CHARM: PROBABILISTIC FORECASTING WITH UNCERTAIN HISTORICAL CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Under changing predictive dynamics, a forecaster must make predictions while the relevance of its historical context remains uncertain. Our key insight is to reuse learned predictive knowledge across multiple candidate contexts, maintaining uncertainty over which historical context should inform the next prediction. We introduce CHARM (Changepoint-Aware Recurrent Memory), which maintains a recurrent state for each candidate history under a shared, frozen predictor. Bayesian online changepoint detection updates run-length probabilities using predictive likelihoods, and CHARM combines the candidate predictive densities under these probabilities to account for uncertainty over context boundaries. We evaluate CHARM on financial realized volatility, where persistent dependence makes history valuable while changing dynamics make its relevance uncertain. Experiments show that CHARM improves average predictive log scores over the same forecaster with continuous history, with gains varying across markets.

1 INTRODUCTION

"History does not repeat itself, but it often rhymes." For forecasting, this suggests a distinction between predictive knowledge that can be reused and historical context whose relevance must be reassessed. This distinction matters for time series that combine persistent temporal dependence with changing predictive dynamics. Financial realized volatility provides one such example, exhibiting dependence across multiple time scales alongside structural changes in its dynamics (Corsi, 2009; Shen et al., 2020).

Retaining outdated context can delay adaptation, whereas truncating history prematurely can discard information that remains predictive. Determining where the relevant context begins is difficult because change locations are unobserved, and limited evidence after a possible change may leave several context boundaries plausible (Adams & MacKay, 2007; Harrison et al., 2020a). A forecaster must therefore issue predictions before this ambiguity is resolved. In this paper, we study how a forecaster can represent, update, and use multiple plausible historical contexts for probabilistic prediction.

Changing predictive dynamics may make the current context outdated even when learned predictive knowledge remains useful. Our key insight is that this knowledge can be shared across multiple hypotheses about the current context. When evidence is insufficient to identify a single relevant boundary, these hypotheses provide alternative bases for prediction, and uncertainty about their relevance should be reflected in the forecast.

We introduce CHARM (Changepoint-Aware Recurrent Memory), which implements this idea by combining a shared recurrent density model with Bayesian online changepoint detection (BOCPD) (Adams & MacKay, 2007). Each candidate history is represented by a recurrent state indexed by its run length. At test time, continuing states absorb new observations, while fresh states represent possible new context boundaries. BOCPD updates run-length probabilities using the candidates' predictive likelihoods, and CHARM combines their predictive densities under these probabilities. The recurrent network and density-head parameters remain fixed, so online adaptation operates through state updates and inference over context boundaries.

Contributions. Our contributions are threefold: **(i) Forecasting with multiple context hypotheses.** We develop CHARM, which represents candidate historical contexts as recurrent states under a

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.

shared predictor. Likelihood-based run-length inference updates their relative support, and predictive marginalization incorporates context uncertainty into the forecast. **(ii) Controlled evaluation of context inference.** We compare continuous-history, fixed-window, static-mixture, and MAP prediction under the same frozen recurrent network and density head, assessing the roles of context choice, online likelihood feedback, and predictive marginalization. **(iii) Forecasting gains and reset opportunities.** Experiments on Oxford-Man equity indices and Binance cryptocurrencies show improved average predictive log scores over continuous-history forecasting, with gains varying across markets. Controlled interventions further reveal opportunities to benefit from resetting context and localized gains from online feedback.

2 Background

2.1 Problem Statement

Online probabilistic forecasting. Let $y_{1:T}$ denote a sequence observed online. At time t, let $\mathcal{F}_t$ denote all information available to the forecaster, including observations $y_{1:t}$ and any covariates known at that time. Before observing y_{t+1}, the forecaster outputs a predictive density $q_t(\cdot \mid \mathcal{F}_t)$ (Salinas et al., 2020). After y_{t+1} becomes available, the forecast is evaluated and the observation can inform subsequent predictions. The sequence may exhibit persistent temporal dependence alongside changing predictive dynamics, with change locations unobserved. We consider candidate historical contexts beginning at different boundaries in the observed stream. Their relative support must be assessed using only $\mathcal{F}_t$, and the forecast may combine predictions from multiple plausible contexts.

Prediction objective. We evaluate forecasts using the average negative log-likelihood (NLL):

$$\mathcal{L}_T(q) = -\frac{1}{T-1}\sum_{t=1}^{T-1} \log q_t(y_{t+1} \mid \mathcal{F}_t). \tag{1}$$

Lower NLL indicates better predictive performance under the logarithmic score, which is strictly proper: its expected loss is minimized by the true conditional predictive distribution (Gneiting & Raftery, 2007). The objective concerns prediction throughout the stream, including periods of stable dynamics and transitions between them.

2.2 Bayesian Online Changepoint Detection

Bayesian Online Changepoint Detection (BOCPD) (Adams & MacKay, 2007; Fearnhead & Liu, 2007) represents uncertainty over the most recent context boundary through a distribution over the run length r_t. Suppressing covariates for notation, let

$$\pi_t^{(r)} = p(y_t \mid y_{1:t-1}, r_{t-1} = r)$$

denote the predictive likelihood supplied by an underlying predictive model (UPM) for candidate r. BOCPD propagates the joint distribution according to

$$p(r_t, y_{1:t}) = \sum_{r_{t-1}} p(r_t \mid r_{t-1})\, \pi_t^{(r_{t-1})}\, p(r_{t-1}, y_{1:t-1}). \tag{2}$$

The transition permits either growth to $r_t = r_{t-1}+1$ or a reset to $r_t = 0$, with probabilities governed by a hazard function. Under this indexing convention, the candidate at time $t-1$ scores y_t before the growth/reset transition. Normalization gives the run-length probabilities used for the next forecast:

$$q_t(u \mid \mathcal{F}_t) = \sum_r p(r_t = r \mid \mathcal{F}_t)\, p(u \mid \mathcal{F}_t, r_t = r). \tag{3}$$

The UPM supplies a predictive distribution and an update rule for each candidate context. Conjugate models provide analytical updates, while learned recurrent predictors can provide context-dependent predictions through state updates (Harrison et al., 2020a). Section 3 specifies CHARM's candidate states and transition-based run length, including how a fresh state retains the latest observed lag.

2.3 AMORTIZED PREDICTIVE INFERENCE

A shared predictor can learn to map observed histories to conditional predictive distributions, amortizing the computation of context-dependent predictions (Ortega et al., 2019; Mikulik et al., 2020; Müller et al., 2022). Let P_{tr} denote the population distribution of training sequences, which may contain changes in predictive dynamics. Suppressing covariates for notation, the population training objective is

$$\theta^\star \in \arg\max_\theta \mathbb{E}_{Y_{1:K}\sim P_{\mathrm{tr}}}\left[\sum_{k=1}^{K-1} \log f_\theta(Y_{k+1} \mid Y_{1:k})\right]. \tag{4}$$

Lemma 1 (Amortized predictive optimum)**.** *Assume that the training conditional densities exist, their expected conditional entropies are finite, and the expected log scores are well-defined. Suppose a single parameter vector represents these conditional densities simultaneously at all scored ages. If the population optimum is attained, then, for each $k = 1, \ldots, K-1$ and P_{tr}-almost every history $Y_{1:k}$,*

$$f_{\theta^\star}(\cdot \mid Y_{1:k}) = p_{\mathrm{tr}}(\cdot \mid Y_{1:k}), \tag{5}$$

where p_{tr} denotes the conditional predictive density under the training law.

Proof. At each scored age, the expected negative log-likelihood decomposes into a conditional-entropy term independent of θ and the expected KL divergence from p_{tr} to f_θ. The common realizing parameter makes all KL terms zero. Every population optimum must therefore also make their finite sum zero. Nonnegativity implies that each KL divergence vanishes almost surely, yielding equation equation 5. A detailed derivation is provided in Appendix E.1. □

The lemma characterizes an ideal optimum under the training law; it does not require each training sequence to follow a single fixed mechanism. In a recurrent implementation, the state provides a learned summary of the supplied history, and a conditional density head produces the next-step distribution (Bishop, 1994; Salinas et al., 2020; Rasul et al., 2021). With shared parameters held fixed, adaptation occurs through state updates. Applying the same predictor to different candidate histories yields alternative predictive summaries. Section 3 combines these summaries with run-length inference to propagate uncertainty over historical context.

3 CHARM: CHANGEPOINT-AWARE RECURRENT MEMORY

CHARM represents uncertainty over historical context through recurrent states associated with alternative context boundaries. Each state summarizes a different suffix of the observed stream, while all candidates share the same learned predictive model. At each forecast origin, CHARM combines their predictive densities using the current run-length probabilities. New observations then update both the candidate states and their relative support.

3.1 CANDIDATE CONTEXTS AND RECURRENT STATES

We specialize Section 2.1 to the scalar autoregressive setting used in our experiments. For the recursion, index the observed initial lag by y_0 and write $\mathcal{F}_t = \sigma(y_0, \ldots, y_t)$. At the boundary for forecasting y_{t+1}, $R_t = r$ denotes a context containing the most recent r completed transitions $e_{t-r+1:t}$, where $e_k = [y_{k-1}, y_k]$. Each candidate has a state $h_t^{(r)} \in \mathbb{R}^{d_h}$ obtained by processing these transitions from the same learned initial state h_0. In particular, $h_t^{(0)} = h_0$ represents a possible boundary immediately after y_t. This transition-based convention is detailed in Appendix A.

The shared GRU update is

$$h_{k+1} = G_\theta(h_k, [y_k, y_{k+1}]). \tag{6}$$

The state is a learned predictive summary, without an assumption of exact sufficiency. A mixture-density head produces

$$f_\theta(u \mid y, h) = \sum_{j=1}^{M} \omega_{\theta,j}(y,h)\mathcal{N}\big(u; \mu_{\theta,j}(y,h), \sigma^2_{\theta,j}(y,h)\big), \tag{7}$$

with normalized nonnegative weights and positive scales. Candidate r therefore supplies

$$g_t^{(r)}(u) = f_\theta(u \mid y_t, h_t^{(r)}). \tag{8}$$

All candidates retain the observed lag y_t as a direct input, including the newly initialized candidate. A boundary resets historical context while preserving the input needed for the next transition.

3.2 Learning a shared predictive model

For a training sequence $Y_{0:L}$, we initialize the state at h_0. The population predictive training objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{Y_{0:L} \sim P_{\mathrm{tr}}} \left[-\frac{1}{L} \sum_{k=0}^{L-1} \log f_\theta(Y_{k+1} \mid Y_k, h_k) \right], \tag{9}$$

where h_k is updated using equation 6 only after scoring Y_{k+1}. For real data, we approximate this expectation using contiguous crops from the training period. Crop boundaries are not assumed to be changepoints, and crops may span changes in predictive dynamics. Simulator-based training uses samples from its specified sequence distribution.

Lemma 1 characterizes the ideal population predictive optimum under P_{tr}. Our fitted recurrent model approximates this conditional predictive mapping; the training conditionals need not coincide with the ideal candidate conditionals assumed in Proposition 1. At evaluation, θ, including h_0, is fixed within each fold and shared across all candidate histories. Architecture and training details are given in Appendix D.

3.3 Online context inference and predictive marginalization

For untruncated inference, let

$$w_t(r) = \Pr_\theta(R_t = r \mid \mathcal{F}_t) \tag{10}$$

denote the model-induced run-length weights available before observing y_{t+1}. We use the same notation for their capped approximations below. CHARM issues the predictive mixture

$$q_t(u \mid \mathcal{F}_t) = \sum_r w_t(r) g_t^{(r)}(u). \tag{11}$$

After observing y_{t+1}, the likelihood update gives responsibilities over the preceding candidates:

$$\alpha_{t+1}(r) = \frac{w_t(r) g_t^{(r)}(y_{t+1})}{\sum_j w_t(j) g_t^{(j)}(y_{t+1})}. \tag{12}$$

For a constant hazard H, the next predictive weights and states satisfy

$$w_{t+1}(0) = H, \qquad h_{t+1}^{(0)} = h_0, \tag{13}$$

$$w_{t+1}(r+1) = (1-H)\alpha_{t+1}(r), \qquad h_{t+1}^{(r+1)} = G_\theta(h_t^{(r)}, [y_t, y_{t+1}]). \tag{14}$$

Initialization uses $w_0(0) = 1$ and $h_0^{(0)} = h_0$. The growth branches absorb the completed transition; the new reset branch represents a possible boundary after y_{t+1}. Thus each observation updates the weights used for subsequent forecasts, while density marginalization carries unresolved boundary uncertainty into the predictive distribution. No online gradient updates are required.

3.4 Bounded inference and computational cost

Exact enumeration retains $t+1$ candidates at forecast time t. The evaluated implementation caps history depth at $R_{\max} \geq 1$, giving at most $R_{\max}+1$ predictive candidates, including the fresh state. When the cap is exceeded, let $\widetilde{w}_{t+1}$ denote the weights before truncation. We retain the boundary suffix state and merge the tail mass:

$$w_{t+1}(R_{\max}) = \sum_{r \geq R_{\max}} \widetilde{w}_{t+1}(r), \qquad w_{t+1}(r) = \widetilde{w}_{t+1}(r) \quad (r < R_{\max}). \tag{15}$$

This preserves probability mass but approximates the discarded predictive densities. Appendix C.1 gives the storage convention and Appendix B provides the full algorithm.

For per-candidate evaluation cost C_θ, capped inference uses $\mathcal{O}(R_{\max} C_\theta)$ time and $\mathcal{O}(R_{\max} d_h)$ state memory per observation, with candidate updates batched. Uncapped inference instead takes $\mathcal{O}(T^2 C_\theta)$ total time for a length-T stream.

4 Predictive Analysis

We analyze how marginalizing context hypotheses relates to predictive accuracy. The first result states when CHARM recovers the Bayesian predictive distribution under an ideal changepoint model. The second separates the effects of errors in candidate predictions and run-length weights on expected predictive log loss.

4.1 Exactness under ideal component predictions

Proposition 1 (Exact run-length marginalization)**.** *Suppose the initial run-length law, reset prior, and constant hazard match the generative model, no candidates are discarded, and* $g_t^{(r)} = p^\star(\cdot \mid \mathcal{F}_t, R_t = r)$ *for every candidate with positive posterior probability, almost surely under that model. Then equation 11 is its Bayesian one-step posterior predictive distribution.*

The result follows by alternating Bayes' rule and the hazard transition (Appendix E.3). It is an ideal filtering statement; learned likelihoods and truncation need not satisfy its assumptions. Structural mechanism identification is not required when different mechanisms imply the same forecasts, as illustrated in Appendix E.2.

4.2 Component and run-length approximation

At a fixed history, suppose the true and learned predictive distributions have representations on a common candidate index set:

$$p_t^\star = \sum_r \gamma_t^\star(r) p_t^{\star,(r)}, \qquad q_t = \sum_r w_t(r) g_t^{(r)}. \tag{16}$$

Proposition 2 (Run-length mixture bound)**.** *Whenever the right-hand side is finite,*

$$\mathrm{KL}(p_t^\star \| q_t) \le \mathrm{KL}(\gamma_t^\star \| w_t) + \sum_r \gamma_t^\star(r) \mathrm{KL}(p_t^{\star,(r)} \| g_t^{(r)}). \tag{17}$$

The bound separates weight error from candidate predictive error by applying the KL chain rule before marginalizing the candidate index (Appendix E.4). Averaging over histories and scored times bounds expected excess NLL relative to the true predictive law, connecting this analysis to equation 1; see equation 34. For capped inference, the true tail must be grouped on the same index set using its conditional mixture; the construction is given in Appendix E.5. The terms can interact because candidate densities also determine the weights. The bound therefore does not establish an improvement over any particular history policy. If all candidate densities coincide at a given history, their mixture is independent of the run-length weights. Candidate predictions must therefore differ for online reweighting to change the forecast; whether that change improves prediction is an empirical question.

5 Experiments

We first assess CHARM's probabilistic forecasting performance on real data. We then hold the learned predictor fixed to examine the roles of alternative context hypotheses, online likelihood weighting, and predictive marginalization. Finally, controlled interventions examine when resetting context creates predictive opportunities and whether online feedback captures part of that benefit.

Table 1: Mean NLL, with lower values indicating better predictions. Bold identifies the lowest mean; † marks baselines for which the adjusted paired interval supports an advantage for CHARM. Bonferroni adjustment retains all 20 comparisons, including GRU-global and GRU-window on both datasets.

Method	Oxford-Man	Binance
AR-expanding	0.998358†	1.102496†
AR-EW	0.922053	1.087815
logHAR-expanding	0.937941†	1.117514†
logHAR-EW	0.933450†	1.121220†
SH-global	0.995425†	1.102278†
SH-window	0.970338†	1.096475†
SH-MOCA	0.943219†	1.089924
SH-BAM-128	0.986231†	1.099962†
CHARM	**0.891185**	**1.041625**

5.1 Experimental Setup

Datasets and forecasting task. We evaluate one-step predictive densities for daily log realized volatility on four Oxford-Man indices (S&P 500, FTSE 100, Nikkei 225, and AEX) and two Binance pairs (BTCUSDT and ETHUSDT). Evaluation spans 2012–June 2018 and 2020–2025, respectively, with 6,495 and 3,770 asset-days in the main comparisons. All methods share the same scored targets within each dataset; missing observations trigger state resets, and scoring requires 22 consecutive available historical observations.

Training and evaluation protocol. We separate model fitting, checkpoint selection, validation of forecasting configurations, and evaluation chronologically. Oxford-Man models are trained once, whereas Binance models are refreshed annually; neural parameters remain fixed within each evaluation fold. We use two neural training seeds and GRU training horizons of 128 and 32 steps, respectively. Within each seed, CHARM and its recurrent controls share the network, density head, and fit-only preprocessing. Window lengths, hazards, and calibration scales are selected on validation data.

Metrics and uncertainty. Our primary metric is NLL of the native predictive density on the original log-RV scale; IS90, coverage, and interval width provide complementary diagnostics. We average seed scores per target, then over time within each market, and finally equally across markets. For real-data comparisons, paired intervals use 20,000 synchronous bootstrap draws with three-calendar-month blocks. We report ordinary 95% and family-adjusted intervals, conditional on the fitted models and markets. Full protocols and statistical details are provided in Appendix F.

5.2 Baseline Methods

Linear forecasting baselines. AR predictors use the latest log-RV, while logHAR predictors additionally use its trailing 5- and 22-observation averages. For each specification, expanding and exponentially weighted (EW) variants accumulate or discount historical sufficient statistics, respectively, with matched priors and observation noise.

Shared-feature Bayesian baselines. The SH variants share a frozen neural feature map of logHAR inputs, a coefficient prior, and observation noise. SH-global and SH-window use expanding and sliding histories; SH-MOCA averages over run lengths; SH-BAM-128 performs greedy history selection over the latest 128 completed batches. These are shared-feature adaptations, with implementation details and limitations documented in Appendix F.

Controls for context choice, weighting, and marginalization. All recurrent controls reuse CHARM's frozen predictor. GRU-global represents the continuous history in a single recurrent state, while GRU-window uses a validation-selected fixed context window. Static mixtures retain

Table 2: Mean NLL for prediction strategies sharing a frozen recurrent predictor. Bold marks the lowest mean; † indicates an advantage for CHARM supported by the adjusted paired interval. Bonferroni adjustment retains the full 26-comparison family for each dataset's main NLL panel. The same-hazard and independently validation-selected static mixtures coincide and are shown once.

Prediction strategy	Oxford-Man	Binance
GRU-global	0.911737†	1.041842
GRU-window (validated)	0.908590†	1.042393
Window256	0.910252†	1.041842
Static mixture	0.898672†	1.042156
CHARM-MAP	0.906641†	1.042613
CHARM	**0.891185**	**1.041625**

the same candidate densities as CHARM but combine them using fixed geometric weights, with either CHARM's hazard or an independently validated hazard. CHARM-MAP retains the full filter but predicts with the candidate having the largest predictive run-length weight, isolating the contribution of density marginalization. Together, these controls assess how candidate contexts are chosen, weighted, and combined under shared predictive knowledge. The privileged known-boundary reset reference is introduced separately in the controlled experiments.

5.3 Forecasting Performance on Real-World Data

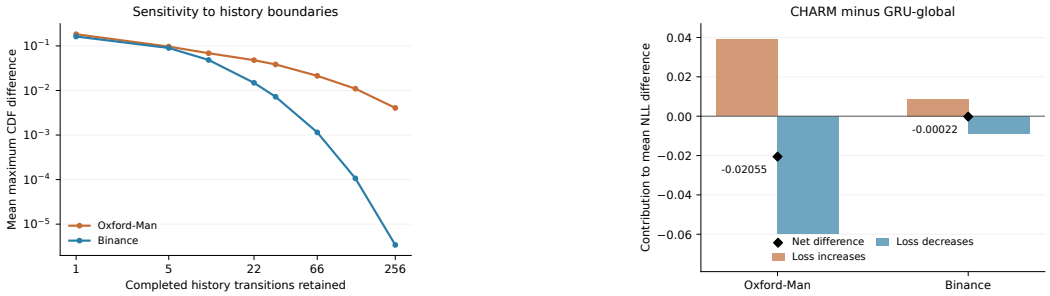
Overall predictive performance. Table 1 compares CHARM with the eight non-GRU baselines. CHARM achieves the lowest mean NLL on both datasets: 0.891185 on Oxford-Man and 1.041625 on Binance. After adjustment for the full comparison family, paired intervals support improvements over seven of the eight baselines on Oxford-Man and six on Binance. The adjusted intervals include zero for AR-EW on both datasets and for SH-MOCA on Binance, so the lowest mean scores do not establish superiority over every baseline.

Variation across markets and scores. The aggregate ranking is not uniform across markets: AR-EW achieves lower NLL than CHARM on the S&P 500 despite CHARM's advantage in the Oxford-Man average. AR-EW also attains the lowest mean IS90 on both datasets, indicating that CHARM's NLL gains do not extend uniformly to interval quality. Complete results for all 11 methods, including market-, year-, and seed-level analyses and auxiliary scores, are reported in Appendix G. These comparisons assess CHARM as a complete forecasting method. We next hold the recurrent predictor fixed to examine the contributions of alternative contexts, online weighting, and predictive marginalization.

5.4 Prediction under Context Uncertainty

Alternative contexts under a shared predictor. Table 2 compares prediction strategies that use the same frozen recurrent predictor with different historical contexts and rules for combining their predictions. On Oxford-Man, CHARM improves on all five controls, including Window256, so limiting the maximum history alone does not reproduce its gains. Relative to GRU-global, the NLL difference is -0.02055, with an adjusted paired interval of $[-0.03111, -0.00962]$. On Binance, the corresponding difference is only -0.00022 $[-0.00303, 0.00238]$. The shared-predictor comparisons therefore support a forecasting benefit on Oxford-Man, while providing no clear evidence of an improvement on Binance.

Online weighting and predictive marginalization. On Oxford-Man, CHARM improves on the same-hazard static mixture by 0.00749 NLL, with a paired difference interval of $[-0.01377, -0.00194]$, supporting the contribution of online likelihood feedback. Independently selecting the static hazard yields the same configuration, so this comparison also tests a validation-tuned competitor. CHARM further improves on CHARM-MAP by 0.01546 NLL $[-0.02178, -0.01036]$, supporting density marginalization over selecting a single candidate while retaining the same filter. These intervals use the adjustment in Table 2.



(a) Sensitivity to context boundaries.

(b) Contributions to the NLL difference.

Figure 1: Post hoc diagnostics of candidate contexts and predictive gains. (a) Candidate–global CDF differences, maximized over 41 points spanning the global mean ± 4 standard deviations. All depths share targets with at least 256 completed transitions (Oxford-Man: 6,260; Binance: 1,906). Finite-grid proximity does not bound tail log scores. (b) Favorable and unfavorable contributions to CHARM minus GRU-global NLL over all main targets (6,495 and 3,770, respectively); diamonds show their sum, and negative values favor CHARM. Both panels follow the original seed, time, and market aggregation.

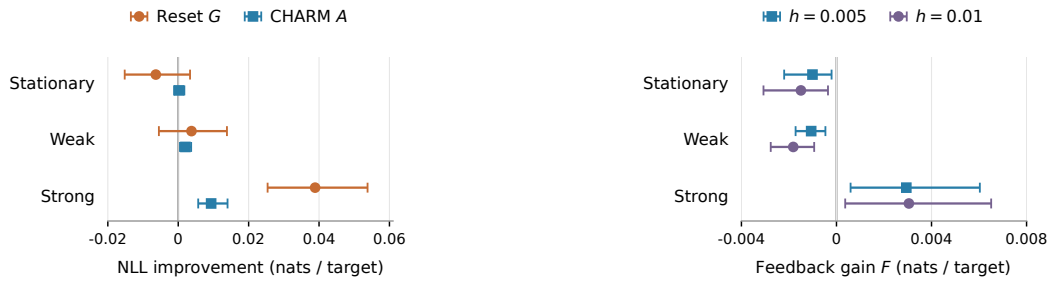
Context sensitivity and limited gains on Binance. Post hoc diagnostics show similar predictions across longer candidate contexts (Figure 1a). On common targets with at least 256 completed history transitions, the mean grid-based CDF difference between the 128-step candidate and GRU-global is 0.000107 on Binance, versus 0.010967 on Oxford-Man. The run-length weights still respond to observations, but many longer candidates yield similar forecasts; very short contexts instead increase average NLL.

Across all main evaluation targets, remaining gains and losses largely cancel: favorable and unfavorable targets contribute -0.00888 and $+0.00866$ to Binance's mean NLL difference (Figure 1b). These findings characterize the fitted predictor and protocol; they do not distinguish sufficient recent information from limitations of learned memory, or establish an absence of long-term dependence.

5.5 Reset Opportunities and Online Feedback

Controlled interventions and diagnostic reference. To examine when alternative contexts offer useful predictive differences, we vary the dynamics while holding the learned predictor fixed. We simulate Gaussian AR(1) log-volatility, $y_t = \mu + \phi_t(y_{t-1} - \mu) + \sigma\varepsilon_t$, with independent standard Gaussian innovations. After a 400-observation prefix with $\phi = 0.85$, the coefficient remains unchanged (stationary), falls to 0.70 (weak change), or to 0.20 (strong change), then returns to 0.85 for 400 targets after an intervention lasting 60, 150, or 400 targets. Innovation noise is fixed and the process is not reinitialized at boundaries. All prediction strategies share one frozen GRU density model (one training seed). Hyperparameters are selected on 16 development source groups and evaluated on 64 independent groups, with initial values and innovations shared across conditions within each group. Known-boundary reset reinitializes only recurrent state, retaining the observed lag; the stationary reset control uses the matched scheduled time. This privileged reference is neither deployable nor an optimal upper bound. Intervals use 20,000 paired source-group bootstrap draws, conditional on the fitted model.

Reset opportunities. The registered reset comparisons use the first 20 targets (age 0–19) after the initial intervention, taking the 150-target condition once to avoid duplicating shared prefixes across durations. Under strong changes, reset reduces NLL by 0.06821, with a ten-comparison Bonferroni-adjusted interval of $[0.04405, 0.09547]$. The corresponding weak-change gain is only 0.00369, while stationary reset increases NLL by 0.01096; both adjusted intervals include zero. Thus, strong changes create a clear opportunity for a freshly initialized context to improve on continuous history, whereas a true change alone does not establish that resetting helps.



(a) Reset opportunity and CHARM improvement.

(b) Incremental value of online feedback.

Figure 2: Exploratory comparisons on matched targets, age 1–19 after the first intervention in the 150-target condition. (a) G and A, using CHARM's selected $H = 0.005$. (b) F compares CHARM and static mixing at each shared hazard. Points and bars show mean gains and pointwise 95% bootstrap intervals from 64 paired source groups; positive values favor the indicated improvement. Horizontal scales differ. Stationary reset uses a scheduled time; no true change occurs.

Exploratory feedback analysis. For mean NLL $L_m(S)$ of method m on the same target set S, define

$$\begin{aligned} G(S) &= L_{\text{global}}(S) - L_{\text{reset}}(S), \\ A(S) &= L_{\text{global}}(S) - L_{\text{CHARM}}(S), \\ F(S) &= L_{\text{static,same-H}}(S) - L_{\text{CHARM}}(S). \end{aligned} \tag{18}$$

Positive values indicate reset opportunity, obtained improvement, and incremental feedback value, respectively. We retrospectively examine age 1–19 in the 150-target condition, excluding age 0, when no new-regime observation is yet available. Under strong changes, $G = 0.03889$ and $A = 0.00935$ (Figure 2a). At CHARM's selected hazard $H = 0.005$, feedback improves NLL by $F = 0.00294$ over the matched static mixture, with a pointwise 95% interval of $[0.00059, 0.00603]$; $H = 0.01$ gives the same direction (Figure 2b). These controls share candidate densities and hazards, differing in likelihood-based weighting throughout the stream. The age-0 strong-change feedback difference is separately -0.00708 at $H = 0.005$. These added comparisons are post hoc and unadjusted for multiplicity, providing exploratory evidence of local feedback value.

Local feedback gains and aggregate performance. At $H = 0.005$, feedback instead hurts under weak changes and stationarity ($F = -0.00107$ and -0.00101), with both pointwise intervals below zero. For strong changes at ages 20–59, feedback intervals include zero. Over all scored targets from the first intervention onward ($t \geq 400$), including the return phase and equally weighting weak/strong changes and all three durations, CHARM improves on global by 0.00109 NLL but is worse than independently tuned static mixing by 0.00134 (adjusted interval $[0.00101, 0.00169]$). Local feedback gains therefore coexist with a stronger overall static competitor; gains from combining candidate contexts need not all arise from online reweighting. Complete duration/noise results, recovery curves, and weight diagnostics appear in Appendix I.

6 Conclusion

CHARM represents alternative contexts as recurrent states under a shared predictor and marginalizes their predictive densities using online run-length weights. Shared-predictor comparisons show gains on Oxford-Man and limited improvements on Binance. Controlled interventions reveal reset opportunities and exploratory local feedback gains, highlighting the potential and limits of forecasting with uncertain historical context. Limitations are discussed in Appendix K.

7 AI USE STATEMENT

We used generative AI tools for literature retrieval and discovery, research ideation and execution, drafting sections of the manuscript, and developing mathematical arguments and proofs. This assistance included refining the problem formulation, discussing experimental designs and results, and drafting and revising proofs. The authors take full responsibility for the final content of this work, including its text, mathematical claims, experimental results, and associated artifacts.We have reviewed all AI-assisted content, checked references against their original sources, and verified the mathematical arguments.

REFERENCES

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007. doi: 10.48550/arXiv.0710.3742.

Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and scalable bayesian online changepoint detection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 642–663. PMLR, 2023. URL https://proceedings.mlr.press/v202/altamirano23a.html.

Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 443–448, 2007. doi: 10.1137/1.9781611972771.42.

Christopher M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Neural Computing Research Group, Aston University, 1994.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009. doi: 10.1093/jjfinec/nbp001.

Paul Fearnhead and Zhen Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007. doi: 10.1111/j.1467-9868.2007.00601.x.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020a.

James Harrison, Apoorva Sharma, and Marco Pavone. Meta-learning priors for efficient online Bayesian regression. In *Algorithmic Foundations of Robotics XIII*, volume 14 of *Springer Proceedings in Advanced Robotics*, pp. 318–337. Springer, 2020b. doi: 10.1007/978-3-030-44051-0_19.

Jeremias Knoblauch and Theodoros Damoulas. Spatio-temporal bayesian on-line changepoint detection with model selection. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2718–2727. PMLR, 2018. URL https://proceedings.mlr.press/v80/knoblauch18a.html.

Vladimir Mikulik, Grégoire Delétang, Tom McGrath, Tim Genewein, Miljan Martic, Shane Legg, and Pedro A. Ortega. Meta-trained agents implement bayes-optimal agents. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d902c3ce47124c66ce615d5ad9ba304f-Abstract.html.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Josue Nassar, Jennifer Rogers Brennan, Ben Evans, and Kendall Lowrey. BAM: Bayes with adaptive memory. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=NdOoQnYPj_.

Pedro A. Ortega, Jane X. Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alexander Pritzel, Pablo Sprechmann, Siddhant M. Jayakumar, Tom McGrath, Kevin Miller, Mohammad Azar, Ian Osband, Neil Rabinowitz, András György, Silvia Chiappa, Simon Osindero, Yee Whye Teh, Hado van Hasselt, Nando de Freitas, Matthew Botvinick, and Shane Legg. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019. doi: 10.48550/arXiv.1905.03030.

Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs Bergmann, and Roland Vollgraf. Multivariate probabilistic time series forecasting via conditioned normalizing flows. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=WiGQBFuVRv.

Yunus Saatçi, Ryan D. Turner, and Carl Edward Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 927–934, 2010.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Dehua Shen, Andrew Urquhart, and Pengfei Wang. Forecasting the volatility of Bitcoin: The importance of jumps and structural breaks. *European Financial Management*, 26(5):1294–1323, 2020. doi: 10.1111/eufm.12254.