

SERA: Semantic Representation Aggregation for Reliable and Interpretable Battery Health Forecasting

Jiawei Li
Infocomm Technology Cluster
Singapore Institute of Technology
Singapore
jiawei.li@singaporetech.edu.sg

Fang Liu
School of Science and Technology
Singapore University of Social Sciences
Singapore
liufang@suss.edu.sg

Wei Zhang*
Infocomm Technology Cluster
Singapore Institute of Technology
Singapore
wei.zhang@singaporetech.edu.sg

Zuming Liu
College of Smart Energy
Shanghai Jiao Tong University
Shanghai, China
zmliu@sjtu.edu.cn

Man-Fai Ng
Institute of Advanced Intelligence and
Computing (IAIC), Agency for Science,
Technology and Research (A*STAR), 1
Fusionopolis Way, #16-16 Connexis,
Singapore 138632, Republic of Singapore
ngmf@a-star.edu.sg

Zhi Wei Seh
Institute of Materials Research and
Engineering (IMRE), Agency for Science,
Technology and Research (A*STAR), 2
Fusionopolis Way, Innovis #08-03,
Singapore 138634, Republic of Singapore
sehzw@a-star.edu.sg

Abstract

Battery state of health (SoH) forecasting is important for battery management, but remains challenging due to nonlinear degradation and heterogeneity across batteries. Existing data-driven approaches primarily use temporal models to learn from numerical battery time series, and higher-level degradation characteristics are often not explicitly represented. These characteristics, however, can provide degradation guidance to support reliable forecasting and make the influence of degradation more interpretable. In this paper, we propose SERA, a semantic representation aggregation framework that complements temporal modelling with degradation semantics. Guided by battery domain expertise, SERA extracts degradation semantics from time series and constructs two complementary representations using rule-based knowledge and LLM-based interpretation. The representations are independently encoded and integrated with the representation learned by temporal models through gated aggregations. Experiments on the mainstream benchmark across multiple prediction horizons and different temporal models show that SERA consistently improves forecasting performance, achieving up to a 37.3% reduction in prediction error over the temporal baseline and enhanced generalizability. Counterfactual analysis examines how forecasts respond to changes in degradation semantics to assess interpretability. The results show that prediction responses are consistent with the meanings of key degradation descriptors across tested horizons. Together, these findings demonstrate that structured degradation semantics and effective aggregation can improve forecasting accuracy and support reliable and interpretable battery health forecasting for advanced battery management.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, Washington, DC, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Keywords

Battery health, semantic representation, large language model, representation learning, rule-based diagnostics.

ACM Reference Format:

Jiawei Li, Fang Liu, Wei Zhang, Zuming Liu, Man-Fai Ng, and Zhi Wei Seh. 2026. SERA: Semantic Representation Aggregation for Reliable and Interpretable Battery Health Forecasting. In . ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Predictive health management is increasingly important in manufacturing and industrial systems, where component health directly affects operational reliability, maintenance planning, and asset utilization. Batteries exemplify this need because their growing use in electric vehicles, energy storage, and industrial equipment requires monitoring and managing performance changes associated with repeated operation and electrochemical aging. Battery degradation reduces available capacity and operational efficiency, shortens service life, and, in severe cases, increases safety risks such as overheating and thermal runaway. These concerns make state of health (SoH), measured as the ratio of a battery's current capacity to its initial capacity, an important indicator for characterizing battery health changes [18, 19]. While SoH estimation characterizes the current battery condition, SoH forecasting predicts its future evolution to inform battery management and lifecycle planning [13].

SoH forecasting is challenging because battery degradation is nonlinear, heterogeneous, and affected by multiple interacting factors. Electrochemical aging mechanisms, operating conditions, temperature, and usage patterns can correspond to different degradation trajectories [5, 17]. The trajectories differ even among batteries of the same model, limiting generalization of SoH models from one set of cells to unseen cells [1, 2]. The challenge becomes more evident when the prediction horizon increases, since the model must infer future degradation from a fixed amount of observed history. Besides, practical forecasting requires not only accurate temporal modelling but also representations that characterize the underlying degradation state to support reliable predictions across batteries and interpretable forecasting behavior.

Existing studies predict battery health using physics based models, data-driven methods, and hybrid solutions [6, 10]. Deep sequential models, such as LSTM and its variants, learn degradation patterns from historical measurements through temporal modelling [12], while Transformer based approaches use attention mechanisms to capture both spatial and temporal dependencies [4, 8]. Transfer learning has also been investigated to improve generalization under domain shift [3, 20]. However, data-driven models often do not explicitly represent higher-level degradation characteristics, such as trend and severity, and possible degradation mechanisms. This limits the ability of the models to explicitly leverage such characteristics for prediction. As such, other approaches incorporate battery domain knowledge into SoH models. For example, empirical observations suggest that SoH generally declines with aging, electrical engineering provides equivalent circuit models to describe battery behavior, and materials science explains mechanisms such as lithium inventory loss and electrode degradation [7, 11, 15]. Such knowledge can be encoded through equations, parameters, or constraints, providing additional guidance and a more comprehensive understanding of battery degradation for SoH prediction.

Although temporal and knowledge-enhanced models have advanced battery health forecasting, existing methods predominantly represent degradation information numerically, which may not explicitly convey higher-level degradation characteristics. Our key idea is to introduce text-based semantic representations as another modality for characterizing degradation and enable more explicit and descriptive expression of degradation knowledge. These representations capture degradation trends and change points, severity and risk, and possible underlying mechanisms. By integrating semantic descriptions with numerical time series, we seek to leverage the strengths of different representations to improve forecasting accuracy and reliability and to enhance interpretability.

Based on this idea, we propose SERA, a semantic representation aggregation framework for reliable and interpretable battery health forecasting with both temporal and semantic representations. Given the battery time series data and domain knowledge, SERA constructs two semantic representations. The rule-based representation provides consistent degradation interpretation through predefined rules, and the LLM-based representation interprets relationships among multiple rules. These representations are encoded separately from the temporal representation and integrated through gated aggregations, which adaptively control the contribution of each representation. The aggregated representation is used to predict future SoH across multiple prediction horizons. Experiments on a mainstream benchmark show that SERA improves forecasting accuracy with up to a 37.3% reduction in mean absolute error (MAE) over the temporal baseline and improved generalization to unseen batteries. Counterfactual analysis further shows that prediction responses are consistent with the meanings of key degradation descriptors, supporting the interpretability of the forecasting behavior.

The main contributions of this work are threefold. First, we introduce a structured semantic representation that translates battery degradation evidence into interpretable knowledge-derived descriptors and makes such information available to SoH models. Second, we develop SERA, which integrates rule- and LLM-based semantic representations with temporal modelling through gated

aggregation, providing a general framework for adaptively incorporating semantic and temporal representations. Third, we evaluate SERA on a real-world dataset across multiple prediction horizons and temporal encoders, and use counterfactual semantic interventions to assess the consistency between forecast responses and the meanings of degradation semantics.

The remainder of this paper is organized as follows. Section 2 defines the forecasting problem. Section 3 presents SERA. Section 4 reports the experimental setup, forecasting results, and interpretability analyses. Section 5 concludes the paper.

2 Battery SoH Forecasting Problem Formulation

In this section, we formulate battery health forecasting as a multi-variate sequence-to-sequence prediction problem and investigate the forecasting task for a future horizon. For a battery with T observed cycles, each comprising charging and discharging time series, let $\mathbf{x}_t \in \mathbb{R}^F$ denote the feature vector extracted from cycle t , where F is the number of input features. These features include cycle-level battery measurements and derived health indicators, such as voltage, current, temperature, and internal resistance. Given the most recent L cycles ending at cycle $t \geq L$, the historical observation window is defined as,

$$\mathbf{X}_{t-L+1:t} = [\mathbf{x}_{t-L+1}, \mathbf{x}_{t-L+2}, \dots, \mathbf{x}_t] \in \mathbb{R}^{L \times F}, \quad (1)$$

where L denotes the observation window length. Let y_t denote the SoH of the battery at cycle t . The forecasting target is the future SoH trajectory over the next H cycles,

$$\mathbf{y}_{t+1:t+H} = [y_{t+1}, y_{t+2}, \dots, y_{t+H}] \in \mathbb{R}^H. \quad (2)$$

The forecasting model therefore learns the mapping,

$$\hat{\mathbf{y}}_{t+1:t+H} = f_{\theta}(\mathbf{X}_{t-L+1:t}), \quad (3)$$

where $f_{\theta}(\cdot)$ denotes a forecasting model parameterized by θ . The model predicts the future SoH trajectory as $\hat{\mathbf{y}}_{t+1:t+H}$, which is expected to closely approximate the ground truth trajectory $\mathbf{y}_{t+1:t+H}$.

3 SERA Methodology

In this section, we introduce the overall architecture of SERA, followed by semantic representation construction, temporal modelling, aggregation, and model training.

3.1 Design Overview of SERA

The overall architecture of SERA is illustrated in Fig. 1, following a four-layer pipeline comprising input, representation, aggregation, and forecasting. At the input layer, a window of time series provides battery measurements that are processed through three branches. The temporal branch learns degradation patterns directly from numerical measurements and provides the baseline for forecasting. The rule-based branch applies battery domain knowledge to extracted degradation evidence to construct semantic descriptions of the battery state. The LLM-based branch interprets the degradation evidence to generate contextual descriptions within the same semantic schema. The representations from the three branches are then adaptively aggregated to predict future SoH.

Semantic representation is one of the key contributions of SERA and is constructed through the rule- and LLM-based branches at

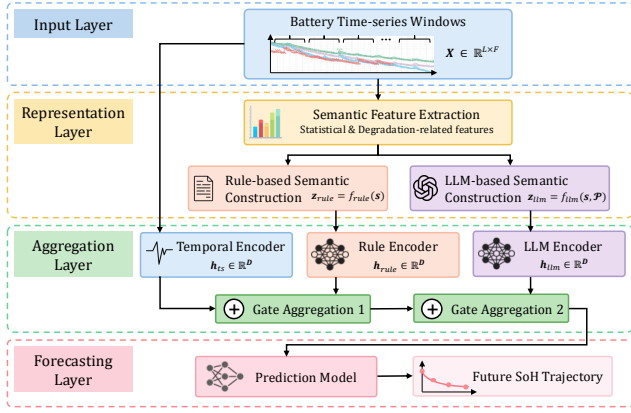


Figure 1: Overall architecture of SERA. Given a window of battery time series, semantic features are extracted to construct rule-based and LLM-based degradation semantics. Temporal and semantic information is encoded separately and then aggregated for SoH forecasting.

the representation layer. The rule-based branch constructs deterministic degradation descriptions using predefined rules grounded in battery domain knowledge. These rules are expected to provide consistent and useful degradation information to guide SoH forecasting. However, they are fixed and not exhaustive, and therefore offer limited flexibility in capturing diverse battery dynamics. To address this limitation, the LLM-based branch leverages the language understanding capability of LLMs to interpret degradation evidence more flexibly and provide richer semantic representations.

At the aggregation layer, the temporal, rule-based, and LLM-based information is independently encoded and integrated through a gated aggregation. The aggregation design enables SERA to adaptively control the contribution of different information sources rather than combining them deterministically. Finally, the aggregated representation is passed to the forecasting layer, where the prediction module generates the future SoH trajectory within the specified prediction horizon. The components of SERA are described in detail below.

3.2 Structured Semantic Representation

Guided by battery domain knowledge, we select seven semantic rules to characterize battery degradation, including trend, risk level, mechanism, severity, change point, key metric, and metric value. The rule definitions, candidate values, and associated degradation features are summarized in Table 1. The rules for trend and change point characterize degradation progression and its changes, while those for risk level and severity indicate degradation risk and severity. The mechanism rule describes possible underlying degradation processes, while key metric and metric value associate the semantic descriptions with quantitative degradation features. Together, these rules define the structured semantic representation as,

$$\mathbf{z} = [z_{\text{trend}}, z_{\text{risk}}, z_{\text{mech}}, z_{\text{sev}}, z_{\text{cp}}, z_{\text{metric}}, z_{\text{value}}]. \quad (4)$$

To assign values to these semantic rules, we preprocess the battery time series into a feature vector $\mathbf{s}$ that captures degradation

characteristics required for semantic construction. These characteristics are derived from temporal changes and statistical variations over the observation window. Specifically, we compute the slope and acceleration ratio of the capacity-based SoH sequence to characterize degradation progression, together with statistical summaries of temperature and internal resistance to characterize operating variability. These derived features provide the quantitative basis for assigning values to the semantic rules. For example, the SoH slope and acceleration ratio are used to distinguish among stable, mild_decay, linear_decay, and accelerating_fade for the trend rule. The resulting $\mathbf{s}$ provides a common input to both semantic construction branches.

The rule-based branch implements the semantic rules using predefined thresholds grounded in domain knowledge. Specifically, the preprocessed features in $\mathbf{s}$ are compared against their corresponding thresholds to determine the candidate values described above. This threshold-based assignment produces a deterministic semantic representation for a given feature vector. The mapping is expressed as,

$$\mathbf{z}^{\text{rule}} = f_{\text{rule}}(\mathbf{s}), \quad (5)$$

where $f_{\text{rule}}(\cdot)$ applies the same predefined rules across all input windows. Although this provides consistent semantic guidance, fixed thresholds may not capture contextual relationships among features, motivating the LLM-based branch.

The LLM branch interprets the same feature vector $\mathbf{s}$ through a prompt specification $\mathcal{P}$ that defines the semantic rules and their candidate values. The prompt instructs the LLM to consider the extracted features jointly when assigning values to the semantic rules, while restricting its output to the common schema, yielding,

$$\mathbf{z}^{\text{llm}} = f_{\text{llm}}(\mathbf{s}, \mathcal{P}), \quad (6)$$

where $f_{\text{llm}}(\cdot)$ denotes the LLM-based semantic mapping process. Drawing on knowledge learned from natural language, this branch provides greater flexibility in capturing contextual relationships among degradation features than predefined threshold-based mappings in Eq. (5). The resulting representation provides complementary semantic guidance while retaining the same semantic rules and candidate values as the rule-based branch.

3.3 Representation Encoding and Aggregation

Given the raw battery time series and the rule- and LLM-based semantic representations, SERA encodes each input along its corresponding branch to obtain temporal and semantic embeddings. These three embeddings are then integrated through gated aggregation for subsequent SoH prediction.

3.3.1 Temporal Encoding. In the first branch, we perform temporal modelling using an encoder that processes the battery time series window $\mathbf{X} \in \mathbb{R}^{L \times F}$ to obtain a temporal embedding. We use PatchTST [9] as the temporal encoder in this study. Its patch-based design groups consecutive observations into local segments, preserving patterns within each patch while reducing the number of tokens processed by attention compared with using one token per cycle. Specifically, the input window is divided into M patches of length P , each projected into a D -dimensional embedding. A learnable classification (CLS) token, used here to summarize the input sequence, is prepended to the patch embeddings, and positional

Table 1: Structured semantic representation schema for battery degradation. Each semantic component is described by its candidate values, supporting source features, and degradation meaning.

Rule	Candidate Values	Source Features	Description
Trend	stable, mild_decay, linear_decay, accelerating_fade	SoH slope, SoH acceleration ratio	Overall degradation tendency observed within the input window.
Risk Level	low, medium, high	SoH slope, temperature std. (T_{std}), IR std. (σ_{IR})	Estimated level of degradation risk based on evidence from multiple battery indicators.
Mechanism	SEI_growth, lithium_plating, cathode_degradation, thermal_aging, impedance_rise, mixed, stable	SoH slope, σ_{IR} , T_{std} , charge time slope	Dominant electrochemical or thermal degradation behaviour inferred from the observed feature patterns.
Severity	mild, moderate, severe	Magnitude of SoH slope, mean SoH level	Strength of the degradation evidence observed within the input window.
Change Point	yes, no	SoH acceleration ratio, SoH acceleration flag	Indication of a possible transition in the observed degradation behaviour.
Key Metric	SoH_slope, Tstd, IRstd, CTslp, QDsIp	Selected diagnostic indicator	Most informative feature supporting the current degradation interpretation.
Metric Value	Numerical ($\mathbb{R}$)	Corresponding raw feature value	Numerical value of the selected diagnostic metric obtained from the extracted battery features.

information is added to retain their temporal order. The resulting sequence of size $(M + 1) \times D$ is processed by L_{enc} Transformer encoder layers. The final CLS token output provides the temporal embedding as,

$$\mathbf{h}_{ts} = \mathbf{E}_{cls}^{(L_{enc})} \in \mathbb{R}^D. \quad (7)$$

Note that SERA is not restricted to PatchTST and can accommodate other temporal encoders to generate temporal embeddings. We evaluate several alternative architectures in the experiments.

3.3.2 Semantic Representation Encoding. The second and third branches independently encode the rule- and LLM-based semantic representations. These representations are generated through different methods and provide different interpretations of the degradation evidence. We use two separate encoders with the same architecture and independent parameters to retain their distinct information. In each branch, the structured semantic text is first encoded into a numerical vector, then processed by an MLP projection, a ReLU activation, and a final linear mapping to obtain the semantic embedding. The resulting semantic embeddings are,

$$\mathbf{h}_{rule} = \text{Linear} \left(\text{ReLU} \left(\text{Proj} \left(\text{Embed} \left(\mathbf{z}^{rule} \right) \right) \right) \right) \in \mathbb{R}^D, \quad (8)$$

$$\mathbf{h}_{llm} = \text{Linear} \left(\text{ReLU} \left(\text{Proj} \left(\text{Embed} \left(\mathbf{z}^{llm} \right) \right) \right) \right) \in \mathbb{R}^D. \quad (9)$$

Here, $\text{Embed}(\cdot)$ converts the structured semantic representation into a numerical vector, and $\text{Proj}(\cdot)$ denotes the MLP projection with intermediate dimension $d_{proj} = 64$. The final linear mapping produces semantic embeddings of dimension D , matching the temporal embedding $\mathbf{h}_{ts}$ for subsequent gated aggregation.

3.3.3 Adaptive Gated Aggregation. SERA integrates the temporal and semantic embeddings through adaptive gated aggregation. The three embeddings are adaptively weighted and aggregated as,

$$\mathbf{h}_{agg} = \mathbf{g}_{ts} \odot \mathbf{h}_{ts} + \mathbf{g}_{rule} \odot \mathbf{h}_{rule} + \mathbf{g}_{llm} \odot \mathbf{h}_{llm}, \quad (10)$$

where $\mathbf{g}_k \in \mathbb{R}^D$ is the input-dependent gating vector for embedding $\mathbf{h}_k$, with $k \in \{ts, rule, llm\}$. Each gating vector is computed as $\mathbf{g}_k = \sigma(\mathbf{W}_k \mathbf{h}_k)$, where $\mathbf{W}_k \in \mathbb{R}^{D \times D}$ contains learned parameters and $\sigma(\cdot)$ denotes the sigmoid function. The operator $\odot$ denotes

element-wise multiplication. Each gate assigns a factor between zero and one to each embedding dimension, allowing the contributions of the three representations to adapt to individual input windows. Factors approaching zero suppress the corresponding dimensions, and factors approaching one retain them. The aggregated embedding $\mathbf{h}_{agg}$ is subsequently used for SoH prediction.

3.4 SoH Forecasting and Model Training

The aggregated embedding $\mathbf{h}_{agg}$ is passed to an MLP prediction head to generate the future SoH trajectory over H cycles as,

$$\hat{\mathbf{y}}_{t+1:t+H} = \mathbf{W}_{out} \text{ReLU}(\mathbf{W}_{hid} \mathbf{h}_{agg} + \mathbf{b}_{hid}) + \mathbf{b}_{out}, \quad (11)$$

where $\mathbf{W}_{hid}$, $\mathbf{W}_{out}$ and their corresponding biases are learnable parameters, and $\hat{\mathbf{y}}_{t+1:t+H} \in \mathbb{R}^H$. The model is trained using a composite objective that combines prediction accuracy with constraints on the temporal behaviour of the predicted trajectory,

$$\mathcal{L} = \mathcal{L}_{base} + \lambda_{mono} \mathcal{L}_{mono} + \lambda_{smooth} \mathcal{L}_{smooth}, \quad (12)$$

where λ_{mono} and λ_{smooth} control the contributions of the two penalty terms. The base term uses a horizon-weighted SmoothL1 (Huber) loss that assigns larger weight to predictions farther into the future, reflecting the increasing difficulty of longer-horizon forecasting.

$$\mathcal{L}_{base} = \frac{1}{BH} \sum_{i=1}^B \sum_{j=1}^H w_j \text{SmoothL1}_{\beta}(\hat{y}_{i,j}, y_{i,j}), \quad w_j = \left(\frac{j}{H} \right)^{\alpha}, \quad (13)$$

where B is the batch size, β is the transition threshold of the SmoothL1 loss, and $\alpha > 0$ controls the weighting across the prediction horizon. Larger weights are assigned to later prediction steps, which encourages the model to maintain forecasting accuracy as the prediction distance increases. Since battery SoH generally declines with aging, a monotonicity penalty discourages increases between consecutive predictions, and we have,

$$\mathcal{L}_{mono} = \frac{1}{B(H-1)} \sum_{i=1}^B \sum_{j=1}^{H-1} \max(0, \hat{y}_{i,j+1} - \hat{y}_{i,j}). \quad (14)$$

Electrochemical aging accumulates slowly over multiple cycles, so that battery SoH typically changes gradually. A smoothness penalty

discourages large changes between consecutive predictions with,

$$\mathcal{L}_{\text{smooth}} = \frac{1}{B(H-1)} \sum_{i=1}^B \sum_{j=1}^{H-1} |\hat{y}_{i,j+1} - \hat{y}_{i,j}|. \quad (15)$$

Together, these loss terms encourage accurate forecasts with monotonically and gradually decreasing SoH trajectories.

4 Experiments

This section evaluates the forecasting performance of SERA. We first describe the experimental setup, followed by the main results and further analyses.

4.1 Experimental Setup

We describe the dataset and sample construction, implementation settings of SERA, and comparison variants.

4.1.1 Dataset and Forecasting Sample Construction. We evaluate SERA on the MIT Battery Degradation Dataset [14], which contains 124 commercial LFP/graphite cells cycled to end-of-life under fast charging conditions at 30°C. The tested cells are A123 Systems APR18650M1A cells with a nominal capacity of 1.1 Ah and a nominal voltage of 3.3 V. The dataset provides cycle-level measurements of capacity, voltage, current, temperature, and internal resistance throughout the battery degradation process.

The batteries are divided by cell identity into 70%, 15%, and 15% for training, validation, and testing, respectively, ensuring that no cell appears in more than one set. Forecasting samples are generated from each battery trajectory using sliding windows. For a battery trajectory of length T , a valid sample requires both an L -cycle observation window and its subsequent H -cycle forecasting target. The i -th sample is constructed as,

$$\mathcal{W}_i = (\mathbf{X}_{i:i+L-1}, \mathbf{y}_{i+L:i+L+H-1}), \quad (16)$$

where $i = 1 + kw$ for nonnegative integers k satisfies $i \leq T - L - H + 1$, and w is the sliding stride. We set $w = 1$, advancing the window by one cycle and producing $N = T - L - H + 1$ samples for each battery with $T \geq L + H$. For example, with $L = 30$ and $H = 90$, the first sample uses cycles 1–30 to predict cycles 31–120.

We fix the look-back length at $L = 30$ cycles and evaluate prediction horizons $H \in \{30, 50, 70, 90\}$, keeping the observed history unchanged. The relative forecasting distance is characterized by the forecast-to-observation ratio,

$$\rho = \frac{H}{L}. \quad (17)$$

The four forecasting settings correspond to $\rho \in \{1.0, 1.67, 2.33, 3.0\}$. Increasing ρ extends the forecast farther beyond the observation window, allowing us to assess forecasting robustness across prediction distances under a fixed history length.

4.1.2 SERA Configurations. For the temporal branch, PatchTST is used with patch length $P = 5$ and representation dimension $D = 128$. The Transformer module of PatchTST L_{enc} contains three encoder layers with four attention heads. For the semantic branches, the rule- and LLM-based representations use separate encoders with the same schema, where the intermediate projection dimension d_{proj} is set to 64. The resulting semantic embeddings are mapped to the same dimension D before adaptive gated aggregation. All

Table 2: Comparison of forecasting performance in terms of test MAE and RMSE across different prediction horizons. Lower values indicate better performance, with the best results highlighted in bold. All values are $\times 10^{-3}$.

Model	$H = 30$		$H = 50$		$H = 70$		$H = 90$	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
TS	1.556	2.309	2.317	2.942	3.043	3.818	3.565	4.727
Rule only	11.384	12.904	12.064	13.578	10.962	12.174	11.641	13.172
LLM only	12.157	14.225	12.806	14.755	11.526	13.296	12.150	14.515
TS + Rule	0.912	1.369	1.883	2.463	2.350	3.135	2.593	3.349
TS + LLM	1.408	2.036	2.108	2.744	2.522	3.202	2.946	3.869
SERA (ours)	1.015	1.323	1.688	2.122	2.142	2.851	2.234	2.971

models are trained using Adam with a learning rate of 10^{-3} and a batch size of 32. Training is performed for up to 80 epochs with gradient clipping at 1.0 and early stopping based on the validation loss with a patience of 10 epochs. For the training objective, the horizon weighting parameter is set to $\alpha = 1.0$, while $\lambda_{\text{mono}} = 0.1$ and $\lambda_{\text{smooth}} = 0.05$. The SmoothL1 threshold β is set to 0.01. The same training configuration is used across all prediction horizons for consistent evaluation.

4.1.3 Comparison Models. We compare SERA with several comparison models to evaluate the contribution of different semantic sources. TS uses only the time series or temporal representation and serves as the forecasting baseline. TS + Rule augments the temporal representation with rule-based semantics, while TS + LLM incorporates LLM-based semantics. SERA represents the complete model, which aggregates both rule- and LLM-based semantic representations with the temporal representation. We also include Rule only and LLM only to examine whether semantic representations can support forecasting without temporal information. All variants use the same input features, data partitions, and training settings for consistent comparison.

To examine whether the proposed semantic representation and aggregation approach depends on a particular temporal model, we further evaluate SERA with PatchTST, a standard Transformer (TFM) [16], and LSTM as the temporal encoders. Forecasting performance is measured using MAE and root mean squared error (RMSE). A single NVIDIA RTX 4090 GPU is used.

4.2 Forecasting Performance Analysis

We report the experimental results and evaluate the forecasting performance of SERA under the following settings.

4.2.1 Overall Forecasting Performance. We first evaluate whether incorporating structured degradation semantics improves SoH forecasting beyond the temporal representation alone. Table 2 presents the results of SERA, the temporal baseline, and different semantic configurations across four prediction horizons, using PatchTST as the temporal encoder. Overall, SERA achieves the lowest RMSE at all tested horizons and the lowest MAE for long horizons with $H \in \{50, 70, 90\}$. Compared with baseline TS, SERA reduces MAE from 1.6 to 1.0, 2.3 to 1.7, 3.0 to 2.1, and 3.6 to 2.2 (all $\times 10^{-3}$) at $H = 30, 50, 70$, and 90, respectively. These results indicate that

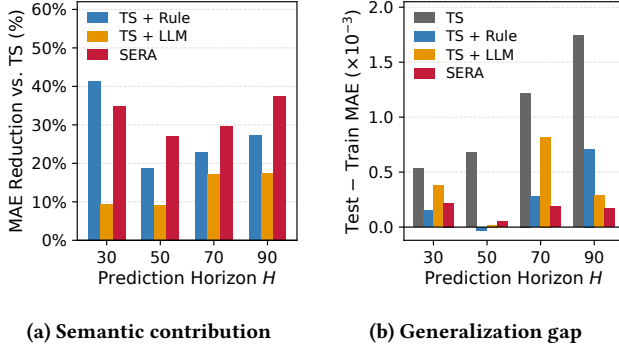


Figure 2: Semantic contribution and generalization analysis across prediction horizons. Fig. 2a: reduction in test MAE relative to the TS baseline. Fig. 2b: generalization gap, measured as the difference between test and training MAE.

structured degradation semantics complement the temporal representation with explicit text-based degradation information, helping the model better capture the evolving battery state and produce more accurate SoH predictions. An exception occurs at $H = 30$, where TS + Rule achieves the lowest MAE of 0.9×10^{-3} , compared with 1.0×10^{-3} for SERA. One possible explanation is that the temporal and rule-based representations already capture the dominant degradation information needed for shorter prediction horizon, while the additional LLM representation affects the efficiency of the forecasting module. As the prediction horizon increases, the comprehensive information provided by both semantic sources becomes more useful, and SERA achieves the best MAE and RMSE.

The results also clarify the role of semantic information. Both TS + Rule and TS + LLM generally improve upon TS, showing that each semantic source provides useful degradation guidance. In contrast, Rule only and LLM only produce MAE values from 0.011 to 0.014, substantially higher than the models that utilize temporal information. This implies that semantic descriptors alone do not retain sufficient information about the absolute SoH level and data-centric trajectory dynamics. They therefore serve more effectively as complementary guidance to the temporal representation than as standalone predictors.

4.2.2 Semantic Contribution and Generalization. We further examine the contribution of semantic information to forecasting and the model’s generalization to unseen batteries. Fig. 2a reports the percentage reduction in test MAE relative to the TS baseline, calculated as $R = (\text{MAE}_{\text{TS}} - \text{MAE}_{\text{variant}}) / \text{MAE}_{\text{TS}} \times 100\%$. A larger positive value indicates a greater relative improvement over TS. Aligned with the results in Table 2, the contribution of semantic information varies with the prediction horizon. At the short horizon $H = 30$, TS + Rule achieves the largest MAE reduction of 41.4%, compared with 34.8% for SERA, suggesting that rule-based semantics already provide effective guidance for short term forecasting. For the relative longer horizons $H \in \{50, 70, 90\}$, SERA consistently outperforms both individual semantic configurations, and its MAE reduction increases from 27.1% to 29.6% and then to 37.3%. Its advantage over TS + Rule is 8.4%, 6.8%, and 10.1% across these horizons, respectively.

Table 3: Comparison of test MAE across temporal encoders at $H = 30$ and $H = 90$. Lower values are better, and the best results are highlighted in bold. All values are $\times 10^{-3}$.

Model	LSTM		TFM		PatchTST	
	$H = 30$	$H = 90$	$H = 30$	$H = 90$	$H = 30$	$H = 90$
TS	1.394	3.578	1.599	3.421	1.556	3.565
TS + Rule	1.505	3.097	1.375	2.708	0.912	2.593
TS + LLM	1.535	3.336	2.013	3.246	1.408	2.946
SERA (ours)	1.229	3.098	1.350	2.636	1.015	2.234

These results suggest that combining rule- and LLM-based semantics and temporal representation becomes increasingly beneficial as the prediction horizon extends.

Fig. 2b provides a different view through the gap between test and training MAE. A gap near zero suggests generalized performance on training and unseen batteries and a positive gap may indicate overfitting. The TS baseline shows a larger gap as H increases, reaching 1.7×10^{-3} at $H = 90$, whereas SERA maintains a relatively small gap across all horizons. A possible reason is that structured semantics capture degradation characteristics shared across batteries, reducing the model’s dependence on battery specific temporal patterns. The results suggest that semantic information supports both forecasting accuracy and generalization to unseen batteries.

4.2.3 Temporal Encoder Sensitivity. We further examine the impact of temporal encoder choice. Table 3 compares the TS baselines and corresponding comparison models using LSTM, TFM, and PatchTST as the temporal encoder at $H = 30$ and $H = 90$. Although incorporating semantic representations improves forecasting across all three temporal encoders, the gains vary with the temporal encoder and prediction horizon. With LSTM as the temporal encoder, SERA achieves the lowest MAE at $H = 30$ and performs nearly identically to TS + Rule at $H = 90$. With TFM, SERA achieves the lowest MAE at both horizons. A similar benefit is observed with PatchTST, where SERA reduces the MAE from 3.6×10^{-3} for TS to 2.2×10^{-3} at $H = 90$. These results suggest that the benefit of semantic representations is not specific to a particular temporal encoder and that semantic information can be effectively aggregated with different temporal representations. Meanwhile, the overall forecasting performance also varies across temporal encoders, with SERA generally achieving better performance with PatchTST than with LSTM and TFM. The varying gains further indicate that the effectiveness of semantic information depends partly on the temporal representations with which it is aggregated.

4.2.4 Case Study. To complement the aggregate forecasting results, Fig. 3 illustrates the full SoH trajectories of two representative batteries with short and long lifespans. The trajectories provide a detailed view of prediction behavior across degradation stages that is not reflected by aggregate error metrics. In both cases, SERA remains close to the ground truth SoH during gradual degradation and accelerated decline stages. Specifically, for the short-lifespan case in Fig. 3a, the trajectories predicted by different SoH models remain close during the early cycles. However, after approximately

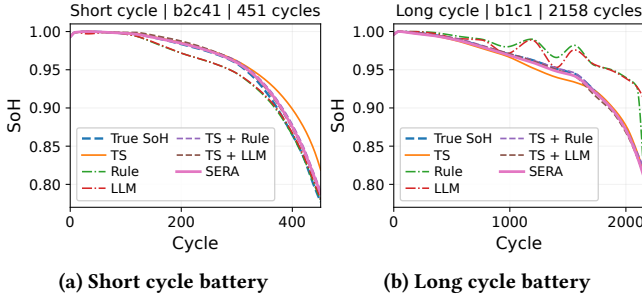


Figure 3: Qualitative comparison of SoH trajectories for representative short- (Fig. 3a) and long-lifespan (Fig. 3b) batteries. The results compare the ground truth trajectories with predictions from the temporal baseline, semantic variants, and models with temporal and semantic representations.

300 cycles, TS gradually deviates from the ground truth and increasingly overestimates SoH, while SERA follows the accelerated decline more closely. For the long-lifespan battery in Fig. 3b, the main difference occurs between approximately 800 and 1,600 cycles, where TS underestimates SoH while SERA better captures the gradual decline. These patterns illustrate how incorporating degradation semantics can help the model capture different degradation behaviors across battery life stages. As also shown in the figures, the Rules and LLM curves fall below the ground truth during the middle stage in Fig. 3a, while exhibiting larger fluctuations and a later transition to accelerated degradation in Fig. 3b. These deviations suggest that semantic descriptors alone do not fully capture the numerical SoH trajectory, supporting their use to guide rather than replace the temporal representation in SERA.

4.3 Interpretability Analysis

Beyond forecasting accuracy, it is important to interpret the influence of structured semantic representations on the predictions of SERA and we perform counterfactual semantic analysis.

4.3.1 Counterfactual Semantic Analysis. Counterfactual analysis examines whether changes in the semantic representation lead to expected changes in the forecast while the observed time series remains unchanged. This allows us to isolate the influence of semantic information from that of the temporal input and assess whether SERA uses the semantic descriptors consistently with their degradation meaning. Given an input window X and its rule- and LLM-based semantic representations z^{rule} and z^{llm} , the original prediction is,

$$\hat{y} = f(X, z^{\text{rule}}, z^{\text{llm}}). \quad (18)$$

We then modify one or multiple rules in both semantic representations while keeping X fixed, producing counterfactual representations $z^{\text{rule}'}$ and $z^{\text{llm}'}$. Applying the same intervention to both representations maintains a consistent hypothetical degradation state. The intervention changes the corresponding prediction to,

$$\hat{y}' = f(X, z^{\text{rule}'}, z^{\text{llm}'}). \quad (19)$$

Here, the intervention effect at prediction step t' is measured by,

$$\Delta_{t'} = \hat{y}'_{t+t'} - \hat{y}_{t+t'}, \quad t' \in \{1, \dots, H\}. \quad (20)$$

A positive value indicates an increase in predicted SoH after the intervention, and a negative value indicates a decrease. For example, changing the value of the Trend rule from stable in z^{rule} and z^{llm} to accelerating_fade in $z^{\text{rule}'}$ and $z^{\text{llm}'}$ is expected to lower predicted SoH, as battery degradation typically accelerates near the battery's end of life with low SoH. If the predicted $\hat{y}'$ with such intervention decreases relative to the predicted $\hat{y}$ without intervention, the response agrees with the expectation based on battery domain knowledge. We call this agreement semantic consistency and use it to interpret the alignment of the model's response with the rule's degradation meaning.

We evaluate 11 interventions covering individual semantic fields, degradation slope, and coherent optimistic or pessimistic combinations. For each prediction horizon, each intervention is evaluated over $n_{\text{cf}} = 200$ randomly sampled counterfactual test windows. We report semantic consistency as the percentage of windows whose predictions change in the expected direction. We also report the mean prediction shift at the final cycle of the horizon as,

$$\bar{\Delta}_H = \frac{1}{n_{\text{cf}}} \sum_{i=1}^{n_{\text{cf}}} \Delta_H^i, \quad (21)$$

where Δ_H^i is the SoH shift for the i -th sampled window. The sign of $\bar{\Delta}_H$ shows the average SoH response direction, while $|\bar{\Delta}_H|$ measures the magnitude of the intervention effect.

4.3.2 Semantic Consistency. Table 4 reports the semantic consistency and final-cycle shift in predicted SoH for each intervention. We first consider the optimistic bundle intervention, which jointly modifies multiple semantic rules toward a healthier degradation state, including a stable trend, low risk, a stable mechanism, and a mild SoH decline. Conversely, the pessimistic bundle intervention modifies these rules toward a worsening degradation state, including an accelerating trend, high risk, Li plating, and a steep SoH decline. These bundled interventions assess whether the resulting prediction changes align with the intended degradation semantics. The optimistic intervention increases predicted SoH and achieves 69.0–90.5% semantic consistency across the prediction horizons, while the pessimistic intervention decreases predicted SoH and achieves 100% consistency. Both interventions generally shift SoH predictions in the expected direction based on domain knowledge, with stronger consistency observed for the pessimistic bundle. One possible explanation is that worsening degradation semantics are more compatible with the degradation trajectory already reflected in the observed time series, whereas imposing healthier rules may partially conflict with this temporal evidence.

To distinguish the contributions of individual rules within the bundles, we next examine interventions that modify one semantic rule at a time. Among the individual interventions, Slope $\rightarrow$ Steep has the largest effect on the predicted SoH and achieves 100% semantic consistency across all horizons. It also yields negative $\bar{\Delta}_H$ values, indicating lower predicted SoH when the slope is intervened toward steeper degradation. The mean SoH decrease grows from 6.9% at $H = 30$ to 25.5% at $H = 90$, indicating a stronger effect at longer prediction horizons as degradation effects accumulate over time. Trend $\rightarrow$ Accelerating also achieves high semantic consistency of 87.0–100.0%, but produces smaller SoH decreases of 0.25–1.23% only.

Table 4: Counterfactual analysis of different semantic interventions. Higher semantic consistency (%) indicates that more predictions change in the expected direction after interventions. Positive (negative) $\bar{\Delta}_H$ values indicate an increase (decrease) in predicted SoH, with larger absolute values indicating stronger effects. $\bar{\Delta}_H$ values are in units of 10^{-3} .

Semantic Edit	H = 30		H = 50		H = 70		H = 90	
	%	$\bar{\Delta}_H$	%	$\bar{\Delta}_H$	%	$\bar{\Delta}_H$	%	$\bar{\Delta}_H$
Optimistic Bundle	90.5	+5.8	69.0	+7.5	86.5	+12.9	70.5	+11.9
Pessimistic Bundle	100.0	-71.3	100.0	-116.8	100.0	-146.1	100.0	-251.3
Slope $\rightarrow$ Mild	75.5	+5.6	84.5	+7.5	82.0	+13.2	71.0	+12.1
Slope $\rightarrow$ Steep	100.0	-68.9	100.0	-129.7	100.0	-142.3	100.0	-255.1
Trend $\rightarrow$ Stable	78.5	+0.9	93.0	+2.2	91.0	+4.2	61.0	+7.7
Trend $\rightarrow$ Accel.	87.0	-2.5	94.5	-3.9	87.5	-5.6	100.0	-12.3
Mech. $\rightarrow$ Stable	86.0	+1.0	64.0	-0.2	44.5	-2.8	26.5	-3.0
Mech. $\rightarrow$ SEI	93.5	-1.1	32.0	+0.0	22.5	+0.6	83.5	-2.5
Mech. $\rightarrow$ Li Plating	99.5	-3.8	51.0	+1.6	99.5	-9.2	99.0	-13.4
Risk $\rightarrow$ Low	50.5	-1.3	39.5	-1.4	85.5	+4.4	66.5	+3.1
Risk $\rightarrow$ High	19.0	+2.6	34.5	-0.1	58.5	+0.2	37.0	-0.3

These results indicate that, under the tested settings, the quantitative slope intervention has a stronger influence on the prediction than the categorical trend intervention. We also test mechanism interventions. For example, Mechanism $\rightarrow$ Li Plating produces negative $\bar{\Delta}_H$ values and high semantic consistency in most cases, as Li plating represents a detrimental degradation mechanism associated with accelerated capacity loss. For risk interventions, Risk $\rightarrow$ High achieves only 19.0–58.5% semantic consistency and produces relatively small $\bar{\Delta}_H$ values, suggesting that its influence may depend more strongly on the battery state encoded in the temporal representation. Overall, the counterfactual analysis demonstrates that the learned semantic information meaningfully influences the SoH forecast and provides an interpretable characterization of the effects of different degradation semantics on the predictions.

5 Conclusion

This paper presents SERA, a framework for battery SoH forecasting that integrates structured text-based degradation semantics derived from battery domain knowledge with numerical battery time series. SERA employs a three-branch architecture that independently learns a temporal representation and rule- and LLM-based semantic representations, which are fused through adaptive gated aggregation for SoH forecasting. Compared with the temporal baseline, SERA reduces SoH forecasting errors by up to 37.3% and improves generalization to unseen batteries. SERA also achieves consistent forecasting gains across different temporal encoders. Counterfactual semantic analysis provides insight into the influence of degradation semantics by examining the direction and magnitude of prediction changes under controlled interventions. The results show more consistent responses to semantic rules that directly characterize degradation progression. Overall, SERA improves SoH forecasting accuracy and interpretability by incorporating explicit degradation knowledge and semantics into temporal modelling effectively.

Acknowledgments

This research is supported by A*STAR under its MTC Individual Research Grants (IRG) (Award M23M6c0113), the Shanghai Sci-tech

Co-research Program (Award 25HB2702600), and MTC Programmatic (Award M23L9b0052 and Award M24N6b0043).

References

- [1] Xingguang Chen, Tao Sun, Xin Lai, Yuejiu Zheng, and Xuebing Han. 2024. Transfer learning strategies for lithium-ion battery capacity estimation under domain shift differences. *Journal of Energy Storage* 90 (2024), 111860.
- [2] Samuel Filgueira da Silva, Mehmet Fatih Ozkan, Faissal El Idriissi, and Marcello Canova. 2026. Conformalized Transfer Learning for Li-ion Battery State of Health Forecasting under Manufacturing and Usage Variability. *arXiv preprint arXiv:2603.24475* (2026).
- [3] Chenglong Duan, Hung Le, and Dazhong Wu. 2025. Lithium-ion battery state-of-health estimation using intra-domain and cross-domain transfer learning: mitigating domain shift based on Wasserstein distance. *Journal of Energy Storage* 132 (2025), 117601.
- [4] Xinyu Gu, Khay Wai See, Penghua Li, Kangheng Shan, Yunpeng Wang, Liang Zhao, Kai Chin Lim, and Neng Zhang. 2023. A novel state-of-health estimation for the lithium-ion battery using a convolutional neural network and transformer model. *Energy* 262 (2023), 125501.
- [5] MM Kabir and Dervis Emre Demirocak. 2017. Degradation mechanisms in Li-ion batteries: a state-of-the-art review. *International Journal of Energy Research* 41, 14 (2017), 1963–1986.
- [6] Dhivya Dharshini Kannan, Wei Li, Wei Zhang, Jianbiao Wang, Zhi Wei Seh, and Man-Fai Ng. 2026. When smaller wins: Dual-stage distillation and pareto-guided compression of liquid neural networks for edge battery prognostics. In *International Conference on Pattern Recognition*. Springer, 414–429.
- [7] Wei Li, Wei Zhang, Yi Xie, Yonggang Wen, and Qingyu Yan. 2026. Operando Thermal Reconstruction for Lithium-Ion Batteries via Physics-Informed Generative Diffusion Models. *IEEE Transaction on Transportation Electrification (TTE)* 12, 1 (2026), 1869–1879.
- [8] Tianfeng Long, Pengcheng Zhang, Xiaoqi Liu, Huaqing Shang, Meiling Yue, Xuesong Shen, and Jianwen Meng. 2025. Improved SOH prediction of lithium-ion batteries based on multi-dimensional feature analysis and transformer framework. *IEEE Open Journal of Vehicular Technology* (2025).
- [9] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).
- [10] Tsuyoshi Oji, Yanglin Zhou, Song Ci, Feiyu Kang, Xi Chen, and Xiulan Liu. 2021. Data-driven methods for battery soh estimation: Survey and a critical analysis. *IEEE Access* 9 (2021), 126903–126916.
- [11] Yuan Qiu, Wei Li, Wei Zhang, Yi Zhou, Fang Liu, Jianbiao Wang, and Zhi Wei Seh. 2026. Interpretable Battery Aging without Extra Tests via Neural-Assisted Physics-based Modelling. *arXiv preprint arXiv:2604.01229* (2026).
- [12] Jiantao Qu, Feng Liu, Yuxiang Ma, and Jiaming Fan. 2019. A neural-network-based method for RUL prediction and SOH monitoring of lithium-ion battery. *IEEE access* 7 (2019), 87178–87191.
- [13] Concetta Semeraro, Mariateresa Caggiano, Abdul-Ghani Olabi, and Michele Dassiti. 2022. Battery monitoring and prognostics optimization techniques: Challenges and opportunities. *Energy* 255 (2022), 124538.
- [14] Kristen A Severson, Peter M Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H Chen, Muratahan Aykol, Patrick K Herring, Dimitrios Fraggedakis, et al. 2019. Data-driven prediction of battery cycle life before capacity degradation. *Nature energy* 4, 5 (2019), 383–391.
- [15] Wen Yang Tan, Jiawei Li, Fang Liu, Wei Zhang, Sumei Sun, Peng Cheng Wang, and Elisa YM Ang. 2026. TIDE: Trustworthy and Interpretable Battery Degradation Estimation with Contextual Learning and Symbolic Distillation. *arXiv preprint arXiv:2607.14640* (2026).
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [17] Jens Vetter, Petr Novák, Markus Robert Wagner, Claudia Veit, K-C Möller, JO Besenhard, Martin Winter, Margret Wohlfahrt-Mehrens, Christoph Vogler, and Abderrezak Hammouche. 2005. Ageing mechanisms in lithium-ion batteries. *Journal of power sources* 147, 1-2 (2005), 269–281.
- [18] Rui Xiong, Linlin Li, and Jinpeng Tian. 2018. Towards a smarter battery management system: A critical review on battery state of health monitoring methods. *Journal of Power Sources* 405 (2018), 18–29.
- [19] Bo Yang, Yucun Qian, Qiang Li, Qian Chen, Jiyang Wu, Enbo Luo, Rui Xie, Ruyi Zheng, Yunfeng Yan, Shi Su, et al. 2024. Critical summary and perspectives on state-of-health of lithium-ion battery. *Renewable and Sustainable Energy Reviews* 190 (2024), 114077.
- [20] Zhuang Ye and Jianbo Yu. 2021. State-of-health estimation for lithium-ion batteries using domain adversarial transfer learning. *IEEE Transactions on Power Electronics* 37, 3 (2021), 3528–3543.