

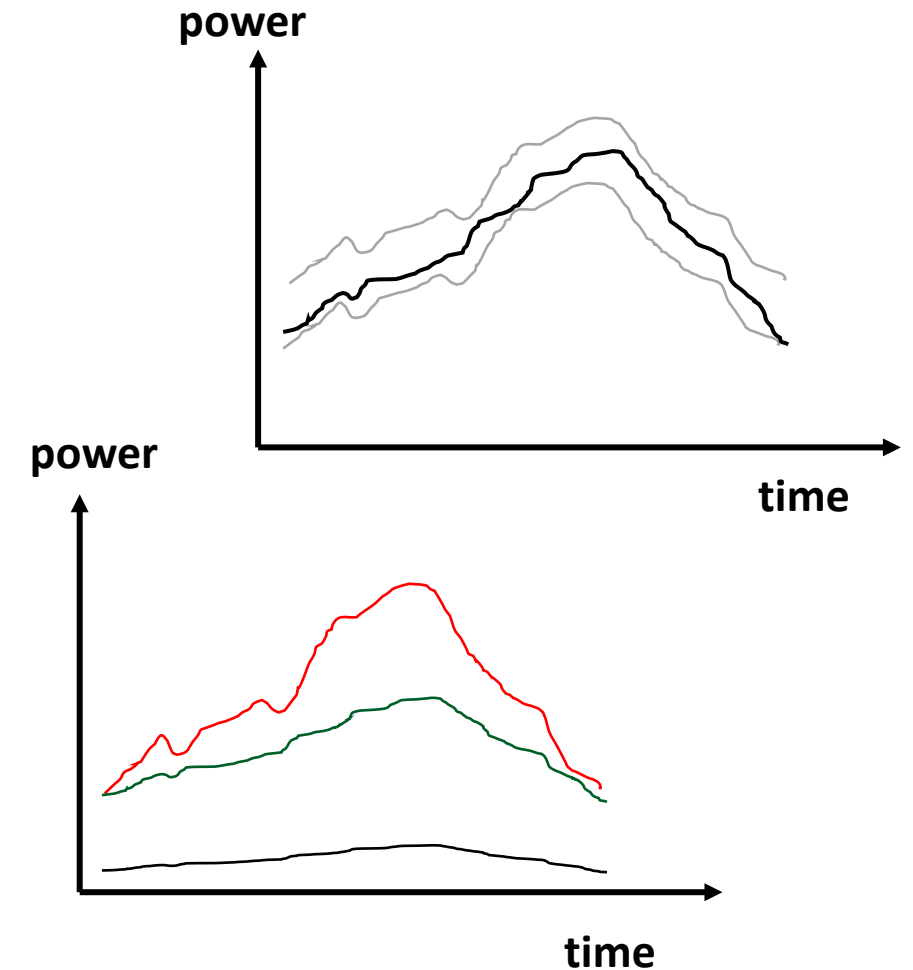


Power System Synthetic Data Generation and Its Application

Ning Lu
North Carolina State University

Outline

- Why do we need synthetic data?
- Approach
 - Directly use real data
 - Statistic-based load profile generation
 - Ours:
 - Generative machine-learning based (GAN-based)
 - Bert-based (Transformer-based)
 - Large-language model (LLM)
- Considerations
 - **Realisticness**
 - Customizable data **resolution**
 - Preserve temporal, spatial, group **correlations**
- Conclusions





1. Using Real-data

Transforming low resolution data to high resolution

Data Resolution

1-Minute Sub-metered data

- End use consumptions of appliances
- Not usually available
- Enabling technologies: IoT sensors

15-minute Smart Meter Data

- Average kWh, kVar, Voltage
- Sensitive information

Hourly

- Temperature, irradiance
- Average kWh

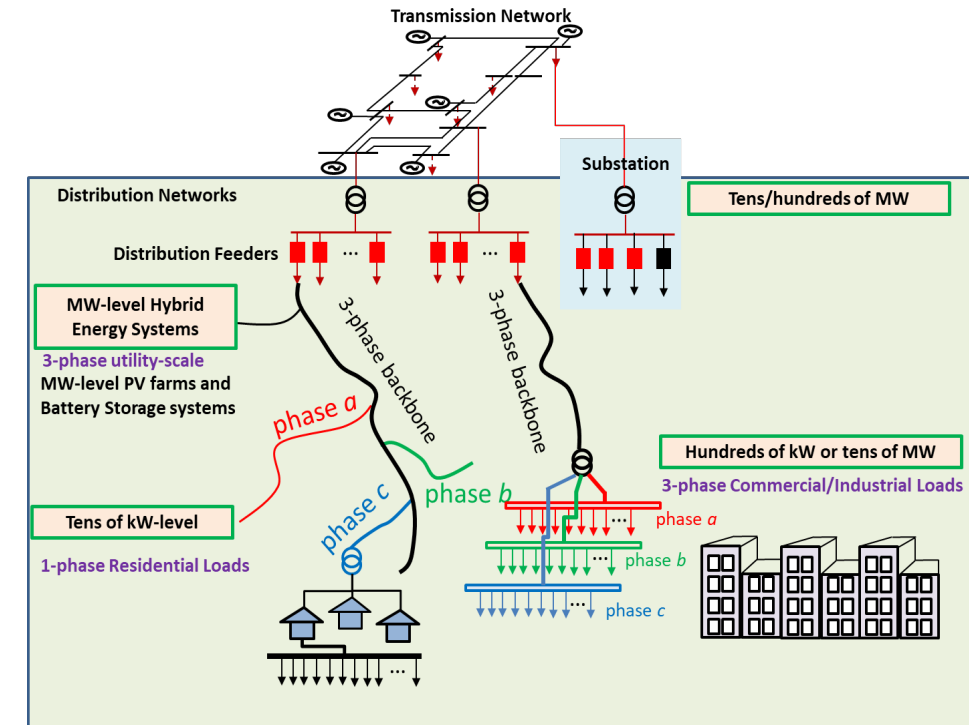
Daily

- Peak hour
- DR events

Monthly

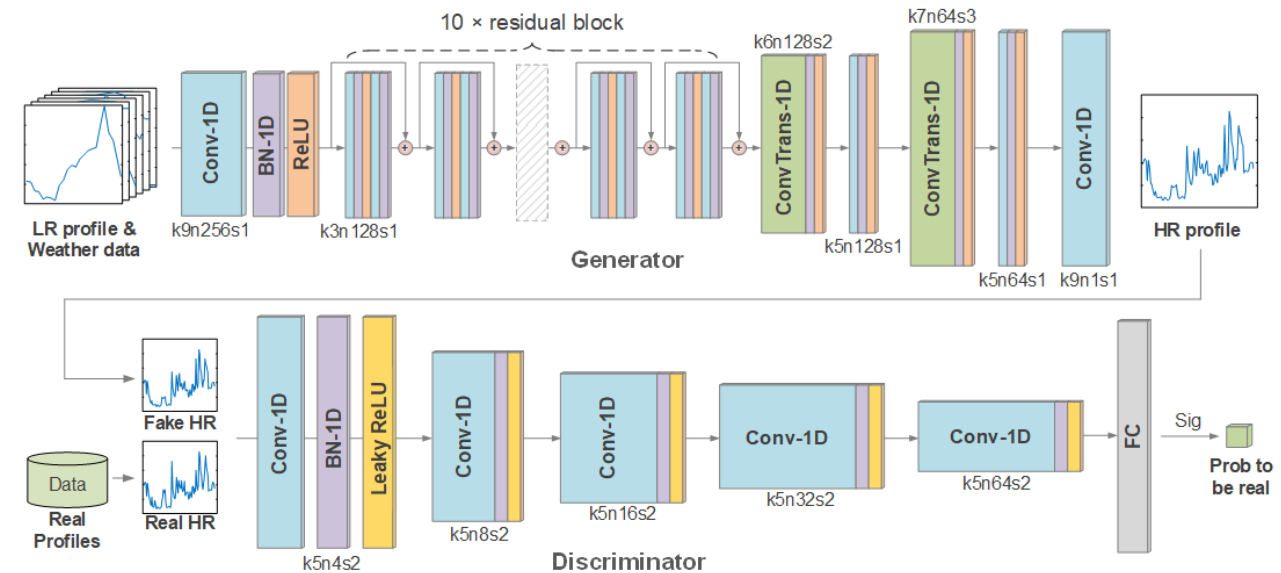
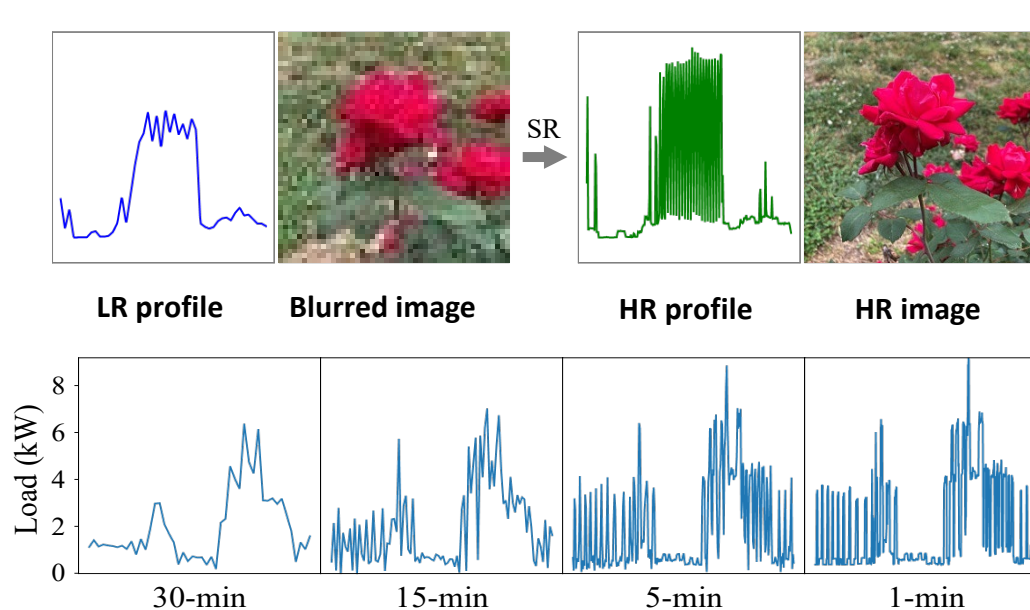
- Utility billing information
- Peak day peak hour

Super resolution



Time

- Generate high-resolution PV and load profiles
- Inputs: **15-min** or **30-min** low resolution (LR)
- **Restore the high-frequency components** from the LR measurements using deep learning methods



Lidong Song, Yiyang Li, and Ning Lu. "ProfileSR-GAN: A GAN based Super-Resolution Method for Generating High-Resolution Load Profiles." *IEEE Transactions on Smart Grid* 13, no. 4 (2022): 3278-3289. [Youtube video](#).



2. Generated from Scratch

Group load profile generation using GAN

Yi Hu, Yiyang Li, Lidong Song, Han Pyo Lee, PJ Rehm, Matthew Makdad, Edmond Miller, and Ning Lu, "MultiLoad-GAN: A GAN-Based Synthetic Load Group Generation Method Considering Spatial-Temporal Correlations," in IEEE Transactions on Smart Grid, vol. 15, no. 2, pp. 2309-2320, Mar. 2024, doi: 10.1109/TSG.2023.3302192.

Load Profile Generation Methods

TABLE I
COMPARISON OF OUR MULTILOAD-GAN MODEL WITH STATE-OF-THE-ART METHODS

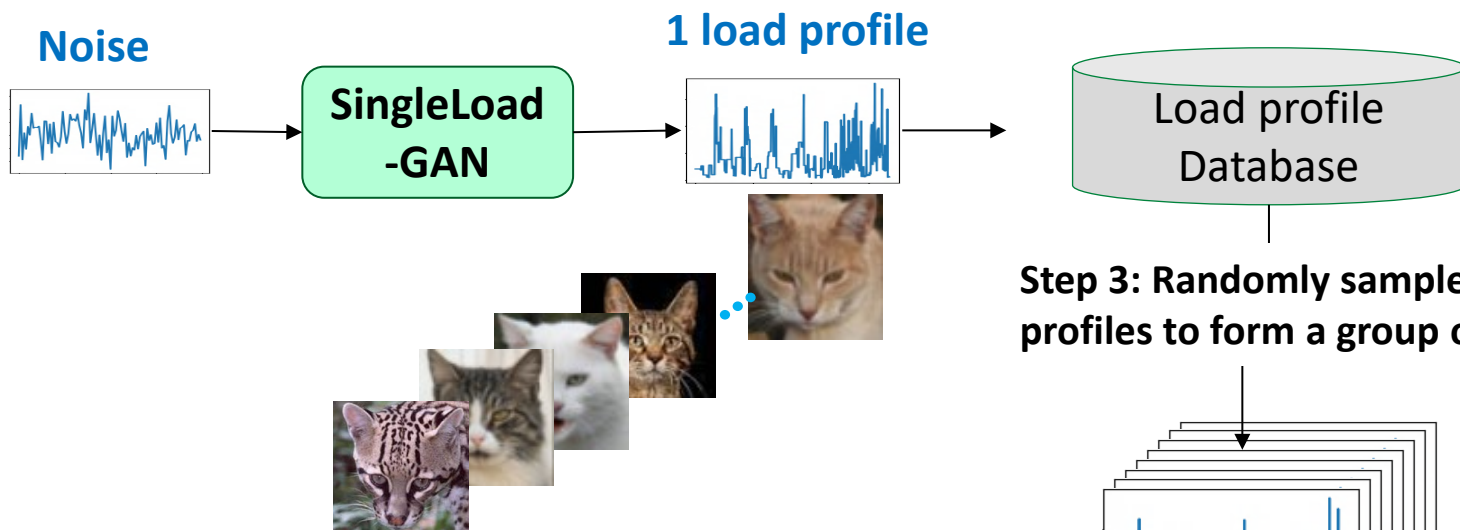
	Description	Advantages	Disadvantages	Model output	
Model-based methods [1][2]	Use physical models, such as building thermodynamics and customer behavioral models, to simulate electricity consumption profiles.	Explainable as the models reflect the laws of physics when describing the behavior behind field measurements	Require detailed physics-based models with many inputs and require parameter tuning.	Single load profile (When generating a load profile, the methods do not consider the spatial-temporal correlations among a group of generated load profiles)	
Data-driven methods	Clustering based [3][4]	Cluster existing load profiles into different categories so that by combining the load profiles across different categories, SLPs are generated.	Easy to implement and can represent some realistic load profile characteristics.		Lack of diversity when using combinations of a limited number of existing profiles.
	Forecasting based [5]-[8]	Generate SLPs based on publicly available load or weather data.	Easy to implement and flexible to generate load profiles with different lengths and granularities.		Depend heavily on historical data. The generated load profiles have similar patterns with historical data, therefore, lack of diversity.
	SingleLoad-GAN-based [10]-[12] (the benchmark method)	GAN-based generative methods to generate the SLP for one customer at a time.	Learn from the real data distribution to generate diversified load profiles with high-frequency details.		Hard to train.
	MultiLoad-GAN (the proposed method)	GAN-based generative methods to generate a group of spatial-temporal correlated load profiles simultaneously. Such load profiles can be loads served by the same transformer or feeder.	Learn from the distribution of real data to generate diversified load profiles with high-frequency details. Preserve the spatial-temporal correlations between loads.	Hard to train.	Multiple spatial-temporal correlated load profiles

Yi Hu, Yiyang Li, Lidong Song, Han Pyo Lee, PJ Rehm, Matthew Makdad, Edmond Miller, and Ning Lu, "MultiLoad-GAN: A GAN-Based Synthetic Load Group Generation Method Considering Spatial-Temporal Correlations," in IEEE Transactions on Smart Grid, vol. 15, no. 2, pp. 2309-2320, Mar. 2024, doi: 10.1109/TSG.2023.3302192.

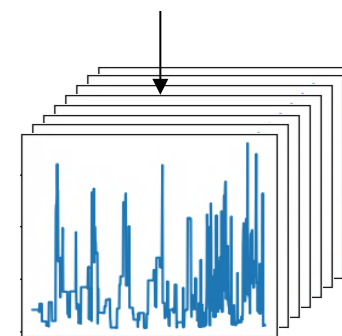
Single-Load GAN Approach

Step 1: Generate one load profile at a time

Step 2: Run step 1 iteratively to obtain a group of load profiles



Step 3: Randomly sample N load profiles to form a group of loads



8 load profiles

Drawbacks:

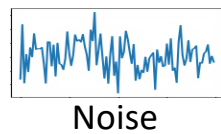
Cannot account for group-level characteristics



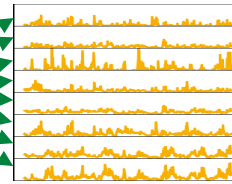
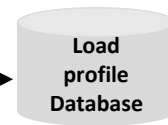
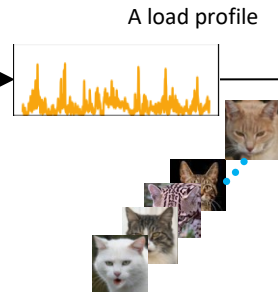
Group-Load GAN Approach → Correlation

Single-Load GAN

Step 1: Generate one load profile at a time



Step 2: Run step 1 for many times to obtain a database of load profiles



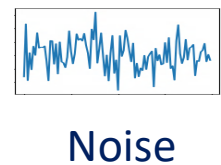
Step 3: Randomly sample N load profiles



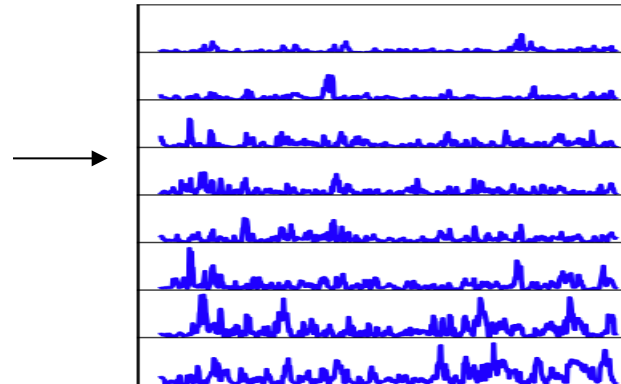
A group of load profiles supplied by the same distribution transformer

Capture group correlation

Group-Load GAN



Generate N load profiles





3. Generated from A set of Given Data

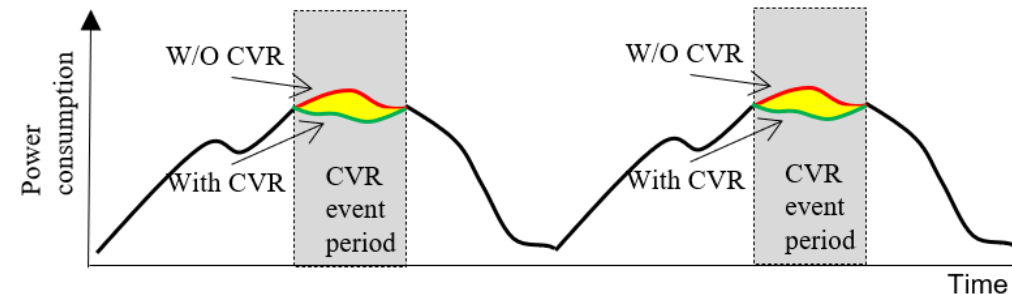
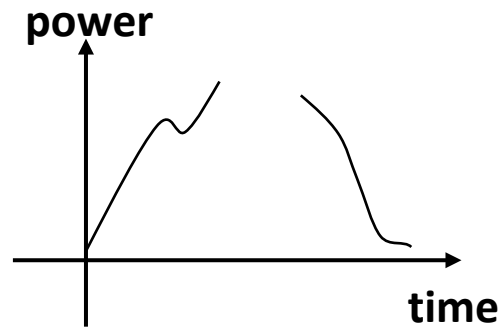
A BERT-based Framework

Yi Hu, K. Ye, H. Kim and N. Lu, "BERT-PIN: A BERT-Based Framework for Recovering Missing Data Segments in Time-Series Load Profiles," in *IEEE Transactions on Industrial Informatics*, doi: 10.1109/TII.2024.3417272.

Condition-based Data Generation

Power System Data Restoration task

Example: Restore baseline load consumption data during a conservation voltage reduction (CVR) event.



Natural Language Processing task

Example: fill-in-the-blank

I have **lived** in _____ for five years,
so I can **speak fluent French**.

I have lived in France for five years, so I can speak fluent French. 90%

I have lived in Quebec for five years, so I can speak fluent French. 8%

I have lived in England for five years, so I can speak fluent French. 1%

⋮

I have lived in apple for five years, so I can speak fluent French. 0.01%

- BERT-PIN: A BERT-based Framework

Prompt



On a Sunday morning, walking on a trail in the woods, describe the scene.



On a Sunday morning, the forest trail is enveloped in a serene tranquility. Sunlight filters through the dense canopy above, casting dappled patterns on the leaf-strewn path. The air is cool and fresh, filled with the earthy scent of moss and damp soil. Birds chirp melodiously from hidden perches, their songs mingling with the subtle rustle of leaves as a gentle breeze whispers through the trees. Along the sides of the trail, wildflowers display bursts of color—vivid blues, yellows, and purples—that contrast with the deep greens of the forest. Every step on the soft, yielding path brings a sense of peace and a disconnect from the hustle of everyday life, making the forest a perfect retreat for a reflective, soothing walk.

Generated content

Generating load profiles or filling in missing data based on known information is similar to using a large language model (LLM) to create an article. In this analogy:

- **Temperature** corresponds to the background or setting of the story.
- **Previous load profiles** are equivalent to previous writings by the same author.
- **Similar load profiles** represent writings from other authors who share a similar storytelling style.

Condition-based Data Generation

- BERT-PIN: A BERT-based Framework

Prompt

On a Sunday morning in a cold winter day, walking on a trail in the woods, describe the scene.



On a cold Sunday morning, the forest trail is transformed into a quiet, wintry wonderland. Snow blankets the ground, muffling footsteps, and clings to the branches of towering trees, creating a soft canopy of white. The air is crisp and invigorating, carrying the clean scent of pine and the subtle hint of a nearby frozen stream. Sunlight, pale and weak, filters through the branches, reflecting off the snow with a gentle glow. The usual chatter of wildlife is subdued, replaced by the occasional distant call of a winter bird or the soft thump of snow falling from a branch. Each breath forms a cloud of vapor in the chilly air, and the path, lined with delicate ice crystals that sparkle in the light, invites a peaceful solitude that refreshes the spirit amidst the stark beauty of the winter woods.

Generated content

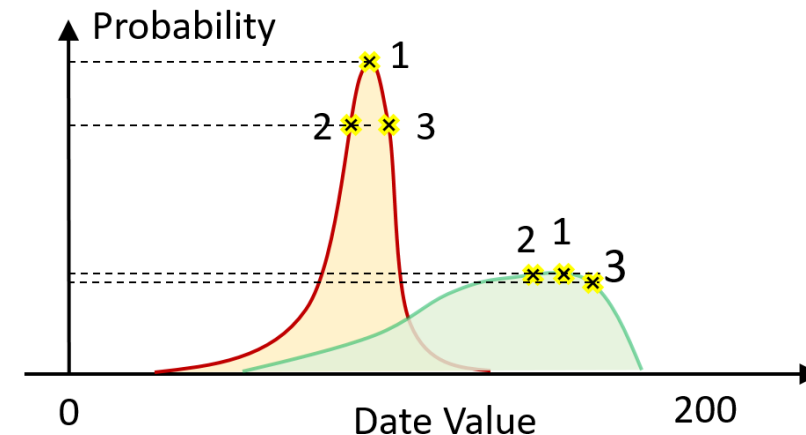
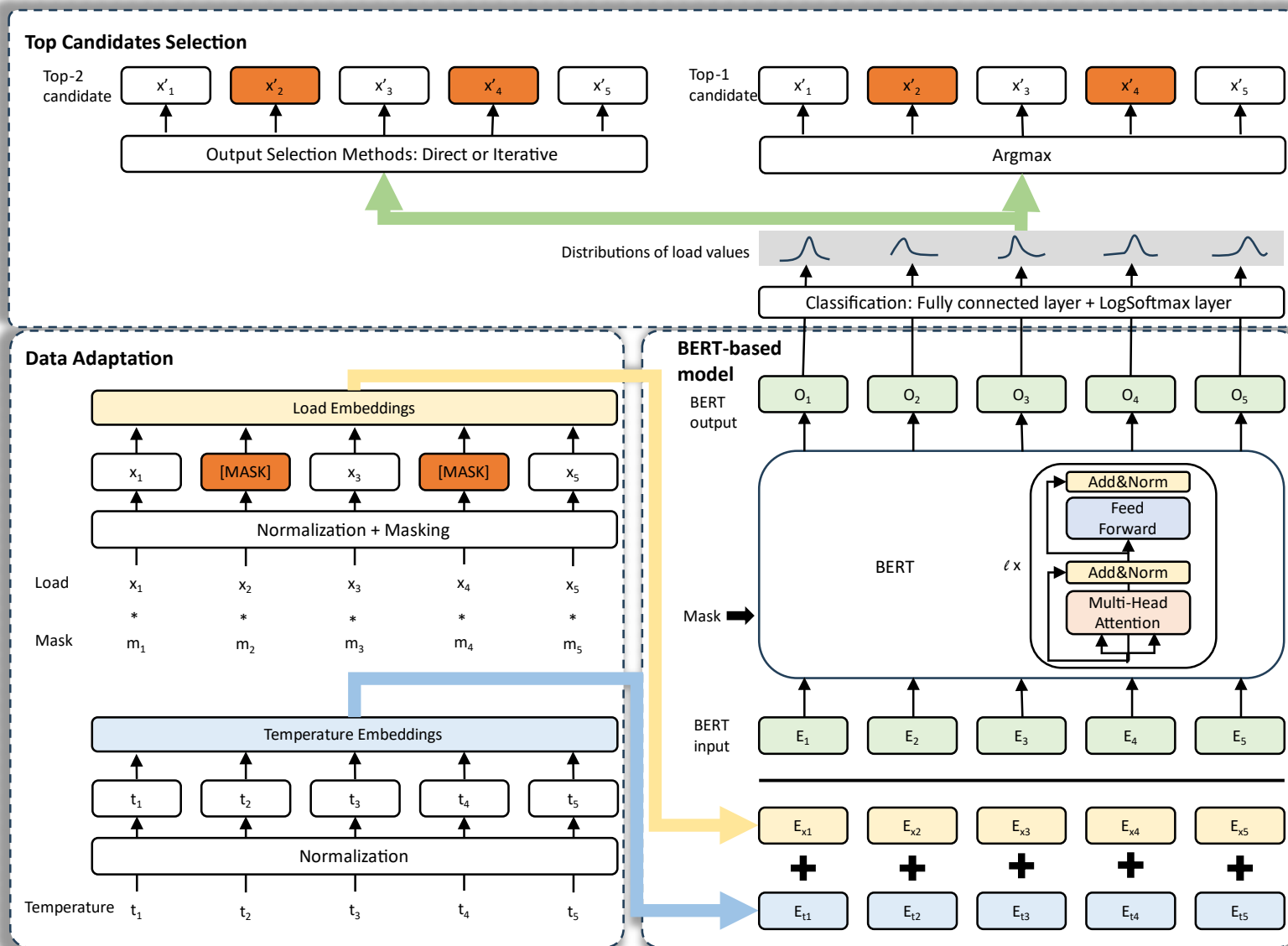
Generating load profiles or filling in missing data based on known information is similar to using a large language model (LLM) to create an article. In this analogy:

- **Temperature** corresponds to the background or setting of the story.
- **Previous load profiles** are equivalent to previous writings by the same author.
- **Similar load profiles** represent writings from other authors who share a similar storytelling style.

Existing methods

		Description	Advantages	Disadvantages
Model-based methods [1]-[5]		Use physical system models to simulate responses to external disturbances for restoring missing data segments.	Explainable as the models reflect the laws of physics.	Require accurate distribution system model.
Data-driven methods	Similarity-based [6]-[9]	Group load profiles by day type, weather conditions, and shape characteristics of load profiles. The missing data segments are restored by referencing to the data on the load profiles having the best similarity match.	Easy to implement and explainable.	Accuracy of the method dependent on selections of similarity metrics and weights.
	Regression-based (the benchmark method)	Use models including linear regression [10], Long Short Term Memory (LSTM) [11][12], Autoencoders [13][14], Gaussian Regression [15], Support Vector Regression (SVR) [16][17], etc. Or combine multiple regression models [18]-[21].	Provide transparent insights into the relationships between input and output. More efficient for small to medium-sized datasets.	Limited complexity and limited contextual understanding. Require manual feature engineering.
	GAN-based [22]-[27]	Use Generative Adversarial Nets solve the missing data restoration problems in power system.	Discover underlying patterns in the data without explicit supervision.	Training instability, mode collapse, and hard to evaluate.
	Load-PIN [28] (the benchmark GAN-based method)	Combine Generative Adversarial Nets with Convolutional layers and multi-head self-attention blocks to improve accuracy.	More accurate than model-, similarity-, regression-, and other GAN based models.	Computationally expensive and require large amount of data. Produce only 1 restoration candidate.
	BERT-PIN (the proposed method)	Bidirectional Encoder can capture long-range dependencies though self-attention mechanisms.	The most accurate method and can produce multiple restoration candidates .	Computationally expensive and require large amount of data.

BERT-PIN Framework

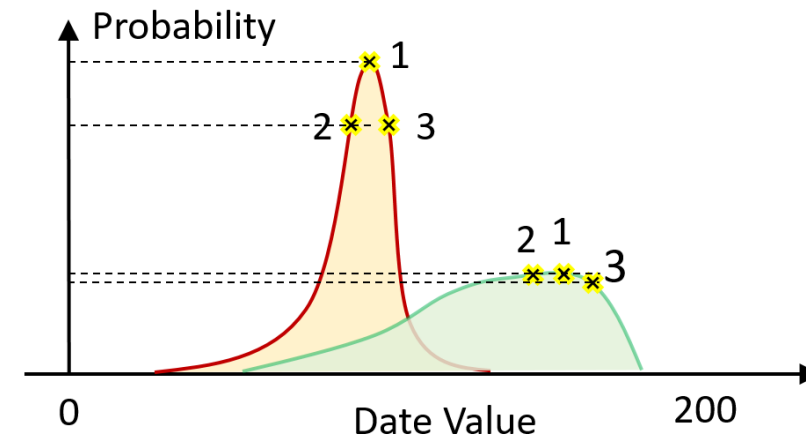
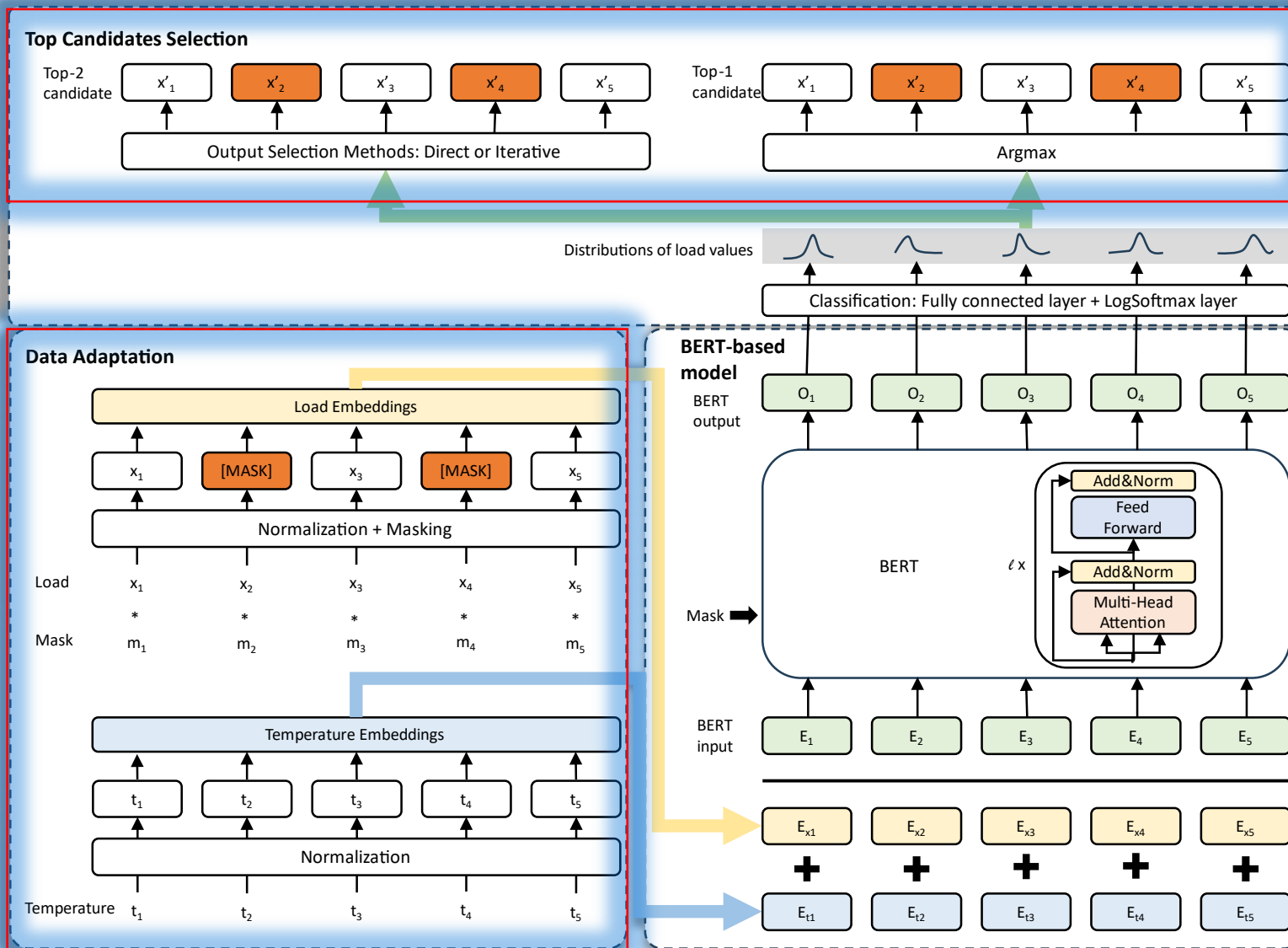


$$X = \frac{1}{P_{MAX}} \sum_{i=1}^{N_{Agg}} P_i (t_{start}: t_{end})$$

It is a classification problem.

$$Loss = (1 - \lambda) * CrossEntropy(X, \hat{X}^1) + \lambda * CrossEntropy(X_m, \hat{X}_m^1)$$

BERT-PIN Framework



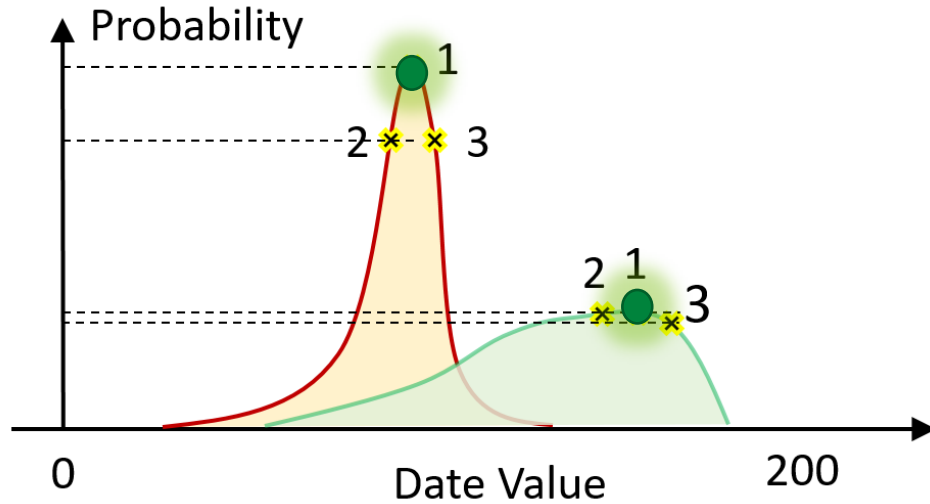
$$X = \frac{1}{P_{MAX}} \sum_{i=1}^{N_{Agg}} P_i (t_{start}: t_{end})$$

It is a classification problem.

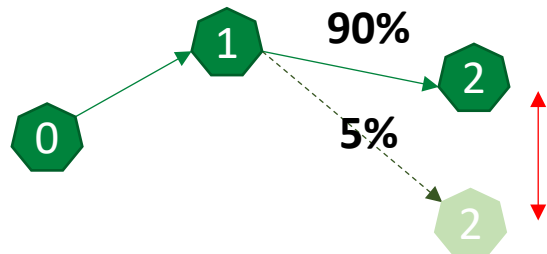
$$Loss = (1 - \lambda) * CrossEntropy(X, \hat{X}^1) + \lambda * CrossEntropy(X_m, \hat{X}_m^1)$$

Top Candidate Selection

Method 1: Directly output the most likely results

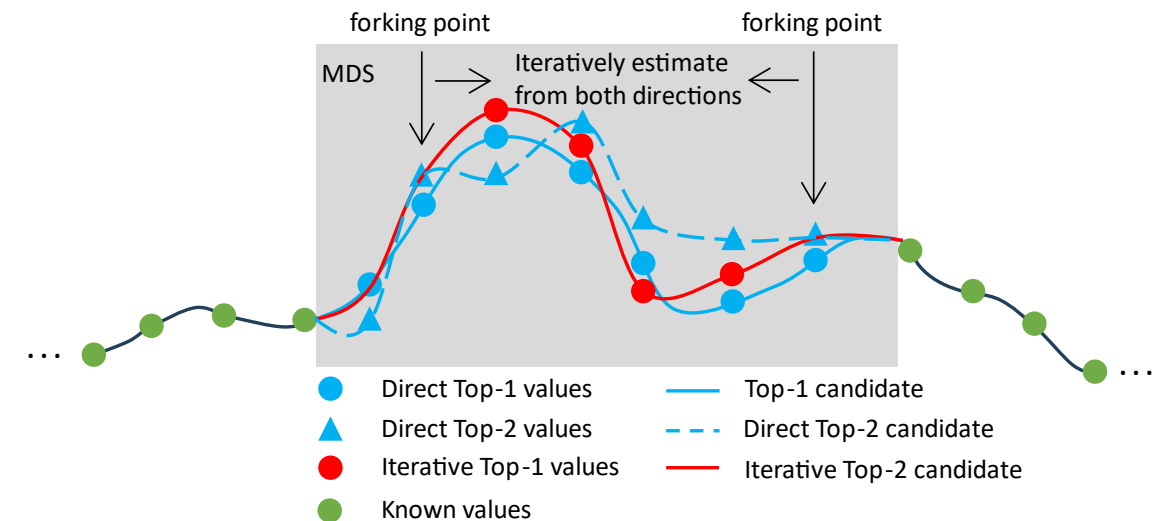


Not a Forking Point



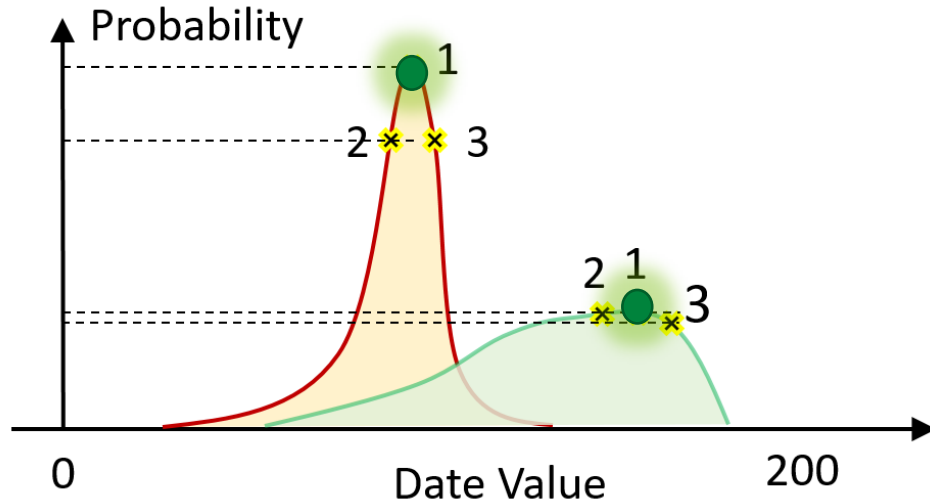
Probability difference
 $e=85% > 20%$

Method 2: Iteratively, output top-2 results

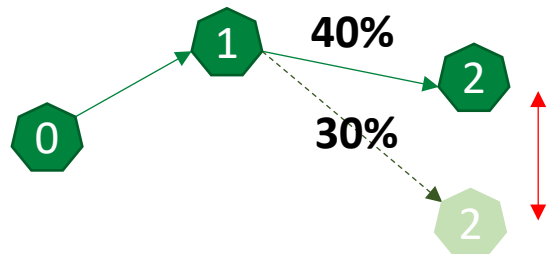


Top Candidate Selection

Method 1: Directly output the most likely results

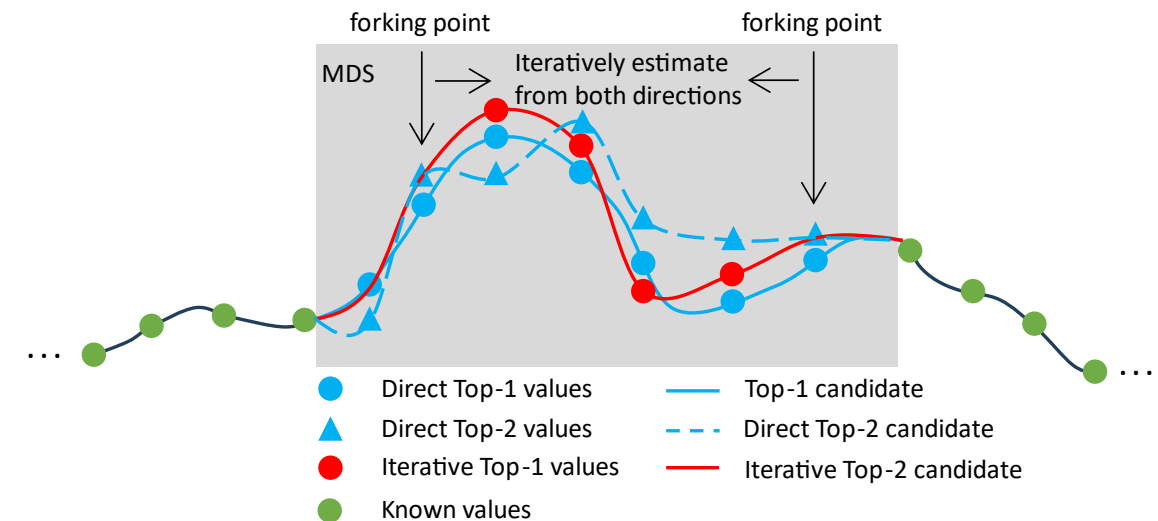


A Forking Point

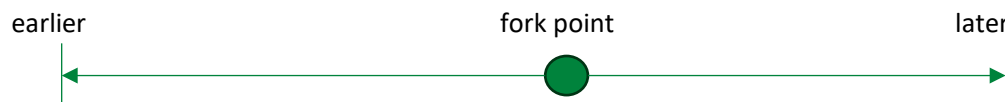
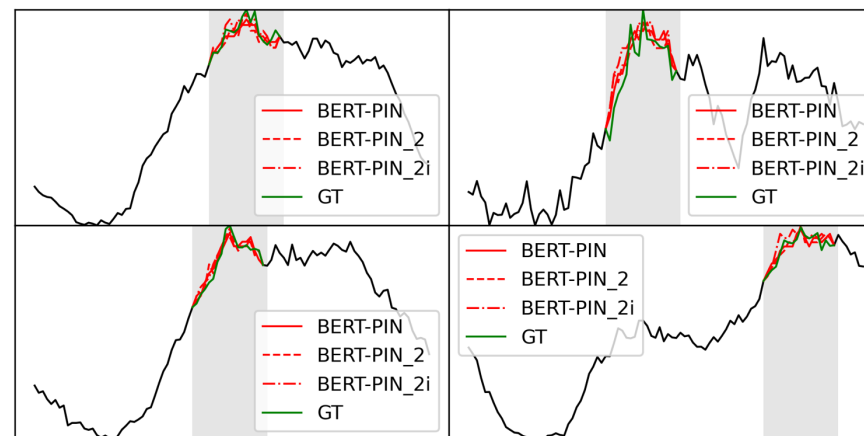
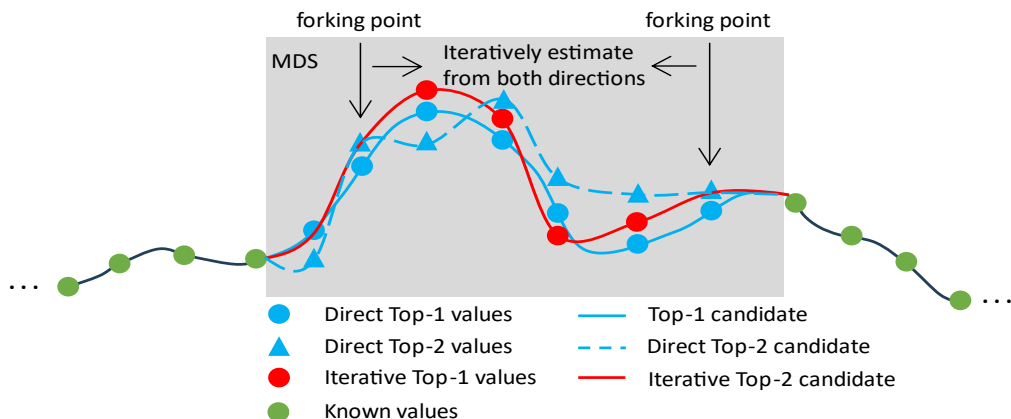


Probability difference
 $e=10% < 20%$

Method 2: Iteratively, output top-2 results



Top-2 Candidate Results



1. Define the "right-side" (or "left-side") fork point as the first point,
2. Count from the rightmost (or leftmost) side of the missing data point,
3. Probability difference between the top-1 and top-2 candidates is less than e .

	Top-1	Top2-direct	Top-2 iterative						Combine
			e=0.8	e=0.5	e=0.3	e=0.1	e=0.05	e=0.02	
MPE	1.523	1.744	2.556	2.433	2.407	2.138	1.87	1.761	1.211
RMSE	0.7404	0.9144	1.317	1.173	1.211	1.071	0.896	0.899	0.577
PKE	0.5130	0.5917	1.044	0.939	0.927	0.871	0.665	0.663	0.426
VLE	0.5870	0.9260	0.827	0.617	0.817	0.573	0.543	0.669	0.407
EGYE	0.8410	0.9582	1.618	1.441	1.412	1.301	1.043	0.986	0.633
FCE	1.509	1.942	2.447	2.18	2.273	2.007	1.727	1.78	1.209
PoCP	-	45.12%	23.88%	24.40%	22.94%	17.29%	12.25%	6.29%	-

- BERT-PIN (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers-based **P**rofile **I**npainting **N**etwork) aims to **restore multiple missing data segments** within load profiles.
- Contributions
 - Offers multiple restoration options within a given confidence level
 - Capable of restoring extensive missing data segments across long time periods
- Results:
 - Tested on different datasets for different aggregation levels and load sizes.
 - BERT-PIN shows **5%-30% greater accuracy** than other methods for missing data recovery tasks.
 - Combining top-1 and top-2 predictions enables **more comprehensive missing data predictions**.
 - BERT-PIN excels in specific tasks like weekly **data gap filling** and **demand response baseline estimation**.

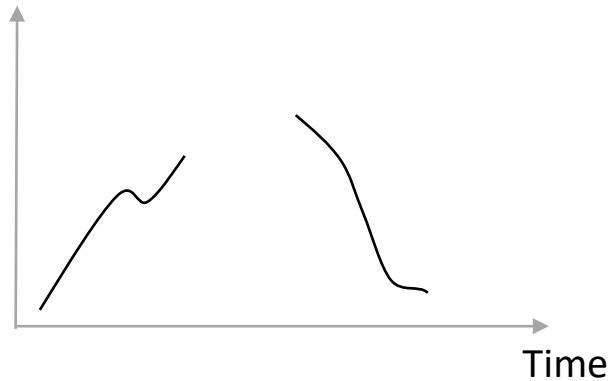


4. Generated from Fine-tuned Large Language Model (LLM)

Fine-tune ChatGPT for Generating Load Profiles

1.Y. Hu, H. Kim, K. Ye, and N. Lu, "Applying Fine-Tuned LLMs for Reducing Data Needs in Load Profile Analysis", arXiv preprint arXiv: 2406.02479, 2024. Available: <https://arxiv.org/abs/2406.02479>

Game With ChatGPT



I have lived in _____ for five years, so I can speak fluent French.

Similar to natural language:

- Time sequence data
- Like a missing word filling task
- Use Large Language Model

Yi "I have lived in _____ for five years, so I can speak fluent French." Please fill in the blank.

ChatGPT "I have lived in France for five years, so I can speak fluent French."

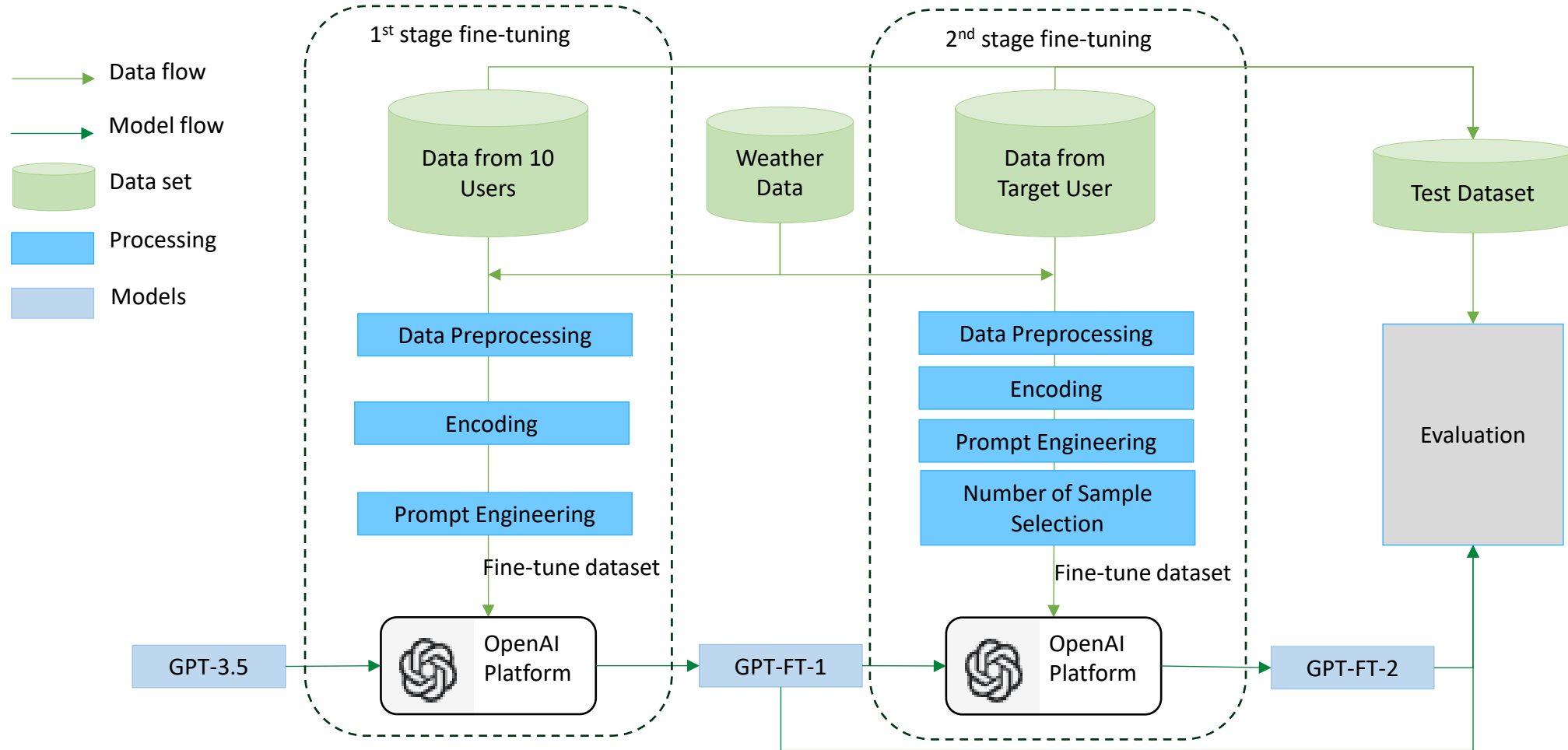
Yi Is there another possible answer?

ChatGPT Certainly! Another possible answer could be:

"I have lived in Quebec for five years, so I can speak fluent French."

In this case, Quebec is a province in Canada where French is one of the official languages, and living there for five years would likely provide ample opportunity to learn and become fluent in French.

Two-stage Fine-tuning Strategy

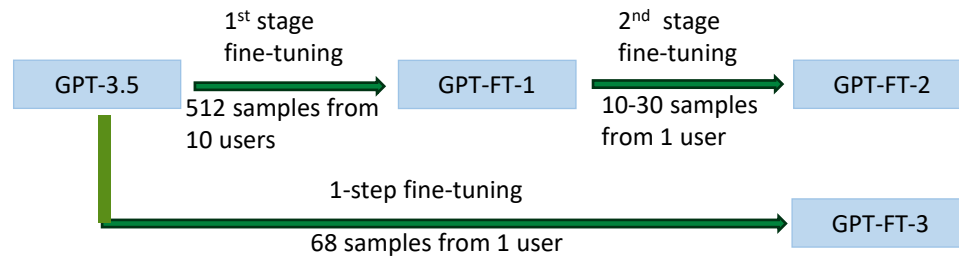
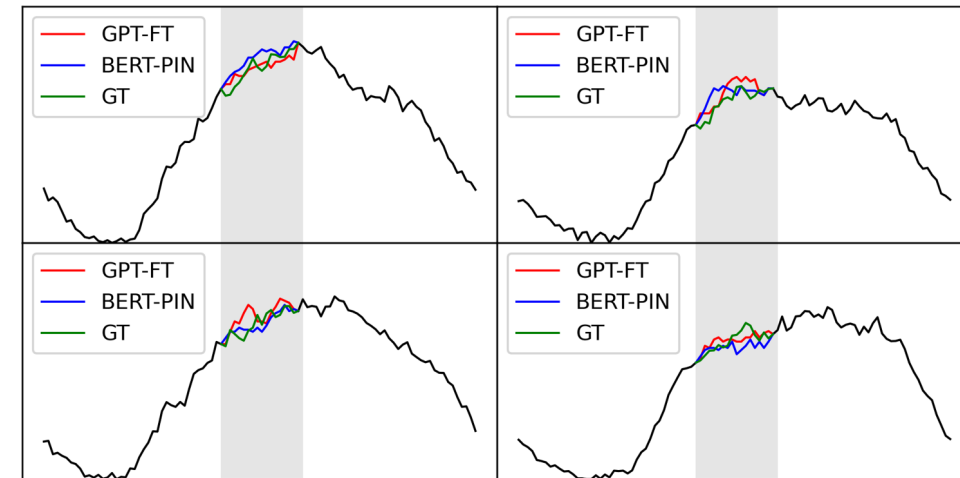
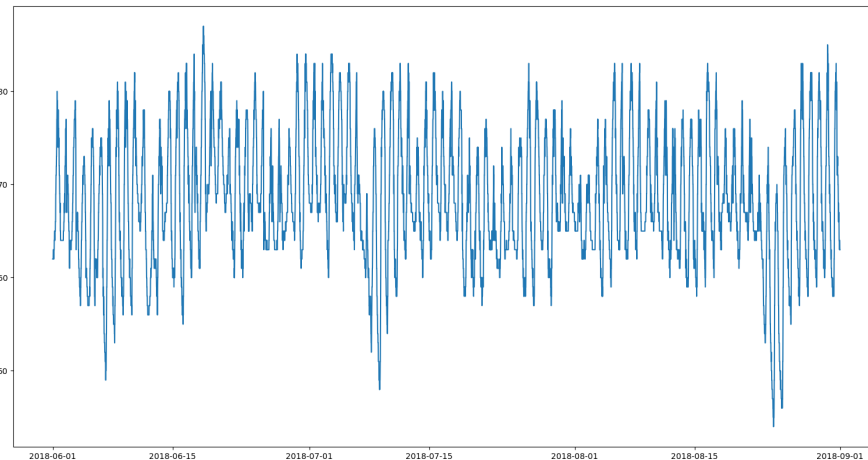


General domain knowledge → Specific **Task** knowledge

General user knowledge → Specific **User** knowledge

Data Availability

86 days in summer, 68 days for training, and 18 days for testing.
20 users



Yi Hu, H. Kim, K. Ye, and N. Lu, "Applying Fine-Tuned LLMs for Reducing Data Needs in Load Profile Analysis", arXiv preprint arXiv: 2406.02479, 2024. Available: <https://arxiv.org/abs/2406.02479>

1st Stage Result

Train and test on user #0-9

	# of samples	Advanced prompt	Separate Load & temperature	Discard Encoding	Remove abnormal days	Errors (%)		
						MPE	RMSE	EGYE
Scenario 1	128	N	N	N	N	5.609	4.513	4.303
Scenario 2	256	N	N	N	N	4.609	3.816	3.652
Scenario 3	512	N	N	N	N	4.59	3.795	3.547
Scenario 4	512	Y	N	N	N	3.806	3.027	2.987
Scenario 5	512	Y	Y	N	N	3.266	2.656	2.469
Scenario 6	512	Y	Y	Y	N	2.48	2.029	1.639
Scenario 7	512	Y	Y	Y	Y	2.221	1.977	1.443
BERT-PIN	~220K	-	-	-	-	1.612	0.699	0.887

- Use **more samples in the first stage fine-tuning** leads to higher accuracy.
- Detailed information about the **task description, encoding strategy, and output constraints** provided in the advanced prompt improves the model's performance.
- Present load encoding and temperature encoding in a two-turn chat helps.
- ChatGPT can **comprehend direct numbers better** than encoded "words".
- The **removal of abnormal days** leads to a further increase in accuracy.

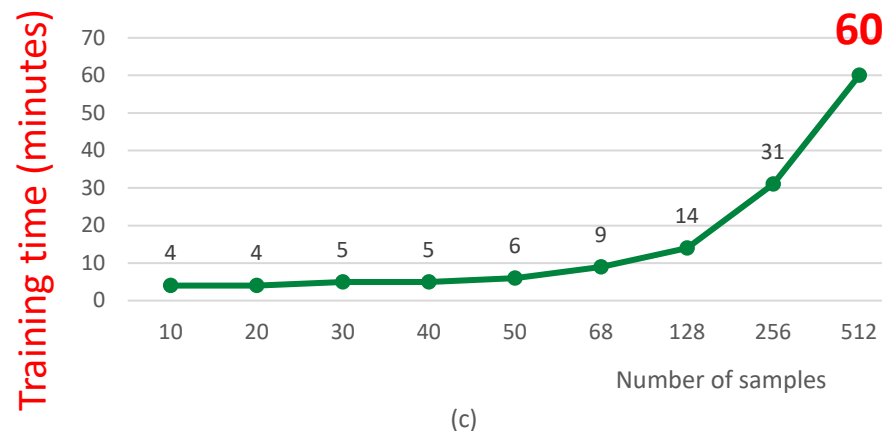
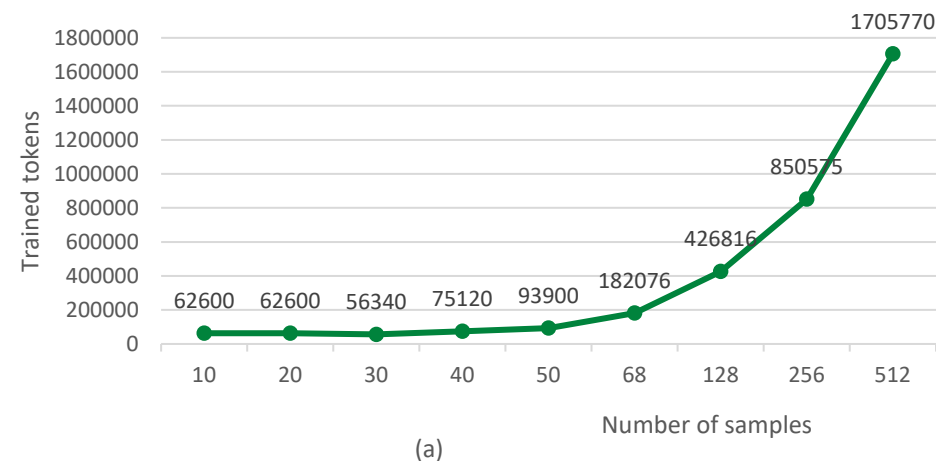
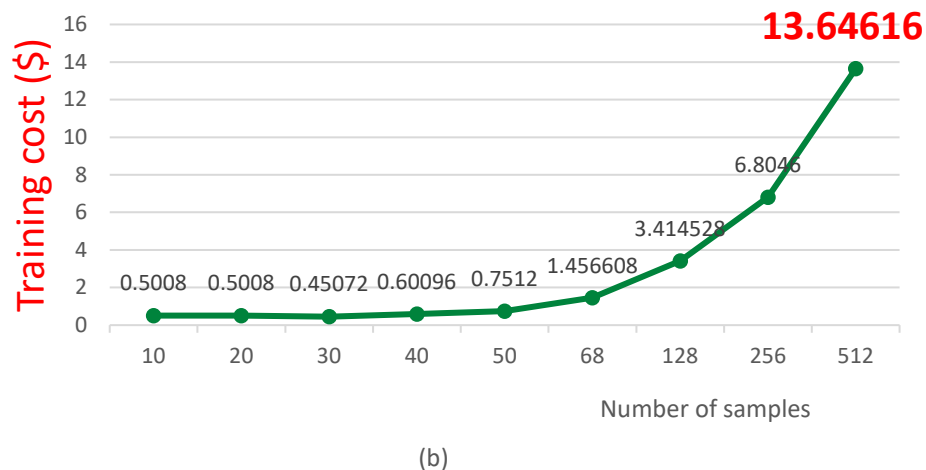
2nd Stage Result

Target Users	Errors (%)	GPT-FT-2 fine-tuned with different numbers of samples					GPT-FT-1 (W/O 2 nd stage)		GPT-FT-3 (W/O 1 st stage)
		10	20	30	40	50	Test on target user	Test on user0-9	
User10	MPE	2.245	2.277	2.428	2.395	2.361	2.497	2.221	2.983
	RMSE	1.988	1.969	2.014	2.094	2.140	2.182	1.977	2.772
	EGYE	1.501	1.513	1.599	1.604	1.637	1.843	1.443	2.461
User11	MPE	2.206	2.168	2.416	2.352	2.549	2.77	2.221	3.216
	RMSE	1.911	1.887	1.957	2.008	2.164	2.438	1.977	2.837
	EGYE	1.499	1.453	1.534	1.649	1.692	2.103	1.443	2.524
User12	MPE	2.276	2.453	2.483	2.533	2.676	2.441	2.221	3.068
	RMSE	1.962	1.997	2.016	2.147	2.163	2.206	1.977	2.649
	EGYE	1.524	1.517	1.567	1.599	1.643	1.784	1.443	2.581
User13	MPE	2.382	2.448	2.565	2.604	2.412	2.503	2.221	2.979
	RMSE	2.009	2.018	2.097	2.105	2.166	2.263	1.977	2.729
	EGYE	1.587	1.589	1.535	1.673	1.749	1.958	1.443	2.441
User14	MPE	2.318	2.248	2.558	2.48	2.336	2.673	2.221	3.110
	RMSE	1.997	1.948	2.021	2.134	2.196	2.273	1.977	2.811
	EGYE	1.472	1.499	1.514	1.731	1.887	2.094	1.443	2.687

- The performance of the second stage fine-tuned model **does not improve with an increase in training data**.
- **Only a small number of samples are required** for effective transfer of knowledge from general users to a specific user.
- The second stage fine-tuned model GPT-FT-2 exhibits improved accuracy for all target users.
- Direct fine-tuning of the original GPT-3.5 model with 68 samples from the target user results in inferior performance.

Cost Analysis

The fine-tuning is executed in **OpenAI server**.
The cost associated with this process varies depending on the number of samples utilized in the fine-tuning process and the base model selection.



Conclusion

- Fine-tuned model restores missing data with accuracy **comparable** to advanced models like BERT-PIN.
- Key insights include the effectiveness of **using a small number of fine-tuning samples**, which underscores the efficiency of few-shot learning.
- Additionally, advanced **prompt engineering** and **separate encoding of load and temperature data** significantly enhance model performance.
- The fine-tuning strategy proved to be **cost-effective** and **time-efficient**, presenting a viable alternative to training models from scratch, especially in **data-limited and resource-scarce** settings.