Supplementary Information

Real-time critical transition discoveries with large language models

Guijun Ma[1,2], Zidong Wang[3], Yuzhe Wang[1], Yong Zhang[4], Haitao Song[5], Han Ding[2,6] & Ye Yuan[1,2,*]

[1]*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China*

[2]*State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*

[3]*Department of Computer Science, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom*

[4]*School of Artificial Intelligence and Automation, Wuhan University of Science and Technology, Wuhan 430081, China*

[5]*Shanghai Artificial Intelligence Research Institute, Shanghai Jiao Tong University, Shanghai 200030, China*

[6]*School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*

1. Complex System Datasets

1.1 HUST-LIB dataset. All HUST-LIB cells were cycled on an 80-channel Neware CT-4008 tester at 30 ℃. Cells underwent a uniform fast-charging protocol but various multi-stage discharging protocols (see Supplementary Information Figure S8). The fast-charging process consisted of a 5C constant-current (CC) charge from 0-80% state of charge (SOC), a 1C CC charge from 80% SOC to 3.6 V and a constant-voltage (CV) hold to 100% SOC with a cutoff of C/20. Discharge comprised four CC steps: 100-60% SOC, 60-40% SOC, 40-20% SOC and 20-0% SOC, with the final stage discharged at 2C to 2 V. For example, a 5C-4C-3C-2C protocol denotes sequential CC discharges of 5C, 4C, 3C and 2C across these intervals. Rest periods of 30 s were imposed between any two stages. Voltage, current and capacity were continuously recorded until capacity first declined to knee point. Of 80 tested cells, 4 were excluded due to abrupt faults or the absence of a pre-defined knee point. In our experiment, 54 cells are used to fine-tune the CT-eProber, and 22 cells serve as test cases for performance evaluation.

1.2 MIT&Stanford-LIB dataset. The MIT&Stanford-LIB dataset contains 124 cells cycled under 72 distinct fast-charging protocols and a uniform discharge protocol at 30 ℃. Charging involved one- or two-step CC profiles to 80% SOC, followed by a 1C CC charge to 3.6 V and a CV step with a cutoff of C/50. Discharge was performed at 4C CC to 2 V. Cycle lives of cells ranged from 150 to 2,300 cycles. In our experiment, 106 cells with well-defined knee points were selected to evaluate the performance of CT-eProber (see Supplementary Information Figure S7). Of these, 86 cells were used to fine-tune the model, and the learned representations were applied to quantitatively predict the knee points of 20 test cells.

1.3 Systemic financial crisis dataset. The systemic financial crisis dataset was derived from the Jordà-Schularick-Taylor Macrohistory Database, one of the most comprehensive long-run resources on advanced economies. It spanned 18 countries (i.e., Australia, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, the United Kingdom, Italy, Japan, the Netherlands, Norway, Portugal, Sweden, the United States and Ireland) covering the period 1870-2016. The database consolidated historical series from archival sources, statistical offices, central banks and prior data-collection initiatives, yielding a unified panel that represents over 90% of advanced-economy output and more than half of global output. In this dataset, the dependent variable was a systemic financial crisis indicator, coded as 1 in the year a systemic financial crisis occurs and 0 otherwise. Across the samples, 88 systemic financial crises and 2,630 normal conditions were recorded. From the 39 available macro-financial variables, we selected five explanatory variables that capture distinct dimensions of economic and financial dynamics: (1)

annual growth in loans to the non-financial private sector relative to GDP; (2) annual growth in real stock prices (see Supplementary Information Note S1); (3) annual growth in real house prices; (4) the current account-to-GDP ratio; and (5) annual growth in real GDP. These indicators have been widely recognized as effective early-warning signals, with credit growth, stock prices and housing markets repeatedly identified as the most reliable predictors of systemic banking crises.

1.4 Robotic accuracy failure dataset. The robotic accuracy failure dataset originated from the National Institute of Standards and Technology (NIST) and was collected from a six-axis Universal Robots UR5 collaborative arm. The dataset comprised 18 groups of experiments, covering six operating conditions that vary in motion speed, payload and starting mode, each repeated three times. Robotic accuracy was quantified by comparing the expected Cartesian coordinates (computed from the commanded joint angles via the kinematic model) with the sensor-measured coordinates of the tool center point. The Euclidean distance between the two coordinates was the true positional error (see Supplementary Information Figure S9). For model inputs, we selected six-dimensional joint current signals, which were highly sensitive to mechanical wear and lubrication failure. Each input sample was constructed as a sequence of 100 discrete time steps (with each step spanning 0.008 s). Labels were defined according to a 4 mm error threshold, a critical tolerance in robotic applications: if the positional error exceeded 4 mm at least once within the subsequent K seconds, the sample was labeled as 1 (failure); otherwise, it was labeled as 0 (normal condition). Of the 18 groups, 15 were used for training, while the remaining 3 groups under identical operating conditions were reserved for testing, ensuring a robust evaluation of model generalizability.

2. Supplemental Figures

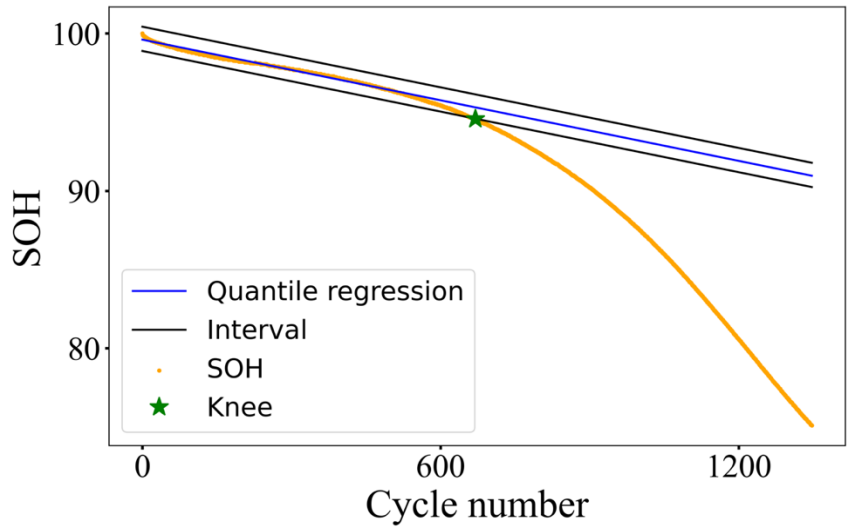

Figure S1 Definition of the knee point in a lithium-ion battery. The ground-truth knee point is determined using quantile regression on the historical state-of-health (SOH) trajectory. A statistically robust safe degradation zone is first established. When five consecutive residuals fall outside this zone, the last of these cycles—and its corresponding SOH value—is identified as the knee point.

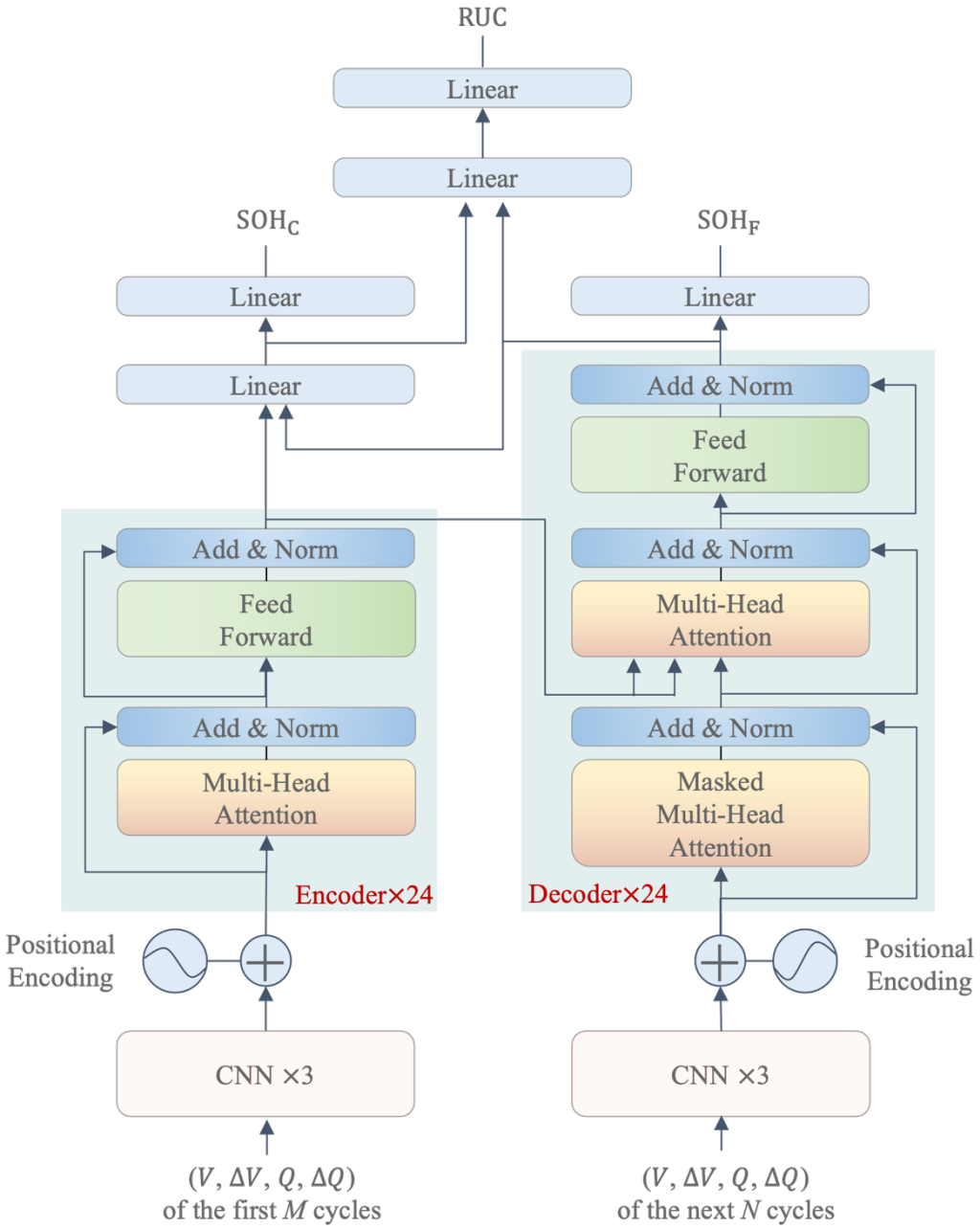

Figure S2 Model architecture of CT-eProber for quantitative early warning of knee points in lithium-ion batteries. Four partial charge curves (V, ΔV, Q, and ΔQ) from the most recent M+N cycles are processed as the prompt data. Data from the most recent N cycles are fed into the decoder, while earlier M cycles are input into the encoder. The primary training label is the number of remaining useful cycles to the knee point (denoted as RUC). Two auxiliary targets—the SOH at the current cycle (denoted as SOH_C) and over the next N cycles (denoted as SOH_F)—are incorporated to enhance model tuning as well as predictive performance.

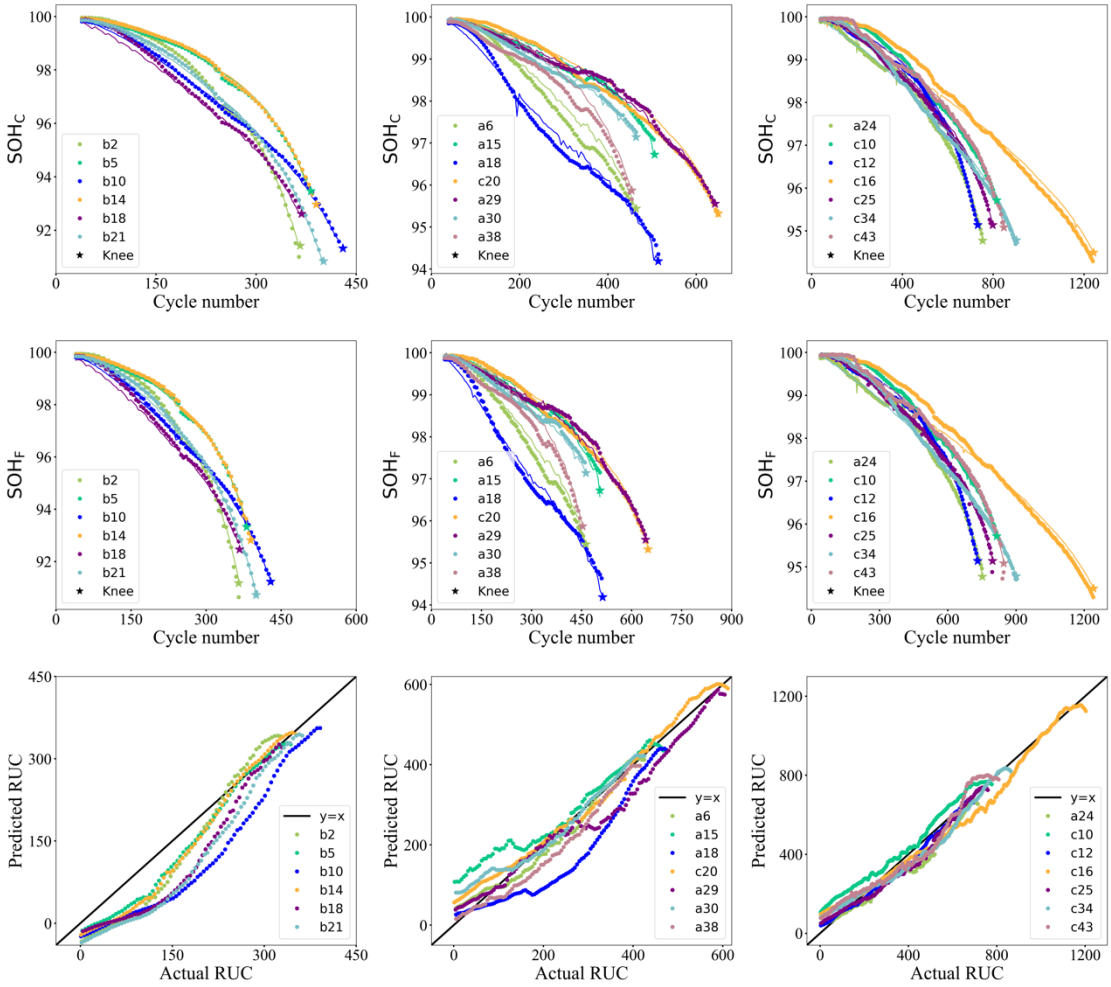

Figure S3 Quantitative early-warning performance of CT-eProber on the MIT&Stanford-LIB dataset [1]. Results for intermediate states: a, SOH estimation at the current cycle—SOH_C versus cycle number, and b, SOH prediction over next 10 cycles—SOH_F versus cycle number. c, Cycle-by-cycle prediction of remaining useful cycles to the knee point (RUC): predicted versus actual values for 20 test cells.

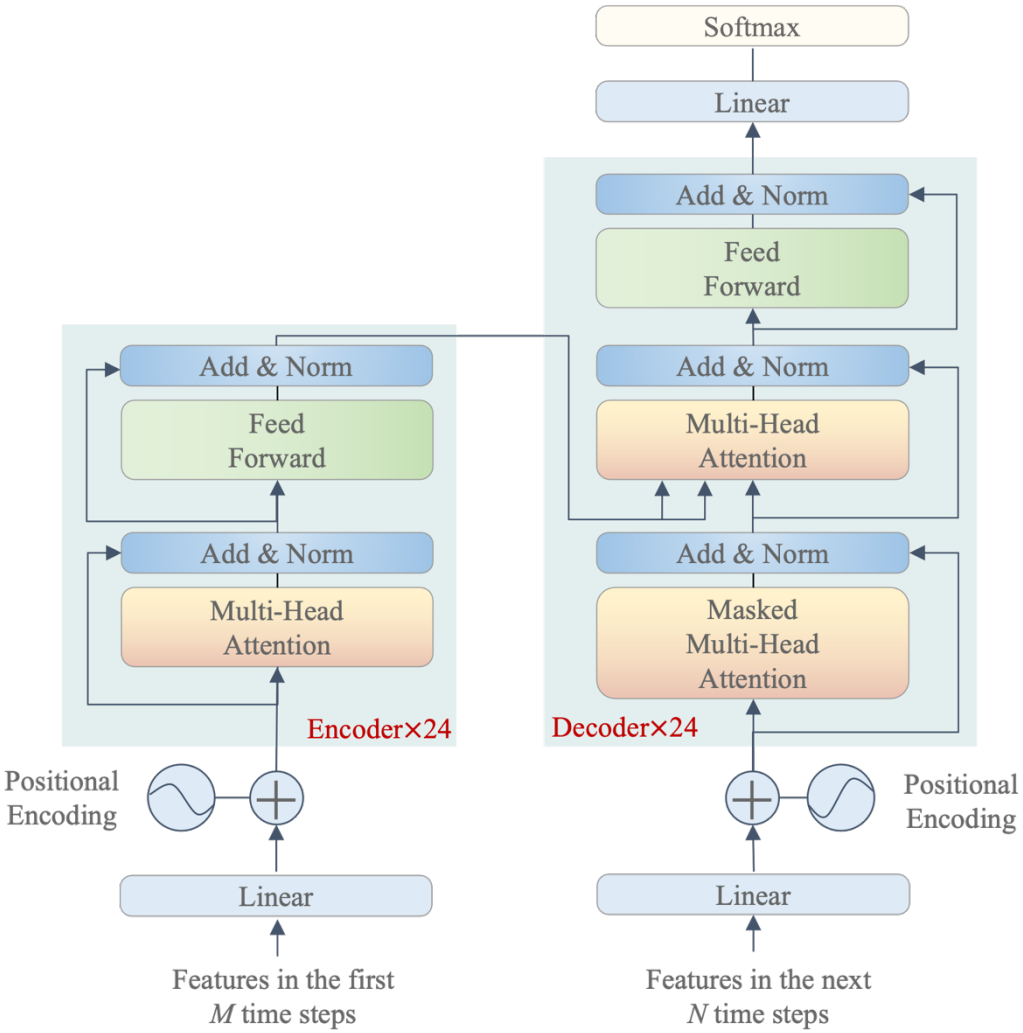

Figure S4 Model architecture of CT-eProber for qualitative early warning. Discrete features from the most recent $M+N$ time steps are processed as the prompt data. Data from the most recent N time steps are fed into the decoder, while earlier M time steps are input into the encoder. The training label is binary, with 0 denoting a normal condition and 1 indicating the occurrence of a critical transition in one specified prediction period.

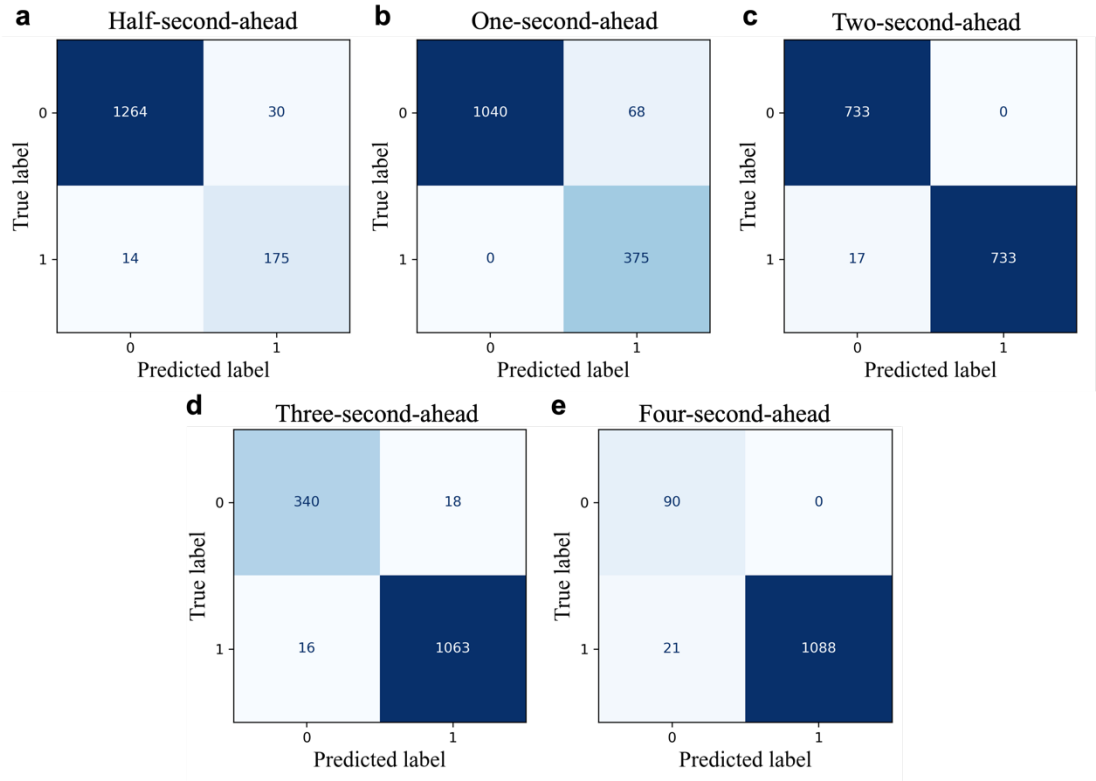

Figure S5 Confusion matrices for robotic accuracy failure prediction of five early warning tasks using CT-eProber: a, half-second-ahead; b, one-second-ahead; c, two-second-ahead; d, three-second-ahead; and e, four-second-ahead.

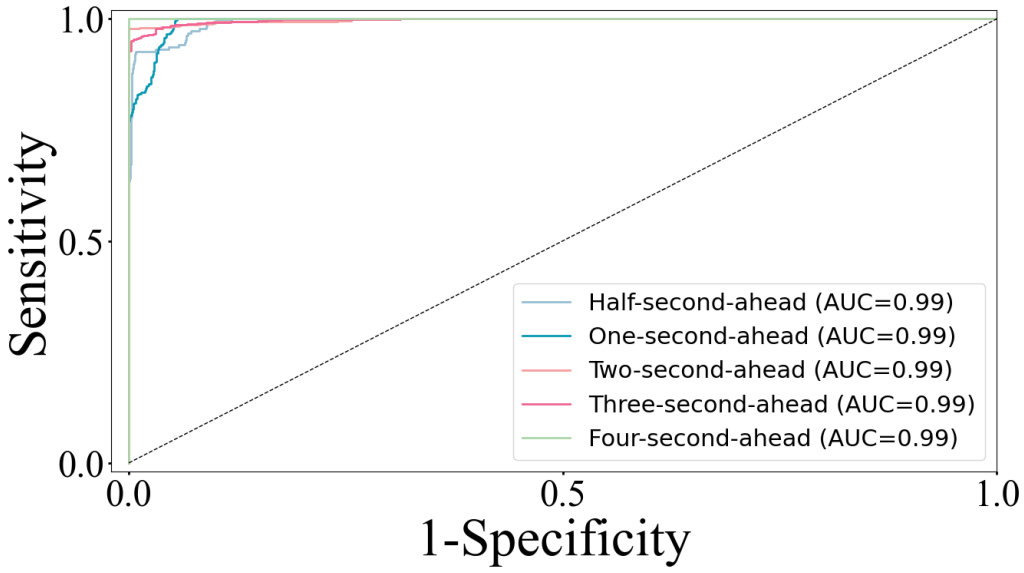

Figure S6 Receiver operating characteristic (ROC) curves for five early-warning tasks in robotic accuracy failure dataset [2]. The area-under-the-curve (AUC) values of the five tasks are provided in detail.

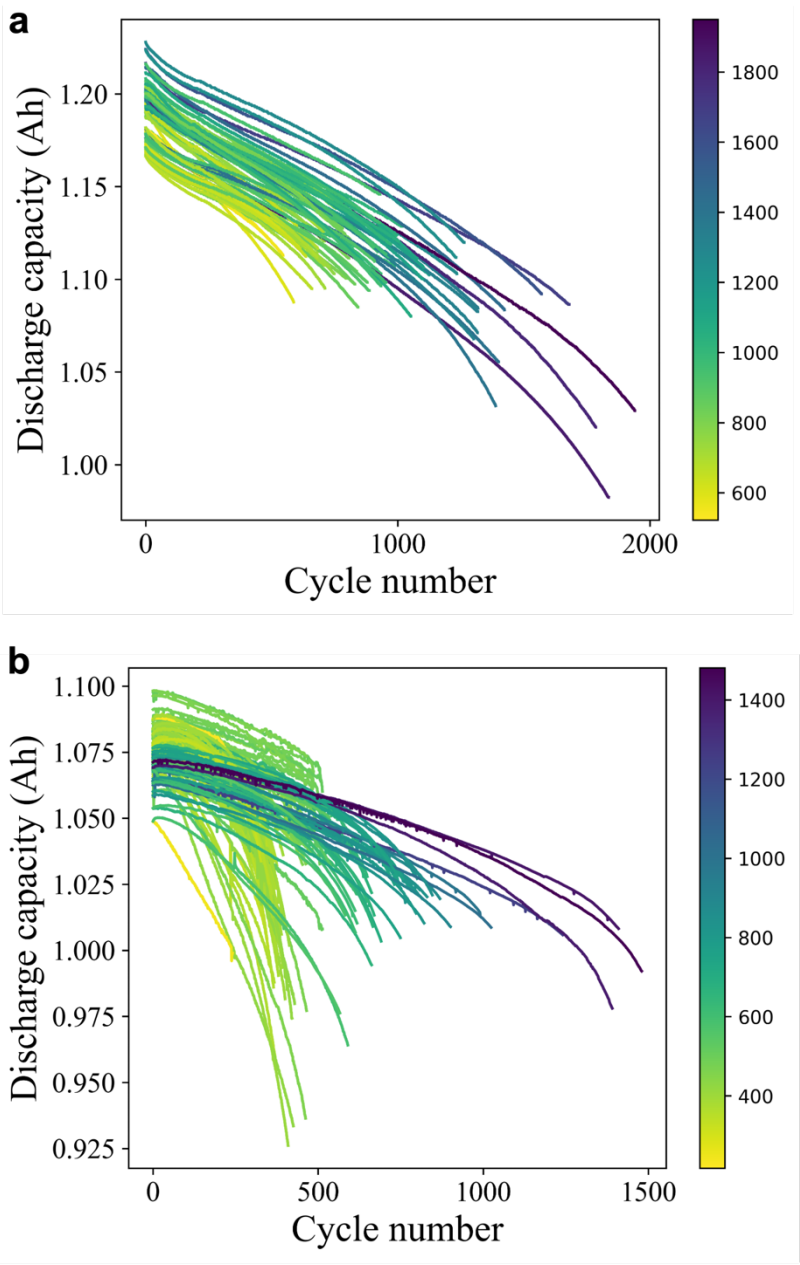

Figure S7 Discharge capacity curves until the occurrence of knee points. a, HUST-LIB dataset. b, MIT&Stanford-LIB dataset. The colour scale denotes the cycle number at which the knee point occurs.

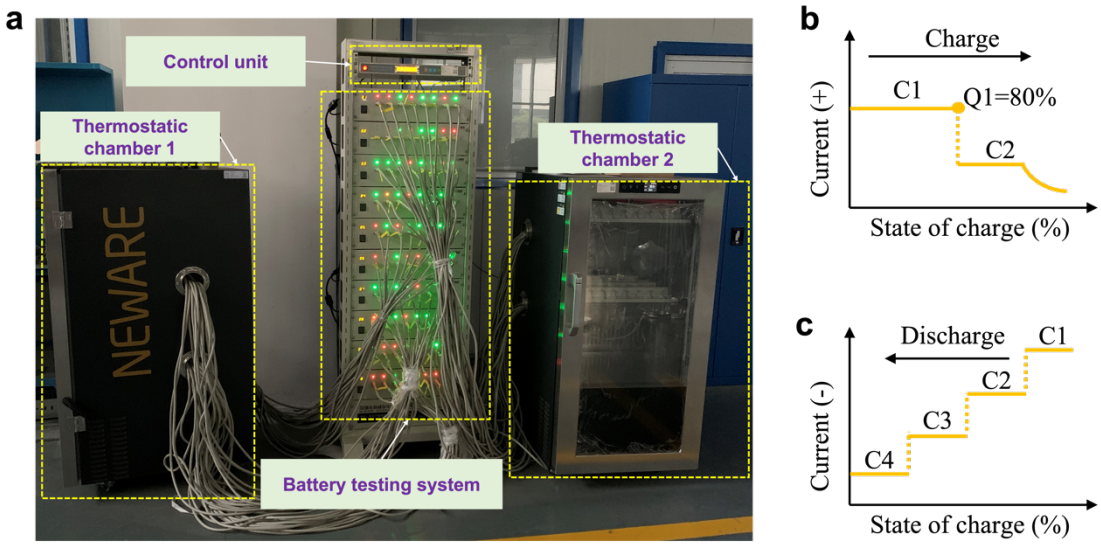

Figure S8 Battery testing configurations of the HUST-LIB dataset [3]. a, Experimental platform comprises a control unit, an 80-channel battery testing system, two thermostatic chambers, and 76 lithium-iron-phosphate/graphite A123 APR18650M1A cells exhibiting pronounced knee points. b, A uniform fast-charging protocol applied to all cells: C1 (5C, 0-80% SOC) → C2 (1C, 80% SOC-3.6 V), followed by a constant-voltage step at 3.6 V until the current tapers to C/20. c, Four-stage discharging protocol from 100% SOC to a cutoff voltage of 2 V. Cells were subjected to diverse discharge profiles: C1 (100-60% SOC) → C2 (60-40% SOC) → C3 (40-20% SOC) → C4 (20% SOC-2 V), where C1-C4 denote the constant-current discharge rates applied in each stage.

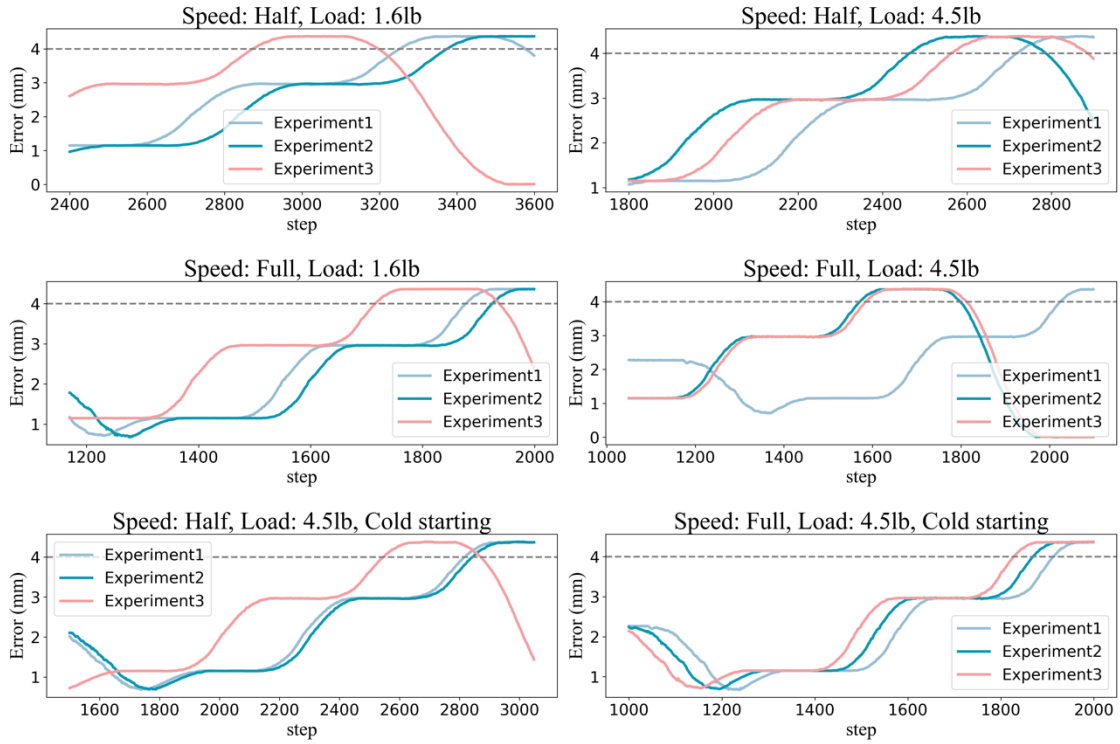

Figure S9 Tool-center-point errors of a six-axis Universal Robots UR5 collaborative arm. The dataset [2] comprises 18 experimental groups under six operating conditions—varying in motion speed, payload, and starting mode—with each condition repeated three times. A 4-mm error threshold is applied to distinguish robotic accuracy failure from normal operation.

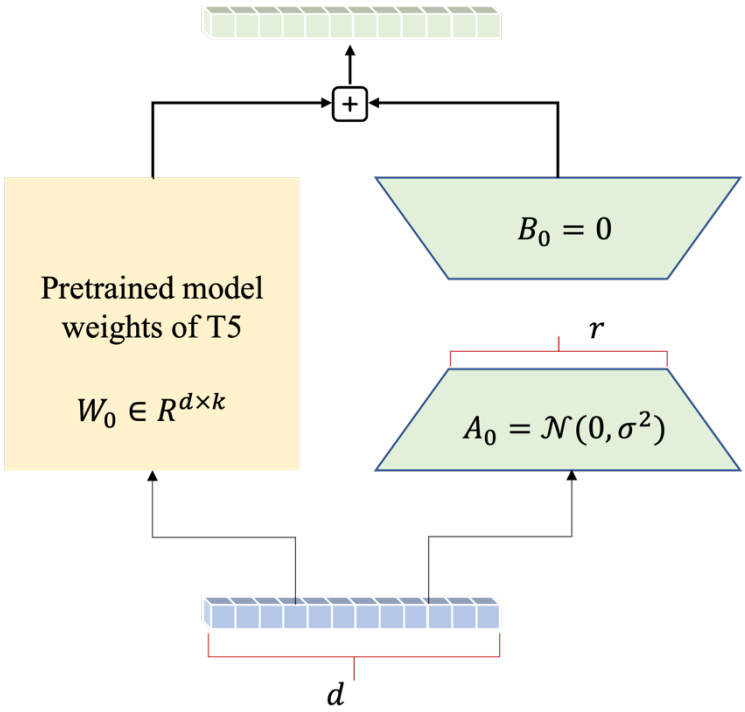

Figure S10 Schematic of low-rank adaptation (LoRA) [4]. LoRA freezes the pretrained weights of the T5 model and injects trainable rank-decomposed matrices (A, B) into its layers, with A initialized from a Gaussian distribution and B initialized to zero. This strategy substantially reduces the number of trainable parameters while maintaining model accuracy.

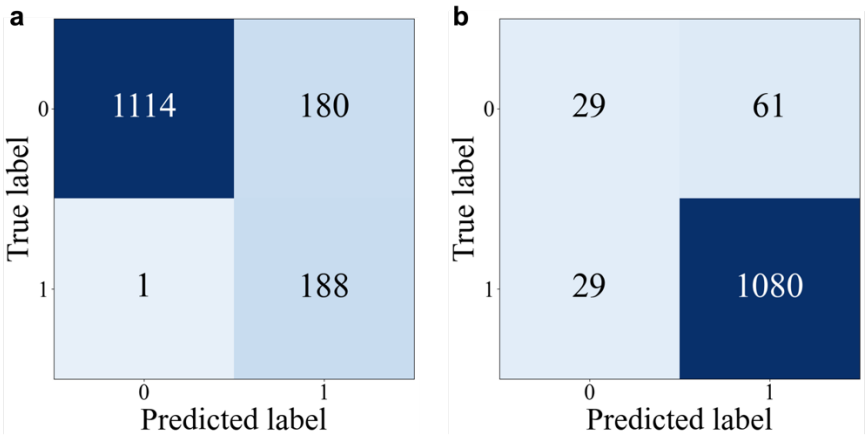

Figure S11 Confusion matrices for robotic accuracy failure prediction of five early warning tasks using non-pretrained CT-eProber: a, half-second-ahead; b, four-second-ahead.

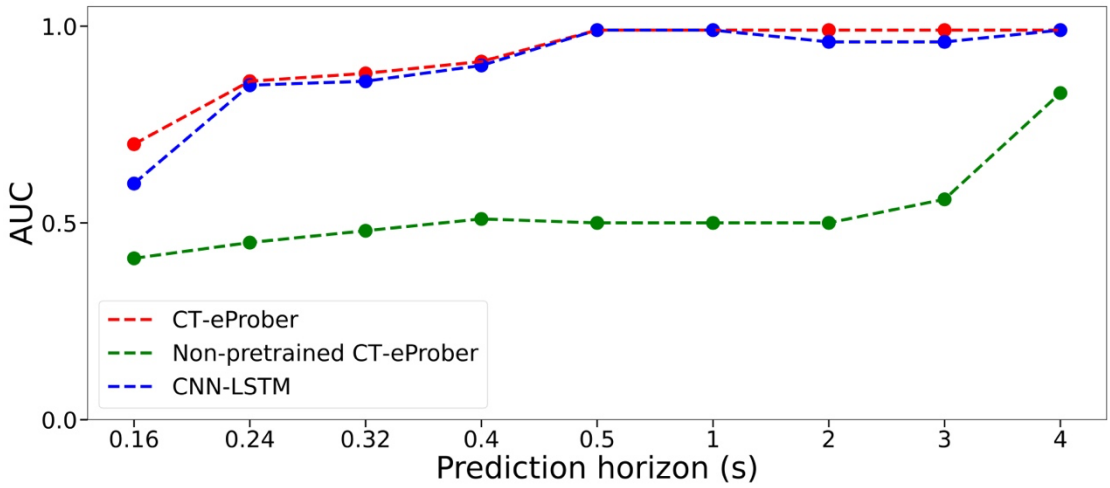

Figure S12 AUC versus prediction horizon for robotic accuracy failure prediction, comparing the CT-eProber, Non-pretrained CT-eProber, and CNN-LSTM models.

3. Supplemental Tables

Table S1 Configuration summary of the CT-eProber embedder for time-series sensor signals. k, s and p represent kernel size, stride size and padding size, respectively. m is the number of neurons in the fully connected layer, used to align with T5. N denotes the chosen number of charge-discharge cycles.

Module	Layer	Configuration	Shape
Input	---	---	(1, 4, 100, N)
CONV1	Convolutional	filters = 8, k = (3, 1), s = (2, 1), p = 0	(1, 8, 49, N)
	Max-pooling	k = (2, 1), s = (2, 1), p = 0	(1, 8, 24, N)
CONV2	Convolutional	filters = 32, k = (3, 1), s = (2, 1), p = 0	(1, 32, 11, N)
	Max-pooling	k = (2, 1), s = (2, 1), p = 0	(1, 32, 5, N)
CONV3	Convolutional	filters = 128, k = (3, 1), s = (2, 1), p = 0	(1, 128, 2, N)
	Max-pooling	k = (2, 1), s = (2, 1), p = 0	(1, 128, 1, N)
Squeeze	Squeeze	---	(1, 128, N)
Reshape	Permute	---	(1, N, 128)
FC	Fully connected	m = 1024	(1, N, 1024)

Table S2 Description of 76 cells in HUST-LIB dataset [3]. The cells were cycled with a uniform charge protocol but completely different discharge protocols (protocols #1-#77). Four-stage discharge from 100% SOC to a cut off voltage of 2 V. Diverse discharge protocols C1 (100% SOC-60% SOC)-C2 (60% SOC-40% SOC)-C3 (40% SOC-20% SOC)-C4 (20% SOC-2 V) are employed for different cells, where C1-C4 represent constant-current discharge rates at the four stages, respectively.

Protocol	Channel	Data attribute	Knee point (cycles)	C1 (100%-60%)	C2 (60%-40%)	C3 (40%-20%)	C4 (20%-2V)
#1	1-1	Train	700	5C	1C	1C	2C
#2	1-2	Train	1,028	5C	1C	2C	2C
#3	1-3	Train	979	5C	1C	3C	2C
#4	1-4	Train	734	5C	1C	4C	2C
#5	1-5	Train	763	5C	1C	5C	2C
#6	1-6	Train	597	5C	2C	1C	2C
#7	1-7	Train	816	5C	2C	2C	2C
#9	2-2	Train	1690	5C	2C	5C	2C
#10	2-3	Train	823	5C	3C	1C	2C
#11	2-4	Train	708	5C	3C	2C	2C
#12	2-5	Train	522	5C	3C	3C	2C
#13	2-6	Train	749	5C	3C	4C	2C
#14	2-7	Test	1,344	5C	3C	5C	2C
#15	2-8	Test	758	5C	4C	1C	2C
#16	3-1	Train	956	5C	4C	2C	2C
#17	3-2	Train	1078	5C	4C	3C	2C
#18	3-3	Train	918	5C	4C	4C	2C
#19	3-4	Train	807	5C	4C	5C	2C
#20	3-5	Train	1,580	5C	5C	1C	2C
#21	3-6	Train	1,266	5C	5C	2C	2C
#22	3-7	Train	996	5C	5C	3C	2C
#23	3-8	Train	1,450	5C	5C	4C	2C
#24	4-1	Train	1,846	5C	5C	5C	2C
#25	4-2	Train	800	4C	1C	1C	2C
#26	4-3	Train	555	4C	1C	2C	2C
#27	4-4	Train	678	4C	1C	3C	2C
#28	4-5	Train	715	4C	1C	4C	2C
#29	4-6	Test	669	4C	1C	5C	2C
#30	4-7	Train	1,241	4C	2C	1C	2C
#31	4-8	Train	723	4C	2C	2C	2C
#32	5-1	Train	1,002	4C	2C	3C	2C
#33	5-2	Test	831	4C	2C	4C	2C
#34	5-3	Train	1,950	4C	2C	5C	2C
#35	5-4	Train	1,398	4C	3C	1C	2C

#36	5-5	Train	646	4C	3C	2C	2C
#37	5-6	Train	1,795	4C	3C	3C	2C
#38	5-7	Train	719	4C	3C	4C	2C
#39	6-1	Train	852	4C	4C	1C	2C
#40	6-2	Test	801	4C	4C	2C	2C
#41	6-3	Test	855	4C	4C	3C	2C
#42	6-4	Test	840	4C	4C	4C	2C
#43	6-5	Train	1,410	4C	4C	5C	2C
#44	6-6	Train	1,173	4C	5C	1C	2C
#45	6-8	Test	1,327	4C	5C	3C	2C
#46	7-1	Train	692	4C	5C	4C	2C
#47	7-2	Train	1,326	4C	5C	5C	2C
#48	7-3	Test	721	3C	1C	1C	2C
#49	7-4	Train	647	3C	1C	2C	2C
#50	7-5	Train	999	3C	1C	3C	2C
#51	7-6	Test	739	3C	1C	4C	2C
#52	7-7	Test	847	3C	1C	5C	2C
#53	7-8	Test	820	3C	2C	1C	2C
#54	8-1	Train	622	3C	2C	2C	2C
#55	8-2	Train	1,326	3C	2C	3C	2C
#56	8-3	Test	1,042	3C	2C	4C	2C
#57	8-4	Test	918	3C	2C	5C	2C
#58	8-5	Train	645	3C	3C	1C	2C
#59	8-6	Train	939	3C	3C	2C	2C
#60	8-7	Train	963	3C	3C	3C	2C
#61	8-8	Test	761	3C	3C	4C	2C
#62	9-1	Train	1,100	3C	3C	5C	2C
#63	9-2	Train	961	3C	4C	1C	2C
#64	9-3	Train	890	3C	4C	2C	2C
#65	9-4	Test	989	3C	4C	3C	2C
#66	9-5	Test	1,128	3C	4C	4C	2C
#67	9-6	Test	1,076	3C	4C	5C	2C
#68	9-7	Train	1,128	3C	5C	1C	2C
#69	9-8	Test	1,041	3C	5C	2C	2C
#70	10-1	Test	889	3C	5C	3C	2C
#71	10-2	Train	960	3C	5C	4C	2C
#72	10-3	Train	963	3C	5C	5C	2C
#73	10-4	Train	941	2C	4C	1C	2C
#74	10-5	Train	1,009	2C	5C	2C	2C
#75	10-6	Train	1,273	2C	3C	3C	2C
#76	10-7	Test	827	2C	2C	4C	2C

#77	10-8	Test	690	2C	1C	5C	2C

Table S3 Quantitative early-warning results for 22 test cells in HUST-LIB dataset.

Cell number	Channel	MAPE for RUC	MAPE for SOH_C	MAPE for SOH_F
#14	2-7	4.74%	0.10%	0.12%
#15	2-8	6.19%	0.29%	0.33%
#29	4-6	2.87%	0.10%	0.12%
#33	5-2	9.79%	0.17%	0.17%
#40	6-2	10.21%	0.14%	0.13%
#41	6-3	3.95%	0.07%	0.08%
#42	6-4	2.82%	0.11%	0.14%
#45	6-8	7.98%	0.40%	0.38%
#48	7-3	4.74%	0.11%	0.14%
#51	7-6	3.08%	0.07%	0.08%
#52	7-7	2.17%	0.14%	0.13%
#53	7-8	11.06%	0.09%	0.09%
#56	8-3	5.00%	0.09%	0.07%
#57	8-4	3.04%	0.09%	0.09%
#61	8-8	12.36%	0.41%	0.42%
#65	9-4	6.84%	0.20%	0.21%
#66	9-5	2.75%	0.14%	0.15%
#67	9-6	4.34%	0.14%	0.14%
#69	9-8	7.60%	0.14%	0.15%
#70	10-1	4.44%	0.10%	0.10%
#76	10-7	7.61%	0.28%	0.28%
#77	10-8	6.35%	0.07%	0.10%

Table S4 Online knee-point detection (KPD) results for 22 test cells in HUST-LIB dataset.

Cell number	Channel	Actual knee point	Detected knee point	Absolute error	Absolute percentage error
#14	2-7	1344	1322	22	1.6%
#15	2-8	758	730	28	3.7%
#29	4-6	669	635	34	5.1%
#33	5-2	831	785	46	5.5%
#40	6-2	801	774	27	3.4%
#41	6-3	855	836	19	2.2%
#42	6-4	840	830	10	1.2%
#45	6-8	1327	1311	16	1.2%
#48	7-3	721	710	11	1.5%
#51	7-6	739	691	48	6.5%
#52	7-7	847	831	16	1.9%
#53	7-8	820	802	18	2.2%
#56	8-3	1042	1031	11	1.1%
#57	8-4	918	906	12	1.3%
#61	8-8	761	733	28	3.7%
#65	9-4	989	971	18	1.8%
#66	9-5	1128	1055	73	6.5%
#67	9-6	1076	1060	16	1.5%
#69	9-8	1041	998	43	4.1%
#70	10-1	889	887	2	0.2%
#76	10-7	827	744	83	10.0%
#77	10-8	690	635	55	8.0%

Table S5 Quantitative early-warning results for 20 test cells in MIT&Stanford-LIB dataset.

Cell number	Actual knee point	MAPE for RUC	MAPE for SOH_C	MAPE for SOH_F
a6	464	3.83%	0.20%	0.19%
a15	506	7.69%	0.19%	0.23%
a18	514	13.22%	0.18%	0.18%
a24	753	6.67%	0.16%	0.18%
a29	643	4.84%	0.13%	0.15%
a30	464	4.26%	0.11%	0.14%
a38	454	6.57%	0.26%	0.22%
b2	366	11.44%	0.20%	0.25%
b5	382	7.32%	0.11%	0.13%
b10	430	16.54%	0.11%	0.13%
b14	390	8.34%	0.10%	0.14%
b18	368	14.71%	0.17%	0.22%
b21	401	15.03%	0.14%	0.17%
c10	816	6.84%	0.15%	0.18%
c12	732	3.48%	0.12%	0.11%
c16	1242	3.77%	0.23%	0.24%
c20	650	2.90%	0.16%	0.15%
c25	798	4.73%	0.09%	0.09%
c34	902	4.70%	0.17%	0.17%
c43	847	5.38%	0.09%	0.12%

Table S6 Online knee-point detection (KPD) results for 20 test cells in MIT&Stanford-LIB dataset.

Cell number	Actual knee point	Detected knee point	Absolute error	Absolute percentage error
a6	464	445	19	4.1%
a15	506	508	2	0.4%
a18	514	506	8	1.6%
a24	753	740	13	1.7%
a29	643	630	13	2.0%
a30	464	465	1	0.2%
a38	454	432	22	4.8%
b2	366	343	23	6.3%
b5	382	366	16	4.2%
b10	430	410	20	4.7%
b14	390	372	18	4.6%
b18	368	367	1	0.3%
b21	401	382	19	4.7%
c10	816	796	20	2.5%
c12	732	715	17	2.3%
c16	1242	1046	196	15.8%
c20	650	638	12	1.8%
c25	798	778	20	2.5%
c34	902	875	27	3.0%
c43	847	798	49	5.8%

Table S7 Comparison of MAPE across different methods for battery knee point prediction on HUST-LIB dataset.

Task	CNN-LSTM	Non-pretrained CT-eProber	CT-eProber
RUC	7.10%	8.21%	5.90%
SOH_C	0.13%	0.14%	0.12%
SOH_F	0.13%	0.14%	0.13%
KPD	3.40%	3.40%	3.40%

Table S8 Comparison of MAPE across different methods for battery knee point prediction on MIT&Stanford-LIB dataset.

Task	CNN-LSTM	Non-pretrained CT-eProber	CT-eProber
RUC	55.39%	51.36%	7.60%
SOH_C	8.15%	7.14%	0.13%
SOH_F	8.22%	7.85%	0.13%
KPD	46.00%	45.99%	3.70%

Table S9 Comparison of AUC across different methods for systemic financial crisis prediction.

Task	CNN-LSTM	Non-pretrained CT-eProber	CT-eProber
One-year-ahead	0.63	0.69	0.87
Two-year-ahead	0.82	0.83	0.96
Three-year-ahead	0.77	0.75	0.87
Four-year-ahead	0.77	0.66	0.84
Five-year-ahead	0.72	0.63	0.83

Table S10 Comparison of AUC across different methods for robotic accuracy failure prediction.

Task	CNN-LSTM	Non-pretrained CT-eProber	CT-eProber
Half-second-ahead	0.99	0.50	0.99
One-second-ahead	0.99	0.50	0.99
Two-second-ahead	0.96	0.50	0.99
Three-second-ahead	0.96	0.56	0.99
Four-second-ahead	0.99	0.83	0.99

4. Supplemental Notes

Note S1 The computational process of stock price

Stock price can be decomposed into two elements: capital appreciation, reflected in stock prices, and dividend distributions, which return part of corporate earnings to shareholders. Financial datasets often report these separately as the total return series and the dividend yield series, from which a dividend-free stock price index can be reconstructed.

Let p_t denote the stock price index at time t, d_t the dividend paid in period t, eq_tr_t the reported equity total return, and eq_dp_t the dividend yield. By definition, the total return incorporates both price appreciation and dividends

$$1 + \text{eq\_tr}_t = \frac{p_t + d_t}{p_{t-1}}$$

whereas the dividend yield is given by

$$\text{eq\_dp}_t = \frac{d_t}{p_t}$$

Eliminating d_t yields the pure stock price growth factor

$$\frac{p_t}{p_{t-1}} = \frac{1 + \text{eq\_tr}_t}{1 + \text{eq\_dp}_t}$$

Iterating this recursion produces a consistent price index normalized to a base value

$$\text{PRICE\_INDEX}_t = \text{base} \cdot \prod_{\tau \le t} \frac{1 + \text{eq\_tr}_\tau}{1 + \text{eq\_dp}_\tau}$$

This formulation isolates the trajectory of stock valuations from dividend policy, providing a clean representation of equity price dynamics across time and markets.

Note S2 Experimental settings of comparative methods

- **Convolutional neural network-Long short-term memory (CNN-LSTM)**: For comparison with the CT-eProber, we implement a CNN-LSTM model in which the T5 encoder-decoder is replaced by a 6-layer LSTM encoder and a 6-layer LSTM decoder. The CNN module remains identical to CT-eProber to ensure a fair comparison. The resulting architecture relies solely on recurrent mechanisms instead of self-attention to capture temporal dependencies.

- **Non-pretrained CT-eProber**: To further investigate the contribution of T5 pretraining within CT-eProber, we construct a non-pretrained variant, referred to as non-pretrained CT-eProber. This version keeps the overall architecture identical to CT-eProber consisting of the same CNN feature extractor, positional encoding, and T5-style encoder-decoder structure but the Transformer is initialized from scratch using a custom T5 configuration rather than loading any pretrained weights. It is worth noting that LoRA is not applied in non-pretrained CT-eProber, since LoRA is ineffective when the base weights lack meaningful pretrained information. Consequently, non-pretrained CT-eProber relies solely on task-specific supervised learning to acquire temporal representations, without leveraging any linguistic or structural priors learned during large-scale T5 pretraining.

Note S3 Predictive analysis of non-pretrained CT-eProber on robotic accuracy failure

We provide a mechanistic interpretation of the counter-intuitive trend observed in Table S10, where the non-pretrained CT-eProber achieves higher accuracy at longer prediction horizons (e.g., 4 s) than at shorter horizons (e.g., 0.5 s). This behavior stems from the specific formulation of the early-warning task. The early-warning task is formulated as an interval-based failure detection problem, in which the model determines whether a failure will occur at any time within a future horizon. Unlike point-wise prediction, which requires identifying the precise onset time of a fault, interval-based detection relaxes the requirement to identifying a "high-risk future state". A short horizon (0.5 s) forces the model to detect abrupt, high-frequency, and often subtle precursors that immediately precede failure. These transient cues are inherently difficult to learn without strong prior knowledge or pretraining. In contrast, a longer horizon (4 s) enables the model to rely on cumulative degradation patterns and lower-frequency structural changes, which are more stable and easier to capture. This broader window lowers task granularity, increases tolerance to temporal uncertainty, and ultimately reduces the complexity of the detection objective. Consequently, the non-pretrained model performs better at longer horizons.

5. Supplemental References

1 Severson KA, Attia PM, Jin N, *et al.* Data-driven prediction of battery cycle life before capacity degradation. ***Nat Energy*** 2019; **4**: 383–391.

2 Qiao G, Weiss BA. Accuracy degradation analysis for industrial robot systems. In: ***12th International Manufacturing Science and Engineering Conference***, Los Angeles, 2017.

3 Ma G, Xu S, Jiang B, *et al.* Real-time personalized health status prediction of lithium-ion batteries using deep transfer learning. ***Energy Environ Sci*** 2022; **15**: 4083–4094.

4 Hu EJ, Shen Y, Wallis P, *et al.* LoRA: Low-rank adaptation of large language models. In: ***10th International Conference on Learning Representations (ICLR)***, Online, 2022.