

STUDY REPORT · OCTOBER 18, 2026

Timestral versus bitcoin

What a time-series foundation model can and cannot forecast:
price and traded volume, one hour and one day ahead.

Jean-Christophe Budin, ask this guy (ATG)

BTC/USDT on Binance, 2018 to October 17, 2026 · all results are out-of-sample

Companion article: askthisguy.com/en/blog/timestral-bitcoin-prevision



ask this guy

Summary

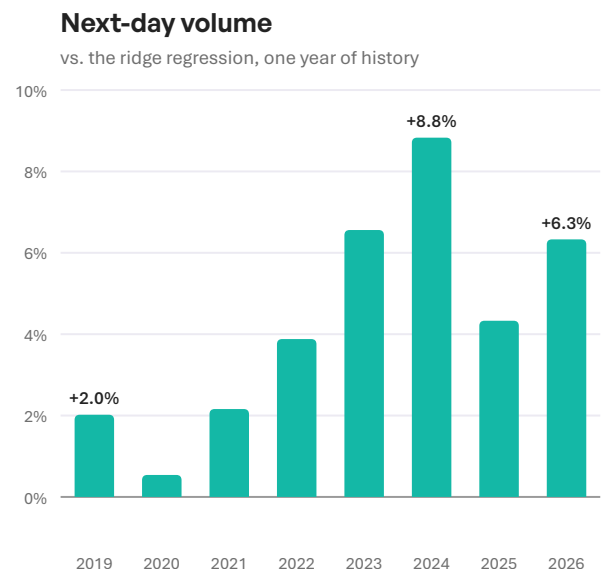
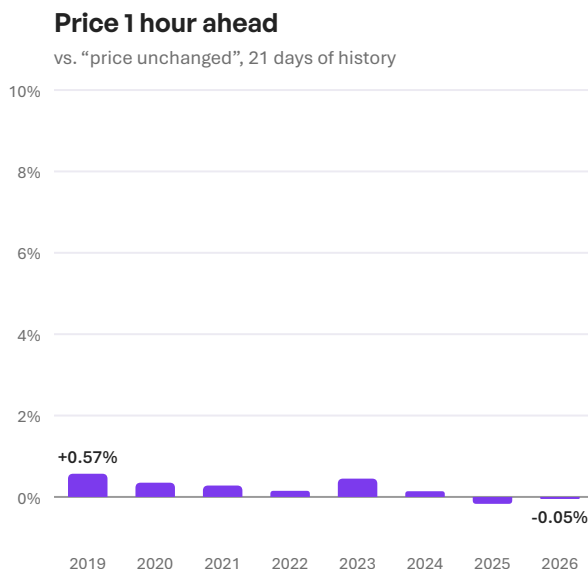
Used zero-shot, Timestral cannot predict where the bitcoin price is going, but it is the best model tested for how much will trade. The edge is clearest at the daily horizon, with the largest model and one year of context.

	Price, 1 h	Price, 24 h	Volume, 1 h	Volume, 1 day
Is the target predictable at all?	Barely: rank autocorrelation of returns -0.01 to -0.11, linear $\approx$ 0	No	Yes: R^2 of log volume $\approx$ 0.84	Yes: $R^2 \approx$ 0.88
Best point forecast in 2026	"Price unchanged": no model beats it	"Price unchanged"	Timestral-209M, +2.7% vs. ridge	Timestral-209M, +6.3% vs. ridge
Timestral vs. best classical model, point, 2019-2026	Small edge in 2019-2024 (long context), gone since 2025	Worse every year (-0.2% to -2.8%)	+2.1% ($p < 0.001$), every year since 2022	+4.6% ($p < 0.001$), positive every year
Timestral vs. best classical model, distribution	1.5% to 2.1% lower quantile loss over 2019-2025, 0.6% to 1.0% in 2026	1.0% to 1.3% lower (long context only), not significant in 2026	Tie ($\pm 1\%$)	7.2% lower ($p < 0.001$)
Economic value of direction forecasts	None: break-even cost 1.3 bp per trade	None	—	—

bp: basis point, 0.01%. A positive gain means a lower error than the baseline. The ridge is a regularized linear regression, refit every year.

Price cannot be forecast, volume can

Accuracy gain of Timestral-209M, year by year (above 0: better than the baseline)



Source: ATG test, BTC/USDT on Binance, out-of-sample forecasts, 2019 to 2026-10-17

1. Questions

1. **Price.** Can a state-of-the-art time-series foundation model, used zero-shot, forecast the bitcoin price, or its direction, 1 hour or 24 hours ahead, better than trivial and classical baselines? Is any edge economically meaningful?
2. **Uncertainty.** Are its predictive distributions (21 quantiles) better than standard volatility models?
3. **Volume.** Is traded volume, a series with strong persistence and seasonality, a better fit for such a model?
4. **Settings.** What matters in practice: model size, context length, input transformations, covariates?

2. Data and protocol

Data

- Binance BTC/USDT spot 15-minute candles, January 1, 2018 to October 17, 2026: **307,811 candles**.
- Source: a Kaggle export up to March 15, 2026, extended with the public Binance API; identical on overlapping candles.
- 29 short exchange outages, all before 2026, handled as missing values.

Out-of-sample design

- **Origins.** Every forecast origin is scored once, on a regular grid: every full hour for price and hourly volume, every UTC midnight for daily volume, from 2019 to October 17, 2026. Each forecast only uses data available at that time.
- **Windows.** 2019-2025 (about 61,000 hourly origins) and 2026 (about 7,000).
- **2026 is the cleanest window.** It follows all training data of the classical models, and is the least likely to overlap Timestral's pre-training corpus (released October 15, 2026; corpus unpublished).

Models

- **Timestral** (Mistral AI; 5M, 28M, 123M and 209M parameters), zero-shot: package `timestral 1.0.0`, CPU inference, library defaults, 21 output quantiles. The median is the point forecast.
- **Trained classical models**, all refit **every year on all earlier data** (walk-forward), so every year is genuinely out-of-sample. Price: SGD linear regressions on normalized price and volume windows. Volume: a ridge regression on lagged log volumes, a seasonal mean and calendar dummies.
- **Untrained baselines:** random walk ("price unchanged"), momentum, seasonal naives, exponentially weighted moving averages (EWMA) with and without time-of-day or time-of-week profiles, and their empirical-quantile versions (filtered historical simulation). All are causal.

Statistics

- **Point accuracy:** Diebold-Mariano tests with Newey-West variance.
- **Probabilistic accuracy:** scaled quantile loss over 21 levels, Diebold-Mariano tests, interval coverage.
- **Trading checks** (price only): non-overlapping trades, 10 bp cost per unit of turnover, a *timing*-t-statistic that removes constant-exposure gains, and the break-even cost.
- **Multiple testing:** p-values are unadjusted. Conclusions rely on $p < 0.001$ or on effects with a consistent sign across years.

Compute: about 12 hours on one Apple M3 laptop, CPU only (4 threads per job, 2 jobs in parallel). The study code is not published; this report describes the method.

3. Price, 1 hour ahead

Task: at every full hour, forecast the open price four 15-minute candles (one hour) later. 68,252 origins from 2019 to October 17, 2026, including 6,957 in 2026.

3.1 Point and direction

2026, 6,957 origins	Mean absolute error	Skill vs. "price unchanged"	DM p	Hit rate
Random walk ("price unchanged")	0.2944%	0	–	–
Timestral-209M, 21 days	0.2946%	-0.05%	0.44	50.2%
Timestral-209M, 7 days	0.2952%	-0.28%	0.008	50.4%
Timestral-28M, 21 days	0.2952%	-0.28%	0.005	50.2%
Timestral-123M, 7 days	0.2955%	-0.37%	0.003	49.9%
Timestral-5M, 7 days	0.2957%	-0.43%	0.006	50.5%
Timestral-5M, 84 days	0.2972%	-0.93%	< 0.001	49.7%
SGD walk-forward (7 days, daily step)	0.2953%	-0.29%	0.026	50.8%
Momentum 1 h	0.4319%	-46.7%	< 0.001	47.9%

Share of up-moves in 2026: 50.2%. Timestral-28M (7 days) and 209M (7 days with volume as a covariate) score -0.45% and -0.32%.

Skill vs. "price unchanged", by year	2019	2020	2021	2022	2023	2024	2025	2026
Timestral-209M, 21 days	+0.57	+0.35	+0.28	+0.15	+0.45	+0.14	-0.17	-0.05
Timestral-209M, 7 days	+0.27	+0.10	+0.26	+0.04	+0.17	-0.12	-0.23	-0.28
SGD walk-forward (1 day, hourly step)	-3.09	-2.65	-2.97	-0.88	-1.00	-1.28	-2.41	-1.61

In %, about 8,750 hourly origins per year (6,957 in 2026).

- **2019-2024: a tiny real edge.** Long-context configurations (21 to 84 days) beat the random walk by a hair; the 209M with 21 days scores +0.26% over 2019-2025 ($p < 0.001$), with a 53.0% hit rate.
- **Mechanism.** It exploits a short-term reversal of small moves: the rank correlation of the next-hour return with the last 15-minute return is -0.11 in 2019 and -0.015 in 2026. The effect fades in 2025-2026, along with Timestral's edge.
- **The walk-forward linear models are worse than the random walk every year.** Model size matters less than context length; the volume covariate changes nothing.

3.2 Economics

Signal	Period	Hit rate	Gross return	Timing t	Break-even cost	Net of 10 bp
Timestral-209M, 21 days	2019-2025	53.0%	+9,436%	3.36	1.32 bp	-100%
Timestral-209M, 21 days	2026	50.2%	-24%	-0.54	< 0	-99%
Buy and hold	2019-2025 / 2026	–	+2,152% / -6.8%	–	–	–

- The best signal is statistically real (timing t-statistic of 3.4 over 61,295 hourly trades), but it earns 1.3 bp per unit of turnover on average, while a regular account pays [0.1% per order on Binance](#), 10 bp, plus the spread.
- It is gone: timing t-statistic of -0.38 in 2025 and -0.54 in 2026.
- **Caveat.** A binomial hit-rate test against 0.5 overstates significance for signals that lean to the majority side. An SGD model with a 52.6% hit rate in 2026 ($p < 10^{-4}$) has a timing t of only 1.05.

3.3 Uncertainty

2026	80% coverage	90% coverage	Quantile loss vs. Gaussian EWMA	vs. time-of-week EWMA
Gaussian EWMA	86.5%	91.7%	–	+1.93%
Time-of-week EWMA, empirical quantiles	79.8%	89.6%	-1.90%	–
Timestral-209M, 7 days	81.7%	90.8%	-2.51%	-0.63% (p = 0.009)
Timestral-209M, 21 days	80.8%	90.2%	-2.88%	-1.01% (p < 0.001)
Timestral-5M, 7 days	82.0%	91.1%	-2.16%	-0.27% (n.s.)

- Timestral's quantiles are calibrated and clearly beat a Gaussian EWMA (-2.0% to -2.9% in 2026, -2.6% to -3.2% over 2019-2025).
- Against the strongest classical model (EWMA with a time-of-week profile and empirical quantiles), the edge falls from about 3.5% in 2019 to 0.5-1% in 2025-2026. After March 2026, only the 209M with 21 days still beats it (-0.60%, p = 0.011).
- What explains most of the gain: volatility clusters ($|return|$ autocorrelation ≈ 0.23), follows the trading clock (peak at 14:00 UTC, 1.56 times the daily mean in 2026), drops 30-50% on weekends, and has fat tails (excess kurtosis ≈ 9 after volatility standardization).

4. Price, 24 hours ahead

Same protocol with a 96-candle (24-hour) horizon: 61,295 hourly origins in 2019-2025 and 6,934 in 2026. Targets overlap, so Diebold-Mariano tests use a Newey-West lag of 24.

Model	Skill 2019-2025 (p)	Timing t	Skill 2026 (p)
Timestral-209M, 21 days	-0.57% (< 0.001)	-1.18	-0.77% (0.020)
SGD walk-forward	-0.70% (< 0.001)	0.59	-0.97% (0.079)
Timestral-5M, 84 days	-0.72% (< 0.001)	-1.47	-1.53% (< 0.001)
Timestral-28M, 21 days	-1.07% (< 0.001)	0.32	-0.88% (0.22)
Timestral-209M, 7 days	-1.36% (< 0.001)	-1.32	-2.45% (0.002)
Momentum 24 h	-50.6% (< 0.001)	-1.61	-42.6% (< 0.001)

Skill vs. “price unchanged”. Random-walk error: 2.165% over 2019-2025, 1.593% in 2026.

- **Every Timestral configuration is worse than “price unchanged” in every year (-0.2% to -2.8%).** None has a significantly positive timing t-statistic.
- Next-24-hour returns are almost uncorrelated with the past: rank correlation with the previous 24 hours of -0.04 to -0.07 per year over 2019-2025, +0.001 in 2026.

- **Uncertainty:** only long-context Timestral beats the Gaussian EWMA over 2019-2025, by about 1%; in 2026 none does significantly. Seasonal profiles do not help at this horizon.
- **Side observation, not a Timestral result.** The sign of the median daily return over the past 30 days (a plain trend rule) has a timing t-statistic of 2.8 over 2019-2025, near 0 since 2024. This is the documented time-series momentum of crypto assets; it was not pre-specified and does not survive a multiple-testing correction.

5. Traded volume, 1 hour and 1 day ahead

5.1 Design

- **Targets:** quote volume (USDT) of the next hour (66,487 targets) and of the next UTC day (2,587 targets), February 1, 2019 to October 17, 2026.
- **Metric:** volume spans three orders of magnitude (0.2 to 3,000 M\$ per hour), so the main metric is the mean absolute log error (MALE). A MALE of 0.32 means a typical error of about 32%.
- **Reference designated before scoring:** the walk-forward ridge regression, the most complete classical forecaster. Every classical model gets a distribution from the empirical quantiles of its own past errors.

5.2 Hourly volume

Model	MALE	vs. last hour	vs. ridge (p)	R ² (log)	Cov. 80 / 90%
Same as last hour	0.384	0	-16.4%	0.785	79.7 / 89.8
Same hour yesterday	0.590	-53.5%	-78.6%	0.500	79.7 / 89.8
EWMA × weekly profile, fast	0.397	-3.2%	-20.1%	0.783	79.8 / 89.7
Ridge walk-forward	0.330	+14.1%	–	0.839	79.7 / 89.7
Timestral-5M, 28 days	0.330	+14.2%	+0.14% (0.29)	0.835	80.1 / 90.1
Timestral-28M, 28 days	0.326	+15.3%	+1.40% (< 0.001)	0.836	80.1 / 90.3
Timestral-123M, 28 days	0.324	+15.8%	+2.04% (< 0.001)	0.837	78.9 / 89.2
Timestral-209M, 28 days	0.323	+15.8%	+2.06% (< 0.001)	0.838	80.9 / 90.7
Timestral-209M, 7 days	0.340	+11.6%	-2.91% (< 0.001)	0.825	80.5 / 90.5

By year, the 209M with 28 days vs. the ridge: +1.1%, -0.5%, +0.6%, +3.9%, +3.0%, +3.1%, +2.1%, +2.7% from 2019 to 2026 (p < 0.001 for the last five values).

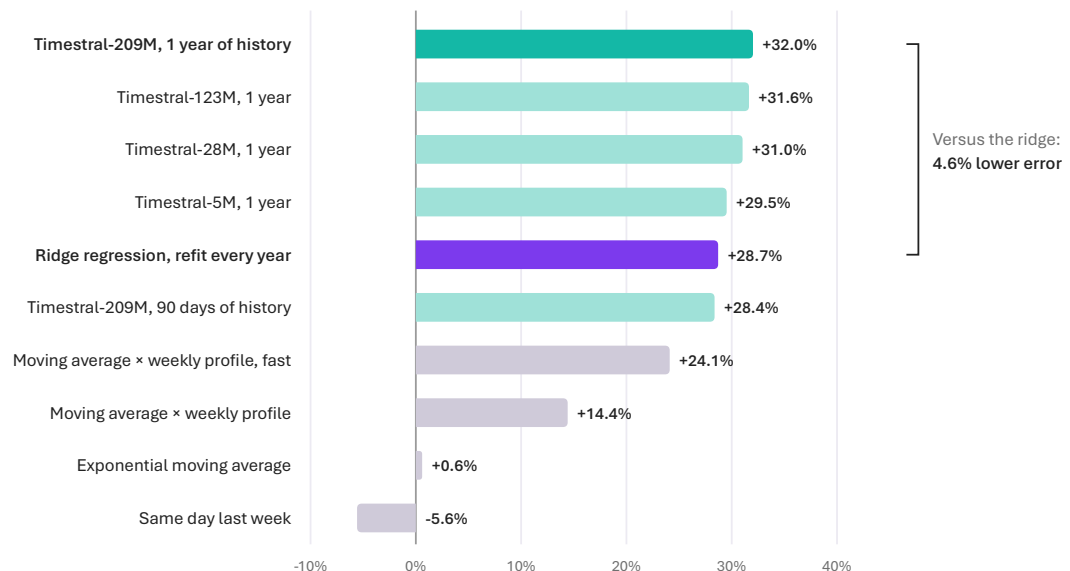
- Hourly volume is highly forecastable: about 84% of the variance of log volume explained, typical error of about 32%.
- **Persistence dominates seasonality:** “same as last hour” is far better than “same hour yesterday”. Seasonal profiles only help when combined with recent levels.
- With 28 days of context, Timestral is the most accurate model; with 7 days it cannot see enough weeks and is 3% worse than the ridge.
- Distributions: a tie with the ridge over the full period. Timestral's median is almost unbiased; the ridge over-predicts by about 7% at the median.

5.3 Daily volume

Model	MALE	vs. last day	vs. ridge (p)	Cov. 80 / 90%	Quantile loss vs. ridge
Same as last day	0.365	0	-40.3%	78.4 / 87.6	+56.5%
Same weekday last week	0.385	-5.6%	-48.1%	78.4 / 88.3	+60.8%
EWMA × weekly profile, fast	0.277	+24.1%	-6.5%	77.8 / 87.7	+4.5%
Ridge walk-forward	0.260	+28.7%	–	78.4 / 88.4	–
Timestral-5M, 1 year	0.257	+29.5%	+1.13% (0.19)	80.6 / 91.0	-2.10%
Timestral-28M, 1 year	0.252	+31.0%	+3.25% (< 0.001)	81.3 / 92.2	-5.29%
Timestral-123M, 1 year	0.249	+31.6%	+4.10% (< 0.001)	81.7 / 91.3	-6.27%
Timestral-209M, 1 year	0.248	+32.0%	+4.64% (< 0.001)	81.1 / 91.2	-7.16%
Timestral-209M, 90 days	0.261	+28.4%	-0.48% (0.63)	80.8 / 91.0	+0.37%

The gain depends on the baseline

Next-day volume: error reduction vs. “same as yesterday”, all years



Source: ATG test, BTC/USDT on Binance, out-of-sample forecasts, 2019 to 2026-10-17

- **Daily volume is the clearest success of the study.** With one year of context, the 209M beats a well-specified, yearly-refit regression by 4.6% in accuracy and 7.2% in quantile loss, every year: +2.0%, +0.5%, +2.2%, +3.9%, +6.6%, +8.8%, +4.3%, +6.3% from 2019 to 2026 ($p < 0.05$ from 2023; 2026: $p = 0.009$). Its intervals are calibrated.
- **Size and context both matter:** 5M +1.1%, 28M +3.3%, 123M +4.1%, 209M +4.6%. With a 90-day context the whole edge disappears: the model needs to see the yearly regime.

6. Cross-cutting lessons

1. **Foundation models forecast structure, not efficiency.** Price direction in a liquid market has no persistent structure: every regularity found (a 15-minute reversal, a 30-day trend) is small, decays, and costs more to trade than it yields. Volume and volatility have strong, stable structure: persistence, clustering, intraday

and weekly seasonality, regime shifts. Timestral captures that structure zero-shot, as well as or better than tailored models.

2. **The right question is “better than which baseline?”**. Against textbook baselines (Gaussian EWMA, seasonal naives), Timestral looks impressive everywhere. Against the strongest classical baseline of each task, its advantage is nil for price, small and shrinking for price uncertainty, and real for volume (about 2% hourly, 5% daily).
3. **Context length is the most important setting**. It is the one that helps in every successful case: 21 days for hourly price risk, 28 days for hourly volume, 365 days for daily volume. Model size helps on volume, not on price.
4. **Evaluation details change conclusions**. With origins every 4 hours instead of every hour, the 1-hour price skill of the 209M was overstated (+0.23% instead of +0.09%), and the best configuration was missed. Binomial hit-rate tests overstate significance. One-off train/test splits hide year-to-year instability.
5. **Possible contamination cuts the other way for price**. If bitcoin history was in Timestral's pre-training corpus, it would favor 2019-2025. The price results are negative or decaying anyway; the volume results hold in 2026, the least exposed year (+2.7% hourly, +6.3% daily).

7. Limitations

- Single asset, single venue: BTC/USDT on Binance. No order-book, derivatives (funding, open interest), on-chain or cross-asset information.
- Zero-shot Timestral only: fine-tuning (Mistral Forge) was not tested. Covariates were tested only as volume for price forecasting, with no effect.
- Classical baselines are well specified but not tuned: no GARCH estimation, no realized-volatility measures, fixed ridge regularization.
- Idealized trading: trades at the candle open without slippage. This flatters strategies, which still fail.
- Price at 24 hours with a volume covariate: only 2026 was run (-2.43%), consistent with the absence of effect at 1 hour.
- Unadjusted p-values; the main claims rely on $p < 0.001$ and on year-by-year consistency.

8. Practical recommendations

Use case	Recommendation
Predict bitcoin price direction from its own history	Do not. Neither Timestral nor classical models beat “unchanged” in 2026, and no historical edge exceeded trading costs.
Risk sizing, stop or alert levels, short-horizon VaR	Timestral (209M, at least 21 days of context) or a seasonal EWMA with empirical quantiles: comparable, both far better than a Gaussian model.
Forecast traded volume (liquidity, execution scheduling, capacity, fee revenue)	Timestral-209M with a long context (28 days hourly, one year daily): the most accurate model tested, calibrated, no training needed. A walk-forward ridge is a close second at 1 hour.

Beyond bitcoin. Business series that behave like volume (orders, collections, support workload, consumption) have the persistence and calendar that Timestral can extend. Series that behave like a price (commodity prices, exchange rates, interest rates) are better handled with scenarios. In every case: a naive baseline, the best classical method, out-of-sample results year by year, checked intervals and a long history.

Appendix: metrics and tests

- **Mean absolute error (price)**: mean of $|\text{forecast} - \text{actual}| / \text{actual}$. **Skill** = $1 - \text{model error} / \text{baseline error}$; positive when the model is better.
- **MALE (volume)**: mean of $|\log(\text{forecast}) - \log(\text{actual})|$, scale-free and symmetric.
- **Diebold-Mariano test**: compares two error series on the same origins; Newey-West variance (lag equal to horizon divided by origin spacing). Two-sided.
- **Quantile loss**: mean over the 21 levels τ of $2 p_{\tau}(\text{actual} - \text{quantile}) / \text{actual}$, with $p_{\tau}(u) = \max(\tau u, (\tau - 1)u)$.
Coverage: share of outcomes inside the central 80% and 90% intervals.
- **Timing t-statistic**: Student t of $(s_t - \bar{s}) \cdot r_t$, where s_t is the position and r_t the realized return. It removes what a constant exposure would earn, so a mostly-long signal in a bull market does not pass for skill.
- **Break-even cost**: mean gross return / mean turnover, in basis points; above this fee level the strategy loses.
- **Filtered historical simulation (FHS)**: empirical quantiles of past returns standardized by estimated volatility, rescaled by the current volatility.

References

- Mistral AI, [Timestral announcement](#), October 15, 2026, and [code repository](#).
- R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd edition, [§ 1.1, “What can be forecast?”](#).
- S. Makridakis, E. Spiliotis and V. Assimakopoulos, [“Statistical and Machine Learning forecasting methods: Concerns and ways forward”](#), PLOS ONE, 2018.
- F. X. Diebold and R. S. Mariano, “Comparing Predictive Accuracy”, *Journal of Business & Economic Statistics*, 1995.
- Binance, [fee schedule](#), accessed October 18, 2026.

© 2026 ask this guy (ATG). This report may be quoted and shared freely with attribution. It is not investment advice.