SELF-EVOLVING TIME-SERIES FORECASTING AGENTS WITH EPISODIC MEMORY AND ONLINE POLICY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-based agents are increasingly used for time-series forecasting because they can organise contextual information, perform multi-step analysis, and guide the sequence of actions required to complete forecasting tasks. Most existing agents focus only on the current forecasting instance. However, in real-world deployments, forecasting commonly operates online, with new forecasts issued from the currently available history as the forecast origin advances and the ground-truth targets of earlier instances progressively become available. These targets provide feedback on the actions taken in earlier instances, yet existing agents generally do not preserve or utilise this information to adapt their subsequent actions. To address this limitation, we introduce **FASE**, a **F**eedback-**A**ware **S**elf-**E**volving forecasting agent that converts such feedback into task-specific experience for subsequent forecasting instances. FASE combines episodic memory, which retrieves relevant completed instances, with online policy learning, which summarises the feedback accumulated across instances into ranking guidance. The proposed framework is evaluated on 29 dataset configurations selected from the GIFT-Eval benchmark. Across these 29 configurations, FASE attains the strongest aggregate point forecasting performance among the evaluated methods and reduces the normalised MAE by 9.1% relative to the best individual foundation model baseline. The results further indicate that the cumulative advantage of FASE increases as delayed feedback accumulates. Together, these findings demonstrate that FASE can continually self-evolve through feedback from completed forecasting instances without updating the parameters of the LLM.

1 INTRODUCTION

Time series forecasting supports a wide range of real-world applications, including power systems, transportation, and financial markets. Recent advances in LLMs and agentic systems have inspired numerous studies that employ LLM-based agents for time-series forecasting. By reasoning over contextual information and coordinating analytical and forecasting tools, these agents extend forecasting beyond a direct mapping from historical observations to future values (Tao et al., 2026b; Ang et al., 2025). Existing work has focused only on the current forecasting instance. However, in real-world deployments, forecasting is typically performed online over a sequence of instances. Each forecast is based on the history available at that time, while the targets of earlier instances are revealed as time progresses. These newly observed targets not only extend the history available for future forecasts but also reveal how well earlier actions of the agent performed. Most forecasting agents use the delayed targets only as additional history, but rarely retain the feedback to improve subsequent actions.

To fill this gap, we introduce **FASE**, a **F**eedback-**A**ware **S**elf-**E**volving forecasting agent that makes feedback from completed forecasting instances available when later actions are selected. FASE combines episodic memory with online policy learning to capture feedback at two levels. Episodic memory incorporates each completed instance immediately and retrieves relevant instances as detailed evidence of tool performance under similar contexts. Online policy learning periodically aggregates feedback across completed instances into a predicted ranking of forecasting tools for the current history. For each forecasting instance, the LLM integrates the current time-series context

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with the retrieved evidence and ranking guidance to determine which forecasting tool to invoke. When the corresponding ground-truth target becomes available, both components are updated so that the completed instance can influence later actions.

Across 29 dataset configurations selected from GIFT-Eval, FASE achieves the strongest aggregate point forecasting performance among all evaluated methods, including a 9.1% reduction in normalised MAE relative to the best individual foundation model baseline. The growth of these gains as delayed feedback accumulates provides empirical evidence that the agent self-evolves through experience from completed forecasting instances.

The main contributions of this work are summarised as follows.

- We introduce **FASE**, a feedback-aware self-evolving forecasting agent that turns delayed feedback from completed forecasting instances into experience for later actions. Without updating the LLM parameters, FASE enables continual adaptation of forecasting tool usage through accumulated experience.
- We develop a two-level feedback-driven adaptation mechanism. Episodic memory retrieves relevant completed instances to provide context-specific evidence, while online policy learning periodically distils feedback accumulated across instances into ranking guidance for the current history. Together, these components combine detailed evidence from individual instances with broader performance patterns learned from accumulated feedback.

2 Related Work

2.1 Time-Series Foundation Models

Time-series foundation models are pretrained on large and heterogeneous collections of time series to capture transferable temporal patterns, enabling zero-shot forecasting on previously unseen tasks. Representative models include Chronos, TimesFM, Moirai, TimeGPT, Timer, MOMENT, and Time-MoE, which explore diverse architectures and pretraining strategies for scalable time-series modelling (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Garza et al., 2023; Liu et al., 2024; Goswami et al., 2024; Shi et al., 2025). Recent studies have further extended this paradigm through continuous-valued generative pretraining in Sundial, billion-parameter serial scaling in Timer-S1, and sampling-rate-equivariant forecasting in FlowState (Liu et al., 2025b; 2026; Graf et al., 2026). FASE builds upon these forecasting foundation models by treating them as numerical forecasting tools and investigating how delayed feedback from completed instances can improve future tool selection and usage.

2.2 Time-Series Forecasting Agents

Recent work has formulated forecasting as an agentic process in which an LLM reasons over contextual information and coordinates forecasting models and analytical tools. KairosAgent combines an LLM-based reasoner with a time-series foundation model and dynamically invokes analytical tools to improve numerical understanding and semantic reasoning (Feng et al., 2026). MemCast incorporates hierarchical memory derived from training data and adapts during inference by adjusting the confidence assigned to retrieved experiences (Tao et al., 2026a). EpiEvolve further explores feedback-driven adaptation by updating episodic experiences and strategic rules as delayed labels become available in a streaming forecasting scenario (Lu et al., 2026). However, its evaluation is conducted in a specific application domain with categorical forecasting targets. Our work instead studies self-evolution through delayed feedback for numerical forecasting across heterogeneous time-series configurations. FASE uses feedback from completed instances to continuously refine how the agent uses forecasting tools in future instances.

2.3 Self-Evolving LLM Agents

Recent approaches enable LLM agents to improve over time by using feedback from prior interactions to update persistent external state. Reflexion converts environmental feedback into verbal reflections, while ExpeL extracts reusable knowledge from accumulated experience (Shinn et al., 2023; Zhao et al., 2024). More recent methods extend this idea through other forms of adaptation.

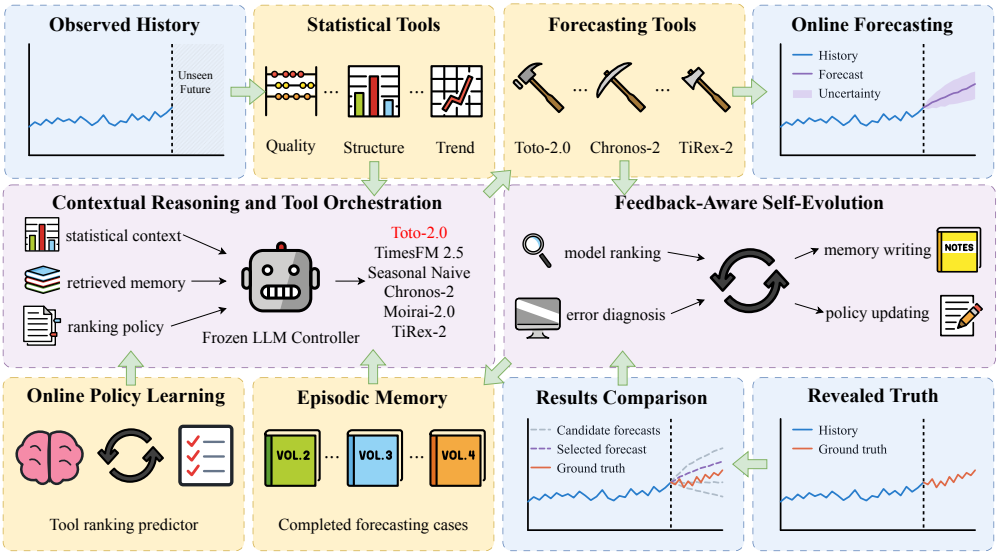


Figure 1: Overview of FASE. For each forecasting instance, the LLM uses statistical context, retrieved experience, and policy guidance to invoke a forecasting tool. After the ground-truth target is revealed, FASE compares the candidate forecasts with the target and uses the resulting feedback to update episodic memory and periodically update the online policy for subsequent actions.

ReasoningBank distils generalisable strategies from successful and failed experiences, ACE incrementally updates a contextual playbook, and GEPA uses reflection to optimise prompt components (Ouyang et al., 2026; Zhang et al., 2026; Agrawal et al., 2026). These approaches primarily focus on adapting textual knowledge, reusable strategies, or prompt-level behaviours. FASE extends this paradigm to delayed numerical feedback in online forecasting. Specifically, episodic memory preserves completed forecasting instances for retrieval according to the current context, whereas online policy learning periodically consolidates feedback across instances into a predicted ranking of forecasting tools.

3 FASE

3.1 Overview

FASE operates over a sequence of forecasting instances with access to a fixed library of forecasting tools. To use information from completed instances in subsequent actions, FASE maintains two forms of external state: episodic memory and an online policy. Both are initialised empty for each task configuration and updated as ground-truth targets become available. Figure 1 summarises how FASE processes each forecasting instance and uses delayed feedback to update episodic memory and the online policy.

For each forecasting instance, the forecasting tools and the LLM receive different representations of the observed history. Each forecasting tool operates directly on the original numerical values. In contrast, providing raw numerical values to the LLM requires the time series to be serialised into a token sequence, resulting in substantial input overhead for long histories containing thousands of observations. Moreover, serialised numerical values provide limited explicit semantic information, making it difficult for the LLM to directly interpret the temporal patterns and structural characteristics of the time series. FASE therefore represents the observed history using a fixed collection of statistical features, which summarise properties such as scale, trend, periodicity, temporal dependence, and missingness. This representation provides the LLM with a compact and structured description of the time series while avoiding the token overhead associated with representing the full numerical history.

Based on the current statistical representation, episodic memory retrieves relevant completed forecasting instances. After the warm-up phase, the online policy additionally produces a ranking of candidate tools. FASE provides the LLM with the task context, statistical representation, retrieved experience, available policy guidance, and descriptions of the forecasting tools. The LLM selects a single tool to produce the point forecast and corresponding quantile forecasts for the current forecasting instance. The remaining tools are run separately on the same observed history before the target is revealed, and all forecasts are stored for evaluation when the target becomes available. After the ground-truth target is revealed, FASE computes the point and probabilistic forecast errors of all tools. The statistical representation, task context, invoked tool, and observed errors form the record of a completed forecasting instance. This record enters episodic memory and supplies feedback to the online policy.

3.2 Episodic Memory

After the ground-truth target for a forecasting instance becomes available, FASE organises the statistical representation, task context, invoked tool, and observed errors into an episodic memory entry for subsequent retrieval. For each candidate tool, the entry records the point forecast error and its corresponding rank, together with the probabilistic forecast error as auxiliary feedback.

FASE organises the active memory, which contains all entries currently eligible for retrieval, into a recent pool and a long-term pool. After a forecasting instance is completed, FASE first updates the retention values of the memory entries included in the context used to select the corresponding action and then inserts the newly completed instance directly into the recent pool. If this insertion causes the recent pool to exceed its capacity, FASE removes the oldest entry and considers it for promotion to the long-term pool. If the long-term pool has available capacity, the entry is promoted directly. Otherwise, whether the candidate enters the long-term pool is determined by its retention value.

Let s_i denote the statistical representation of forecasting instance i. To compute retention values, FASE measures the similarity between the current representation s_i and the representation s_j stored in an active entry using a normalised statistical distance. Let $\mathcal{G}_{ij}$ denote the statistical feature groups containing observed features in both s_i and s_j, and let $\mathcal{Q}_{ij,g}$ denote the corresponding features in group g. The distance is defined as

$$d(s_i, s_j) = \frac{1}{|\mathcal{G}_{ij}|} \sum_{g \in \mathcal{G}_{ij}} \frac{1}{|\mathcal{Q}_{ij,g}|} \sum_{q \in \mathcal{Q}_{ij,g}} \rho\left(\frac{|s_{i,q} - s_{j,q}|}{\sigma_q}\right), \qquad \rho(z) = \frac{z}{1+z}, \tag{1}$$

where σ_q is the standard deviation of feature q across active entries. The transformation ρ bounds the influence of features with large normalised discrepancies.

For each forecasting instance, active entries are partitioned according to the tool used in each stored instance and the set of candidate tools tied for the best rank based on point forecast error. Let $\mathcal{E}_i$ denote the entries included in the context used to select the action for instance i. For each partition C, let $j_C^{(1)}$ and $j_C^{(2)}$ denote the nearest and second-nearest entries, respectively. The contribution of entry j is defined as

$$\delta_{i,j} = \begin{cases} d(s_i, s_{j_C^{(2)}}) - d(s_i, s_{j_C^{(1)}}), & j = j_C^{(1)} \in \mathcal{E}_i,\ |C| > 1, \\ 1 - d(s_i, s_{j_C^{(1)}}), & j = j_C^{(1)} \in \mathcal{E}_i,\ |C| = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where C denotes the partition containing entry j. The retention value of each active entry is the running average of its contributions across retention updates. Thus, a large retention value indicates that the entry is consistently a more distinctive representative of the associated tool use and observed candidate performance than nearby alternatives.

Given the current statistical representation s_i, FASE ranks all active entries according to $d(s_i, s_j)$, retrieves those with the smallest distances, and presents their contextual statistics and observed tool performance to the LLM as evidence for selecting the current action.

3.3 Online Policy Learning

Episodic memory incorporates feedback from each completed forecasting instance immediately and makes relevant instances available when subsequent actions are selected. Online policy learning instead aggregates feedback over multiple completed instances and periodically updates the policy state based on accumulated experience. FASE implements this policy state using a lightweight predictor g_{θ_i}, which is updated online, and a first-in, first-out feedback pool.

Let $\mathcal{F}$ denote the set of available forecasting tools. Let $x_i \in \mathbb{R}^N$ denote the N most recent visible values in the current history. FASE normalises the observed values and combines them with a binary observation mask, producing $z_i = [\mathrm{Normalise}(x_i), \mathrm{Mask}(x_i)] \in \mathbb{R}^{2N}$. The predictor is a lightweight MLP $g_{\theta_i} : \mathbb{R}^{2N} \to \mathbb{R}^{|\mathcal{F}|}$ that maps z_i to relative performance scores $u_i = g_{\theta_i}(z_i)$. A lower score $u_{i,m}$ indicates better expected point forecasting performance for tool m. Sorting the scores in ascending order produces a tool ranking, which FASE converts into a textual list and provides to the LLM as additional policy guidance for the current action.

When the ground-truth target for forecasting instance i becomes available, FASE adds the input representation and observed tool errors to the feedback pool. After sufficient feedback has accumulated, the predictor is initialised using the available feedback pool and subsequently updated periodically as additional instances are completed. Each update samples minibatches uniformly from the current feedback pool. The corresponding hyperparameters are given in Section 4.2. For a record (z, ℓ) drawn from such a minibatch, let $u = g_\theta(z)$ and $\bar{\ell} = |\mathcal{F}|^{-1} \sum_{m \in \mathcal{F}} \ell_m$. The predictor minimises the following pairwise loss, averaged over all tool pairs and then over the minibatch:

$$\mathcal{L}(z, \ell) = \frac{1}{\binom{|\mathcal{F}|}{2}} \sum_{p<q} \frac{|\ell_p - \ell_q|}{\bar{\ell}} \,\mathrm{softplus}(\mathrm{sgn}(\ell_q - \ell_p)(u_p - u_q)) \,, \tag{3}$$

The sign term orients each pair so that the tool with lower point error is encouraged to receive the lower score. The loss assigns larger weights to pairs with greater relative differences in point forecast error.

4 Experiments

4.1 Datasets and Metrics

We evaluate FASE on 29 dataset configurations selected from the 97 configurations in GIFT-Eval (Aksu et al., 2024), retaining those with 500 to 5,000 forecasting instances in the test split. The lower bound provides enough forecasting instances to examine self-evolution as feedback accumulates, while the upper bound limits computational cost. The selected configurations cover all seven GIFT-Eval domains: Econ/Fin, Energy, Healthcare, Nature, Sales, Transport, and Web/CloudOps. Some source datasets contribute multiple configurations under different sampling frequencies or forecast terms. Together, the configurations cover sampling frequencies from 10 seconds to one month, short, medium, and long forecast terms, and prediction lengths ranging from 8 to 720 time steps. Detailed information about the evaluated datasets is provided in Appendix A.

We follow the GIFT-Eval protocol for test splits and the construction of forecasting instances. Each configuration may include multiple series, with up to 20 forecasting instances per series. Episodic memory and the online policy accumulate at the task level rather than per series: the agent processes all forecasting instances of one series in chronological order, then moves to the next series, following the item order used by GIFT-Eval.

We evaluate point forecasts using MASE, MAE, and RMSE, and probabilistic forecasts using CRPS. CRPS is estimated from the nine predicted quantiles at levels 0.1 through 0.9. For every task configuration, each metric is normalised by the corresponding score of Seasonal Naive. We report the geometric mean of these normalised scores across the 29 configurations, giving each configuration equal weight. For the rank columns in Table 1, systems are ranked within each configuration for each metric. Tied systems receive the mean of their ranks, and the resulting ranks are then averaged arithmetically across configurations.

Table 1: Aggregate results over the 29 evaluated GIFT-Eval configurations. Bold and underlined values indicate the best and second-best results within each column, respectively.

	MASE ↓		MAE ↓		RMSE ↓		CRPS ↓	
Method	Norm.	Rank	Norm.	Rank	Norm.	Rank	Norm.	Rank
Seasonal Naive	1.000	7.79	1.000	7.72	1.000	7.86	1.000	7.97
Chronos-2 (Ansari et al., 2025)	0.752	4.55	0.729	5.03	0.720	4.59	0.545	4.90
TimesFM 2.5 (Google Research, 2025)	0.753	4.59	0.737	4.62	0.705	3.86	0.528	4.17
Moirai-2.0-R-Small (Liu et al., 2025a)	0.774	5.97	0.739	5.21	0.756	6.24	0.542	5.38
TiRex-2 (Podest et al., 2026)	<u>0.728</u>	4.34	<u>0.685</u>	3.97	<u>0.677</u>	3.72	**0.489**	3.38
Toto-2.0-313m (Khwaja et al., 2026)	0.745	4.03	0.699	4.34	0.716	4.93	0.500	3.93
Uniform Ensemble[a]	0.731	<u>2.86</u>	0.697	<u>3.41</u>	0.681	<u>2.86</u>	0.501	**3.10**
FASE	**0.679**	**1.86**	**0.622**	**1.69**	**0.643**	**1.93**	<u>0.497</u>	<u>3.17</u>

[a] Uniform Ensemble averages the forecasts of the five foundation models with equal weights.

4.2 Implementation Details

FASE uses a fixed tool pool that comprises Seasonal Naive and five time-series foundation models: Chronos-2 (Ansari et al., 2025), TimesFM 2.5 (Google Research, 2025), Moirai-2.0-R-Small (Liu et al., 2025a), TiRex-2 (Podest et al., 2026), and Toto-2.0-313m (Khwaja et al., 2026). These forecasters also serve as the individual model baselines. To isolate the effects of episodic memory and online policy learning, we define Base Agent as an additional baseline. Its LLM receives only the context of the current forecasting instance, including the statistical summary and descriptions of the available forecasting tools, without retrieved instances or policy guidance. Qwen3.8-27B-FP8 (Qwen Team, 2026) serves as the LLM controller, with the temperature set to 0.1, top-p set to 1.0, and the reasoning effort set to xhigh to support consistent and deliberate action selection. The prompt templates used by the controller are provided in Appendix E.

For each forecasting instance, the observed history is summarised by the 18 statistical features that define s_i. This representation is provided to the LLM as a compact description of the current history and is also stored with each episodic memory entry. The features characterise temporal dynamics, missingness, periodicity, covariate relationships, and other properties of the history. Appendix D lists the individual definitions of these features.

The recent and long-term memory pools have capacities of 100 and 900 entries, respectively, and FASE retrieves ten entries for each forecasting instance. The online policy instead operates on the most recent $8H$ history values, which are normalised separately for each forecasting instance and paired with a binary observation mask. Completed instances are inserted into a first-in, first-out feedback pool with a capacity of 3,000 records. The ranker is initialised after feedback from 100 instances has been collected, using 10 epochs, and is subsequently updated after every 25 additional valid instances, using 25 minibatches of 256 uniformly sampled records. For optimisation, we use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 10^{-3} and a weight decay of 10^{-5}.

5 Results and Discussion

5.1 Main Results

Table 1 compares FASE with the fixed forecasting tools and with Uniform Ensemble over the 29 evaluated GIFT-Eval configurations. Appendix F reports the corresponding results for each configuration. FASE achieves the strongest overall point forecasting performance, with the lowest normalised errors for MASE, MAE, and RMSE. TiRex-2 is the strongest individual foundation model on all three point metrics. Relative to TiRex-2, FASE reduces MASE, MAE, and RMSE by 6.8%, 9.1%, and 5.0%, respectively. The improvements across complementary error measures show that the advantage of FASE is not tied to a single definition of point forecast quality.

The average ranks provide a complementary view by comparing systems within each configuration and then weighting all configurations equally. FASE also ranks first on all three point metrics, showing that its aggregate gains are accompanied by consistently strong relative performance rather

Table 2: Results aggregated by prediction length. Bold and underlined values indicate the best and second-best results within each column, respectively.

	Long		Medium		Short	
Method	MASE ↓	CRPS ↓	MASE ↓	CRPS ↓	MASE ↓	CRPS ↓
Seasonal Naive	1.000	1.000	1.000	1.000	1.000	1.000
Chronos-2 (Ansari et al., 2025)	0.782	0.547	0.782	0.563	0.729	0.538
TimesFM 2.5 (Google Research, 2025)	0.763	0.519	0.773	0.546	0.741	0.525
Moirai-2.0-R-Small (Liu et al., 2025a)	0.823	0.560	0.831	0.567	0.733	0.525
TiRex-2 (Podest et al., 2026)	<u>0.748</u>	<u>0.475</u>	<u>0.760</u>	**0.488**	<u>0.708</u>	<u>0.496</u>
Toto-2.0-313m (Khwaja et al., 2026)	0.753	**0.474**	0.786	0.500	0.727	0.512
Uniform Ensemble	0.749	0.490	0.762	0.509	0.711	0.503
FASE	**0.703**	0.515	**0.713**	<u>0.492</u>	**0.656**	**0.491**

than being driven by a small number of configurations with unusually large improvements. FASE also outperforms Uniform Ensemble, the strongest fixed baseline by average rank, reducing its normalised MASE, MAE, and RMSE by 7.1%, 10.8%, and 5.6%, respectively. This improvement over a fixed equal-weight combination indicates that FASE learns from accumulated feedback which forecasting tool is better suited to each forecasting instance.

The gains are less consistent for probabilistic forecasting. FASE obtains the second-lowest aggregate CRPS of 0.497, behind TiRex-2 at 0.489, while its average CRPS rank is also slightly higher than that of Uniform Ensemble. This difference is consistent with the feedback mechanism that guides tool selection. The retrieved instances emphasise point forecasting errors and the associated ranks, and the online policy is optimised to order tools according to point forecasting performance. Probabilistic error is incorporated only as auxiliary feedback. Accordingly, FASE delivers more consistent gains in point forecasting than in probabilistic forecasting.

Table 2 examines whether the point forecasting advantage is concentrated at a particular prediction length. FASE obtains the lowest normalised MASE in the long, medium, and short groups, reducing MASE relative to TiRex-2 by approximately 6.0%, 6.2%, and 7.3%, respectively. The similar gains across the three groups show that the aggregate improvement persists over the evaluated range of prediction lengths. The CRPS results are less consistent across groups: FASE performs best on the short configurations and second best on the medium configurations, but does not lead on the long configurations. This contrast again shows that the main benefit of the current feedback design lies in point forecasting.

5.2 Self-Evolution over Accumulated Feedback

We examine whether FASE can continually self-evolve through episodic memory and online policy learning. For each configuration, we calculate the mean MASE of FASE and the Base Agent over all forecasting instances processed up to each point. The reduction achieved by FASE is expressed as a percentage of the Base Agent MASE and then averaged equally across configurations. Figure 2 presents the resulting trajectory over the first 500 processed instances.

The cumulative gain exhibits an overall upward trend, rising from 3.8% after 25 processed instances to 6.7% after 100 instances and 8.6% after 500 instances. Despite minor local fluctuations, the relative advantage remains positive and continues to increase as more forecasting instances are completed. Because each point includes all preceding forecasting instances, this sustained rise shows that later instances strengthen rather than dilute the advantage accumulated earlier. The trend provides evidence that the feedback incorporated into episodic memory and the online policy continues to improve subsequent actions as experience grows.

5.3 Ablation Study

Table 3 evaluates the separate and joint contributions of episodic memory and online policy learning. Memory Only and Policy Only both improve every reported metric over the Base Agent. In particular, they reduce normalised MASE from 0.758 to 0.683 and 0.687, respectively, showing that

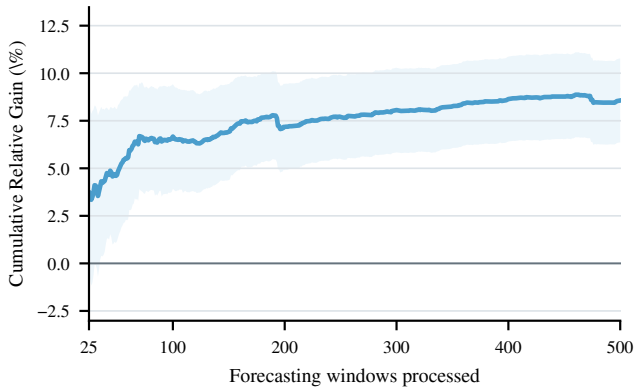


Figure 2: Cumulative relative MASE gain of FASE over the Base Agent across the first 500 forecasting instances in the replay. The curve begins at 25 processed instances, and the shaded band denotes the standard error across configurations.

Table 3: Ablation of episodic memory and online policy learning over the same 29 evaluated GIFT-Eval configurations. Base Agent uses neither mechanism; Memory Only uses episodic memory; Policy Only uses online policy learning; FASE combines both. Bold and underlined values indicate the best and second-best results within each column, respectively.

Method	MASE ↓	MAE ↓	RMSE ↓	CRPS ↓
Base Agent	0.758	0.718	0.713	0.543
Memory Only	<u>0.683</u>	0.631	0.655	**0.483**
Policy Only	0.687	<u>0.629</u>	<u>0.649</u>	0.505
FASE	**0.679**	**0.622**	**0.643**	<u>0.497</u>

each mechanism provides useful guidance even when used independently. Episodic memory contributes evidence from completed instances with contexts similar to the current one, whereas the online policy periodically consolidates feedback from a broader set of instances into a ranking of the forecasting tools. Their similar point forecasting performance suggests that these two forms of feedback are individually useful despite operating at different levels of aggregation.

FASE combines both mechanisms and achieves the lowest MASE, MAE, and RMSE among the four variants, with relative reductions ranging from 9.8% to 13.4% relative to the Base Agent. It also performs consistently better on all three point metrics than either variant using only one mechanism. Although the additional gains are modest, their consistency supports the complementary roles of evidence retrieved from similar contexts and broader performance patterns captured by the policy. Memory Only retains the lowest CRPS, while FASE ranks second on this metric. This exception is consistent with the design of both feedback mechanisms, which prioritise point forecasting errors and use probabilistic error only as auxiliary information.

5.4 Analysis of Memory and Policy

Episodic memory influences tool use by retrieving completed instances with contexts similar to the current forecasting instance and presenting the observed performance of each forecasting tool in those instances. The LLM combines this evidence with the current context to assess whether a previous tool advantage is likely to persist. As shown in Figure 3, Memory Only performs better than the Base Agent on 61.4% of forecasting instances and worse on 20.0%, indicating that local performance patterns often transfer to similar contexts but do not always persist. Online policy learning provides a more direct signal: the ranker converts accumulated feedback into a predicted ordering of the tools, which the LLM considers when choosing its action. The policy therefore improves forecasting performance when this ordering captures the current dynamics, but can degrade performance when the learned pattern does not apply.

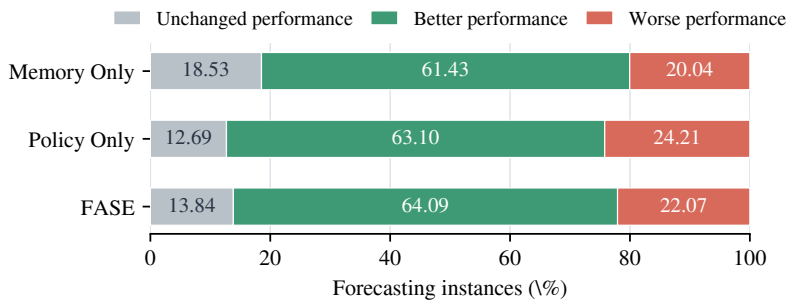


Figure 3: Distribution of forecasting performance relative to the Base Agent. Better and worse performance correspond to lower and higher MASE, respectively. Unchanged performance includes forecasting instances in which the evaluated method selects the same tool as the Base Agent or selects a different tool that yields the same MASE; the latter account for less than 0.3% of all evaluated instances. Results are averaged equally across configurations.

Table 4: MASE changes relative to the Base Agent. Improvement is the MASE reduction from forecasting instances with better performance. Degradation is defined analogously for instances with worse performance. Net improvement is their difference. Results are then averaged equally across configurations.

	MASE change relative to Base Agent (%)		
Method	Improvement	Degradation	Net improvement
Memory Only	11.4	2.2	9.2
Policy Only	11.7	2.6	9.1
FASE	11.9	2.2	9.7

FASE jointly uses these two sources of information. Table 4 shows that the resulting improvement of 11.9% outweighs the degradation of 2.2%, yielding the largest net MASE improvement of 9.7%. Compared with the variants using only one mechanism, FASE achieves a larger aggregate improvement without increasing degradation beyond the level observed for Memory Only. The positive balance shows how combining evidence from similar completed instances with the broader ordering learned by the policy improves tool use overall, even though either signal can occasionally lead to a worse action.

6 Conclusion

LLM-based agents are increasingly used for time-series forecasting, with existing approaches focusing only on the current forecasting instance. In real-world deployments, however, forecasting commonly operates online, and the ground-truth targets of earlier instances become available over time. Existing agents rarely retain the feedback provided by these targets to improve subsequent actions. To address this gap, we introduced FASE, a feedback-aware self-evolving forecasting agent that combines episodic memory with online policy learning. Episodic memory retrieves relevant experience from completed forecasting instances, while online policy learning summarises accumulated feedback into ranking guidance for subsequent actions. Experiments on 29 GIFT-Eval configurations show that FASE achieves the strongest aggregate point forecasting performance among the evaluated methods. Its performance gains increase as feedback accumulates, and further analysis shows that memory and policy guidance change how the agent uses forecasting tools in subsequent instances. Together, these findings demonstrate that FASE can continually self-evolve through feedback from completed forecasting instances.

AI USE STATEMENT

In this work, we used generative AI tools for drafting a section of this manuscript, assisting with writing and language polishing, identifying related work, analysing experimental data, and interpreting the reported results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artefacts produced with the aid of generative AI.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Yihao Ang, Yifan Bao, Lei Jiang, Jiajie Tao, Anthony K. H. Tung, Lukasz Szpruch, and Hao Ni. Structured agentic workflows for financial time-series modelling with LLMs and reflective feedback. In *Proceedings of the 6th ACM International Conference on AI in Finance*, 2025.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Kun Feng, Ziwei Shan, Yuchen Fang, Yiyang Tan, Sihan Lu, Shuqi Gu, Xingyu Lu, Lintao Ma, and Kan Ren. KairosAgent: Agentic time series forecasting with fused semantic reasoning. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1. *arXiv preprint arXiv:2310.03589*, 2023.

Google Research. TimesFM 2.5, 2025. URL `https://github.com/google-research/timesfm`.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Lars Graf, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. FlowState: Sampling-rate-equivariant time-series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025b.

Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-S1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yiming Lu, Sihang Zeng, Zhengxu Tang, Max Lau, Fei Liu, and Wei Jin. EpiEvolve: Self-evolving agents for streaming pandemic forecasting under regime shifts. *arXiv preprint arXiv:2606.05513*, 2026.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex-2: Generalizing TiRex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, 2026.

Qwen Team. Qwen3.8-27B-FP8, 2026. URL `https://huggingface.co/Qwen/Qwen3.8-27B-FP8`.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Xiaoyu Tao, Mingyue Cheng, Ze Guo, Shuo Yu, Yaguo Liu, Qi Liu, and Shijin Wang. MemCast: Memory-driven time series forecasting with experience-conditioned reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a.

Xiaoyu Tao, Mingyue Cheng, Chuang Jiang, Tian Gao, Huanjian Zhang, and Yaguo Liu. Cast-R1: Learning tool-augmented sequential decision policies for time series forecasting. *arXiv preprint arXiv:2602.13802*, 2026b.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

A EVALUATION CONFIGURATIONS

Table 5 provides detailed information about the 29 GIFT-Eval configurations used in our evaluation. Configurations from the same source dataset are grouped together.

Table 5: GIFT-Eval configurations used for evaluation. Rows are grouped by source dataset. Prediction length is measured in time steps.

Dataset	Domain	Frequency	Term	Pred. Length	Instances
bitbrains_fast_storage	Web/CloudOps	5T	medium	480	5,000
bitbrains_fast_storage	Web/CloudOps	5T	long	720	5,000
bitbrains_fast_storage	Web/CloudOps	H	short	48	5,000
bitbrains_rnd	Web/CloudOps	5T	medium	480	2,000
bitbrains_rnd	Web/CloudOps	5T	long	720	2,000
bitbrains_rnd	Web/CloudOps	H	short	48	2,000
bizitobs_service	Web/CloudOps	10S	short	60	630
car_parts_with_missing	Sales	M	short	12	2,674
electricity	Energy	D	short	30	1,850
electricity	Energy	H	medium	480	2,960
electricity	Energy	H	long	720	1,850
electricity	Energy	W-FRI	short	8	1,110
hierarchical_sales	Sales	D	short	30	826
hospital	Healthcare	M	short	12	767
kdd_cup_2018_with_missing	Nature	D	short	30	540
kdd_cup_2018_with_missing	Nature	H	medium	480	540
kdd_cup_2018_with_missing	Nature	H	long	720	540
LOOP_SEATTLE	Transport	5T	long	720	4,845
LOOP_SEATTLE	Transport	D	short	30	646
LOOP_SEATTLE	Transport	H	medium	480	646
LOOP_SEATTLE	Transport	H	long	720	646
m4_daily	Econ/Fin	D	short	14	4,227
M_DENSE	Transport	H	short	48	600
restaurant	Sales	D	short	30	807
solar	Energy	10T	short	48	2,740
solar	Energy	10T	medium	480	1,507
solar	Energy	10T	long	720	1,096
solar	Energy	H	short	48	2,603
SZ_TAXI	Transport	15T	short	48	1,092

B Online Forecasting Protocol

We evaluate all methods by replaying the forecasting instances of each GIFT-Eval configuration in a fixed order. Within each series, the instances are processed chronologically; after all instances of one series have been processed, evaluation proceeds to the next series in the item order provided by GIFT-Eval. Because the benchmark does not define a shared clock across different series, no additional temporal ordering is imposed between series. Episodic memory and the online policy persist across series within the same configuration, but both are reinitialised for every new configuration.

At each forecast origin, tool inputs are constructed only from the information available at that time. All evaluated forecasting instances are univariate. GIFT-Eval configurations with multivariate targets are converted into separate univariate series before inference, so every forecasting tool receives a single target channel. Seasonal Naive uses the complete available history and a seasonal period inferred from the sampling frequency. For Chronos-2, TimesFM 2.5, Moirai-2.0-R-Small, TiRex-2, and Toto-2.0-313m, we retain the most recent 8,192, 15,360, 4,000, 8,192, and 4,096 time steps, respectively. Where available, past and known future covariates are supplied to TiRex-2, while Toto-2.0-313m receives past covariates; the remaining tools receive only the target history. Chronos-2, Moirai-2.0-R-Small, TiRex-2, and Toto-2.0-313m receive missing target values directly through their native interfaces. For Seasonal Naive and TimesFM 2.5, missing values are linearly interpolated over timestamps within each target channel. Leading and trailing missing values are replaced by the nearest observed value, and a channel with no observed values is filled with zeros.

The LLM chooses its action using the current statistical context and any memory or policy information derived from previously completed instances. The selected tool produces the forecast used by the agent, while the remaining tools are run separately on the same available history before the target is revealed. The candidate forecasts are not provided to the LLM, and none of the tools has access to the target during inference. TimesFM 2.5 is run with input normalisation, positivity inference, flip invariance, a continuous quantile head, and quantile crossing correction enabled. Each adapter returns a point forecast and nine quantile forecasts at levels 0.1 through 0.9, all arranged in the same horizon-by-channel format. No additional scaling, clipping, or calibration is applied after the adapter output. Uniform Ensemble is constructed by averaging the point forecasts of the five foundation models with equal weights and applying the same averaging operation separately at each quantile level.

After the target becomes available, we compute the errors of all six candidate forecasts. The resulting feedback is then added to episodic memory and to the feedback pool of the online policy, and any state update takes effect from the next forecasting instance.

C Complete Implementation Settings

Table 6 summarises the implementation settings used in all reported experiments. The online policy is maintained independently for each GIFT-Eval configuration, as is episodic memory.

Table 6: Complete implementation settings used in the experiments.

Component	Setting
LLM controller	Qwen3.8-27B-FP8; temperature 0.1; top-p 1.0; reasoning effort xhigh; sampling seed 1234.
Statistical context	18 features computed from up to the 15,360 most recent history positions.
Episodic memory	Recent pool capacity 100; long-term pool capacity 900; 10 retrieved instances per action.
Policy input	Most recent $8H$ observations, normalised within each forecasting instance using the median and standard deviation, together with a binary observation mask.
Policy predictor	MLP with dimensions $2(8H) \to 2048 \to 128 \to 6$ and ReLU activations after the hidden layers.
Feedback pool	First-in, first-out pool with capacity 3,000; uniform minibatch sampling.
Initial policy training	Begins after 100 completed instances and runs for 10 epochs.
Periodic policy update	Performed after every 25 additional completed instances using 25 minibatches of size 256.
Policy optimisation	AdamW; learning rate 10^{-3}; weight decay 10^{-5}; random seed 7.

D STATISTICAL FEATURES

Table 7 defines the 18 features used to summarise the observed history of each forecasting instance.

Table 7: Statistical features used to form the representation s_i.

Group	Statistic	Description
Data quality	Missing ratio	Fraction of missing target values.
Data quality	Zero ratio	Fraction of observed target values equal to zero.
Temporal dynamics	Global trend strength	Standardised slope over the full history.
Temporal dynamics	Recent trend strength	Standardised slope over the recent history.
Temporal dynamics	Recent level shift strength	Standardised shift between recent and prior means.
Temporal dynamics	Recent scale change strength	Normalised change between recent and prior scales.
Structural change and complexity	Change strength	Largest standardised mean shift across candidate splits.
Structural change and complexity	Linear fit error ratio	Linear one-step error relative to persistence.
Structural change and complexity	Turning behaviour ratio	Rate of direction changes in successive differences.
Structural change and complexity	Direction entropy	Entropy of adjacent change directions.
Spectral structure	Spectral concentration	Detrended energy in the strongest frequency bins.
Spectral structure	Spectral entropy	Normalised spectral entropy.
Spectral structure	Stability	Variance of standardised block means.
Missingness	Longest missing block	Longest consecutive missing segment.
Periodicity	Recent period strength	Strongest recent detrended periodic component.
Covariates	Strongest covariate relation	Largest absolute correlation between the target and any covariate.
Covariates	Covariate relation stability	Agreement between correlations of the target with covariates in early and late segments.
Intermittency	Recent event rate change	Change in recent versus historical nonzero rate.

E PROMPT TEMPLATES

All controller variants use the same shared system and user templates. Memory Only and FASE additionally include the episodic memory instruction and retrieved memory block, whereas Policy Only and FASE include the online policy instruction and ranking block. The definitions of the statistical features in Table 7 are prepended to every user message. The serialised values inserted into the templates vary across forecasting instances, but the surrounding instructions remain fixed.

Shared system instruction.

```
You are selecting a forecasting model for a time-series prediction task.
Choose one model from available_model_ids that you expect to achieve the
best point-forecast accuracy for the current case.
The supplied statistics describe the historical observations available at
the current forecast origin.
Use these statistics to understand the sequence's trend, periodicity,
variability, missingness, and other measured characteristics.
Consider how these characteristics affect forecasting over the requested
prediction horizon.
Use the model capability descriptions to assess each model's suitability
for the current input, including its supported variables, covariates,
missing-value handling, and context length.
Base your selection on the information provided for this case and your
knowledge of time-series forecasting.
Interpret null as an unavailable statistical measurement or a missing
historical observation.
```

Episodic memory instruction. The following instruction is appended when episodic memory is enabled.

```
The retrieved memories provide additional observations from
previous cases, including their historical statistics and the
performance of every evaluated model.
Consider these observations together with the current statistics, model
capabilities, and your time-series forecasting knowledge.
When assessing each available model for the current case, examine
how that model's performance varied across the retrieved cases as
their historical statistics varied.
Use these past relationships to refine how well you expect each model to
perform for the current case.
Lower mase and lower mase_rank indicate better point-forecast
performance.
Lower quantile_loss_rank and lower
mean_weighted_sum_quantile_loss indicate better probabilistic
performance.
Use the probabilistic results as supporting information when the
point-forecast evidence for multiple models is similar.
```

Online policy instruction. Policy Only uses the following instruction.

```
Use the online ranking as additional information when estimating the
performance of each available model for the current case. Consider it
together with the current statistics and the model capabilities. Compare
all six models and select the model with the best expected point-forecast
accuracy.
```

When episodic memory and the online policy are both enabled, FASE uses the following version.

```
Use the online ranking as additional information when estimating the
performance of each available model for the current case. Consider it
together with the current statistics, the retrieved cases, the model
capabilities, and the probabilistic forecasting results. Compare
all six models and select the model with the best expected
point-forecast accuracy.
```

Every system message ends with the following output instruction.

```
Return your selection as a JSON object with the field
"selected_model_id".
```

User template.

```
[STATISTICAL FEATURE DEFINITIONS]

The following statistics describe the historical observations for the
current case.
<statistics>
[CURRENT STATISTICS]
</statistics>

[RETRIEVED MEMORY BLOCK, IF ENABLED]
[ONLINE MODEL RANKING BLOCK, IF ENABLED]

The following forecasting models are available for this prediction.
<available_model_ids>
[AVAILABLE MODEL IDENTIFIERS]
</available_model_ids>

The following model cards describe the supported inputs and
outputs of the forecasting models.
<model_capabilities>
[MODEL CAPABILITY DESCRIPTIONS]
</model_capabilities>
```

Statistical feature definitions. The following guide replaces the first placeholder in the user template. Line breaks are added only for presentation.

```
The following definitions apply to both the current statistics and the
retrieved cases' statistics.
L is task.statistics_history_length, the number of historical time
positions used for statistics, up to the last 15360 positions including
the latest history.
task.forecast_horizon is the number of future positions to predict, and
task.frequency describes the sampling frequency.
task.past_covariate_count describes the number of historical covariate
channels.
Recent trend, level, and scale statistics use the last max(4, L // 4)
positions, limited by the available history.
Recent periodicity uses the last max(12, L // 3) positions, and recent
nonzero activity uses the last max(8, L // 4) positions, each limited by
the available history.
The separately supplied recent_history observations cover the last
forecast_horizon historical positions when that input is enabled.
quality_profile.missing_ratio is the fraction of missing positions across
all historical target values.
quality_profile.zero_ratio is the fraction of observed historical target
values equal to zero.
global_recent_dynamics.global_trend_strength is the fitted historical
slope multiplied by L and divided by the historical standard deviation,
with positive values indicating an upward trend and negative values
indicating a downward trend, summarized by the median across valid
channels.
global_recent_dynamics.recent_trend_strength is the fitted recent slope
multiplied by the recent segment length and divided by the historical
standard deviation, summarized by the median across valid channels.
global_recent_dynamics.recent_level_shift_strength is the absolute
difference between recent and preceding historical means divided by the
historical standard deviation, summarized by the maximum across valid
channels.
global_recent_dynamics.recent_scale_change_strength is the absolute
```

```
difference between recent and preceding historical standard deviations
divided by their sum, summarized by the maximum across valid channels.
change_regime_analysis.change_strength is the largest absolute difference
between means before and after a split in approximately the middle 20% to
80% of history, divided by the historical standard deviation and
maximized across valid channels.
nonlinear_complexity.linear_fit_error_ratio compares the mean squared
residual of a fitted one-step linear model with that of repeating the
previous observation, using adjacent observed pairs and averaging across
valid channels.
nonlinear_complexity.turning_behavior_ratio is the fraction of observed
consecutive triples whose two successive changes have different signs,
including zero as a sign and averaging across valid channels.
nonlinear_complexity.direction_entropy is the binary entropy of upward
versus non-upward changes in adjacent observed pairs, averaged across
valid channels.
dependence_screen.spectral_concentration is the fraction of
nonzero-frequency energy in the three strongest frequency bins after
removing a quadratic trend from complete historical channels, averaged
across valid channels.
background.entropy is normalized power-spectrum entropy of standardized
complete nonconstant historical channels, with higher values indicating
more dispersed spectral energy, summarized by the median across valid
channels.
background.stability is the sample variance of block means in
standardized complete nonconstant historical channels, with higher values
indicating greater changes in historical mean levels, summarized by the
median across valid channels.
The block width for background.stability is task.seasonal_period when
greater than 1, otherwise 10, and at least two complete blocks are
required.
missingness_structure.longest_missing_block is the longest consecutive
historical missing block across target channels, measured in time steps
and provided when missing values occur.
time_varying_periodicity.recent_period_strength is the strongest
nonzero-frequency energy fraction in the recent periodicity segment after
quadratic detrending, averaged across valid complete channels and
provided when L is at least 36 and a candidate period is detected.
covariate_relationship.strongest_covariate_relation is the maximum
absolute Pearson correlation between historical targets and covariates
over time-aligned observed pairs, provided when covariates are present.
covariate_relationship.covariate_relation_stability compares
target-covariate correlations in the first and second halves of history
using max(0, 1 - abs(first-half correlation - second-half correlation)),
averaged across valid target-covariate pairs.
intermittent_structure.recent_event_rate_change is the absolute
difference between recent and whole-history observed nonzero fractions,
maximized across valid channels and provided when L is at least 8 and the
historical zero ratio is at least 0.45.
null denotes an unavailable statistical value, and failed_tool_ids
identifies statistical tools whose computation failed.
```

Model capability descriptions. The following object replaces the model capability placeholder in every controller variant.

```
{
  "seasonal_naive": {
    "multivariate_input": "independent_channels",
    "past_covariates": "ignored",
    "future_covariates": "ignored",
    "missing_values": "requires_linear_interpolation",
    "context_limit": {
      "type": "no_adapter_limit",
      "maximum_history_points": null
    }
```

```
  },
  "chronos2": {
    "multivariate_input": "joint_channels",
    "past_covariates": "ignored",
    "future_covariates": "ignored",
    "missing_values": "native",
    "context_limit": {
      "type": "finite",
      "maximum_history_points": 8192
    }
  },
  "timesfm_2p5": {
    "multivariate_input": "independent_channels",
    "past_covariates": "ignored",
    "future_covariates": "ignored",
    "missing_values": "requires_linear_interpolation",
    "context_limit": {
      "type": "finite",
      "maximum_history_points": 15360
    }
  },
  "moirai_2_small": {
    "multivariate_input": "independent_channels",
    "past_covariates": "ignored",
    "future_covariates": "ignored",
    "missing_values": "native",
    "context_limit": {
      "type": "finite",
      "maximum_history_points": 4000
    }
  },
  "tirex2_gifteval_zs": {
    "multivariate_input": "joint_channels",
    "past_covariates": "consumed",
    "future_covariates": "consumed",
    "missing_values": "native",
    "context_limit": {
      "type": "finite",
      "maximum_history_points": 8192
    }
  },
  "toto_2_313m": {
    "multivariate_input": "joint_channels",
    "past_covariates": "consumed",
    "future_covariates": "ignored",
    "missing_values": "native",
    "context_limit": {
      "type": "finite",
      "maximum_history_points": 4096
    }
  }
}
```


The optional episodic memory and online policy blocks are formatted as follows.

```
The following retrieved cases contain historical statistics and observed
model performance.
<memories>
[RETRIEVED CASES AND MODEL OUTCOMES]
</memories>

<online_model_ranking>
The following ranking is generated by an online model ranker. It is
updated over time using the observed point-forecast performance of
the six models on previously completed cases from this task.
```

```
It provides an additional estimate of the relative point-forecast
performance of the six models for the current case:

[ORDERED MODEL IDENTIFIERS]
</online_model_ranking>
```

F PER-CONFIGURATION RESULTS

Tables 8 and 9 report normalised MASE and CRPS for each evaluated configuration. All values are normalised by the corresponding score of Seasonal Naive, which is therefore omitted from the tables.

Table 8: Normalised MASE for each evaluated GIFT-Eval configuration. Lower values indicate better point forecasting performance.

Dataset	Setting	Chronos-2	TimesFM 2.5	Moirai-2.0-R-Small	TiRex-2	Toto-2.0-313m	Uniform Ensemble	**FASE**
bitbrains_fast_storage	5T / medium	0.777	0.815	0.813	0.797	0.751	0.777	0.732
	5T / long	0.762	0.795	0.807	0.790	0.719	0.758	0.709
	H / short	0.754	0.818	0.822	0.823	0.717	0.774	0.723
bitbrains_rnd	5T / medium	0.959	0.969	0.969	0.967	0.949	0.958	0.948
	5T / long	0.933	0.948	0.952	0.945	0.926	0.934	0.921
	H / short	0.964	0.969	0.962	0.979	0.956	0.962	0.950
bizitobs_service	10S / short	0.597	0.593	0.713	0.625	0.581	0.591	0.572
car_parts_with_missing	M / short	0.694	0.698	0.688	0.695	0.729	0.696	0.699
electricity	D / short	0.692	0.704	0.715	0.707	0.718	0.683	0.685
	H / medium	0.761	0.760	0.779	0.763	0.766	0.744	0.744
	H / long	0.781	0.752	0.828	0.769	0.798	0.763	0.756
	W-FRI / short	0.680	0.716	0.794	0.727	0.775	0.713	0.695
hierarchical_sales	D / short	0.654	0.657	0.659	0.656	0.660	0.652	0.662
hospital	M / short	0.839	0.824	0.831	0.830	0.817	0.816	0.825
kdd_cup_2018_with_missing	D / short	0.806	0.796	0.829	0.792	0.796	0.797	0.787
	H / medium	0.780	0.726	0.788	0.739	0.735	0.724	0.717
	H / long	0.804	0.755	0.802	0.759	0.767	0.750	0.746
LOOP_SEATTLE	5T / long	0.785	0.676	0.742	0.653	0.661	0.665	0.607
	D / short	0.516	0.498	0.516	0.500	0.510	0.498	0.489
	H / medium	0.607	0.540	0.635	0.499	0.616	0.563	0.499
	H / long	0.560	0.532	0.585	0.497	0.567	0.534	0.496
m4_daily	D / short	1.031	1.006	0.937	1.003	0.970	0.969	0.941
M_DENSE	H / short	0.526	0.517	0.533	0.527	0.509	0.510	0.481
restaurant	D / short	0.678	0.678	0.692	0.673	0.689	0.667	0.676
solar	10T / short	0.720	0.982	0.595	0.931	0.776	0.741	0.466
	10T / medium	0.855	0.908	1.078	0.889	0.956	0.870	0.709
	10T / long	0.916	0.979	1.161	0.928	0.899	0.918	0.767
	H / short	1.035	0.943	0.924	0.435	0.924	0.788	0.425
SZ_TAXI	15T / short	0.711	0.712	0.715	0.712	0.709	0.708	0.711

Table 9: Normalised CRPS for each evaluated GIFT-Eval configuration. Lower values indicate better probabilistic forecasting performance.

Dataset	Setting	Chronos-2	TimesFM 2.5	Moirai-2.0-R-Small	TiRex-2	Toto-2.0-313m	Uniform Ensemble	**FASE**
bitbrains_fast_storage	5T / medium	0.519	0.634	0.579	0.497	0.479	0.514	0.607
	5T / long	0.636	0.706	0.685	0.584	0.500	0.593	0.666
	H / short	0.798	0.756	0.601	0.654	0.572	0.651	0.679
bitbrains_rnd	5T / medium	1.015	0.722	0.508	0.533	0.472	0.640	0.533
	5T / long	0.861	0.632	0.484	0.505	0.454	0.573	0.881
	H / short	0.830	0.490	0.538	0.480	0.472	0.546	0.548
bizitobs_service	10S / short	0.264	0.255	0.339	0.278	0.257	0.261	0.244
car_parts_with_missing	M / short	0.558	0.547	0.544	0.553	0.634	0.559	0.595
electricity	D / short	0.528	0.519	0.520	0.524	0.544	0.505	0.492
	H / medium	0.586	0.577	0.631	0.586	0.605	0.568	0.571
	H / long	0.570	0.531	0.635	0.548	0.594	0.545	0.542
	W-FRI / short	0.478	0.525	0.669	0.507	0.565	0.520	0.510
hierarchical_sales	D / short	0.331	0.331	0.333	0.333	0.335	0.329	0.348
hospital	M / short	0.863	0.811	0.829	0.827	0.791	0.801	0.832
kdd_cup_2018_with_missing	D / short	0.565	0.562	0.583	0.558	0.566	0.563	0.581
	H / medium	0.597	0.563	0.666	0.571	0.576	0.567	0.562
	H / long	0.510	0.480	0.566	0.476	0.493	0.484	0.479
LOOP_SEATTLE	5T / long	0.672	0.597	0.685	0.569	0.571	0.575	0.531
	D / short	0.417	0.396	0.418	0.405	0.414	0.401	0.392
	H / medium	0.389	0.342	0.427	0.306	0.388	0.353	0.308
	H / long	0.318	0.305	0.351	0.275	0.319	0.300	0.276
m4_daily	D / short	0.927	0.891	0.823	0.881	0.852	0.853	0.830
M_DENSE	H / short	0.466	0.465	0.480	0.476	0.460	0.456	0.435
restaurant	D / short	0.387	0.388	0.394	0.385	0.394	0.381	0.387
solar	10T / short	0.467	0.654	0.408	0.615	0.494	0.478	0.454
	10T / medium	0.445	0.520	0.630	0.495	0.510	0.464	0.442
	10T / long	0.431	0.491	0.606	0.455	0.445	0.439	0.429
	H / short	0.587	0.569	0.579	0.261	0.536	0.461	0.307
SZ_TAXI	15T / short	0.649	0.650	0.652	0.650	0.647	0.645	0.648