Agentic Continual Learning for Adaptive Forecasting in Thermal-Hydraulic Loops

Doyeong Lim, In Cheol Bang*
Department of Nuclear Engineering, Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-gil, Ulju-gun, Ulsan, 44949, Republic of Korea
**Corresponding author: icbang@unist.ac.kr*

Keywords : Thermal-hydraulics, Agentic AI, Continual learning, Adaptive Forecasting, Predictive modeling

1. Introduction

With advances in sensor instrumentation, data-driven forecasting models based on multivariate time-series data have become increasingly useful for proactive anomaly detection, short-term state prediction, and decision support during experimental operations. Deep learning models have shown strong predictive performance for thermal-hydraulic variables under transient conditions [1-2].

Despite this progress, most forecasting models are trained offline on a fixed set of operating conditions and deployed as static inference engines. While initial performance may be satisfactory, prediction errors can grow as facility configurations, boundary conditions, or phenomenological regimes change during experimental campaigns. This condition lock-in problem is compounded by a manual reconstruction bottleneck: engineers must manually identify degraded segments, curate new data, retrain the model, and redeploy it. Retraining models without a systematic governance process can cause them to lose accuracy on previously learned conditions, a problem known as catastrophic forgetting [3]. This motivates the need for a framework that continuously monitors model performance, triggers updates automatically, and ensures that model transitions do not degrade existing capabilities. Several enabling components exist but remain fragmented across communities, for example, multi-agents have demonstrated the ability to autonomously plan and execute multi-step scientific workflows [4].

Despite these individual advances, however, no existing framework integrates monitoring, autonomous model updating, and governance-based version control into a unified system tailored to the operational realities of thermal-hydraulic test loops, where conditions can change frequently across campaigns. This study proposes a preliminary agentic self-evolving forecasting framework for the multivariate time-series data generated by the thermal-hydraulic facility.

2. Methods

2.1. Experimental Facility

As an experimental campaign, a scaled-down integral effect test facility dataset was used, which was designed to simulate APR-1400 thermal-hydraulic behavior [5]. The dataset consists of six experimental cases with 15,807 steps covering temperature, pressure, flow rate, and related measurements. The dataset consists of six experimental case scenarios. The first three cases serve as the training set and the remaining three as the streaming evaluation set.

2.2 Models and Agent Decision Logic

The study evaluates five distinct AI architectures for the foundational forecasting engine: LSTM, GRU, Transformer, GNN, and NODE models. Each architecture is tailored for multi-output, multi-step-ahead prediction and trained for up to 1000 epochs. The core of the proposed framework is the Agent layer, driven by gpt-5.4-mini, which interfaces with the forecasting model through chunk-level monitoring signals rather than direct prediction generation. At each stream chunk, the agent observes model error, uncertainty, and drift status, and then orchestrates an observe-decide-act-govern loop over the streaming data.

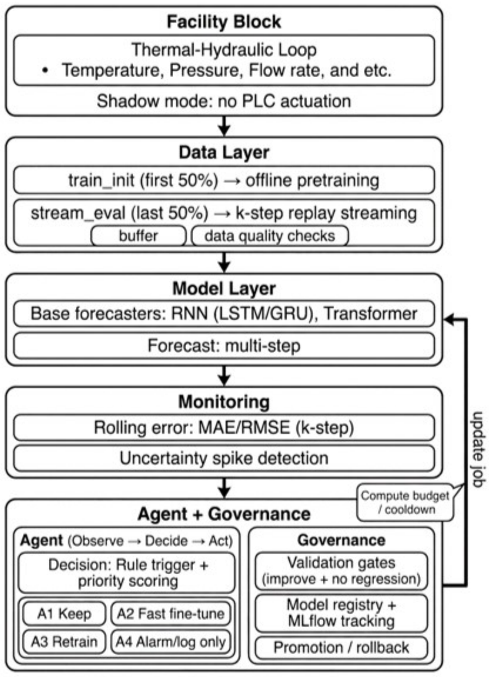

Fig. 1. Architecture of the proposed agentic self-evolving forecasting framework for experimental facility streaming operation.

As shown in Figure 1, The agent follows: If operation remains stable, the agent executes a HOLD command to conserve resources. If uncertainty spikes without clear error degradation, an ALARM_ONLY status is logged. When warning-level degradation persists for consecutive chunks, the agent attempts to FINE_TUNE the incumbent model. If explicit drift is detected or errors exceed a critical threshold, the agent triggers RETRAIN to generate a challenger model. During this process, the incumbent champion continues serving predictions until the challenger is validated and promoted.

3. Results and Discussion

Table I summarizes the offline pretraining results of five candidate forecasting architectures. Among all candidates, Transformer achieved the lowest validation MAE and validation loss of 0.718 and 1.339, respectively, followed by LSTM with a validation MAE of 0.748 and validation loss of 1.381. NODE, GNN, and GRU yielded validation MAEs of 0.805, 0.880, and 0.886, respectively. These results indicate that the Transformer provided the best fit on the train-validation split, while the other architecture remained competitive within a relatively narrow validation range.

Table I: Offline pretraining performance before streaming test.

Model	Validation MAE	Validation Loss
Transformer	0.718	1.339
LSTM	0.748	1.381
NODE	0.805	1.635
GNN	0.880	1.715
GRU	0.886	1.646

The five-model offline benchmark was used to expand challenger diversity, whereas the completed streaming experiment below reports the validated agentic forecasting setup. In the streaming experiment, activating the drift-triggered updates and governance gates (Agent-On) enabled the system to recognize and respond to phenomenological shifts. Across 229 stream chunks, the agent issued HOLD 190 times, FINE_TUNE 16 times, and RETRAIN 23 times over 11 detected drift events. Because the champion-challenger validation gate was intentionally conservative to prevent regression, only 4 of the 23 challengers were promoted, corresponding to a 17.4% replacement rate.

The aggregate impact of these interventions is reflected in the baseline-versus-agent comparison. Agent-On reduced the mean MAE from 14.88 to 13.24, corresponding to an 11.0% relative improvement, reduced the exceedance ratio from 35.4% to 22.7%, and lowered the worst-case MAE from 36.58 to 34.37, as shown in Fig. 2. A local before/after analysis further showed that retraining reduced the mean MAE over the subsequent three chunks by about 3.0 on average, whereas fine-tuning produced mixed gains under continuing regime shifts. The largest instantaneous error peaks were not fully eliminated because abrupt regime transitions create large prediction gaps before sufficient new evidence can be accumulated, while the governance gate intentionally delays promotion until the challenger demonstrates non-regressive performance.

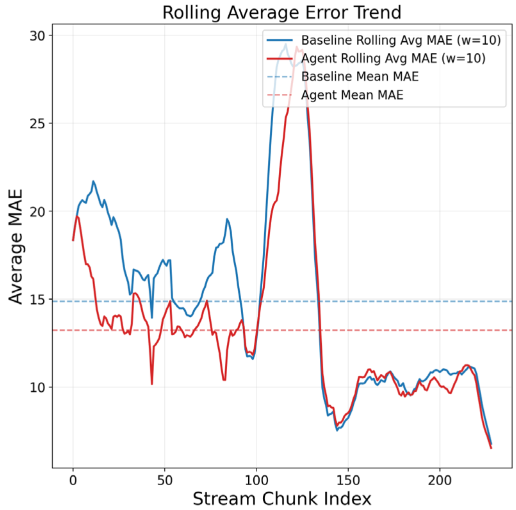

Fig. 2. Rolling average MAE trend on the streaming dataset: Baseline vs Agent-On.

4. Conclusions

This paper presented an preliminary agentic self-evolving forecasting framework designed to address the condition lock-in problem in digital twins for complex thermal-hydraulic loops. The AI agent continuously evaluates prediction confidence and selects the appropriate intervention strategy. Experimental results showed that the agentic approach reduced the overall MAE by 11.0% compared to a static baseline, thereby mitigating performance degradation during operational shifts.

ACKNOWLEDGMENTS

This work was supported by Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea Government (MOTIE) (no. RS-2024-00403194).

REFERENCES

[1] M. Kaminski, A. Diab, Time-series forecasting of a typical PWR undergoing large break LOCA, Sci. Technol. Nucl. Install. 2024.
[2] D. Lim, Z. Ndum Ndum, C. Young, Y. Hassan, Y. Liu, An AI-driven thermal-fluid testbed for advanced small modular reactors: Integration of digital twin and large language models, AI Therm. Fluids, 4, 2025.
[3] A. Cossu, A. Carta, V. Lomonaco, D. Bacciu, Continual learning for recurrent neural networks: An empirical evaluation, Neural Networks, 143, 2021.
[4] Y. Liu, Z. Abulawi, A. Garimidi, D. Lim, Automating data-driven modeling and analysis for engineering applications using large language model agents, arXiv:2510.01398, 2025.
[5] K.M. Kim, I.C. Bang, Design and operation of the transparent integral effect test facility, URI-LO for nuclear innovation platform, Nucl. Eng. Technol. 53, 2021.