EpiEvolve: Self-Evolving Agents for Streaming Pandemic Forecasting under Regime Shifts

Yiming Lu[1], **Sihang Zeng**[2], **Zhengxu Tang**[1], **Max Lau**[1], **Fei Liu**[1], **Wei Jin**[1]
[1]Emory University [2]University of Washington
{yiming.lu, wei.jin}@emory.edu

Abstract

Epidemic LLM forecasters are usually trained and evaluated as static supervised models, whereas operational pandemic forecasting is a streaming process in which labels arrive after predictions and disease regimes shift over time. We study this mismatch in weekly COVID-19 hospitalization trend forecasting across five variant regimes. We introduce EpiEvolve, a self-evolving agent that wraps an LLM forecaster trained on the warm-start period and keeps its weights fixed during streaming. EpiEvolve adapts by storing forecast outcomes in a hierarchical episodic memory, reflecting on delayed labels, retrieving cases relevant to the current regime, and distilling recurring errors into strategic rules. The resulting context lets the forecaster reuse its own past predictions and outcomes in later weeks while following a chronological protocol that prevents future leakage. On the streaming dataset, EpiEvolve reaches 0.629 average accuracy, compared with 0.561 for the static backbone and 0.325 for the external CDC ensemble, and reduces recovery lag after regime shifts from 5 to 2 weeks. Ablations show that reflection, strategic memory, and regime-aware retrieval each contribute to the gains.

1 Introduction

Infectious disease forecasting is central to the response to pandemics, where decisions are made under tight time and resource constraints. Useful forecasts must connect epidemiological trajectories with contextual signals about geography, policy, and viral evolution. Recent epidemic LLM forecasters demonstrate that language models are a promising reasoning interface for this setting, capable of representing both numerical trends and unstructured textual evidence within a unified prediction pipeline (Du et al., 2025; Gong et al., 2025; Li et al., 2025). However, while these systems establish that LLMs can be adapted to epidemic forecasting, they are usually trained and evaluated as static supervised models.

This static evaluation paradigm misses the central property of real-world forecasting that the environment changes after deployment. Public health forecasts are issued iteratively, ground truth arrives only after the predictions are fixed, and the data generating process can shift when new pathogen lineages, immunity profiles, or policy conditions emerge (Cramer et al., 2022; Reich et al., 2022). COVID-19 made this instability visible as successive variants altered transmission and immune escape across regions, so evidence that was predictive in one period could become less reliable in the next (Pham et al., 2025; Wang et al., 2023). Consequently, a static model trained on a historical window can fail even when it receives rich inputs. Without mechanisms to adapt to or learn from the errors, these failures may persist. The critical question is whether an LLM forecaster can recover once the epidemic enters a shifted epidemiological regime.

Answering this question requires changes to both evaluation and modeling. On the evaluation side, delayed labels impose a strict chronological protocol. A method must issue forecasts using only information available at that week, and it may update its state after the corresponding labels arrive. This constraint prevents future leakage and makes post-shift performance recovery measurable. On the modeling side, regime shifts change the relevance of historical evidence. Recent errors may be informative after a shift, while older examples can mislead even when they appear superficially relevant. Streaming fine-tuning can update parameters, and retrieval can surface past cases, but neither by itself specifies how forecast outcomes should be organized, reflected on, and converted into reusable strategies. This motivates an agentic structure that remembers past forecasts, interprets outcome feedback, and chooses evidence for the next prediction.

To address these challenges, we propose EpiEvolve, a self-evolving agent for streaming epidemic forecasting. EpiEvolve treats delayed forecast errors after deployment as the primary signal for adaptation. Instead of updating model parameters, it enables a frozen epidemic LLM backbone to adapt through hierarchical episodic memory, retrieval conditioned on epidemic regimes, and lesson distillation from outcomes. This design converts forecast errors into reusable lessons, allowing the agent to recover from distribution shifts without costly gradient updates.

We evaluate EpiEvolve on weekly COVID hospitalization trend forecasting across five COVID variant regimes, using a chronological streaming setup that reflects delayed outcome feedback in real deployment. To our knowledge, EpiEvolve is the first self-evolving LLM agent for streaming epidemic forecasting under regime shifts. Our contributions are:

- **A self-evolving epidemic forecasting agent.** We introduce EpiEvolve, a streaming LLM agent that adapts after deployment through hierarchical episodic memory, regime-conditioned retrieval, and outcome-informed lesson distillation while keeping the forecasting backbone frozen.
- **A streaming evaluation with delayed feedback.** We formulate COVID hospitalization trend forecasting as a chronological prediction problem across recurring variant regimes, measured by both average accuracy and performance recovery after distribution shifts.
- **An empirical analysis of adaptation mechanisms.** We compare EpiEvolve with static forecasting, retrieval-based memory, reflection-based memory, external LLM in context forecasting, and streaming fine-tuning. EpiEvolve improves average accuracy from 0.561 to 0.629 over a static frozen backbone and reduces recovery lag from 5 to 2 weeks, while ablations isolate the contribution of memory-based self-evolution.

2 Related Work

Epidemic Forecasting, Pandemic LLMs, and Regime Shifts. Operational epidemic forecasting combines iterative prediction with delayed scoring, while the biological, behavioral, and policy context shifts between cycles. The U.S. COVID-19 Forecast Hub and collaborative hubs document both the value of ensembles and the instability of individual models (Cramer et al., 2022; Reich et al., 2022; Howerton et al., 2023). This is a form of concept drift (Gama et al., 2014) now flagged from neural-representation statistics (Greco et al., 2025; Ayers et al., 2025), yet in epidemic streams the cues often surface first in text, such as variant announcements and policy updates. Recent epidemic LLM forecasters unify these signals in a single language interface (Du et al., 2025; Gong et al., 2025; Li et al., 2025), and a parallel line uses LLM agents as planning or policymaking assistants under outbreak scenarios (Mao et al., 2025; Shi et al., 2026; Aoki and Ghaffarzadegan, 2026). Across these settings, however, methods are trained or prompted on a historical window or operate inside a simulator, leaving open how a deployed forecaster should adapt to its own delayed errors across regime shifts.

Self-Evolving LLM Agents and Memory. An emerging line of LLM agent research treats accumulated experience as the unit of adaptation, replacing test-time parameter updates (Wang et al., 2025a; Hu et al., 2025; Chen et al., 2026; Zheng et al., 2025) with memory of past outcomes. Recent systems consolidate verbal reflections into reusable rules (Wu et al., 2025a), distill agent trajectories into retrievable principles (Wu et al., 2025b), and benchmark such memory under continuous test-time streams (Wei et al., 2025). A parallel direction treats the memory store itself as the object of learning (Zhang et al., 2025a,b), including procedural memory from long-horizon trajectories (Wang et al., 2025b) and shared experience across cooperating agents (Weng et al., 2026). These systems mostly target episodic tasks with immediate ground truth; temporal forecasting under delayed labels (Csaba et al., 2024) and recurring regime shifts (Wu et al., 2024) lie outside their evaluation scope, yet the design pattern of a frozen backbone with evolving memory transfers naturally to it.

LLMs and Foundation Models for Time Series. A complementary direction adapts LLM and transformer architectures directly to numerical time-series prediction, either by reprogramming continuous inputs for a frozen language backbone (Jin et al., 2023), tokenizing the series into a completion vocabulary (Ansari et al., 2024), or augmenting sequence modeling with spatial-aware reinforcement learning (Ni et al., 2026) and epidemiology-aware neural ODE dynamics (Wan et al., 2025). This line designs the backbone and representation rather than post-deployment adaptation; the resulting forecasters are precisely the kind of frozen backbone that a self-evolving agent can wrap.

3 Problem Setting

3.1 Streaming Forecasting Task

We study epidemic forecasting as a streaming classification problem over geographical regions. Let $\mathcal{S}$ denote the set of regions and let $t = 1, \ldots, T$ index weekly forecasting rounds. For each region $s \in \mathcal{S}$ and week t, the model observes $x_{s,t}$, which contains only information available before the forecast is issued, including recent epidemic time series, surveillance signals, policy text, and regional context. The target is a trend label $y_{s,t} \in \mathcal{Y}$ from a finite ordinal label set.

At week t, the forecaster uses a deployment memory $\mathcal{M}_t$ containing only information available before week t, such as past cases, textual reflections, or distilled lessons. For state s, a retrieval operator R converts this memory and the current input $x_{s,t}$ into a non-parametric context

$$r_{s,t} = R(\mathcal{M}_t, x_{s,t}).$$

The forecast is then produced as

$$\hat{y}_{s,t} = f_{\theta_t}(x_{s,t}, r_{s,t}), \qquad \hat{y}_{s,t} \in \mathcal{Y}. \tag{1}$$

This notation covers three cases: static baselines, where both θ_t and $r_{s,t}$ are fixed; memory-based agents, where adaptation occurs through updates to $\mathcal{M}_t$ and retrieval context; and streaming fine-tuning baselines, where adaptation occurs through updates to θ_t.

The defining constraint is the delayed feedback. At week t, forecasts for all regions must be generated using only current parameters θ_t and deployment memory $\mathcal{M}_t$. Labels are revealed only after the entire weekly batch has been fixed:

$$\mathcal{B}_t = \{(x_{s,t}, \hat{y}_{s,t}, y_{s,t}) : s \in \mathcal{S}\}. \tag{2}$$

The method may update its state for the next round:

$$(\theta_{t+1}, \mathcal{M}_{t+1}) = U(\theta_t, \mathcal{M}_t, \mathcal{B}_t). \tag{3}$$

This ordering prevents leakage within the week by ensuring that the results of week t cannot inform the predictions of other regions in the same week. EpiEvolve satisfies this constraint with a pretrained backbone, $\theta_{t+1} = \theta_t$, and adapts through non-parametric updates to memory, retrieval, and lessons. Streaming fine-tuning baselines may instead update θ_t after the weekly batch is complete.

3.2 Regime Evaluation and Data Instantiation

We instantiate this streaming strategy on weekly COVID-19 hospitalization trend forecasting across the 50 U.S. states. Before streaming begins, each method receives a warm-start period for offline preparation, such as training or initializing the backbone model. Each weekly state-level example combines spatial attributes, epidemiological time series, public health policy text, vaccination signals, and genomic surveillance information (Du et al., 2025). The prediction target is a five-class ordinal hospitalization trend label: substantial decreasing, moderate decreasing, stable, moderate increasing, or substantial increasing. The details of implementation are provided in Section 5.1.

To assess adaptation under the distribution shift, we partition the streaming period into contiguous regimes, such as dominant variant eras. Let $\mathcal{T}_r$ denote the set of weeks in regime r. We report regime-level accuracy:

$$\mathrm{Acc}(r) = \frac{1}{|\mathcal{T}_r|\,|\mathcal{S}|} \sum_{t \in \mathcal{T}_r} \sum_{s \in \mathcal{S}} \mathbf{1}[\hat{y}_{s,t} = y_{s,t}]. \tag{4}$$

These regimes are post-hoc partitions used only for reporting and are never supplied to the forecaster. The variant text that appears in $x_{s,t}$ is part of the observable input at forecast time, sourced from public surveillance. We additionally examine recovery around regime transitions, since aggregate accuracy may obscure whether a method adapts quickly after a distributional shift.

4 Method: EpiEvolve

We introduce EpiEvolve, a self-evolving agent with external memory for epidemic forecasting. Figure 1 illustrates the overall framework. We organize the discussion as follows: Section 4.1 introduces the forecasting backbone. Sections 4.2 through 4.4 detail the hierarchical episodic memory, the reflection module responsible for memory updates, and the strategic distillation process. Section 4.5 describes the drift detector and the retrieval policy. Finally, Section 4.6 integrates these components into a complete streaming pipeline.

4.1 Forecasting Backbone

EpiEvolve instantiates the streaming protocol of Section 3 by fixing the backbone parameters and routing all adaptation through the deployment memory $\mathcal{M}_t$. We fine-tune an LLM-based forecaster f_θ on the warm-start period and keep its

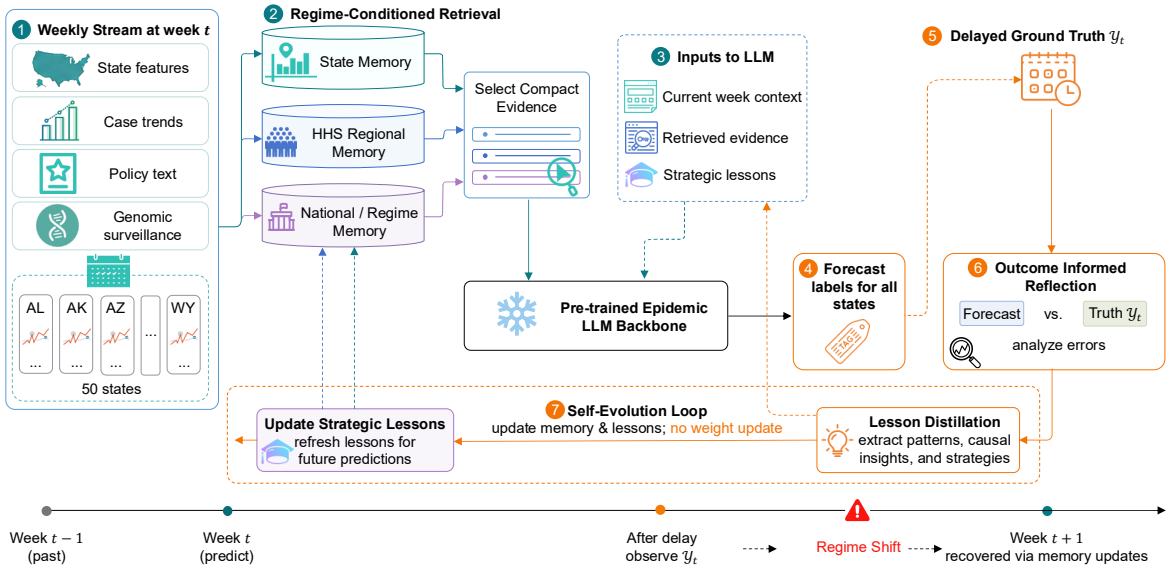


Figure 1: **Overview of EpiEvolve Pipeline.** Each week t, (1) state-level features (epi trends, policy, genomic surveillance) drive (2) regime-conditioned retrieval over episodic memory (state, regional, national), (3) feeding the retrieved cases, distilled lessons, and current context to a pre-trained LLM backbone for (4) hospitalization trend class forecasts. (5) When delayed truth y_t arrives, (6) reflection writes a new episodic entry and triggers lesson distillation. Recovery after regime shifts comes entirely from memory and rule updates.

parameters fixed throughout streaming evaluation, so $\theta_{t+1} = \theta_t$ at every cycle of Equation 3; the architecture and training procedure are described in Section 5.1. Withholding gradient updates at deployment is a deliberate design choice. It isolates memory adaptation as the sole driver of the improvement we report, and matches scenarios where backbone weights cannot be touched after release.

Therefore, all adaptation enters f_θ through the prompt. We extend the prompt template with two structured slots, `<MEMORY>` and `<RULES>`, whose contents constitute the retrieved context $r_{s,t}$. When both slots are empty, f_θ reduces to a static baseline; when populated, it conditions on retrieved evidence without changing its parameters.

4.2 Hierarchical Episodic Memory

The deployment memory $\mathcal{M}_t = (\mathcal{E}_t, \mathcal{L}_t, \rho_t)$ combines a hierarchical episodic store of past forecast outcomes, a strategic memory of distilled rules, and a regime indicator from the drift detector. This subsection describes $\mathcal{E}_t$ and the part of the retrieval operator R that draws from it; Sections 4.4 and 4.5 describe $\mathcal{L}_t$ and ρ_t.

After a regime shift, errors arrive faster than gradients can absorb them, yet cluster systematically across regions with similar epidemiological context. EpiEvolve records each forecast outcome as an episodic entry

$$e = (s,\ t,\ \phi_{s,t},\ \hat{y}_{s,t},\ y_{s,t},\ \ell_{s,t},\ \rho_t), \tag{5}$$

where $\phi_{s,t}$ embeds recent hospitalization and case trends, vaccination summary, dominant-variant indicator, and policy text. The term $\ell_{s,t}$ is a one-sentence reflection from Section 4.3, and ρ_t is the regime indicator at write time. Relative to a query at state s, $\mathcal{E}_t$ partitions into a state view $\mathcal{E}_t^{\mathrm{S}}(s)$ (entries about s), a regional view $\mathcal{E}_t^{\mathrm{R}}(s)$ (other states in the same HHS region), and a national view $\mathcal{E}_t^{\mathrm{N}}(s)$ (other HHS regions).

Retrieval pulls the top-N entries from $\mathcal{E}_t$ under a regime-aware similarity score

$$\mathrm{score}(e,\ s,\ t) = \cos(\phi_{s,t},\ \phi_e) \cdot w(\rho_t,\ \rho_e), \tag{6}$$

where $w(\rho_t, \rho_e) = 1$ if $\rho_t = \rho_e$ and α otherwise, with $\alpha = 0.5$ fixed throughout. The selected entries are rendered into the `<MEMORY>` slot. The three scopes define how the store is organized, not separate retrieval quotas. Early in a new regime, for example, the current state may have no entries from the same regime, so the ranking can fall back to broader regional or national cases that still match the current features. Section 5.4 ablates this design by restricting retrieval to a single tier.

Method	Adaptation strategy	Delta / early	BA.1	BA.2	BA.5	BQ.1	Avg.	Recovery lag
CDC ensemble	None (external)	0.393	0.458	0.469	0.139	0.165	0.325	N/A
PandemicLLM	None	0.587	0.713	0.642	0.431	0.456	0.561	5
Streaming fine-tune	Gradient update	0.587	0.700	0.625	0.460	0.490	0.566	5
Claude-ICL	In-context only	0.563	0.624	0.561	0.479	0.447	0.557	5
Retrieval-only	Episodic retrieval	0.602	0.718	0.663	0.488	0.512	0.591	4
Reflection + episodic	Reflection memory	0.611	0.728	0.672	0.523	0.541	0.604	3
EpiEvolve	Reflection + strategic memory	**0.624**	**0.751**	**0.687**	**0.563**	**0.582**	**0.629**	**2**

Table 1: **Main regime-wise streaming forecasting results.** Regime columns report hospitalization trend classification accuracy for each variant-era slice; Avg. aggregates across the full period; Recovery lag measures how quickly each method recovers after regime transitions. Lower recovery lag is better.

4.3 Reflection-Driven Memory Writing

Episodic entries are useful only when their content is selective and labeled with the correct correction signal. Once the labels for week t arrive in $\mathcal{B}_t$, a reflection module takes $(x_{s,t},\, \hat{y}_{s,t},\, y_{s,t})$ as input and emits both the one-sentence reflection $\ell_{s,t}$ and a candidate rule fragment. The reflection prompt asks the model to identify which features supported the forecast, which should have shifted it toward the true class, and which regime-conditioned cue the error coincides with. The new entry is appended once to $\mathcal{E}_t$, and the candidate rule is queued for the strategic distiller. The fixed schema keeps the memory store compact and prevents reflection text from dominating later prompts as the stream grows.

4.4 Strategic Lesson Distillation

Episodic entries are concrete but fragile because a single trajectory may not transfer across regions or regimes. We promote recurrent reflection patterns into a strategic memory $\mathcal{L}_t$ of predicate-form rules

$$\lambda = (\text{preconds},\, \text{consequent},\, c,\, n,\, \rho_\lambda), \qquad (7)$$

where the preconditions are a conjunction of feature predicates and the consequent names a target class shift. The support count n is the number of forecast weeks in which the preconditions have matched a region-week observation, and the confidence $c \in [0, 1]$ is the empirical fraction of those weeks in which the consequent agreed with the truth, updated on each new match as $c \leftarrow (nc + \mathbf{1}[\text{consequent matched truth}])/(n{+}1)$. The regime tag ρ_λ records the regime in which the rule was first distilled. New rules enter $\mathcal{L}_t$ with n initialized to the number of distillation-window entries that satisfied the preconditions and c to the corresponding in-window precision. The distiller runs on a sliding window of recent episodic entries every K weeks and on every drift event from Section 4.5. At each forecast call, rules whose preconditions match $x_{s,t}$ with confidence above a threshold are rendered into the <RULES> slot. Partially matched rules appear as soft hints.

4.5 Drift Detector and Retrieval

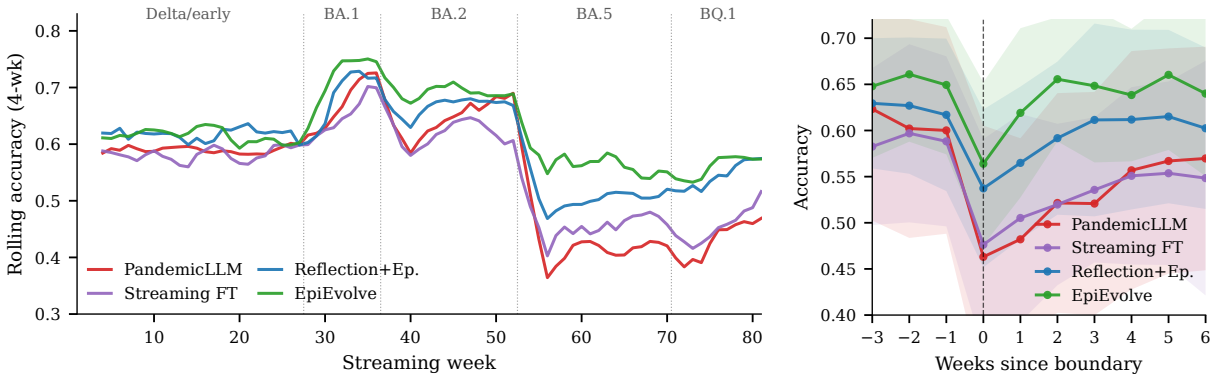
Adaptation must respond when the environment changes faster than memory accumulates new evidence. In the post-batch update step, the drift detector evaluates two triggers: (i) the weekly cross-region average ordinal error from $\mathcal{B}_t$ exceeds the warm-start baseline mean by more than τ_σ standard deviations, or (ii) the dominant-variant field in $x_{s,t}$, observable from public surveillance, names a new variant relative to the previous week. Neither trigger reads the evaluation regime labels of Section 3. On a drift event, the regime indicator advances from ρ_t to ρ_{t+1} and strategic distillation runs on the most recent entries of the previous regime. The new ρ_{t+1} biases the retrieval score in Equation 6 for the next forecast cycle and gates $\mathcal{L}_t$ so that obsolete rules are demoted but not deleted.

4.6 Streaming Agent Loop

The week-t procedure ties the components together. EpiEvolve enters the week with deployment memory $\mathcal{M}_t$. For every region $s \in \mathcal{S}$, the retrieval operator R assembles $r_{s,t}$ from $\mathcal{E}_t$ and $\mathcal{L}_t$ under ρ_t, the backbone produces $\hat{y}_{s,t} = f_\theta(x_{s,t},\, r_{s,t})$, and forecasts for the entire week are emitted before any label is observed, which prevents one region's truth from leaking into another region's prediction in the same week. Once the batched labels $\mathcal{B}_t$ arrive, the reflection module writes new episodic entries, the strategic distiller (if triggered) updates $\mathcal{L}_t$, and the drift detector decides whether to advance the regime indicator from ρ_t to ρ_{t+1}, yielding $\mathcal{M}_{t+1}$.

5 Experiments

We evaluate EpiEvolve as a deployment-time adaptation method for streaming epidemic forecasting.



(a) Adaptation curve over the variant-era stream.

(b) Boundary-centered recovery.

Figure 2: **Adaptation behavior across the variant-era stream.** (a) Rolling 4-week accuracy across the evaluation period for four representative methods; the static backbone collapses at BA.5 and stays depressed through BQ.1, while EpiEvolve dips less at every boundary and returns to its new-regime steady-state level fastest. (b) Boundary-centered recovery: weekly accuracy aligned by weeks since each transition and averaged across the four boundaries; shaded bands show across-boundary standard deviation.

The experiments are organized around three questions: whether memory-based self-evolution improves forecasting under the same chronological information constraint, whether it improves recovery after variant regime transitions, and which mechanisms are responsible for the gains.

5.1 Experimental Setup

Data stream. We use a weekly COVID-19 stream that forecasts hospitalization-trend classes for 50 U.S. states. The warm-start period covers data through 2021-05-31 and is used for fine-tuning the backbone. Streaming evaluation runs from 2021-06-07 to 2022-12-19, giving 81 weekly rounds and 4,050 state-week forecasts. The stream is partitioned into five variant regimes: Late-Alpha to Delta, Omicron BA.1, BA.2, BA.5, and BQ.1.

Metrics. The primary metric is hospitalization-trend classification accuracy over the full stream and within each regime. We additionally report mean squared error on the ordinal label space, a boundary recovery curve, and the recovery lag. For each variant boundary at week t, let $\bar{A}_r$ denote the method's mid-regime steady-state accuracy in the new regime r, computed as the mean weekly accuracy from week $t+4$ to the regime end (after the boundary dip has subsided). Recovery lag is the smallest $k \geq 1$ for which the weekly accuracy first reaches $\bar{A}_r$. We report the average across the four variant boundaries and mark a method N/A when the new regime exhibits sustained collapse rather than a transient dip, defined as $\bar{A}_r$ falling below half of the pre-transition rolling 4-week accuracy at any boundary.

Methods. All methods obey the delayed feedback protocol and share the same backbone (except Claude-ICL and CDC) fine-tuned on the warm-start period following the PandemicLLM setup of Du et al. (2025). Baselines include PandemicLLM without streaming updates, a streaming fine-tuning baseline with periodic gradient updates, an in-context learning forecaster using Claude, the COVID-19 Forecast Hub ensemble of Cramer et al. (2022) mapped to our five-class hospitalization-trend target, and the retrieval-only and reflection-only variants of EpiEvolve. We evaluate the backbone with Qwen3-14B-Base (Appendix A).

5.2 Main Results

Table 1 presents the performance comparison. EpiEvolve outperforms all baselines on both average accuracy and the post-shift regimes where forecasting is hardest. Table 1 reports regime-wise accuracy and recovery lag for each method. PandemicLLM is strong through Delta and BA.1 but collapses when BA.5 emerges in mid-2022, dropping from 0.713 at BA.1 to 0.431 at BA.5 and staying below 0.46 through BQ.1. The external CDC ensemble baseline trails all LLM-based methods, averaging 0.325 and collapsing below 0.17 on BA.5 and BQ.1. EpiEvolve attains the highest average accuracy (0.629 vs. 0.561) and the lowest recovery lag (2 vs. 5 weeks), with its largest margin in the BA.5 and BQ.1 regimes where every other method also struggles. Claude-ICL is comparable to PandemicLLM but below EpiEvolve, indicating

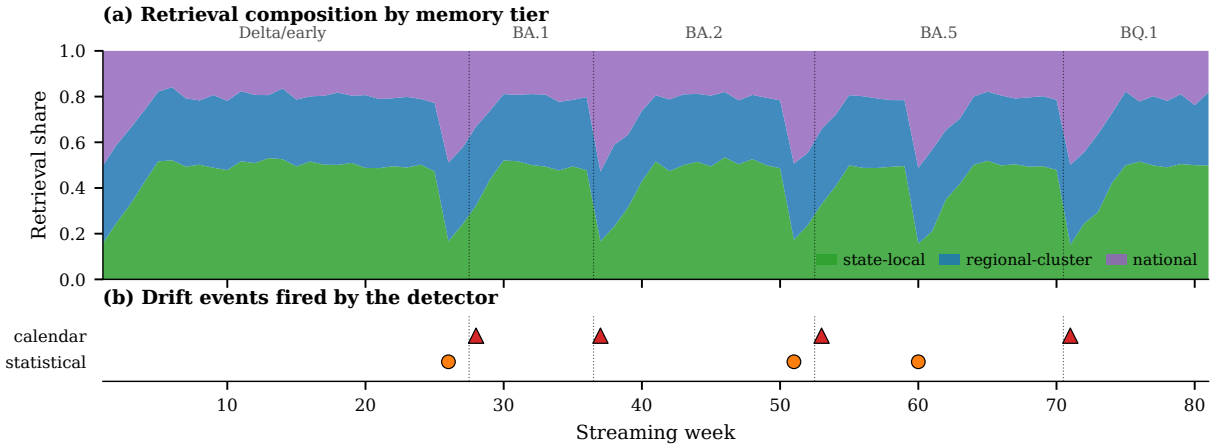


Figure 3: **EpiEvolve's internal state over the variant regime stream.** Panel (a) shows the memory tier of each top N retrieved entry over time. Within a stable regime, entries from the same state gradually account for more of the retrieved context. At variant transitions, the current regime has few local entries, so retrieval shifts toward regional and national cases with similar features. Panel (b) shows drift events. Triangles denote updates triggered by a change in the dominant variant field of $x_{s,t}$, and circles denote statistical triggers from the rolling error signal.

EpiEvolve's gain comes from the memory architecture rather than the access to the proprietary LLM.

5.3 Regime Recovery Analysis

The time view confirms that EpiEvolve's gains come from post-shift behavior, not from averaging across uneventful weeks. Figure 2a plots rolling 4-week accuracy across the streaming weeks: PandemicLLM tracks BA.1 and BA.2 well but drops at BA.5 and never recovers within the BA.5 or BQ.1 windows, while EpiEvolve dips less at every boundary and rises back within two weeks. Figure 2b aligns predictions by weeks since each boundary and averages across the transitions; PandemicLLM falls to roughly 0.45 and barely moves over the next six weeks, while EpiEvolve drops to about 0.58 and stabilizes at its new-regime steady-state level by week $+2$. Streaming fine-tuning and reflection-only memory sit between these extremes, consistent with their partial gains in Table 1.

5.4 Ablation Study

Table 2 reports two kinds of ablations: removing one component of EpiEvolve at a time, and restricting episodic retrieval to a single tier. The largest accuracy drop comes from removing reflection, which leaves no textual lesson to retrieve and also disables the strategic distiller, collapsing the configuration to a retrieval-only baseline; ordinal MSE rises by 0.14, indicating that the missing reflection also lets the remaining errors land further from ground truth. Decomposing against the -0.025 from removing only strategic memory isolates the reflection text's own contribution to -0.013; strategic rules and reflection text therefore split the joint mechanism's gain roughly $2/3$ to $1/3$. Removing strategic memory or restricting retrieval to flat top-N (no regime-aware weighting) each cost between 0.02 and 0.03 in accuracy and similar amounts in MSE. The drift detector is different: removing it shifts recovery lag by one week but barely moves MSE, since the agent eventually adapts to the new regime. The matched comparison against streaming fine-tuning under the same delayed feedback constraint shows that prompt and memory adaptation can match or exceed explicit parameter updates while leaving the backbone weights immutable.

Configuration	Acc. ↑	MSE ↓	Lag ↓
EpiEvolve (full)	**0.629**	**0.65**	**2**
Component ablations			
– strategic memory	0.604	0.71	3
– reflection (retrieval only)	0.591	0.79	4
– drift detector	0.612	0.68	3
– regime-aware retrieval	0.604	0.73	3
Memory scope (single tier)			
state only	0.598	0.74	4
regional only	0.602	0.70	3
national only	0.589	0.78	4

Table 2: Component and memory scope ablations. Each configuration removes or replaces one piece of EpiEvolve while keeping the same backbone and delayed feedback protocol.

Restricting episodic retrieval to a single memory tier costs between -0.028 and -0.041 in average accuracy, and an MSE penalty of similar mag-

Case study: Florida, week of 2022-06-13 (second week of the BA.5 regime).

Input (excerpts; full template in Appendix F).

Variant: "BA.5 emerging since 2022-06-06, rapidly displacing BA.2, with higher antibody evasion than prior subvariants."

Trend (last five weeks): *stable, stable, moderate decreasing, stable, stable*.

Dynamic: Vaccination 67%; no statewide mask mandate.

<MEMORY> (3 of N=8 entries retrieved by regime-aware score)

• FL, 2021-12-20 (ρ= BA.1; cross-regime same state): *stable* → *moderate increasing*. "BA.1 emergence + vaccination plateau; hospitalizations rose despite two stable weeks."

• GA, 2022-06-06 (ρ= BA.5; regional analogue, HHS-4): *moderate increasing* → truth confirmed. "HHS-4 neighbor; BA.5 uptick matched the variant-emergence cue."

• National, 2022-02-21 (ρ= BA.2; cross-regime cross-state): *stable* → *moderate increasing*.

<RULES> (1 of 2 matched rules)

• *IF* the variant text mentions "emerging" *AND* $\geq$ 2 stable weeks in the last five *AND* vaccination $< 75\%$ *THEN* *moderate increasing*. (c=0.71, n=14, regime: variant-shift)

Forecast. $\hat{y}_{s,t}$ = *moderate increasing* (p=0.62); truth: *moderate increasing*. Backbone (with empty <MEMORY> and <RULES>): *stable* (incorrect).

Reflection. "BA.5 emergence with vaccination near 67% and two prior stable weeks confirmed the variant-shift uptick."
Rule update. Matched rule's confidence and support: c: 0.71 → 0.73, n: 14 → 15.

Figure 4: **One forecasting cycle of EpiEvolve walked end to end.** Top block: the model's actual input slots (variant text, recent trend, dynamic features) together with the <MEMORY> and <RULES> populated by hierarchical retrieval and rule matching. Middle block: the model's prediction and a counterfactual from the backbone with empty memory and rules. Bottom block: the agent's writeback for this week, comprising a one-sentence reflection appended to $\mathcal{E}_t$ and accessible to all three memory scopes and the quantitative update to the matched rule's confidence and support.

nitude: state-local retrieval misses cross-regional patterns, regional retrieval loses fine-grained local signal, and national retrieval loses geographic specificity. The full hierarchical configuration outperforms every single tier variant on all metrics.

Figure 3 shows how retrieval and drift detection evolve during the stream. In Panel (a), state-level entries become more common as a regime matures, since the memory accumulates recent outcomes for the same state. After a transition, those local entries are sparse, and retrieval shifts toward regional and national cases with similar features. Panel (b) shows that the detector combines two kinds of signals. Four events follow changes in the dominant variant field of $x_{s,t}$, while three additional events are triggered by spikes in rolling ordinal error. These statistical triggers occur at weeks 26, 51, and 60, capturing disruptions that the variant field alone does not mark. This pattern explains why removing the drift detector mainly increases recovery lag in Table 2. The agent still adapts from later feedback, but it reacts less quickly.

5.5 Case Study

Figure 4 demonstrates the synergy of our method components (Section 4) at a variant boundary. Hierarchical retrieval extracts same-state, regional, and national cross-regime exemplars to capture the "variant emergence + vaccination plateau → hospitalizations rise" pattern, successfully overriding recent local stability. The matched strategic rule reinforces this forecast, while the agent's reflection generates a new insight to inform future retrievals.

6 Conclusion

In this paper, we propose EpiEvolve, a self-evolving agent for streaming epidemic forecasting under variant regime shifts that wraps a pre-trained LLM and turns each week's delayed errors into reusable knowledge for the next forecast. The agent uses a hierarchical episodic memory at state, regional, and national scopes, distills recurring patterns into predicate-form rules, and gates retrieval on a regime indicator from a drift detector. Experiments on a multi-regime COVID hospitalization-trend stream show that this self-evolution improves average accuracy and shortens post-shift recovery over baselines, with the largest gains in the most out-of-distribution regimes. Organizing a system's own past errors can therefore substitute for retraining the backbone in this streaming hospitalization-trend setting, suggesting that this style of adaptation may transfer to other forecasting deployments where model weights cannot be touched.

Limitations

This work is limited by the scope and granularity of the processed COVID dataset, the unavoidable ambiguity of hard regime boundaries, and the fact that hospitalization trend classification is not a full epidemiological forecasting task. The memory and reflection components may also be sensitive to prompt design, retrieval budget, and the quality of generated strategic lessons. External LLM baselines may introduce cost and reproducibility concerns.

Acknowledgments

References

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, and 1 others. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Goshi Aoki and Navid Ghaffarzadegan. 2026. Ai agents as policymakers in simulated epidemics. *arXiv preprint arXiv:2601.04245*.

Jacob Glenn Ayers, Buvaneswari A Ramanan, and Manzoor A Khan. 2025. Detecting concept drift in neural networks using chi-squared goodness of fit testing. *arXiv preprint arXiv:2505.04318*.

Arthur Chen, Zuxin Liu, Jianguo Zhang, Akshara Prabhakar, Zhiwei Liu, Shelby Heinecke, Silvio Savarese, Victor Zhong, and Caiming Xiong. 2026. Test-time adaptation for llm agents via environment interaction. In *The Fourteenth International Conference on Learning Representations*.

Estee Y Cramer, Evan L Ray, Velma K Lopez, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Tilmann Gneiting, Katie H House, Yuxin Huang, and 1 others. 2022. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119.

Botos Csaba, Wenxuan Zhang, Matthias Müller, Ser-Nam Lim, Mohamed Elhoseiny, Philip H Torr, and Adel Bibi. 2024. Label delay in online continual learning. *Advances in Neural Information Processing Systems*, 37:119976–120012.

Rituparna Datta, Zihan Guan, Baltazar Espinoza, Yiqi Su, Priya Pitre, Srini Venkatramanan, Naren Ramakrishnan, and Anil Vullikanti. 2026. Agentic framework for epidemiological modeling. *arXiv preprint arXiv:2602.00299*.

Hongru Du, Yang Zhao, Jianan Zhao, Shaochong Xu, Xihong Lin, Yiran Chen, Lauren M Gardner, and Hao 'Frank' Yang. 2025. Advancing real-time infectious disease forecasting using large language models. *Nature Computational Science*, 5(6):467–480.

Yuxin Fan, Yuxiang Wang, Lipeng Liu, Xirui Tang, Na Sun, and Zidong Yu. 2025. Research on the online update method for retrieval-augmented generation (rag) model with incremental learning. In *2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, pages 1740–1744. IEEE.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37.

Chenghua Gong, Rui Sun, Yuhao Zheng, Juyuan Zhang, Tianjun Gu, Liming Pan, and Linyuan Lv. 2025. Epillm: unlocking the potential of large language models in epidemic forecasting. *arXiv preprint arXiv:2505.12738*.

Salvatore Greco, Bartolomeo Vacchetti, Daniele Apiletti, and Tania Cerquitelli. 2025. Unsupervised concept drift detection from deep learning representations in real-time. *IEEE Transactions on Knowledge and Data Engineering*.

Emily Howerton, Lucie Contamin, Luke C Mullany, Michelle Qin, Nicholas G Reich, Samantha Bents, Rebecca K Borchering, Sung-mok Jung, Sara L Loo, Claire P Smith, and 1 others. 2023. Evaluation of the us covid-19 scenario modeling hub for informing pandemic response under uncertainty. *Nature communications*, 14(1):7260.

Jinwu Hu, Zhitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. 2025. Test-time learning for large language models. *arXiv preprint arXiv:2505.20633*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and 1 others. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chenxiang Li, Qiqiao Zhang, Yue Zhang, Bowen Zhao, Jule Yang, Li Qi, Jun Ding, and Dechao Tian. 2025. Fine-tuned large language models enhance influenza forecasting. *medRxiv*, pages 2025–03.

Jiaye Lin, Yifu Guo, Yuzhen Han, Sen Hu, Ziyi Ni, Licheng Wang, Mingguang Chen, Hongzhang Liu,

Ronghao Chen, Yangfan He, and 1 others. 2025. Seagent: Self-evolution trajectory optimization in multi-step reasoning with llm-based agents. *arXiv preprint arXiv:2508.02085*.

Zewen Liu, Juntong Ni, Bohan Wang, Max SY Lau, and Wei Jin. 2025. Pre-training epidemic time series forecasters with compartmental prototypes. *arXiv preprint arXiv:2502.03393*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Kangkun Mao, Fang Xu, Jinru Ding, Yidong Jiang, Yujun Yao, Yirong Chen, Junming Liu, Xiaoqin Wu, Qian Wu, Xiaoyan Huang, and 1 others. 2025. Epiplanagent: Agentic automated epidemic response planning. *arXiv preprint arXiv:2512.10313*.

Juntong Ni, Shiyu Wang, Ming Jin, Qi He, and Wei Jin. 2026. Streasoner: Empowering llms for spatio-temporal reasoning in time series via spatial-aware reinforcement learning. *arXiv preprint arXiv:2601.03248*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Kien Pham, Chrispin Chaguza, Rafael Lopes, Ted Cohen, Emma Taylor-Salmon, Melanie Wilkinson, Volha Katebi, Nathan D Grubaugh, and Verity Hill. 2025. Large-scale genomic analysis of sars-cov-2 omicron ba. 5 emergence, united states. *Emerging infectious diseases*, 31(Suppl 1):S45.

Qitao Qin, Yucong Luo, Yihang Lu, Zhibo Chu, Xiaoman Liu, and Xianwei Meng. 2025. Towards adaptive memory-based optimization for enhanced retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7991–8004.

Nicholas G Reich, Justin Lessler, Sebastian Funk, Cecile Viboud, Alessandro Vespignani, Ryan J Tibshirani, Katriona Shea, Melanie Schienle, Michael C Runge, Roni Rosenfeld, and 1 others. 2022. Collaborative hubs: making the most of predictive epidemic modeling.

Mohammad Hosseini Samaei, Faryad Darabi Sahneh, Lee W Cohnstaedt, and Caterina M Scoglio. 2026. Epidemiqs: Prompt-to-paper llm agents for epidemic modeling and analysis. *IEEE Transactions on Artificial Intelligence*.

Ziyi Shi, Xusen Guo, Hongliang Lu, Mingxing Peng, Haotian Wang, Zheng Zhu, Zhenning Li, Yuxuan Liang, Xinhu Zheng, and Hai Yang. 2026. Coordinated pandemic control with large language model agents as policymaking assistants. *arXiv preprint arXiv:2601.09264*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Guancheng Wan, Zewen Liu, Xiaojun Shan, Max SY Lau, B Aditya Prakash, and Wei Jin. 2025. Earth: Epidemiology-aware neural ode with continuous disease transmission graph. In *Forty-second International Conference on Machine Learning*.

Qian Wang, Sho Iketani, Zhiteng Li, Liyuan Liu, Yicheng Guo, Yiming Huang, Anthony D Bowen, Michael Liu, Maple Wang, Jian Yu, and 1 others. 2023. Alarming antibody evasion properties of rising sars-cov-2 bq and xbb subvariants. *Cell*, 186(2):279–286.

Yu Wang, Xinshuang Liu, Xiusi Chen, Sean OBrien, Junda Wu, and Julian McAuley. 2025a. Self-updatable large language models by integrating context into model parameters. In *International Conference on Learning Representations*, volume 2025, pages 16961–16979.

Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. 2025b. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H Chi, and 1 others. 2025. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*.

Zhaotian Weng, Antonis Antoniades, Deepak Nathani, Zhen Zhang, Xiao Pu, and Xin Eric Wang. 2026. Group-evolving agents: Open-ended self-improvement via experience sharing. *arXiv preprint arXiv:2602.04837*.

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*.

Cheng-Kuang Wu, Zhi R Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2024. Streambench: Towards benchmarking continuous improvement of language agents. *Advances in Neural Information Processing Systems*, 37:107039–107063.

Chunlong Wu, Ye Luo, Zhibo Qu, and Min Wang. 2025a. Meta-policy reflexion: Reusable reflective memory and rule admissibility for resource-efficient llm agent. *arXiv preprint arXiv:2509.03990*.

Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, and 1 others. 2025b. Evolver: Self-evolving llm agents through an experience-driven lifecycle. *arXiv preprint arXiv:2510.16079*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2026. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. 2025a. Memgen: Weaving generative latent memory for self-evolving agents. *arXiv preprint arXiv:2509.24704*.

Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, Junhao Wang, He Zhu, Wangchunshu Zhou, and Shuicheng Yan. 2025b. Memevolve: Meta-evolution of agent memory systems. *arXiv preprint arXiv:2512.18746*.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. Spurious forgetting in continual learning of language models. *arXiv preprint arXiv:2501.13453*.

A EpiEvolve Implementation Details

We instantiate the backbone with Qwen3-14B-Base, adapted to the streaming task on the warm-start period; backbone training and checkpoint selection are described in Section 5.1. The agent retrieves N=8 episodic entries per forecast call from $\mathcal{E}_t$, ranked by the regime-aware score in Equation 6 with cross-regime weight fixed at α=0.5; the embedding ϕ is a 256-dimensional vector from a frozen sentence encoder applied to a structured rendering of $x_{s,t}$. Reflection runs every week; strategic distillation runs every K=4 weeks and on every drift event; the drift threshold is τ_σ=2. $\mathcal{E}_t$ is capped and trimmed by a salience score that combines entry recency with the magnitude of the ordinal forecast error, so that informative mistakes are retained longer than uneventful correct predictions. Strategic-rule preconditions are conjunctions of feature comparisons over $x_{s,t}$ fields (numeric thresholds, substring presence, categorical equality). New rules enter $\mathcal{L}_t$ with n and c initialized to the in-window match count and precision over the distillation window, and are pruned when support, confidence, or recency falls below configured thresholds. Episodic entries and rules are persisted as JSON-Lines under `runs/epievolve/memory/` so that runs are restartable and ablations can replay or perturb the memory state. Sensitivity to N, K, and τ_σ is reported in the robustness analysis of Section C.

B Additional Baseline Details

CDC ensemble. The COVID-19 Forecast Hub ensemble (Cramer et al., 2022) aggregates probabilistic weekly hospitalization forecasts from approximately forty independent teams into a consensus distribution of twenty-three quantile values per (state, week). We use the publicly archived `COVIDhub-ensemble` forecasts for the streaming window 2021-06-07 to 2022-12-19, restricted to issued dates that precede the corresponding truth week to satisfy the delayed-feedback protocol of Section 3. Each quantile is mapped to a five-class hospitalization-trend label by computing the per-100k weekly rate change against the same 3-week-smoothed baseline used for the ground-truth labels and binning with the fixed thresholds of Du et al. (2025); the final label for each (state, week) is the mode across the twenty-three binned quantiles.

C Robustness

Configuration	Avg. Acc.	Rec. lag
$N = 4$	0.612	2
$N = 8$ *(default)*	**0.629**	**2**
$N = 12$	0.616	2
$K = 2$	0.622	2
$K = 4$ *(default)*	**0.629**	**2**
$K = 8$	0.618	3
$\tau_\sigma = 1$	0.620	2
$\tau_\sigma = 2$ *(default)*	**0.629**	**2**
$\tau_\sigma = 3$	0.625	3

Table 3: Hyperparameter sensitivity. Each block varies one parameter while others remain the default. N: total retrieval budget across the three memory tiers. K: distillation cadence in weeks. τ_σ: drift threshold in standard deviations. Recovery lag is in weeks; lower is better.

Table 3 varies the three hyperparameters while holding the rest fixed. The results are stable across these settings. Average accuracy ranges from 0.612 to 0.629, and recovery lag changes by at most one week. The default values give the best overall trade-off among the tested settings. Smaller retrieval budgets leave too little evidence in the prompt, less frequent distillation slows rule updates, and stricter drift thresholds delay regime detection. We observe similar stability when shifting each evalua-

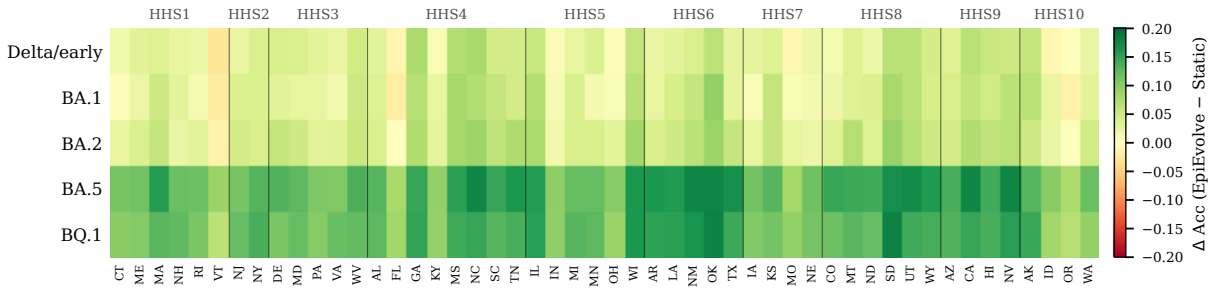


Figure 5: **Per-state EpiEvolve gain over the backbone.** Each cell is the accuracy of EpiEvolve minus the backbone for one state in one regime; states are grouped by HHS region. Gains concentrate in the BA.5 and BQ.1 rows where the backbone is furthest from its training distribution, and they correlate within HHS regions, since states that share federal coordination structure also share evidence available to EpiEvolve's regional memory tier.

tion regime boundary by one or two weeks; regime accuracy and recovery lag remain unchanged.

D Qualitative Analysis

Figure 5 complements the ablations by breaking the EpiEvolve gain down per (state, regime) cell. The BA.5 and BQ.1 rows are systematically the brightest, with per state improvement averaging around 0.13. Within HHS correlation reflects the regional memory $\mathcal{E}_t^{\mathrm{R}}$ pooling evidence across states that share a federal coordination structure. A handful of cells are slightly negative, mostly in early regimes where the backbone already does well and the additional memory context offers little upside.

E Per-Class Behavior

To verify that EpiEvolve's gains are broad rather than concentrated in the majority class, Figure 6 reports the row-normalized confusion matrix. The truth label distribution in our streaming window is moderately imbalanced: stable 29%, moderate increasing 20%, substantial increasing 18%, substantial decreasing 17%, moderate decreasing 16%. The matrix is diagonal-dominant with off-diagonal mass concentrated on adjacent ordinal classes, and per-class recall stays above 0.55 on every class; the dominant stable class is only slightly easier (0.75) than the harder transition classes (moderate increasing and decreasing at 0.55), indicating that the memory architecture helps the harder classes as much as the easy ones.

F Prompt Templates

EpiEvolve issues three types of LLM calls per week. The backbone f_θ is fine-tuned on the warm-start period and answers each (state, week) forecast call. Once the delayed labels for week t arrive, the

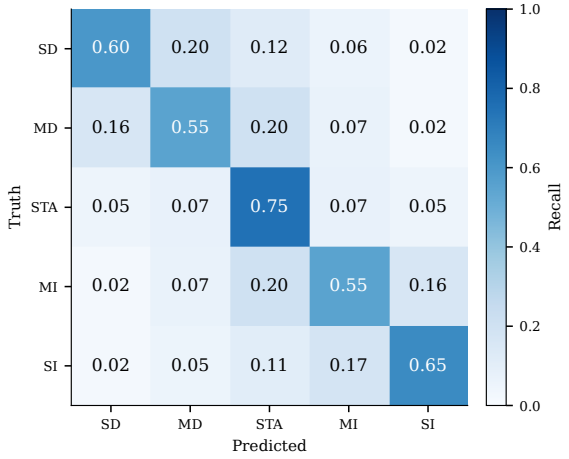


Figure 6: **EpiEvolve confusion matrix (normalized).** Rows are truth classes, columns are predictions.

reflection module emits a one-sentence lesson and a candidate rule per (state, week), and the strategic distiller consolidates the most recent reflections into new strategic rules every $K{=}4$ weeks and on each drift event. Bracketed fields are substituted from the agent state at call time.

The lesson is appended to the new episodic entry e of Equation 5; the candidate rule is queued for the next distillation pass.

Backbone forecasting prompt: one call per (state, week)

```
You are an epidemic forecaster. Predict the next week's hospitalization-trend class for one US
state.

[State]: <state> (FIPS <fips>, HHS region <hhs>)
[Week of]: <target_date>; population <pop>
[Variant]: "<variant_text>"
[Trend] last 5 weeks (oldest first):
  <c_{t-4}>, <c_{t-3}>, <c_{t-2}>, <c_{t-1}>, <c_{t}>
[Dynamic]: vaccinated <vax>%; <policy_text>

<MEMORY> top-N retrieved episodic entries (regime-aware similarity)
  - <state_e>, <date_e> (rho=<regime_e>; <scope>):
    <y_hat_e> -> <y_e>. "<reflection_text>"
  ... (N entries)

<RULES> matched strategic rules; (c, n) = confidence, support
  - IF <precond_conjunction> THEN <class>. (c=<c>, n=<n>, regime=<rho_lambda>)
  ...

OUTPUT
  forecast: <one of: substantial decreasing | moderate decreasing |
            stable | moderate increasing | substantial increasing>
  prob: <float in [0,1]>
  why: <one sentence, <= 25 words>
```

Reflection prompt: one call per (state, week) after labels arrive

```
You are reviewing one weekly forecast post-hoc to extract a reusable lesson.

[State]: <state>; [Week of]: <date>; [Regime]: <rho_t>
[Variant]: "<variant_text>"
[Trend used]: <c_{t-4}>, ..., <c_{t}>
[Dynamic]: vaccinated <vax>%; <policy_text>

[Forecast]: <y_hat> (prob <p>) -> [Truth]: <y_true>
[Outcome]: <correct | incorrect, ordinal offset <k>>

Write a one-sentence lesson (<= 30 words) that:
  (1) names features that supported the forecast,
  (2) names features that should have shifted it toward the truth,
  (3) identifies the regime-conditioned cue the error coincides with.

Also propose a candidate rule whose preconditions are a conjunction of 1-4 predicates over {
    variant text contains "<keyword>", trend pattern, vaccination<%>, policy mention}.

OUTPUT
  lesson: <sentence>
  candidate_rule: IF <conjunction> THEN <class>
```

Strategic distillation prompt: every K=4 weeks and on each drift event

```
You are distilling a sliding window of weekly reflections into strategic rules.

[Window]: weeks <t-K+1>..<t> (K=<K>); regime <rho_t>
[Episodic entries] (state, date, forecast -> truth, lesson):
  - <state_1>, <date_1>: <y_hat_1> -> <y_1>. "<lesson_1>"
  - ...
  (<M> total)

[Existing rules in regime <rho_t>] (do not repeat):
  - IF <preconds> THEN <class>. (c=<c>, n=<n>)
  ...

Propose 0 to 3 NEW strategic rules. Each rule must:
  - have 1-4 precondition predicates;
  - be supported by >= 3 entries in the window;
  - differ from every existing rule by >= 1 predicate.

OUTPUT (one per line, or "no_new_rules"):
  - IF <conjunction> THEN <class>.
```