead. The lightweight probabilistic module introduces negligible additional parameters and can be integrated into existing operational forecasting pipelines. Since the LLM encoder remains frozen, training is confined to the projection and decoder layers, ensuring scalability for real-time grid applications where retraining frequency is high. This characteristic makes LLMformer a practical solution for energy utilities seeking both accuracy and interpretability under resource constraints.

4.3. Impact of forecast horizon on model performance

To assess the robustness of each model across varying forecast lengths, we evaluate performance from short-term (24 hours) to long-term (720 hours, i.e., one month) forecast horizons. Fig. 10 illustrates the MAE and CRPS trends as the forecast horizon increases.



(a) MAE (b) CRPS

Figure 10: Forecasting performance across different forecast horizons (input horizons: 72).

As expected, all models exhibit a clear degradation in forecasting accuracy as the prediction horizon becomes longer. This trend is particularly steep in the short-term range (up to 240 hours), indicating that most models struggle to maintain precision as the target moves further into the future. However,

beyond this range, the performance decline begins to plateau, suggesting that the models adapt to long-term uncertainty with stabilized errors.

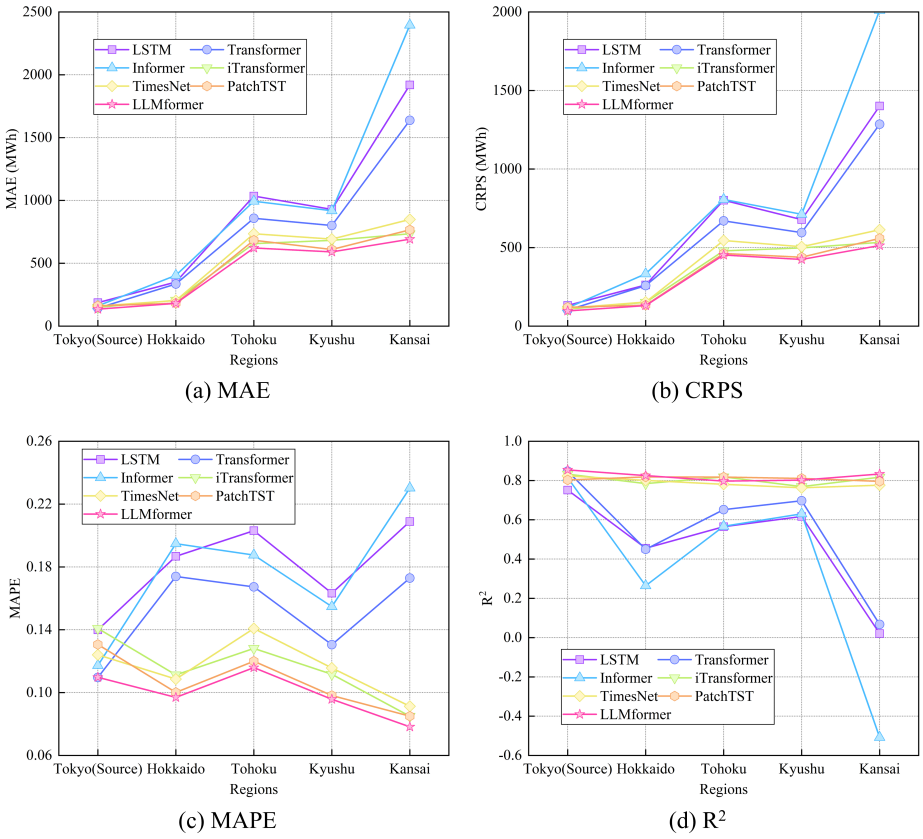
Throughout the entire range of forecast horizons, the proposed LLMformer consistently outperforms all benchmark models in both MAE and CRPS. This demonstrates the model's superior ability to retain accuracy and probabilistic calibration across both short- and long-term forecasting tasks. The advantage of LLMformer becomes increasingly pronounced at longer horizons, highlighting its strong generalization capacity and robustness to accumulated forecast uncertainty.

4.4. Generalization performance across regional datasets

To evaluate the cross-domain generalization ability of different models, we conduct zero-shot learning experiments where models are trained using data from one region and tested on all five target regions. Fig. 11 shows the results when the models are trained on data from Tokyo.

As shown in Fig. 11(a) and (b), which present the dimensioned metrics MAE and CRPS, the prediction error tends to increase with the scale of the target region's electricity demand. This is expected, as larger absolute values in the target series naturally lead to larger absolute prediction errors. Among all the regions, Hokkaido exhibits the most similar demand scale and generation structure to Tokyo, resulting in the smallest forecasting errors. In contrast, the Kansai region differs most significantly from Tokyo in both energy scale and supply composition, leading to the highest prediction errors across all models.

In contrast, the dimensionless metrics shown in Fig. 11(c) and (d)—MAPE and R^2 —reveal the models' normalized performance and generalization trends. Most advanced models, such as iTransformer, TimesNet, PatchTST, and LLMformer, exhibit relatively consistent performance across all regions, with MAPE around 12% and R^2 values close to 0.8. This suggests strong robustness and transferability in handling regional variability. In particular, the proposed LLMformer outperforms all other models on both deterministic and probabilistic metrics, showing the best generalization performance across regions.



(a) MAE (b) CRPS

(c) MAPE (d) R^2

Figure 11: Forecasting performance across different regions (input horizons: 72; forecast horizons: 24;).

Table 6 further summarizes the zero-shot forecasting results averaged over all region pairs, where each region is used once as the source domain. Regardless of the training region, the proposed LLMformer consistently achieves the best results across all test regions, demonstrating its strong adaptability to diverse regional patterns in electricity demand and generation.

Table 6: Average forecasting performance across target domains using different source regions (Results are averaged over target domains for each source region; input horizons: 72; forecast horizons: 24; MAE and CRPS are reported in MWh).

Target region	Metrics	Model						
		LSTM	Transformer	Informer	iTransformer	TimesNet	PatchTST	**LLMformer**
Tokyo	MAE	232	195	228	181	183	163	**153**
	MAPE	0.20	0.19	0.21	0.21	0.17	0.14	**0.14**
	R^2	0.66	0.73	0.66	0.76	0.79	0.80	**0.82**
	CRPS	171	142	169	130	130	123	**111**
	PICP	0.72	0.80	0.72	0.92	0.90	0.93	**0.85**
	PIW	615	602	608	816	785	850	**571**
Hokkaido	MAE	321	285	358	210	203	179	**175**
	MAPE	0.17	0.15	0.18	0.12	0.11	0.10	**0.10**
	R^2	0.52	0.56	0.38	0.77	0.80	0.82	**0.83**
	CRPS	239	215	279	152	149	131	**127**
	PICP	0.67	0.71	0.60	0.81	0.75	0.89	**0.81**
	PIW	852	776	774	773	679	812	**619**
Tohoku	MAE	979	852	983	709	749	628	**603**
	MAPE	0.19	0.17	0.19	0.15	0.15	0.12	**0.12**
	R^2	0.60	0.66	0.57	0.78	0.77	0.82	**0.82**
	CRPS	756	648	767	511	545	461	**442**
	PICP	0.58	0.67	0.59	0.85	0.78	0.91	**0.80**
	PIW	2034	2026	2038	2822	2636	3013	**2182**
Kyushu	MAE	944	841	953	752	725	594	**617**
	MAPE	0.16	0.14	0.16	0.13	0.12	0.10	**0.10**
	R^2	0.61	0.67	0.59	0.72	0.75	0.81	**0.81**
	CRPS	695	619	721	544	523	443	**436**
	PICP	0.67	0.73	0.65	0.86	0.82	0.92	**0.82**
	PIW	2468	2320	2386	3093	2766	3060	**2245**
Kansai	MAE	2144	1510	2214	808	848	765	**699**
	MAPE	0.24	0.16	0.24	0.09	0.09	0.09	**0.09**
	R^2	-0.30	0.25	-0.33	0.78	0.77	0.79	**0.82**
	CRPS	1656	1154	1755	581	612	560	**518**
	PICP	0.67	0.65	0.49	0.87	0.83	0.92	**0.85**
	PIW	5955	3733	3970	3289	3274	3578	**2775**

Although the current analysis focuses on Japan, the proposed framework is data-agnostic and can be directly applied to other electricity markets or climatic regions. Since the LLM encoder captures abstract temporal dependencies

rather than region-specific features, the model exhibits strong transferability across countries with minimal retraining. Future studies may extend this framework to international datasets or integrated energy systems, further verifying its scalability and robustness.

4.5. Performance under anomalous conditions

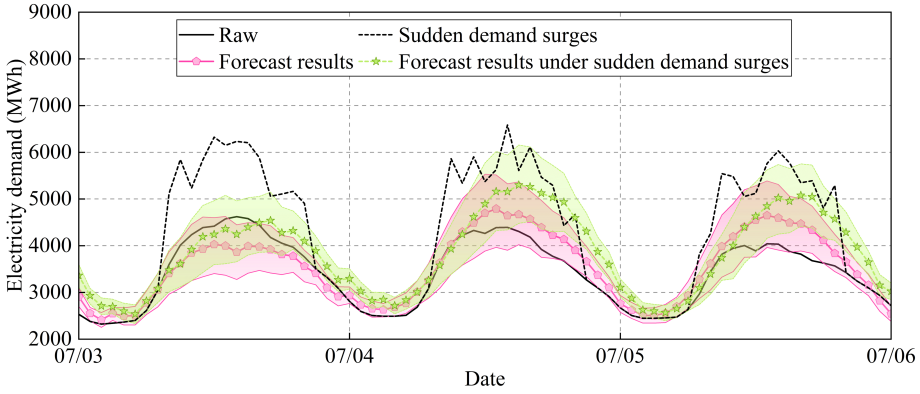
4.5.1. Robustness under sudden demand surges

To evaluate the robustness of the models under abrupt changes, we simulate scenarios in which electricity demand experiences sudden surges due to extreme weather or unexpected load events. Table 7 summarizes the performance of each model under both normal and sudden demand surge conditions, while Fig. 12 visualizes the predictive distributions across three representative models: LSTM, iTransformer, and the proposed LLMformer.

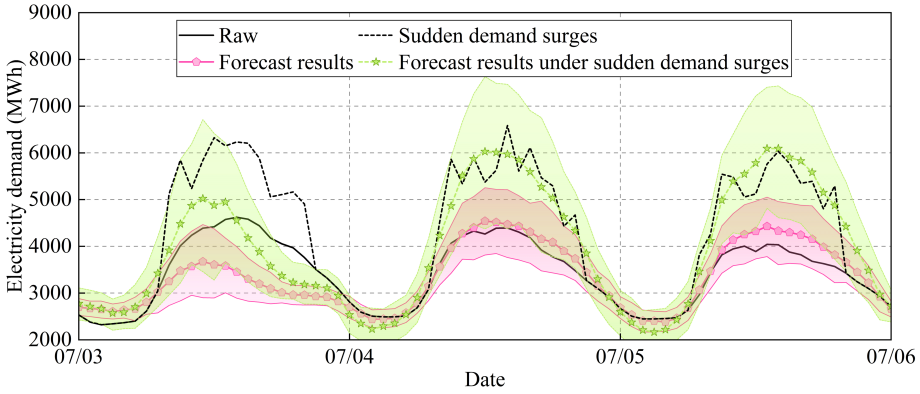
Table 7: Forecasting performance under raw and sudden demand surge conditions (input horizons: 72; forecast horizons: 24; MAE and CRPS are reported in MWh).

Prediction conditions	Metrics	Model						
		LSTM	Transformer	Informer	iTransformer	TimesNet	PatchTST	**LLMformer**
Raw	MAE	149	132	129	118	131	121	**102**
	MAPE	0.81	0.78	0.80	0.93	0.93	0.86	**0.86**
	R^2	692	535	578	772	833	636	**523**
	CRPS	209	180	179	159	183	169	**141**
	PICP	0.12	0.11	0.11	0.12	0.11	0.10	**0.09**
	PIW	0.71	0.77	0.78	0.83	0.80	0.82	**0.86**
Sudden surges	MAE	468	365	433	289	346	268	**264**
	MAPE	0.43	0.53	0.47	0.87	0.84	0.89	**0.79**
	R^2	718	618	647	1489	1589	1704	**1028**
	CRPS	566	446	523	396	480	357	**357**
	PICP	0.22	0.17	0.18	0.17	0.18	0.14	**0.15**
	PIW	0.50	0.67	0.58	0.73	0.62	0.76	**0.78**

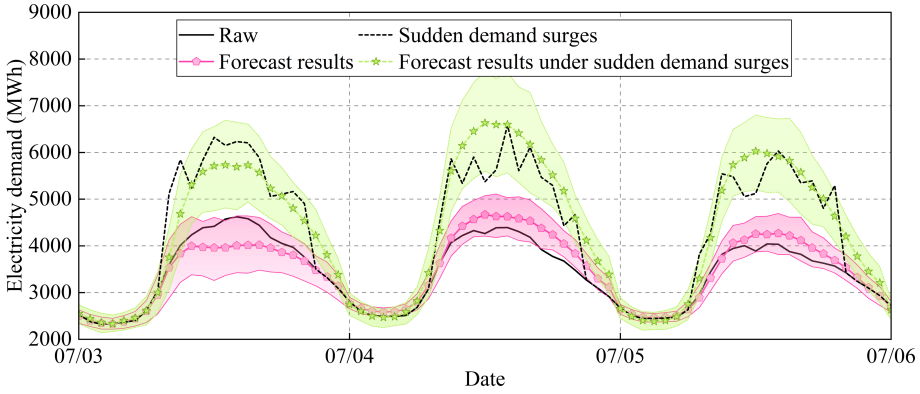
The results reveal that all models experience a noticeable drop in performance when exposed to conditions not seen during training. This is particularly evident for conventional models such as LSTM and iTransformer (Fig. 12(a) and (b)), which tend to either maintain unchanged forecasts despite demand shifts or expand prediction intervals excessively. These behaviors lead to reduced accuracy in both point and probabilistic forecasts, as reflected by significantly increased MAE and CRPS, and deteriorated PICP values.



(a) LSTM



(b) iTransformer



(c) LLMformer

Figure 12: Probabilistic forecasting results of LSTM, iTransformer and LLMformer under raw and sudden demand surge conditions (input horizons: 72; forecast horizons: 24).

In contrast, the proposed LLMformer (Fig. 12(c)) demonstrates strong adaptability to such distributional shifts. It maintains accurate forecast trajectories and well-calibrated uncertainty bands, even when demand spikes deviate significantly from historical patterns. As shown in Table 7, LLMformer achieves the lowest MAE and CRPS under sudden surge conditions while preserving competitive PICP and PIW values. These results highlight the model's superior capacity to generalize beyond the training distribution and maintain reliable performance in real-world operational scenarios.

4.5.2. Performance under gradual distribution shifts

To further evaluate the robustness of each model, we examined their forecasting performance under gradual shifts in the data distribution, which may arise from seasonal transitions, long-term demand growth, or changes in consumption behavior. As shown in Table 8, we compare model accuracy on the raw test set with performance under a shifted data distribution.

Table 8: Forecasting performance under raw and gradual shifts conditions (input horizons: 72; forecast horizons: 24; MAE and CRPS are reported in MWh).

Prediction conditions	Metrics	Method						
		LSTM	Transformer	Informer	iTransformer	TimesNet	PatchTST	**LLMformer**
Raw	MAE	132	113	117	108	114	107	**97**
	MAPE	0.83	0.79	0.79	0.93	0.91	0.87	**0.86**
	R^2	614	456	479	712	684	578	**494**
	CRPS	186	156	162	149	160	149	**135**
	PICP	0.14	0.11	0.12	0.14	0.12	0.12	**0.11**
	PIW	0.75	0.82	0.81	0.83	0.82	0.83	**0.85**
Gradual shifts	MAE	173	140	145	125	133	130	**115**
	MAPE	0.74	0.72	0.72	0.92	0.89	0.83	**0.82**
	R^2	645	503	531	795	771	675	**547**
	CRPS	237	190	198	174	186	182	**161**
	PICP	0.14	0.11	0.12	0.14	0.13	0.12	**0.11**
	PIW	0.71	0.81	0.80	0.83	0.82	0.82	**0.85**

The results indicate that gradual shifts lead to relatively mild degradation in performance compared to the sudden demand surges analyzed previously. Most advanced models, including iTransformer, TimesNet, PatchTST, and the proposed LLMformer, successfully maintain stable forecasting accuracy. In contrast, simpler architectures such as LSTM and Transformer show noticeable

performance drops under distributional shift, with increases in MAE and CRPS and decreases in R^2 .

Among all models, LLMformer consistently achieves the best performance across all metrics under both raw and shifted conditions. For instance, LLMformer yields the lowest MAE (115 MWh), the lowest CRPS (161 MWh), and the highest R^2 (0.82) under gradual shifts. These results suggest that the proposed model is not only accurate but also robust to moderate changes in the data distribution, highlighting its suitability for deployment in real-world scenarios where operational conditions are subject to drift over time.

5. Conclusion

Deterministic forecasting, while widely used, often suffers from limitations such as overconfidence and lack of uncertainty quantification, which can lead to suboptimal decisions in energy system control and planning. In contrast, probabilistic forecasting provides more robust and informative guidance for operational and strategic decision-making. Meanwhile, LLMs have recently demonstrated strong capabilities in time series forecasting. However, their application to probabilistic forecasting remains challenging, as it typically requires task-specific prompt engineering or costly fine-tuning.

To address these challenges, this study proposed a novel forecasting architecture based on an encoder–decoder framework, where an LLM is used as the encoder to leverage its contextual understanding of time series data, and a Transformer-based decoder is employed to convert LLM embeddings into forecasted sequences. Additionally, we introduced a lightweight probabilistic forecasting module along with an adaptive weighted reward function, enabling conventional models originally designed for point forecasting to produce probabilistic outputs. The proposed method is applicable to both deterministic and probabilistic forecasting tasks. The main findings are summarized as follows:

1) The proposed probabilistic output module and adaptive weighted reward function enable various forecasting models to produce distributional out-

puts without altering their core architecture. This method achieves comparable accuracy to standard MSE and NLL loss-based models while allowing seamless switching between deterministic and probabilistic forecasting tasks.

2) The proposed LLMformer consistently outperforms benchmark models including State-of-the-Art models across all three forecasting targets—electricity demand, renewable energy generation, and fossil energy generation—under various forecasting horizons. Compared with LSTM, LLMformer achieves up to a 31.82% reduction in forecasting error. In particular, forecasting tasks with less variability (e.g., electricity demand and fossil generation) show errors of approximately 10%, while those involving higher volatility (e.g., renewable generation) exhibit errors around 20%.

3) Zero-shot evaluation on datasets from five different Japanese regions indicates that models generally maintain high accuracy when the data scale and composition are similar to the source domain. However, performance degrades when the target domain significantly differs. Among all models, LLMformer demonstrates the best generalization ability, reducing errors by up to 63.96% compared to LSTM across regions.

4) All models experience notable performance drops under abrupt distributional shifts (e.g., sudden demand surges due to extreme weather). Traditional models tend to either ignore the shift or overcompensate by enlarging prediction intervals. In contrast, LLMformer maintains accurate and calibrated probabilistic forecasts even under such conditions. Under mild, gradual distribution shifts, all models perform relatively stably, with LLMformer again achieving the most robust results.

Despite the promising results, this study has several limitations that warrant further investigation. First, although the proposed LLMformer demonstrates strong performance and generalization ability, it relies on a frozen pretrained LLM as the encoder. This may limit its adaptability to domain-specific charac-

teristics, especially when the temporal patterns in the target data significantly differ from the general knowledge captured by the LLM. Incorporating fine-tuning or domain adaptation strategies may further enhance its effectiveness. Second, the probabilistic forecasting module introduced in this study focuses on Gaussian assumptions. Extending it to accommodate non-Gaussian or multi-modal distributions could improve its applicability to more complex real-world scenarios. Third, the evaluation was conducted using historical weather and electricity data from a limited number of regions in Japan. Future work could explore larger-scale, multi-country datasets, as well as real-time deployment and integration into actual energy management systems.

Acknowledgment

This work was supported by project No. 23KJ0766 funded by the Japan Society for The Promotion of Science. All the code of models and datasets are open sourced in LLMformer GitHub Repository.

References

[1] A. Aghahosseini, A. Solomon, C. Breyer, T. Pregger, S. Simon, P. Strachan, A. Jäger-Waldau, Energy system transition pathways to meet the global electricity demand for ambitious climate targets and cost competitiveness, Applied energy 331 (2023) 120401.

[2] G. S. Seck, E. Hache, J. Sabathier, F. Guedes, G. A. Reigstad, J. Straus, O. Wolfgang, J. A. Ouassou, M. Askeland, I. Hjorth, et al., Hydrogen and the decarbonization of the energy system in europe in 2050: A detailed model-based analysis, Renewable and Sustainable Energy Reviews 167 (2022) 112779.

[3] U. A. Saari, S. Damberg, L. Frömbling, C. M. Ringle, Sustainable consumption behavior of europeans: The influence of environmental knowledge and

risk perception on environmental concern and behavioral intention, Ecological Economics 189 (2021) 107155.

[4] A. Ahmed, T. Ge, J. Peng, W.-C. Yan, B. T. Tee, S. You, Assessment of the renewable energy generation towards net-zero energy buildings: A review, Energy and buildings 256 (2022) 111755.

[5] D. Adu, D. Jianguo, S. N. Asomani, A. Abbey, Energy generation and carbon dioxide emission—the role of renewable energy for green development, Energy Reports 12 (2024) 1420–1430.

[6] A. Román-Portabales, M. López-Nores, J. J. Pazos-Arias, Systematic review of electricity demand forecast using ann-based machine learning algorithms, Sensors 21 (13) (2021) 4544.

[7] S. Mertens, Design of wind and solar energy supply, to match energy demand, Cleaner Engineering and Technology 6 (2022) 100402.

[8] M. Sharma, N. Mittal, A. Mishra, A. Gupta, Survey of electricity demand forecasting and demand side management techniques in different sectors to identify scope for improvement, Smart Grids and Sustainable Energy 8 (2) (2023) 9.

[9] Z. Hu, Y. Gao, S. Ji, M. Mae, T. Imaizumi, Improved multistep ahead photovoltaic power prediction model based on lstm and self-attention with weather forecast data, Applied Energy 359 (2024) 122709.

[10] T. Gao, D. Niu, Z. Ji, L. Sun, Mid-term electricity demand forecasting using improved variational mode decomposition and extreme learning machine optimized by sparrow search algorithm, Energy 261 (2022) 125328.

[11] R. V. Klyuev, I. D. Morgoev, A. D. Morgoeva, O. A. Gavrina, N. V. Martyushev, E. A. Efremenkov, Q. Mengxu, Methods of forecasting electric energy consumption: A literature review, Energies 15 (23) (2022) 8919.

[12] N. Mounir, H. Ouadi, I. Jrhilifa, Short-term electric load forecasting using an emd-bi-lstm approach for smart grid energy management system, Energy and Buildings 288 (2023) 113022.

[13] H. Khajeh, H. Laaksonen, Applications of probabilistic forecasting in smart grids: A review, Applied Sciences 12 (4) (2022) 1823.

[14] M. H. Ahmed, L.-S. Lin, Dissolved oxygen concentration predictions for running waters with different land use land cover using a quantile regression forest machine learning technique, Journal of Hydrology 597 (2021) 126213.

[15] A. Saeed, C. Li, Z. Gan, Y. Xie, F. Liu, A simple approach for short-term wind speed interval prediction based on independently recurrent neural networks and error probability distribution, Energy 238 (2022) 122012.

[16] S. Ghimire, R. C. Deo, D. Casillas-Perez, S. Salcedo-Sanz, S. A. Pourmousavi, U. R. Acharya, Probabilistic-based electricity demand forecasting with hybrid convolutional neural network-extreme learning machine model, Engineering Applications of Artificial Intelligence 132 (2024) 107918.

[17] N. Wei, C. Yin, L. Yin, J. Tan, J. Liu, S. Wang, W. Qiao, F. Zeng, Short-term load forecasting based on wm algorithm and transfer learning model, Applied Energy 353 (2024) 122087.

[18] H. Panamtash, S. Mahdavi, Q. Zhou, Probabilistic solar power forecasting: A review and comparison, in: 2020 52nd North American Power Symposium (NAPS), IEEE, 2021, pp. 1–6.

[19] M. Shamsi, P. Cuffe, Prediction markets for probabilistic forecasting of renewable energy sources, IEEE Transactions on Sustainable Energy 13 (2) (2021) 1244–1253.

[20] H. Wen, P. Pinson, J. Ma, J. Gu, Z. Jin, Continuous and distribution-free probabilistic wind power forecasting: A conditional normalizing flow approach, IEEE Transactions on Sustainable Energy 13 (4) (2022) 2250–2263.

[21] T. Han, X. Gu, D. Li, K. Chen, R.-G. Cong, L.-T. Zhao, Y.-M. Wei, Causal neural network for carbon prices probabilistic forecasting, Applied Energy 397 (2025) 126343.

[22] X. Huang, D. Wu, B. Boulet, Metaprobformer for charging load probabilistic forecasting of electric vehicle charging stations, IEEE Transactions on Intelligent Transportation Systems 24 (10) (2023) 10445–10455.

[23] J. J. Quiñones, L. R. Pineda, J. Ostanek, L. Castillo, Towards smart energy management for community microgrids: Leveraging deep learning in probabilistic forecasting of renewable energy sources, Energy Conversion and Management 293 (2023) 117440.

[24] N. Zhou, X. Xu, Z. Yan, M. Shahidehpour, Spatio-temporal probabilistic forecasting of photovoltaic power based on monotone broad learning system and copula theory, IEEE Transactions on Sustainable Energy 13 (4) (2022) 1874–1885.

[25] S. Dang, L. Peng, J. Zhao, J. Li, Z. Kong, A quantile regression random forest-based short-term load probabilistic forecasting method, Energies 15 (2) (2022) 663.

[26] T. B. Brown, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[27] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

[28] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al., A survey on large language models: Applications, challenges, limitations, and practical usage, Authorea Preprints (2023).

[29] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, R. McHardy, Challenges and applications of large language models, arXiv preprint arXiv:2307.10169 (2023).

[30] Z. Hu, Y. Gao, L. Sun, M. Mae, A novel attention-enhanced llm approach for accurate power demand and generation forecasting, Renewable Energy (2025) 123465.

[31] A. Nazir, A. K. Shaikh, A. S. Shah, A. Khalil, Forecasting energy consumption demand of customers in smart grid using temporal fusion transformer (tft), Results in Engineering 17 (2023) 100888.

[32] J. Su, C. Jiang, X. Jin, Y. Qiao, T. Xiao, H. Ma, R. Wei, Z. Jing, J. Xu, J. Lin, Large language models for forecasting and anomaly detection: A systematic literature review, arXiv preprint arXiv:2402.10350 (2024).

[33] S. Lim, R. Schmälzle, Artificial intelligence for health message generation: an empirical study using a large language model (llm) and prompt engineering, Frontiers in Communication 8 (2023) 1129082.

[34] X. Lin, W. Wang, Y. Li, S. Yang, F. Feng, Y. Wei, T.-S. Chua, Data-efficient fine-tuning for llm-based recommendation, in: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval, 2024, pp. 365–374.

[35] Y. Xia, J. Kim, Y. Chen, H. Ye, S. Kundu, C. C. Hao, N. Talati, Understanding the performance and estimating the cost of llm fine-tuning, in: 2024 IEEE International Symposium on Workload Characterization (IISWC), IEEE, 2024, pp. 210–223.

[36] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, et al., Time-llm: Time series forecasting by reprogramming large language models, arXiv preprint arXiv:2310.01728 (2023).

[37] C. Chang, W.-Y. Wang, W.-C. Peng, T.-F. Chen, Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters, arXiv preprint arXiv:2308.08469 (2023).

[38] H. Xue, F. D. Salim, Utilizing language models for energy load forecasting, in: Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2023, pp. 224–227.

[39] F. J. Lara-Abelenda, D. Chushig-Muzo, P. Peiro-Corbacho, A. M. Wägner, C. Granja, C. Soguero-Ruiz, Personalized glucose forecasting for people with type 1 diabetes using large language models, Computer Methods and Programs in Biomedicine 265 (2025) 108737.

[40] T. Wu, Q. Ling, Stellm: Spatio-temporal enhanced pre-trained large language model for wind speed forecasting, Applied Energy 375 (2024) 124034.

[41] R. Chen, H. Jiang, T. Guo, C. Fan, Can large language models forecast carbon price movements? evidence from chinese carbon markets, Research in International Business and Finance (2025) 102951.

[42] T. Guo, E. Hauptmann, Fine-tuning large language models for stock return prediction using newsflow, arXiv preprint arXiv:2407.18103 (2024).

[43] Z. Lai, T. Wu, X. Fei, Q. Ling, Bert4st:: Fine-tuning pre-trained large language model for wind power forecasting, Energy Conversion and Management 307 (2024) 118331.

[44] Y. Wang, H. A. Karimi, Exploring large language models for climate forecasting, arXiv preprint arXiv:2411.13724 (2024).

[45] G. Liu, Y. Bai, K. Wen, X. Wang, Y. Liu, G. Liang, J. Zhao, Z. Y. Dong, Lfllm: A large language model for load forecasting, Authorea Preprints (2024).

[46] Z. Duan, C. Bian, S. Yang, C. Li, Prompting large language model for multi-location multi-step zero-shot wind power forecasting, Expert Systems with Applications (2025) 127436.

[47] W. Wang, Y. Luo, M. Ma, J. Wang, C. Sui, A novel forecasting framework leveraging large language model and machine learning for methanol price, Energy 320 (2025) 135123.

[48] X. Liu, W. Wang, Deep time series forecasting models: A comprehensive survey, Mathematics 12 (10) (2024) 1504.

[49] N. Gruver, M. Finzi, S. Qiu, A. G. Wilson, Large language models are zero-shot time series forecasters, Advances in Neural Information Processing Systems 36 (2023) 19622–19635.

[50] M. Tan, M. Merrill, V. Gupta, T. Althoff, T. Hartvigsen, Are language models actually useful for time series forecasting?, Advances in Neural Information Processing Systems 37 (2024) 60162–60191.

[51] H. Xue, F. D. Salim, Promptcast: A new prompt-based learning paradigm for time series forecasting, IEEE Transactions on Knowledge and Data Engineering 36 (11) (2023) 6851–6864.

[52] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, Y. Lu, Temporal data meets llm–explainable financial time series forecasting, arXiv preprint arXiv:2306.11025 (2023).

[53] Y. Hu, Q. Li, D. Zhang, J. Yan, Y. Chen, Context-alignment: Activating and enhancing llm capabilities in time series, arXiv preprint arXiv:2501.03747 (2025).

[54] S. Hochreiter, Long short-term memory, Neural Computation MIT-Press (1997).

[55] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 11106–11115.

[56] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, M. Long, itransformer: Inverted transformers are effective for time series forecasting, arXiv preprint arXiv:2310.06625 (2023).

[57] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, M. Long, Timesnet: Temporal 2d-variation modeling for general time series analysis, arXiv preprint arXiv:2210.02186 (2022).

[58] Y. Nie, N. H. Nguyen, P. Sinthong, J. Kalagnanam, A time series is worth 64 words: Long-term forecasting with transformers, arXiv preprint arXiv:2211.14730 (2022).

[59] R. Koenker, K. F. Hallock, Quantile regression, Journal of economic perspectives 15 (4) (2001) 143–156.

[60] A. C. Gonçalves, X. Costoya, R. Nieto, M. L. Liberato, Extreme weather events on energy systems: a comprehensive review on impacts, mitigation, and adaptation measures, Sustainable Energy Research 11 (1) (2024) 4.

[61] S. Wang, M. W. Zafar, D. G. Vasbieva, S. Yurtkuran, Economic growth, nuclear energy, renewable energy, and environmental quality: Investigating the environmental kuznets curve and load capacity curve hypothesis, Gondwana Research 129 (2024) 490–504.

[62] M. Bilgili, S. Tumse, S. Nar, Comprehensive overview on the present state and evolution of global warming, climate change, greenhouse gasses and renewable energy, Arabian Journal for Science and Engineering 49 (11) (2024) 14503–14531.

[63] M. Shanker, M. Y. Hu, M. S. Hung, Effect of data standardization on neural network training, Omega 24 (4) (1996) 385–397.

[64] N. Fei, Y. Gao, Z. Lu, T. Xiang, Z-score normalization, hubness, and few-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 142–151.

[65] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[66] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).

[67] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).