

GPT4Battery: An LLM-driven Framework for Adaptive State of Health Estimation of Raw Li-ion Batteries

Yuyuan Feng^{*}, Guosheng Hu, Zhihong Zhang[†]

Abstract

State of health (SOH) is a crucial indicator for assessing the degradation level of batteries that cannot be measured directly but requires estimation. Accurate SOH estimation enhances detection, control, and feedback for Li-ion batteries, allowing for safe and efficient energy management and guiding the development of new-generation batteries. Despite the significant progress in data-driven SOH estimation, the time and resource-consuming degradation experiments for generating lifelong training data pose a challenge in establishing one large model capable of handling diverse types of Li-ion batteries, e.g., cross-chemistry, cross-manufacturer, and cross-capacity. Hence, this paper utilizes the strong generalization capability of large language model (LLM) to propose a novel framework for adaptable SOH estimation across diverse batteries. To match the real scenario where unlabeled data sequentially arrives in use with distribution shifts, the proposed model is modified by a test-time training technique to ensure estimation accuracy even at the battery's end of life. The validation results demonstrate that the proposed framework achieves state-of-the-art accuracy on four widely recognized datasets collected from 62 batteries. Furthermore, we analyze the theoretical challenges of cross-battery estimation and provide a quantitative explanation of the effectiveness of our method.

1 Introduction

Rechargeable Li-ion batteries (LIBs) are crucial in many modern-day applications ranging from portable electronics and medical devices to renewable energy integration in power grids and electric vehicles [Genikomsakis *et al.*, 2021]. The entire LIB chain could reach a value of more than \$400 billion and a market size of 4.7 TWh by 2030 [Jakob Fleischmann, 2023].

State of health (SOH) is a critical state evaluating the degradation level of batteries, which cannot be measured directly but requires estimation [Lu *et al.*, 2023]. Obtaining an accurate SOH is vital to ensure safe and efficient battery management and guide the design and manufacturing of new-generation batteries [Severson *et al.*, 2019]. As a multidisciplinary issue, battery state estimation presents a challenge for researchers in both battery and machine learning fields. Existing data-driven SOH estimation studies require lifelong data with precise SOH labels to establish the mapping relationship for every LIB [Roman *et al.*, 2021]. Obtaining the training data necessitates degradation experiments that often require many months to years. However, with the rapid development of new-generation batteries, waiting this long to develop separate models for each LIB greatly hinders the deployment of battery management systems (BMS) [Ng *et al.*, 2020; Roman *et al.*, 2021].

To lighten the burden of data collection, several transfer learning-based methods for SOH estimation have emerged recently [Shu *et al.*, 2021; Tian *et al.*, 2021]. For instance, [Tan and Zhao, 2020] developed an LSTM-FC network with a fine-tuning strategy to predict SOH using only the first 25 % of the target LIB dataset. Besides, [Lu *et al.*, 2023] integrated a swarm of DNNs with domain adaptation to enable SOH estimation without target LIB labels. While these approaches alleviate the data collection difficulties to some extent, they still cannot follow the battery upgrading pace and pose a barrier to developing battery technologies. Another issue is that existing transfer learning methods are effective on well-collected datasets for fine-tuning or feature alignment, but they fail to match the real-world scenario where a raw battery continues to age, bringing unlabeled data incrementally over months. Hence, developing a large model that can handle diverse types of LIBs (e.g., cross-chemistry, cross-manufacturer, and cross-capacity) while having a good match on raw, new-generation batteries in real life is crucial.

This paper firstly proposes an LLM-driven framework for SOH estimation that does not require to be trained by target data in advance. Since creating massive datasets and training a large LIB model is rarely sustainable, we explore the cross-modal migratability potential of the language model's powerful generalization capability. We convert [Radford *et al.*, 2019] GPT2's modality from language to battery data by retraining the input and output layers and plug-in 'battery-

^{*}Email: ehmppmonkey@outlook.com

[†]Email: zhihong@xmu.edu.cn

specific adaptors’ to create our backbone. Secondly, a test-time training strategy is employed to fully leverage the incrementally acquired unlabeled data, adapting to real-world scenarios where a raw battery undergoes temporal aging through charging and discharging. This strategy effectively mitigates cumulative errors resulting from temporal distribution shifts of LIBs in use over months. The proposed framework demonstrates comparable accuracy among extensive baseline methods under regular settings and achieves state-of-the-art accuracy under zero-shot settings. This work also presents a preliminary analysis of the theoretical challenges associated with cross-battery estimation and elucidates our framework’s underlying principles.

In summary, our main contributions are as follows:

- We fully utilize the strong generalization capability of LLM to establish a framework for cross-battery SOH estimation, lighting the burden of months-to-year degradation experiments for data collection.
- We employ a test-time training strategy to match the real-world scenario where a raw battery continues to age, bringing unlabeled data incrementally but often involves temporal distribution shifts. This strategy ensures estimation accuracy even at the battery’s end of life.
- The validation results demonstrate that the proposed framework achieves state-of-the-art zero-shot accuracy on four widely recognized LIB datasets, and two of them are even comparable to the latest domain adaptation methods.

2 Related Work

2.1 Data-driven battery SOH estimation

Data-driven approaches for battery SOH estimation display greater benefits in accuracy and online computation efficiency than traditional mechanism based models such as equivalent circuit models (ECMs) and physics-based models (PBMs) [Ng *et al.*, 2020]. In [Roman *et al.*, 2021], the author proposed a machine learning pipeline for battery state of health estimation involving four algorithms: Bayesian Ridge Regression (BPR), Gaussian process regression (GPR), Random Forest (RF), and Deep ensemble of neural networks (dNNe).

While accurate health state estimation methods have progressed significantly, the time- and resource-consuming degradation experiments needed to generate lifelong time training data make target battery agnostic approaches attractive [Ye and Yu, 2021; Han *et al.*, 2022]. For instance, [Tan and Zhao, 2020] developed an LSTM-FC network to predict SOH by fine-tuning only the first 25 % of the target dataset. Similarly, [Wang *et al.*, 2023] retrained LSTM using only two target battery cells during the transfer learning process. Besides, [Lu *et al.*, 2023] integrated a swarm of deep neural networks with domain adaptation to enable SOH estimation in the absence of target battery labels, which inspired our work.

2.2 Capability of LLM

In-modality generalization capability. Extensive research work has verified the in-modality generalization capability of

large language models. Bert [Devlin *et al.*, 2018] used transformer encoders and employed a masked language modeling task to recover the random masked tokens within a text. Furthermore, OpenAI proposed GPT [Radford *et al.*, 2018] and GPT2 [Radford *et al.*, 2019], verifying that scaling up language models significantly improves zero-shot performance on various downstream language tasks. By few-shot instruction tuning, GPT3 [Brown *et al.*, 2020] even reached competitiveness with prior state-of-the-art fine-tuning approaches on NLP tasks that even required on-the-fly reasoning or domain adaptation.

Cross-modality capability. Pre-trained knowledge of language models is generally used for downstream tasks with a different modality. Indeed, [Kao and Lee, 2021] verified the cross-modality power of Bert in transferring from natural language to ammonia acid, DNA, and music. Additionally, [Lu *et al.*, 2022] investigated the capability of a transformer pre-trained on natural language to generalize to other modalities with minimal fine-tuning. Their so-called Frozen Pretrained Transformer (FPT) shows superiority over the randomly initialized same architecture in improving performance and compute efficiency on cross-modality downstream tasks spanning numerical computation, vision, and protein fold prediction. Furthermore, [Zhou *et al.*, 2023] also leveraged a modified GPT2 on general time series analysis and achieved results comparable to SOTA. [Pang *et al.*, 2023] reveals that LLMs trained solely on textual data are surprisingly strong encoders for purely visual tasks.

Inter-modality migration of generalization capability. Relatively, exploring whether the powerful generalization capability of LLM can be migrated to another modality is an ongoing research direction. Note that [Zhou *et al.*, 2023] discovered the superior few-shot/zero-shot capability of LLM on general time series analysis experimentally despite a little lack of explainability.

2.3 Test-Time Training

Test-Time Training (TTT) is a general approach for improving the performance of predictive models when there are distribution shifts between training and testing data [Sun *et al.*, 2020]. A typical TTT model has a supervised main task head for labeled training data and a self-supervised head for unlabeled testing data. When testing data arrives in an online stream, the online version of TTT also updates incrementally via self-supervision. This method operates well when the self-supervised task propagates gradients that correlate with the main task, demonstrating an improved generalization ability on many visual benchmarks for distribution shifts [Gandelsman *et al.*, 2022].

3 Methodology

In this section, we introduce our proposed framework, GPT4Battery, which leverages LLM’s generalization capability for cross-battery state of health (SOH) estimation, along with a more practical setting—test-time training strategy as shown in Figure 1.

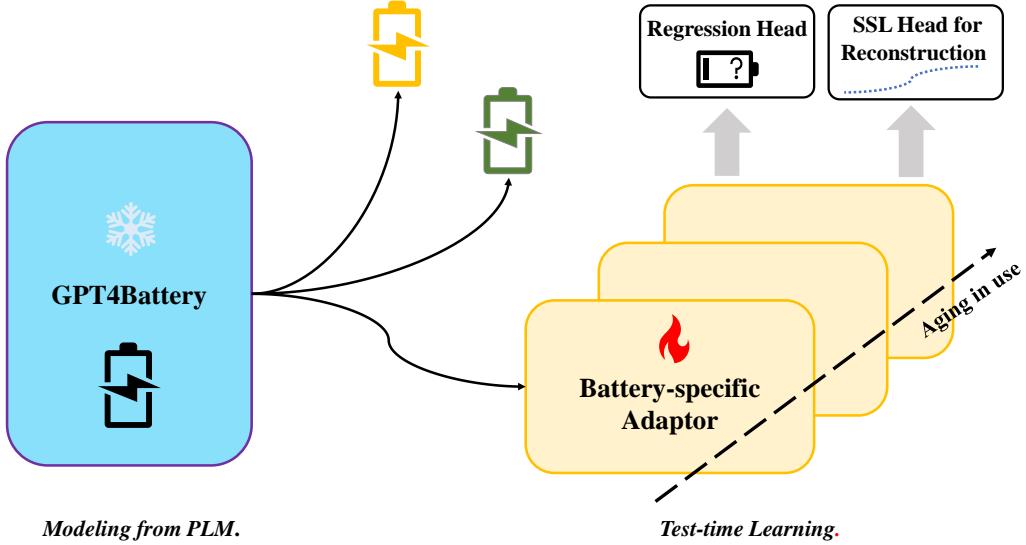


Figure 1: Proposed GPT4Battery architecture. We highlight the leverages of LLM’s generalization capability for cross-battery SOH estimation and the scenario-fit setting of test-time learning strategy. For each different LIB, we freeze the most learned knowledge from language models and only fine tune the adaptor. The two head structure are designed for test-time learning.

Dataset	Electrode Material (Cathode)	Nominal capacity	Voltage range	Samples	Collector
CALCE	LCO	1.1(Ah)	2.7-4.2(V)	2807	University of Maryland
SANYO	NMC	1.85(Ah)	3.0-4.1(V)	415	RWTH Aachen University
KOKAM	LCO/NCO	0.74(Ah)	2.7-4.2(V)	503	University of Oxford
PANASONIC	NCA	3.03(Ah)	2.5-4.29(V)	2770	Beijing Institute of Technology
GOTION	LFP	27(Ah)	2.0-3.65(V)	4262	Beijing Institute of Technology

Table 1: Main specifications of selected LIB datasets.

3.1 Problem Formulation

Using the battery’s raw voltage-time charging curve is a popular and easy-to-implement approach to accurately estimate multiple states over battery life [Tian *et al.*, 2022; Roman *et al.*, 2021]. Determining SOH can be considered a univariate supervised regression problem in this case. In addition to traditional SOH estimation, which requires training data over the battery’s full life cycle, the proposed framework is designed to perform well under a zero-shot scenario. In this scenario, the pre-trained model is tested on a raw battery with different chemistry and manufacture without training in advance. Since developing a battery degradation dataset takes 644–8473 hours of degradation experiments [Lu *et al.*, 2023], models with such adaptive capability are of great value, as described above.

Problem 1: Regular SOH Estimation

We use partial charging data to determine the battery state of health (SOH) [Tian *et al.*, 2021]. Battery SOH is generally defined as:

$$SOH_s = \frac{Q_s}{Q} \quad (1)$$

where Q is the initial maximum battery capacity, and Q_s is the maximum battery capacity at the s^{th} cycle.

In a constant-current charging process, the voltage $V(t)$ and current $I(t)$ are stored by BMS at every time step, and a voltage sampling window is applied to capture the partial charging window at the s^{th} cycle:

$$q_i^s(V) = \int_{V_{\min}}^{V_{\min} + i\Delta V} |I(t)| dt, i \in \{0, 1, \dots, K\} \quad (2)$$

The cycle’s charging feature $q_i(V)$ is obtained by gridding the voltage sampling window with step ΔV from V_{\min} to $V_{\min} + K\Delta V$. Then, we expand it to a 1-d vector and normalize it with the initial capacity Q to make it adaptive to different LIBs.

$$\mathbf{q}^s(\mathbf{V}) = [q_0(V), q_1(V), \dots, q_K(V)]/Q \quad (3)$$

Next, the sampled data sequence is mapped to the SOH:

$$SOH_s = f_{DNN}(\mathbf{q}^s(\mathbf{V})) \quad (4)$$

Problem 2: Zero-shot Adaptivity to Raw Battery

In this work, we propose a new problem setting: to predict SOH values for unseen batteries without prior training. More specifically, given a set of L -labeled samples $\mathcal{L} = \{(x_1, y_1), (x_2, y_2), \dots, (x_L, y_L)\}$ for pre-training, where x denotes the partial charging curve $q(V)$ and y is the target SOH label from the experimental lifelong battery dataset.

When a raw battery comes in the real scenario, the fine-tuning dataset comprises only U unlabeled data $\mathcal{U} = \{x_1, x_2, \dots, x_U\}$ that arrive sequentially since the degradation process during usage typically spans from months to years. Meanwhile, the label is inaccessible due to incomplete discharging under real-world usage, making it challenging to modify by supervised fine-tuning. Nevertheless, obtaining the partial charging curves of the first cycle (i.e., in a fresh state) is easily achievable (e.g., through LIB formation or factory testing), and their state of health can be considered 100%.

3.2 Pre-trained language model for cross-battery SOH estimation

Since it is rarely sustainable to create massive datasets and train a large LIB model, we explore the migratability potential of language model’s generalization capability. A few modifications are made to suit the pre-trained language model (PLM) for battery data. We aim to emphasize the inherent internal computation within the language model by freezing its main components. As a result, the following three adjustments are made to a pre-trained GPT-2 model [Radford *et al.*, 2019].

Frozen Pre-trained Language Model. As self-attention layers and FFN (Feedforward Neural Networks) contain the most learned knowledge from pre-trained language models [Lu *et al.*, 2022], freezing these components would be fair to deploy the migration of PLM’s generalization capability from language to downstream tasks. The positional embeddings and layer normalization layer are trainable as standard practice.

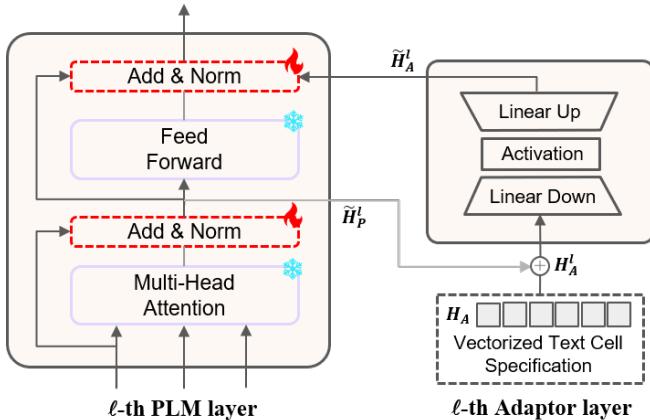


Figure 2: Knowledgeable ‘battery specific adaptor’ with plug-in cell text specification.

Algorithm 1 GPT4Battery Pipeline

Input: A set of L labeled samples

$\mathcal{L} = \{(x_1, y_1), (x_2, y_2), \dots, (x_L, y_L)\}$ from an experimental dataset and an unlabeled target dataset \mathcal{U} with $\{(x_1, 1), x_2, x_3, \dots, x_u\}$ coming in sequentially.

Parameter: Frozen pre-trained GPT2 parameters f , main task head parameters h , and self-supervised head g .

Output: The label for every x in \mathcal{U} coming out sequentially.

- 1: **Stage 1: Joint pre-training**
 - 2: **for** x_i, y_i in L **do**
 - 3: $f_0 \leftarrow \min_f l_m(x_i, y_i; f, h) + l_s(x_i; f, g)$
 - 4: $h_0 \leftarrow \min_h l_m(x_i, y_i; f, h)$
 - 5: $g_0 \leftarrow \min_g l_s(x_i; f, g)$
 - 6: **end for**
 - 7: **Stage 2: Joint fine-tune with 1st cycle label**
 - 8: Assign a stopping loss;
 - 9: **while** $l_m(x_1, 1; f, h) >$ stopping loss **do**
 - 10: $f_1 \leftarrow \min_f l_m(x_1, 1; f_0, h_0) + l_s(x_1; f_0, g_0)$
 - 11: $h_1 \leftarrow \min_h l_m(x_1, 1; f_0, h_0)$
 - 12: $g_1 \leftarrow \min_g l_s(x_1; f_0, g_0)$
 - 13: **end while**
 - 14: **Stage 3: Self-supervised Test-time training**
 - 15: **for** x_i in U **do**
 - 16: $f_x \leftarrow \min_f l_s(x_i; f_1, g_1)$
 - 17: $g_x \leftarrow \min_g l_s(x_i; f_1, g_1)$
 - 18: Predict: $f_x \circ h_1$
 - 19: **end for**
-

Trainable input layers. Re-initializing a new input layer to query the transformer is important since the model operates in a new modality . We use linear probing to minimize the amount of computation outside the transformer.

Knowledgeable battery-specific adaptor. To enhance the pre-trained model’s performance on downstream tasks with minimal effort, PEFT methods, such as LoRA [Hu *et al.*, 2021], VPT [Jia *et al.*, 2022], and Prefix Tuning [Li and Liang, 2021], have been widely employed in CV and NLP. This work elaborates on our knowledge of battery adapters, incorporating each LIB cell’s vectorized text specifications for battery-specific fine-tuning. For the l -th $l \in [1, L]$ adapter layer, the input $H_A^l \in \mathbb{R}^{(m+n) \times d}$ is formed by vertically concatenating the hidden features $\tilde{H}_P^l \in \mathbb{R}^{m \times d}$ from the l -th PLM layer and the vectorized text specification of LIB cell $H_A^l \in \mathbb{R}^{n \times d}$ from knowledge. m and n denote the length of the PLM input sequence and knowledge piece, respectively, and d is the hidden size. A learnable gate function is used to obtain crucial query information by filtering the hidden features of PLM. Specifically,

$$H_A^l = [\tilde{H}_P^l \odot \sigma(G); H_A] \quad (5)$$

Now given the input H_A^l , the adaptor layer projects it down to the r dimension with a linear down layer. Then a self-attention layer is employed to fuse the battery knowledge and the query information from the PLM. After that, another linear projection layer is applied to project it up to the original dimension d , and finally, this output is input to the PLM

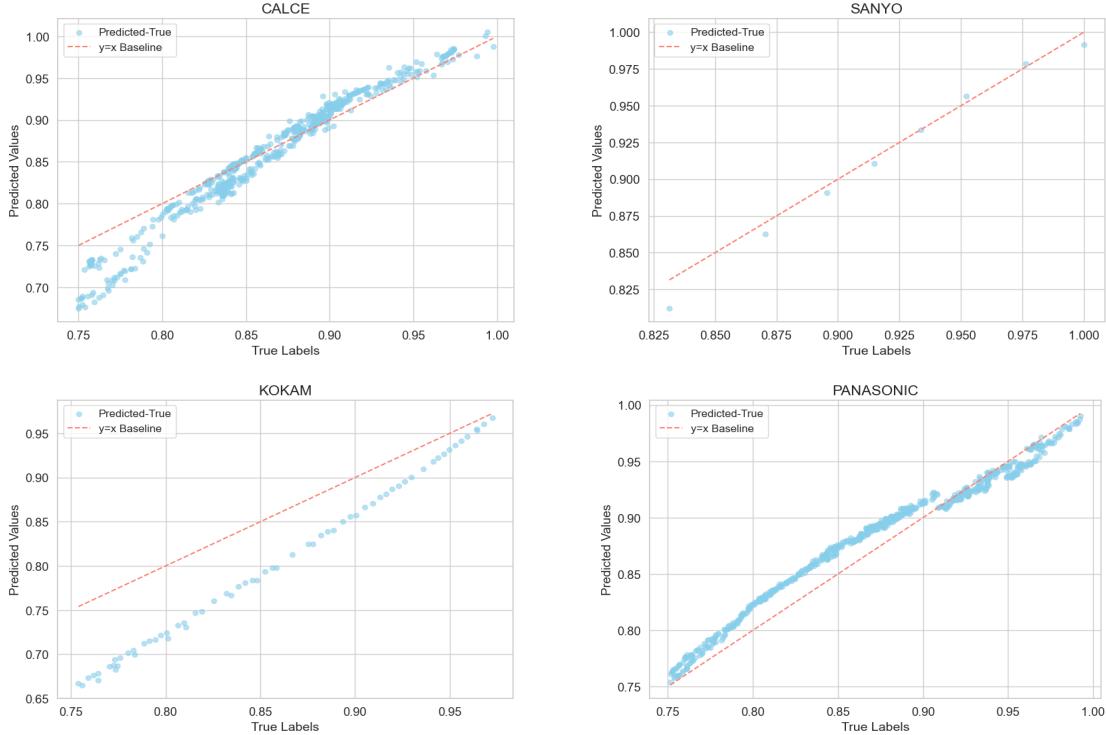


Figure 3: Visual results of adaptation to four different kinds of LIBs under zero-shot setting using GPT4Battery, pre-trained on the GOTION Dataset.

through a residual connection. For each different LIB, a separate battery-specific adaptor is employed to enhance the fine-tuning performance.

3.3 Adaption to Raw Battery with Test-time Training

In a real-world scenario, a raw battery undergoes continuous aging with incrementally acquired unlabeled data, sampled over months. This renders existing transfer learning-based methods unadaptable. In this section, we introduce the theory of test-time training, which aligns more closely with practical settings.

Two Head Structure

We consider a Y-shaped structure with two heads, similar to [Gandelsman *et al.*, 2022]: a feature extractor f simultaneously followed by a self-supervised head g and a main task head h . For the main task (SOH estimation) h and self-supervised head (charging curve reconstruction) g , we use linear down projection and linear up projection respectively for clear comparisons. Here the self-supervised task is designed by recovering the input charging curve from the hidden output features of the modified PLM. It should be noted that f is exactly the modified PLM, as presented in the last section.

Pipeline

Joint Pre-Training. An experimental labeled dataset \mathcal{L} containing $\{(x_1, y_1), \dots, (x_n, y_n)\}$ is used to train the main

task network $f \circ h$. Meanwhile, the reconstruction task $f \circ g$ is trained, conducting a joint training which adopts a multi-task learning strategy [Caruana, 1997]. Losses for both tasks are added together, and gradients are taken to collect all parameters. The joint training problem is therefore formulated as follows:

$$\mathbf{f}_0, \mathbf{h}_0, \mathbf{g}_0 = \min_{\mathbf{f}, \mathbf{h}, \mathbf{g}} \frac{1}{n} \sum_{i=1}^n l_m(x_i, y_i; \mathbf{f}, \mathbf{h}) + l_s(x_i; \mathbf{f}, \mathbf{g}) \quad (6)$$

where l_m denotes the main SOH regression loss defined by MSE and l_s is the reconstruction loss of charging curve x .

Fine-tuning using the ‘labeled’ 1st cycle . The partial charging curves of the first cycle (i.e., in fresh status) are easily obtained by LIB formation or factory test, and their SOH labels can be treated as 1. We take full advantage of this sample to cover the distribution gap caused by the battery mechanism. An early stopping technique is applied to prevent overfitting on this label as presented in Algorithm 1.

$$\mathbf{f}_1, \mathbf{h}_1, \mathbf{g}_1 = \min_{\mathbf{f}, \mathbf{h}, \mathbf{g}} l_m(x_1, 1; \mathbf{f}_0, \mathbf{h}_0) + l_s(x_1; \mathbf{f}_0, \mathbf{g}_0) \quad (7)$$

Test-Time Training. At test time, we start from the PLM pre-trained encoder f_1 , as well as two task heads g_1, h_1 fine-tuned in the last step. The main task head h_1 is frozen, and the following loss is optimized as each test sample x arrives:

Datasets	CALCE		SANYO		KOKAM		PANASONIC		Average	
Methods/Metrics	MAE(%)	Time	MAE(%)	Time	MAE(%)	Time	MAE(%)	Time	MAE(%)	Time
GPR	23.58	-	21.00	-	31.83	-	30.70	-	26.78	-
RD	8.74	-	13.33	-	9.02	-	10.52	-	10.40	-
SVR	4.27	-	5.62	-	6.44	-	5.53	-	5.47	-
CNN	10.31	-	17.90	-	14.64	-	25.46	-	17.08	-
Benchmark1	4.13	4.86	2.78	4.25	4.00	4.14	4.57	5.5	3.87	4.69
Benchmark2	3.35	11.35	0.88	25.3	6.21	13.18	1.44	11.16	2.97	15.25
GPT4Battery	1.43	43.73	0.87	68.72	5.56	255.24	0.81	61.17	2.17	107.22
swarm	1.12	-	1.21	-	1.76	-	2.09	-	1.55	-

Table 2: Comparison of methods under zero-shot setting. We calculate the MAE (as %) for each dataset and average four datasets as well as inference time (ms) for three benchmarks. A lower MAE score indicates better performance. **Red:** best, **Black:** second best. Note that swarm is tested under domain adaptation setting where target data is accessible.

$$\mathbf{f}_x, \mathbf{g}_x = \min_{\mathbf{f}, \mathbf{g}} l_s(x_i; \mathbf{f}_1, \mathbf{g}_1) \quad (8)$$

after TTT, we make a prediction on each x_i as $\mathbf{f}_x \circ \mathbf{h}_1$.

4 Experiments

This section evaluates the proposed GPT4Battery on five publicly recognized datasets collected from 65 battery cells to demonstrate our framework’s effectiveness on this challenging problem.

4.1 Experiment Settings

Datasets. The experiments employ the following datasets manufactured by CALCE [He *et al.*, 2011], SANYO [Li *et al.*, 2021], KOKAM [Birkl, 2017], PANASONIC, and GOTION HIGH-TECH [Lu *et al.*, 2023]. These datasets cover widely-used cathode active materials, a capacity ranging from 0.74Ah to 27Ah, and five different manufacturers, emphasizing our method’s potential to work directly on new-generation batteries under a zero-shot setting. This paper adopts [Lu *et al.*, 2023] for data pre-processing. The differences between these five LIBs are compared in Table 1 .

Baselines. The GPT4Battery is compared against three categories of methods:

- Four popular data-driven SOH estimation methods include Gaussian process regression (GPR), support vector regression (SVR), Random Forest (RD) [Roman *et al.*, 2021], and CNN [Tian *et al.*, 2022].
- One latest domain adaptation based transfer learning method for estimation without the label of target LIBs [Lu *et al.*, 2023].
- We created Benchmark1 by substituting PLM with a regular CNN architecture and Benchmark2 by disabling the test-time training technique.

4.2 Results

Evaluation under zero-shot setting

Under a zero-shot setting, we employed GOTION [Lu *et al.*, 2023] as the pre-training dataset for joint training because it

contains more samples relatively. After obtaining a foundation model, test-time training is employed to adapt this model to four different types of batteries. Fig. 3 illustrates the visual accuracy performance in the absence of any data of the four LIBs except for the 1st cycle with SOH label considered as 1.

Table 2 reports the comparative results of the mean absolute error (MAE) on baselines and inference time on three benchmarks. Without the target training data, the existing methods fail to provide reliable estimation with their MAEs over 10% except for SVR. In contrast, the proposed GPT4Battery framework achieves accurate SOH estimation of an average MAE of 2.17% on four datasets. On SANYO and PANASONIC, our method even outperforms the latest domain adaptation method where the target LIB features are accessible by MAE of 0.34% and 0.28%, respectively.

The expense of a little inference time (within 1 seconds per cycle) is certainly acceptable considering one charging cycle could take hours. It should be noted that the least desirable performance is observed on KOKAM, mainly attributed to high linearity of this specific dataset.

Datasets	#1	#2	#3	#4	#5
Methods/Metrics	MAE(%)				
GPR	6.42	5.84	8.06	7.63	6.39
RD	0.48	0.13	0.11	0.52	0.29
SVR	4.23	5.87	6.36	5.50	4.66
CNN	0.72	1.90	2.65	0.72	0.28
Benchmark2	2.22	0.61	1.13	1.40	0.25

Table 3: Comparison of methods under regular setting. We calculate the MAE (as %) for each dataset. **Black:** best.

Evaluation under regular setting

Under regular settings, we disabled TTT and compared this LLM-driven model (Benchmark2) to the existing methods. LIBs within one dataset are divided into training and testing sets to fit this traditional machine learning setting. Unsurprisingly, our model shows no superior performance over baseline methods, as a model with strong generalization capability may not converge as effectively on one specific dataset.

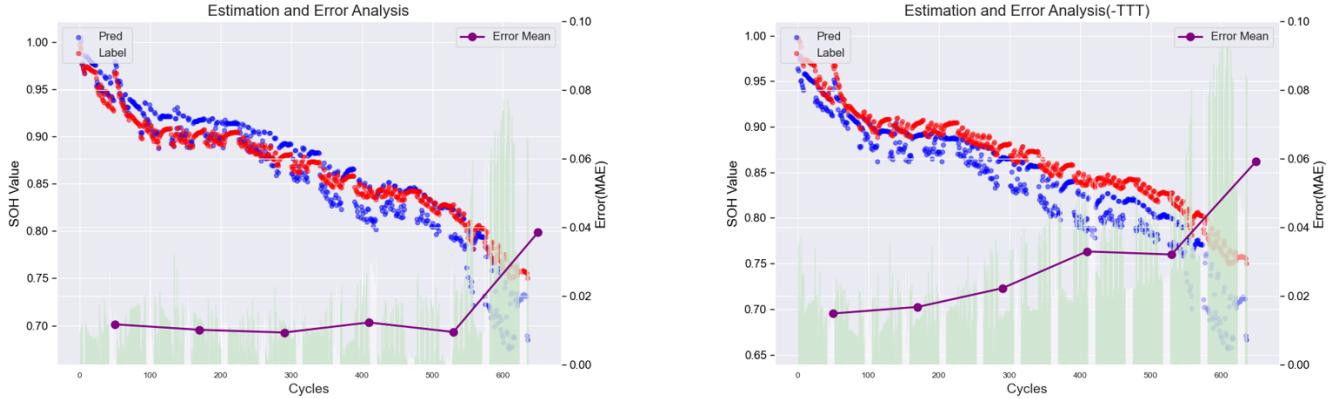


Figure 4: Quantitative comparison of GPT4Battery (**left**) and Benchmark2 (**right**, with TTT disabled) on CALCE under zero-shot setting. Despite both methods demonstrating accurate estimation early in the life cycles, GPT4Battery with TTT effectively reduces errors accumulated due to temporal distribution shifts, ensuring accuracy in the mid-to-late life cycles.

4.3 Analysis

The challenge for zero-shot estimation originates from the *spatial domain shifts* across batteries with different chemistry, capacities, or manufacturers. Meanwhile, since the degradation process often spans over months to years, variations in the mechanisms within the battery (such as side reactions and stability of solid electrolyte interface) [Brousseau *et al.*, 2005] are more likely to induce *temporal distribution shifts*, making a fixed model fall in the mid-to-late life cycle. The excellent performance of our framework can be attributed to the powerful generalization capability migrated from LLM and test-time learning strategy, both effectively working in dealing with this out-of-distribution problem.

Here, we provide a qualitative explanation of a crucial strategy in our proposed framework, i.e., test-time training. We present two estimation and error analysis maps to visualize the contributions of TTT to SOH estimate. Compared with Benchmark2, where TTT is disabled, GPT4Battery achieves significantly lower Mean Absolute Error (MAE) throughout the life cycle. This stands in contrast to the accumulating estimation error observed in Benchmark2, confirming the assumption of *temporal distribution shifts*.

Theoretically, an intuitive explanation for TTT is that the self-supervised task happens to propagate gradients that correlate with those of the main task [Sun *et al.*, 2020]. In our framework, we believe that the charging curve reconstruction task finds a better *bias-variance trade-off* under the temporally accumulated distribution shifts. The fixed model is biased because it is completely based on biased training data, and the migrated generalization ability from language models is not powerful enough to handle this non-trivial change. The other extreme is completely discarding the pre-trained knowledge and training a new model from scratch on each test input. However, this is undesirable because of the high variance of each input, and the reconstruction task does not always contribute to the main regression task according to our trial and error.

5 Conclusion

We proposed an LLM-driven framework equipped with time-time learning for cross-battery state of health (SOH) estimation, addressing the real-world scenario where raw battery data arrives sequentially without labels. Theoretically, we analyzed that the challenges for cross-battery tasks come from the spatial distribution shifts between diverse LIBs and temporal distribution shifts during the months-to-year aging process, emphasizing the necessity of LLM-driven and test-time learning. To our knowledge, we are the first to explore the iter-modal migratability potential of large language model’s generalization capability and deploy it successfully. Practically, our proposed GPT4Battery achieved state-of-art results in zero-shot SOH estimation across various LIBs.

References

- [Birkl, 2017] Christoph Birkl. Oxford battery degradation dataset 1. 2017.
- [Broussely *et al.*, 2005] Michel Broussely, Ph Biensan, F Bonhomme, Ph Blanchard, S Herreyre, K Nechev, and RJ Staniewicz. Main aging mechanisms in li ion batteries. *Journal of power sources*, 146(1-2):90–96, 2005.
- [Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.
- [Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [Gandelsman *et al.*, 2022] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.
- [Genikomsakis *et al.*, 2021] Konstantinos N. Genikomsakis, Nikolaos-Fivos Galatoulas, and Christos S. Ioakimidis. Towards the development of a hotel-based e-bike rental service: Results from a stated preference survey and techno-economic analysis. *Energy*, 215:119052, 2021.
- [Han *et al.*, 2022] Te Han, Zhe Wang, and Huixing Meng. End-to-end capacity estimation of lithium-ion batteries with an enhanced long short-term memory network considering domain adaptation. *Journal of Power Sources*, 520:230823, 2022.
- [He *et al.*, 2011] Wei He, Nicholas Williard, Michael Osterman, and Michael Pecht. Prognostics of lithium-ion batteries based on dempster–shafer theory and the bayesian monte carlo method. *Journal of Power Sources*, 196(23):10314–10321, 2011.
- [Hu *et al.*, 2021] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.
- [Jakob Fleischmann, 2023] Martin Linder Mikael Hanicke Evan Horetsky Dina Ibrahim Sören Jautelat ukas Torscht Alexandre van de Rijt Jakob Fleischmann, Patrick Schaufluss. Battery 2030: Resilient, sustainable, and circular. 2023. January 16, 2023.
- [Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.
- [Kao and Lee, 2021] Wei-Tsung Kao and Hung-yi Lee. Is bert a cross-disciplinary knowledge learner? a surprising finding of pre-trained models’ transferability. *arXiv preprint arXiv:2103.07162*, 2021.
- [Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.
- [Li *et al.*, 2021] Weihan Li, Neil Sengupta, Philipp Dechent, David Howey, Anuradha Annaswamy, and Dirk Uwe Sauer. One-shot battery degradation trajectory prediction with deep learning. *Journal of Power Sources*, 506:230024, 2021.
- [Lu *et al.*, 2022] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Frozen pretrained transformers as universal computation engines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7628–7636, 2022.
- [Lu *et al.*, 2023] Jiahuan Lu, Rui Xiong, Jinpeng Tian, Chenxu Wang, and Fengchun Sun. Deep learning to estimate lithium-ion battery state of health without additional degradation experiments. *Nature Communications*, 14(1):2760, 2023.
- [Ng *et al.*, 2020] Man-Fai Ng, Jin Zhao, Qingyu Yan, Gareth J Conduit, and Zhi Wei Seh. Predicting the state of charge and health of batteries using data-driven machine learning. *Nature Machine Intelligence*, 2(3):161–170, 2020.
- [Pang *et al.*, 2023] Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. *arXiv preprint arXiv:2310.12973*, 2023.
- [Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [Roman *et al.*, 2021] Darius Roman, Saurabh Saxena, Valentin Robu, Michael Pecht, and David Flynn. Machine learning pipeline for battery state-of-health estimation. *Nature Machine Intelligence*, 3(5):447–456, 2021.
- [Severson *et al.*, 2019] Kristen A Severson, Peter M Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H Chen, Muratahan Aykol, Patrick K Herring, Dimitrios Fragedakis, et al. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5):383–391, 2019.
- [Shu *et al.*, 2021] Xing Shu, Jiangwei Shen, Guang Li, Yuanjian Zhang, Zheng Chen, and Yonggang Liu. A flexible state-of-health prediction scheme for lithium-ion battery packs with long short-term memory network and

transfer learning. *IEEE Transactions on Transportation Electrification*, 7(4):2238–2248, 2021.

[Sun *et al.*, 2020] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[Tan and Zhao, 2020] Yandan Tan and Guangcai Zhao. Transfer learning with long short-term memory network for state-of-health prediction of lithium-ion batteries. *IEEE Transactions on Industrial Electronics*, 67(10):8723–8731, 2020.

[Tian *et al.*, 2021] Jinpeng Tian, Rui Xiong, Weixiang Shen, Jiahuan Lu, and Xiao-Guang Yang. Deep neural network battery charging curve prediction using 30 points collected in 10 min. *Joule*, 5(6):1521–1534, 2021.

[Tian *et al.*, 2022] Jinpeng Tian, Rui Xiong, Weixiang Shen, Jiahuan Lu, and Fengchun Sun. Flexible battery state of health and state of charge estimation using partial charging data and deep learning. *Energy Storage Materials*, 51:372–381, 2022.

[Wang *et al.*, 2023] Yixiu Wang, Jiangong Zhu, Liang Cao, Bhushan Gopaluni, and Yankai Cao. Long short-term memory network with transfer learning for lithium-ion battery capacity fade and cycle life prediction. *Applied Energy*, 350:121660, 2023.

[Ye and Yu, 2021] Zhuang Ye and Jianbo Yu. State-of-health estimation for lithium-ion batteries using domain adversarial transfer learning. *IEEE Transactions on Power Electronics*, 37(3):3528–3543, 2021.

[Zhou *et al.*, 2023] Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained lm. *arXiv preprint arXiv:2302.11939*, 2023.