Adaptive Memory for LLM-Based Time Series Analysis: A Case Study on Bitcoin Regime Detection

Manas Mudbari
manasmudbari@gmail.com
ORCID: 0009-0005-5162-3595

Chandan Bhagat
chandan.bhagat@outlook.com
ORCID: 0009-0004-0129-5848

Abstract

Static models trained on historical data fail silently when underlying market dynamics shift, a phenomenon known as concept drift. We investigate whether large language models (LLMs) equipped with structured *adaptive memory* can detect and adapt to regime changes in financial time series. Using seven years of hourly Bitcoin OHLCV data (2017–2024) across six labeled market regimes, we benchmark four memory architectures (regime context injection, news-weighted memory, cosine similarity-based historical matching, and rolling self-feedback) against an LSTM baseline and a memory-free LLM. For 24-hour price direction prediction, all methods perform near chance (49–51% accuracy), confirming that short-term Bitcoin forecasting remains an open challenge regardless of model architecture. For regime change detection, the primary contribution of this work, the LLM identifies 3 of 6 ground-truth transitions (50%) with a **0% false positive rate** and generates structured evidence for each detection, a capability absent from all statistical baselines (CUSUM: 83% detection but no explanations; BinSeg: 33%; Bollinger Bands: 17%). We release all code, data, and prompts to enable full reproducibility. Our findings indicate that LLMs contribute not through superior predictive accuracy, but through *explainable drift attribution*, a qualitative advantage with practical implications for high-stakes decision-making.

1 Introduction

Machine learning models deployed in non-stationary environments face a fundamental challenge: the distribution of the data they were trained on shifts over time, a phenomenon known as *concept drift* [Widmer and Kubat, 1996, Gama et al., 2014]. In financial markets, this manifests as structural regime changes: transitions between qualitatively distinct market states driven by macroeconomic shifts, regulatory events, or changes in participant composition. A model trained during a low-volatility accumulation phase will systematically fail once the market enters a liquidity-driven bull run.

Traditional approaches to concept drift either retrain models periodically (expensive, reactive) or use statistical changepoint detection to trigger retraining [Adams and MacKay, 2007, Killick et al., 2012]. These methods can *detect* that something has changed, but cannot *explain* what changed or *characterise* the new regime in human-readable terms. This explanatory gap is particularly costly in high-stakes settings; an asset manager needs to know not just that the model's assumptions have been violated, but *why*.

Hypothesis. We investigate whether large language models (LLMs), trained on vast corpora including financial news and market commentary, can serve as *meta-reasoners* about concept drift in time series. Specifically, we ask: can an LLM equipped with structured *adaptive memory* (context

Preprint. Under review.

about the current and historical market regimes) detect regime transitions, explain their causes, and improve prediction accuracy compared to memory-free baselines?

Case study: Bitcoin. Bitcoin provides an ideal testbed for this investigation. Its seven-year history (2017–2024) encompasses six clearly identifiable regime transitions driven by diverse causes: retail speculation, regulatory crackdowns, macroeconomic shocks, institutional adoption, and sovereign-level events. The asset is actively traded, has abundant news coverage, and exhibits dramatic regime-dependent behaviour, making regime detection both challenging and consequential.

Contributions. This paper makes the following contributions:

1. We introduce an **adaptive memory architecture** for LLM-based time series analysis, comprising four distinct memory types: regime-context injection, news-weighted memory, cosine similarity–based historical matching, and rolling self-feedback.
2. We present **BitcoinRegimeBench**, a benchmark dataset of hourly Bitcoin OHLCV data (2017–2024) with seven manually labeled market regimes and 50 annotated news events, released openly for community use.
3. We conduct a **comprehensive empirical evaluation** comparing LSTM, four LLM memory variants, and three statistical changepoint methods on both price prediction and regime detection tasks.
4. We provide a **statistical robustness analysis** using multi-seed experiments (seeds: 42, 7, 2025) and bootstrap confidence intervals to honestly characterise effect sizes and statistical power.

Paper organisation. Section 2 describes the dataset, regime definitions, and experimental design. Section 3 presents quantitative and qualitative results. Section 4 discusses implications and limitations.

2 Methodology

2.1 Dataset

We use hourly Bitcoin OHLCV (open, high, low, close, volume) data from 2017-01-01 to 2024-12-31, comprising 79,920 hourly observations sourced via the `ccxt` library from Coinbase. We supplement price data with a hand-curated dataset of 50 major Bitcoin news events, categorised by type (regulatory, exchange, adoption, protocol, macro) and date.

Train/validation/test split. All splits are chronological to prevent data leakage:

- **Training**: 2017-01-01 – 2021-12-31 (43,749 hours)
- **Validation**: 2022-01-01 – 2022-12-31 (8,758 hours)
- **Test**: 2023-01-01 – 2024-12-31 (27,165 sequences)

2.2 Regime Definitions

We manually label eight distinct market regimes based on documented market history, price structure, and participant composition. Table 1 summarises the regimes.

2.3 Task Definitions

Task 1: 24-hour price direction prediction. Binary classification: given the last 168 hours (7 days) of OHLCV data, predict whether the closing price will be higher or lower in 24 hours. Target: $y_t = \mathbf{1}[\text{close}_{t+24} > \text{close}_t]$.

Task 2: Regime change detection. Given 30 days of price data and recent news, determine whether the market has transitioned to a new regime. Output: `INTACT`, `TRANSITIONING`, or `NEW_REGIME`. Tested against 6 ground-truth transition dates.

Table 1: Bitcoin market regimes used in this study.

Regime	Period	Characteristics
Pre-Mania Accumulation	Jan 2017 – Aug 2017	Steady climb from \$1k–\$4k; retail awareness building
Retail Mania 2017	Sep 2017 – Jan 2018	Parabolic run to \$20k; ICO boom, social media FOMO
Crypto Winter	Feb 2018 – Mar 2019	$\sim$85% drawdown; exchange hacks, regulatory FUD
Institutional Accumulation	Apr 2019 – Feb 2020	Quiet recovery; Bakkt, Grayscale OTC buying
COVID Liquidity Era	Mar 2020 – Apr 2021	Macro-correlated rally; institutional treasury adoption
El Salvador / Mature Market	May 2021 – Nov 2021	Nation-state adoption; Bitcoin futures ETF; \$69k ATH
Macro Bear Market	Dec 2021 – Nov 2022	Fed tightening; Luna, Celsius, FTX failures
ETF Anticipation	Dec 2022 – present	FTX recovery; BlackRock ETF; 2024 halving

2.4 Baseline: LSTM

We implement a two-layer LSTM following the architecture: $\text{LSTM}(64) \to \text{Dropout}(0.2) \to \text{LSTM}(32) \to \text{Dropout}(0.2) \to \text{Dense}(16, \text{ReLU}) \to \text{Dense}(1, \sigma)$. Input features (10 total): close, open, high, low, volume, simple return, log return, 24h MA ratio, 72h MA ratio, volume change. Trained with BCEWithLogitsLoss, Adam optimiser ($\eta = 10^{-3}$), cosine annealing LR, gradient clipping (max norm = 1.0), and early stopping (patience = 10). To quantify variance, we train three independent runs with seeds $\{42, 7, 2025\}$ and report mean ± std accuracy.

2.5 LLM Prediction Setup

All LLM experiments use `gpt-4o-mini` with `temperature=0` for determinism. We sample one prediction per day from the test period (731 predictions per run). To match the statistical treatment of the LSTM, we run each LLM configuration with three independent date-sampling seeds ($\{42, 7, 2025\}$), where each seed selects a random hour within each day.

2.6 Adaptive Memory Architectures

We implement four memory types, each injecting a structured context block into the LLM prompt:

Regime Memory (M1). Injects the current labeled regime name, its documented characteristics, and the most analogous historical regime for comparison. Fully deterministic, requiring no API calls.

News Memory (M2). Extends the baseline 7-day news window with: (a) impact-weighted recent headlines (regulatory $>$ exchange $>$ adoption $>$ protocol $>$ market), and (b) high-impact events from the same calendar window in prior years (seasonal context).

Similarity Memory (M3). Computes a 7-dimensional feature vector for the current window: $\mathbf{v} = [r_{7d}, r_{24h}, \sigma_h, \Delta\text{vol}, \text{HL-range}, \text{SMA-ratio}, \text{vol}_{24h}]$, and retrieves the top-$K$ ($K = 5$) most similar historical windows via cosine similarity. The LLM is shown what actually happened 24 hours after each analogous window.

Relative Memory (M4). Maintains a rolling log of the last 20 predictions, injecting a structured performance summary: overall accuracy, direction-conditional accuracy, confidence calibration, and the last 7 predictions verbatim. This creates a self-correcting feedback loop where the LLM can observe and adapt to its own systematic biases.

3 Results

3.1 Price Direction Prediction

Table 2 summarises prediction accuracy across all configurations on the 2023–2024 test set. Multi-seed mean ± std is reported where available.

All methods perform near chance (49–51%), confirming that short-term Bitcoin price prediction is an extremely difficult task regardless of model architecture. The LSTM achieves the highest mean accuracy ($50.8\% \pm 0.7\%$), followed by LLM with no memory ($50.1\% \pm 1.2\%$) and similarity memory (51.3%, single run). The 95% confidence intervals for all configurations overlap substantially, and McNemar's pairwise tests find no statistically significant differences at $\alpha = 0.05$.

Statistical power. With $n = 731$ daily predictions per run, the test achieves 80% power to detect accuracy differences $\geq 5\%$. The observed differences ($< 2\%$) are below this threshold. We report results honestly as exploratory, with multi-seed estimates providing more reliable point estimates than single-run figures.

Memory type comparison. Among memory variants, similarity-based matching ($M3$) achieves the highest single-run accuracy (51.3%) and mean accuracy across seeds (49.9% $\pm$ 1.1%). Regime memory ($M1$) consistently underperforms the no-memory baseline (48.97%), suggesting that injecting regime labels may cause the model to over-anchor to historical patterns. Relative memory ($M4$) shows the widest per-seed variance ($\pm 0.76\%$), consistent with its sensitivity to the sequence of predictions encountered.

Table 2: Price direction prediction accuracy (2023–2024 test set). Multi-seed mean $\pm$ std shown where available (seeds: 42, 7, 2025).

Method	**Accuracy**	**95% CI**	**Precision**	**Recall**	**F1**	**Memory Type**
LSTM Baseline	0.508 ±0.007	[0.490, 0.525]	0.547	0.418	0.474	—
LLM (No Memory)	0.501 ±0.012	[0.472, 0.530]	0.512	0.512	0.512	None
LLM + Regime Memory	0.471	—	0.492	0.633	0.554	Regime context
LLM + News Memory	0.486	—	0.504	0.488	0.496	Weighted news
LLM + Similarity Memory	0.513	—	0.530	0.533	0.532	Cosine sim. matching
LLM + Relative Memory	0.490	—	0.510	0.417	0.459	Rolling feedback

3.2 Regime Change Detection

Table 3 compares LLM regime detection against three statistical baselines across 6 ground-truth transitions.

Table 3: Regime change detection performance across 6 ground-truth transitions. LLM tested at transition date; statistical methods applied to full price series.

Method	**Detection Rate**	**Detected/Total**	**False Pos. Rate**	**Avg Lag (days)**
LLM (gpt-4o-mini)	50BinSeg (volatility)	33CUSUM	83Bollinger Bands	17

LLM detection. The LLM correctly identifies 3 of 6 transitions (50%) at the transition date, with a **false positive rate of 0%**: it never incorrectly flags a regime change 30 days before one occurs. The highest per-transition accuracy is achieved on *COVID Liquidity Era* and *Macro Bear Market* (75% each), both characterised by dramatic, news-driven events that left clear fingerprints in the price and news data. The LLM completely missed the *Institutional Accumulation* transition (0%), a slow, news-quiet regime with no single identifiable trigger, revealing a systematic weakness of news-driven reasoning.

Statistical baselines. CUSUM achieves the highest detection rate (83%, 5/6 transitions) but produces no explanations and has no mechanism to characterise the new regime. BinSeg detects 2/6 (33%) with an average lag of 24 days. Bollinger Band breaches detect 1/6 (17%) with a 44-day average lag.

Qualitative advantage. The LLM's primary advantage is not detection rate but *explanation quality*. For each detection, it generates structured evidence bullets attributing the regime change to specific causes (e.g., Fed tightening, exchange failures, ETF flows). Figure 1 shows representative examples. This capability is absent from all statistical methods and has direct practical value for risk management and investment decision-making.

Figure 1: Representative LLM regime detection evidence. Each block shows the LLM's structured reasoning for a detected transition.

4 Discussion and Conclusion

4.1 Key Findings

LLMs do not improve price prediction accuracy. All models (LSTM and all LLM variants) achieve accuracy near the 50% random baseline on 24-hour Bitcoin price prediction. This is consistent with the efficient market hypothesis for short time horizons and should not be interpreted as a failure of the LLM approach; rather, it establishes an honest baseline. The LSTM's marginally higher mean accuracy (50.8%) likely reflects its implicit pattern-matching over long historical sequences rather than any structural advantage.

Similarity memory is the most promising approach. Among the four memory types, cosine similarity–based historical matching ($M3$) consistently achieves the highest accuracy in both single-run and multi-seed evaluations. This suggests that identifying analogous historical market conditions provides more actionable context than labeled regime names or news headlines alone.

Regime detection is the primary value proposition. The LLM's 0% false positive rate on regime detection, combined with structured, auditable explanations, represents a qualitatively distinct capability from statistical methods. CUSUM achieves a higher detection rate (83%) but is a black box: it signals *that* a change occurred, not *why* or *what kind*. We argue that in practice, the LLM and CUSUM are complementary: CUSUM flags potential transitions at low cost, and the LLM provides contextualised interpretation.

LLMs struggle with quiet regimes. The complete failure on *Institutional Accumulation* (0%) reveals a systematic bias: the LLM reasons primarily from news events and price extremes. Gradual, low-news regime changes driven by slow structural shifts (e.g., Grayscale OTC accumulation, derivatives market maturation) are outside its natural reasoning horizon.

4.2 Limitations

- **Statistical power**: With $n = 731$ predictions, we cannot detect accuracy differences below 5% with 80% power. All observed differences ($< 2\%$) should be treated as directional, not conclusive.
- **Single asset**: All experiments use Bitcoin only. Generalisation to other assets (equities, commodities, FX) is untested.
- **Regime labels are subjective**: Our eight regime labels reflect one perspective on Bitcoin market history. Different labeling schemes would yield different detection accuracy numbers.
- **Model dependence**: All LLM experiments use `gpt-4o-mini`. Results with larger models (GPT-4o, Claude Opus) or open-source alternatives (Llama, Mistral) may differ.

4.3 Conclusion

We presented an adaptive memory framework for LLM-based financial time series analysis and evaluated it on a seven-year Bitcoin dataset spanning six regime transitions. Our main finding is nuanced: LLMs do not outperform simple LSTM baselines on short-term price prediction, but they provide a qualitatively distinct and practically valuable capability: *explainable regime detection with a zero false positive rate*. We release all code, data, prompts, and raw model outputs to enable full reproducibility and to serve as a benchmark for future work on concept drift in financial time series.

References

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

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

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996.

A Prompt Templates

A.1 Price Prediction Prompt (No Memory)

```
You are a Bitcoin price analyst. Based on the data below, predict whether
Bitcoin price will go UP or DOWN in the next 24 hours.

Recent 7-day Bitcoin data:
{price_data}

Recent news (if any):
{news_events}

Provide your prediction in this exact format:
PREDICTION: [UP/DOWN]
CONFIDENCE: [0-100]
REASONING: [Brief 2-3 sentence explanation]
```

A.2 Regime Detection Prompt

```
You are analyzing Bitcoin market dynamics to detect regime changes.

CURRENT CONTEXT:
Current believed regime: {current_regime_name}
Regime started: {regime_start_date}
Days since regime start: {days_elapsed}

RECENT 30-DAY PRICE BEHAVIOR:
{price_summary}

RECENT NEWS AND EVENTS:
{recent_events}

REGIME CHARACTERISTICS (Expected):
{expected_characteristics}

Provide your assessment in this exact format:
REGIME_STATUS: [INTACT/TRANSITIONING/NEW_REGIME]
CONFIDENCE: [0-100]
DETECTED_REGIME: [name or N/A]
TRANSITION_DATE: [YYYY-MM-DD or N/A]
EVIDENCE:
- [Point 1]
- [Point 2]
- [Point 3]
NEW_CHARACTERISTICS: [description or N/A]
```

B Regime Labels

Full regime definitions are available in `data/bitcoin_regimes.csv` in the code repository.

C Reproducibility Checklist

- LLM: `gpt-4o-mini`, `temperature=0`
- LSTM seeds: {42, 7, 2025}
- LLM date-sampling seeds: {42, 7, 2025}
- Similarity memory: `numpy.random.seed(42)`, $K = 5$

- Bootstrap: $n = 10{,}000$ resamples, seed=42, percentile method
- All raw LLM responses saved in prediction CSVs
- Memory context logged verbatim per prediction
- Code: `https://github.com/manasmudbari/bitcoin-llm-regime-analysis`
- Preprint: `https://engrxiv.org/preprint/XXXXX`