Does a Time-Series Foundation Model Add Economic Value to Systematic Trading? A Six-Study Falsification Test of TimesFM

Ogdn Ames

Ames Investment Systems

Author Note

Ogdn Ames is affiliated with Ames Investment Systems. Analyses were run on a single six-core CPU workstation, with the QuantConnect LEAN cloud engine used for replication. No external funding was received, and the author reports no conflicts of interest. Code, intermediate data, and a provenance ledger for every reported number are available from the author. Correspondence concerning this article should be addressed to Ogdn Ames, Ames Investment Systems.

Abstract

Time-series foundation models promise strong zero-shot forecasts, but their value to systematic trading depends on incremental information beyond conventional models and on whether forecast gains survive implementation costs. We evaluate Google's TimesFM 2.5 across five distinct designs and one deployment study spanning 2000–2026, using walk-forward estimation, same-slot controls, explicit trading costs, and later-date evaluation windows. For next-day equity volatility, TimesFM outperforms persistence, EWMA, and GARCH (Diebold–Mariano $t = -5.0$ to -7.5), is statistically indistinguishable from HAR, and underperforms HAR augmented with implied volatility ($t = 4.3$ to 6.9). These forecast gains do not translate into reliable incremental returns against the strongest controls. The pre-registered multi-sleeve test yields ΔSharpe = −0.030 (one-sided $p = .561$); the regime-adaptive system loses 2.7 percentage points of annual return with TimesFM (90% interval [−4.7, −0.8]); and the deployed system gains only +0.007 Sharpe from TimesFM (90% interval [−0.007, +0.021]). The descriptive pooled estimate is +0.012 Sharpe (90% interval [−0.004, +0.027]). Relative forecast loss worsens in 17 of 19 later-window comparisons, a pattern consistent with distribution shift or possible pretraining overlap but not diagnostic of contamination. In a conservative sensitivity analysis, using 16,440 program-wide trials and pooled trial dispersion reduces the deployed system's deflated Sharpe ratio from .997 to .50. Overall, TimesFM improves some forecasts but provides no robust evidence of incremental economic value in these designs.

Keywords: time-series foundation models, TimesFM, volatility forecasting, volatility targeting, backtest overfitting, deflated Sharpe ratio, pretraining exposure

Does a Time-Series Foundation Model Add Economic Value to Systematic Trading? A Six-Study Falsification Test of TimesFM

Foundation models pretrained on large collections of time series can forecast unseen series without task-specific fitting. TimesFM uses a decoder-only transformer for patched time-series inputs (Das et al., 2024); its open 2.5 checkpoint is a 200-million-parameter model with continuous quantile forecasts (Google Research, 2025). Chronos, Moirai, and Lag-Llama pursue related generalist forecasting approaches (Ansari et al., 2024; Rasul et al., 2023; Woo et al., 2024). For a quantitative investor, however, benchmark accuracy is not the endpoint. The relevant question is whether a foundation-model forecast adds information beyond competent conventional controls and improves a trading system after realistic implementation costs.

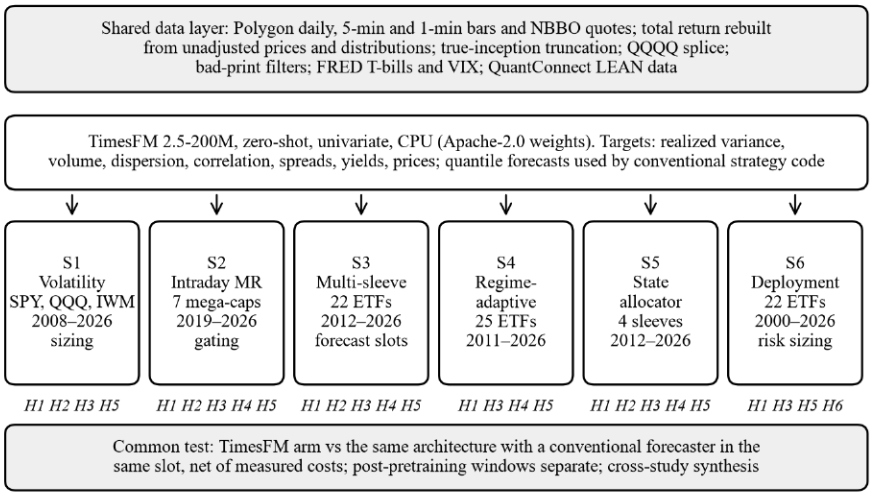
Expected returns are a difficult first target because excess equity returns have proved hard to forecast out of sample (Welch & Goyal, 2008). Second moments are more persistent: realized variance is strongly serially dependent (Andersen et al., 2003), and volatility forecasts are routinely used for position sizing (Harvey et al., 2018; Moreira & Muir, 2017). A useful foundation-model signal therefore has to clear two hurdles. It must first add information beyond strong conventional forecasters of volatility, liquidity, or market state, and that increment must then survive sizing, turnover, financing, and transaction costs. Existing evidence is concentrated on forecast loss, often on public datasets that may overlap with pretraining corpora (Aksu et al., 2024), while several studies find that simpler baselines can match substantially heavier architectures (Tan et al., 2024; Zeng et al., 2023). The forecast-to-portfolio link is therefore the central empirical question.

This paper tests that link in one research program. TimesFM 2.5 is evaluated as a forecaster of second-moment and state variables in five distinct designs: volatility sizing (S1),

intraday trade gating (S2), a multi-sleeve portfolio's forecast slots (S3), a regime-adaptive levered system (S4), and a state-variable allocator (S5). A sixth study (S6) was conducted under a mandate to deploy a TimesFM-inclusive system and is used to isolate the model's incremental contribution. In every design, the TimesFM arm is compared with the same trading architecture using a conventional forecaster in the same slot, net of modeled implementation costs, with later-date windows reported separately (Figure 1). The paper adds three cross-study analyses: a uniform re-estimate of the TimesFM Sharpe effect, a program-wide deflated-Sharpe sensitivity analysis, and a comparison of forecast skill across pre- and post-release or post-2023 windows.

Figure 1

Structure of the Research Program



Note. The six studies share one data layer and one forecasting model. The hypotheses tested in each study are listed below its box. MR = mean reversion.

Literature Review

Time-Series Foundation Models

TimesFM predicts future patches of a series with a decoder-only transformer (Das et al., 2024); Chronos tokenizes scaled values for language-model architectures (Ansari et al., 2024), Moirai uses any-variate attention (Woo et al., 2024), and Lag-Llama is a probabilistic decoder on

lagged features (Rasul et al., 2023). Two cautions motivate the present design. First, simple linear models can match elaborate transformers on many long-horizon benchmarks (Zeng et al., 2023), and the language-model component of LLM-based forecasters can add little (Tan et al., 2024). Second, benchmark observations can enter pretraining corpora, a problem documented in adjacent machine-learning domains (Sainz et al., 2023). The official TimesFM 2.5 model card lists GIFT-Eval Pretrain, Wikimedia Pageviews through November 2023, Google Trends through 2022, and synthetic or augmented data among its pretraining sources (Google Research, 2025). Because public documentation does not establish series-level overlap or a single cutoff for every source, the later-date tests in this paper are contamination-risk checks, not guaranteed contamination-free holdouts.

Volatility Forecasting and Volatility Management

Realized variance from intraday returns proxies latent variance precisely (Andersen et al., 2003). GARCH models are the standard parametric benchmark (Bollerslev, 1986), and the HAR model captures long memory by regressing future realized variance on daily, weekly and monthly averages (Corsi, 2009). Because the target is noisy, forecasts are scored with QLIKE, which is robust to proxy noise (Patton, 2011). Volatility management raises Sharpe ratios in many factors (Harvey et al., 2018; Moreira & Muir, 2017). The strategies here combine it with trend following (Moskowitz et al., 2012) and cross-sectional momentum (Jegadeesh & Titman, 1993), so a forecast layer must improve a system that already has positive expected return.

Backtest Overfitting and Forecast Comparison

Repeated backtesting inflates the best observed Sharpe ratio (Bailey et al., 2014); the deflated Sharpe ratio corrects for the number and dispersion of trials (Bailey & López de Prado, 2014), and the same concern motivates higher hurdles (Harvey et al., 2016) and false-discovery-

rate control (Benjamini & Hochberg, 1995). We compare Sharpe ratios with the stationary bootstrap (Politis & Romano, 1994), forecasts with Diebold–Mariano tests (Diebold & Mariano, 1995), encompassing regressions (Harvey et al., 1998) and the model confidence set (Hansen et al., 2011), quantiles with the pinball loss (Gneiting & Raftery, 2007), and estimate walk-forward with purging and embargo (López de Prado, 2018).

Hypotheses

The program decomposes the central claim into six hypotheses (Table 1). H1 asks whether TimesFM clears naive and standard forecasting benchmarks. H2 is the more demanding forecasting test: whether TimesFM adds information beyond strong, format-matched controls such as HARX, gradient boosting, or AR(1). H3–H5 test whether any increment survives trading costs, placebo controls, and later-date evaluation windows. H6 addresses implementation parity inside QuantConnect.

Table 1

Hypotheses, Null Hypotheses, Test Statistics and Decision Rules

	Hypothesis (alternative)	Null	Test statistic	Decision rule
H1	TimesFM forecasts of second-moment and state variables beat persistence, EWMA and GARCH	Equal loss	QLIKE, MSE of log variance, pinball loss; Diebold–Mariano t	Reject null if DM $t < -1.96$ in favour of TimesFM
H2	TimesFM carries information that HAR, HARX, gradient boosting and AR(1) in identical format do not	No incremental information	DM t vs strongest baseline; encompassing weight; 90% MCS membership	Support requires DM gain vs the strongest baseline or MCS membership that the baseline lacks
H3	The information raises net-of-cost out-of-sample risk-adjusted returns	ΔSharpe ≤ 0 (or Δ return ≤ 0)	Paired stationary-bootstrap ΔSharpe; paired return difference	S3 (pre-registered): one-sided $p < .10$ and placebos below TimesFM; S2: C > B in validation and lockbox; elsewhere 90% interval excluding 0
H4	Real TimesFM features beat time-shifted or shuffled TimesFM features	Placebo $\geq$ real	ΔSharpe real − placebo; share of placebos beaten	Real arm must exceed every placebo arm
H5	Any forecast edge persists in later-date or post-release evaluation windows	Edge vanishes in the later window	Relative loss in earlier vs later windows	Support requires the edge to persist in the later window
H6	A TimesFM system can run entirely inside QuantConnect with forecast parity	Parity fails	Max $\lvert\Delta\rvert$ log-variance forecast; full-run completion	$\lvert\Delta\rvert$ small relative to forecast error; full backtest completes

Note. H1–H5 are tested in five distinct designs (S1–S5); H6 and the deployment questions only in S6. Decision rules are those written in the project protocols where one existed (S2 config/protocol.yaml, S3 PROTOCOL.md); for S1, S4, S5 and S6 the rule

is the 90% interval stated here, applied after the fact. DM = Diebold–Mariano; MCS = model confidence set; EWMA = exponentially weighted moving average; HARX = HAR with log VIX, leverage and overnight-gap terms.

Method

Data

S1–S5 use Polygon.io bars (daily, 5-minute, or 1-minute) and quotes; S6 uses QuantConnect's LEAN data. Several vendor-data defects materially distorted uncorrected backtests (Table 12, Appendix A). Before their launch dates, some ETF symbols return bars for unrelated securities, so every series is truncated at a hard-coded inception date; QQQ is spliced with its former symbol QQQQ; the 2016 XLF–XLRE spin-off is mis-adjusted in two different ways, so total returns are rebuilt from unadjusted closes, dividends, and integer splits; and erroneous intraday highs and lows are capped before range-based volatility is computed. The 3-month T-bill rate (FRED DTB3) is the cash rate. Bid–ask spreads are measured from quote data where available; commissions, slippage, and financing are modeled explicitly. In S3's 1,692 sampled sessions, for example, the median opening half-spread is 0.24 basis points for SPY, while HYG averaged 14.3 basis points in 2008 and 0.6 after 2012.

TimesFM Configuration

Every study uses the 200-million-parameter TimesFM 2.5 checkpoint, whose pretrained weights are Apache-2.0 licensed (Google Research, 2025). Inference is CPU-only, zero-shot, and univariate, using the continuous quantile head, with a context of 256 observations (512 in S5) and horizons from 24 five-minute bars (S2) to 4 weeks (S6). S1 also includes a research-only TimesFM 3.0 comparison; its pretrained weights are restricted to non-commercial, non-production use (Google Research, 2026), and that arm is excluded from the trading conclusions.

Study Designs

Table 2 summarizes the six designs and Figure 2 their windows. In each study the TimesFM arm and its control share every other component: S1 sizes a 200-day trend strategy by next-day realized-variance forecasts and compares TimesFM with nine conventional forecasters, each also with a TimesFM feature added; S2 gates two intraday mean-reversion books with no gate (A), a trailing gate (B), a time-of-day-adjusted cheap forecast (B2), TimesFM (C) or learned gates (D); S3 runs 15 arms, including six placebos, through the same walk-forward selection of 900 configurations; S4 chooses among 1,296 TimesFM and 864 control configurations by nested walk-forward; S5 feeds the same allocator conventional features (B), TimesFM state forecasts (C) or AR(1) forecasts in the identical format (C0); and S6 compares volatility from HARX blended with TimesFM (V10), HARX alone (A1) and TimesFM alone (A2) in 14 sequential cloud runs.

Estimation, Leakage Control, Costs and Engines

Every fitted component uses only data available before its application window, with purging and an embargo around refit dates, and each project tests that perturbing future data leaves decisions unchanged. Decisions made at the close execute at the next open (S2: one minute after the decision bar). Design changes made after viewing any result were logged as amendments (Table 14, Appendix C); S3 discloses one full-sample smoke test. Costs combine measured half-spreads, 1 to 2 basis points of slippage and commissions; cash earns the T-bill rate and leverage pays a spread over it. Results were reproduced in Backtrader (S1, S2, S4, S5) and the LEAN cloud engine (S1, S4, S5, S6); the reconciliations are reported under H6.

Table 2

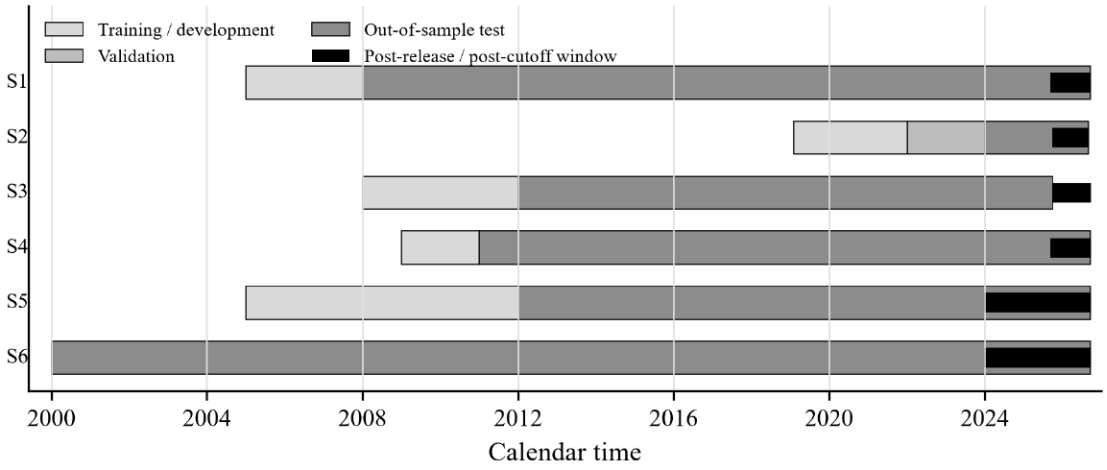
Design of the Six Studies

Study	Universe and data	Out-of-sample window	TimesFM role	Control in the same slot	Primary endpoint
S1 Volatility	SPY, QQQ, IWM; Polygon 5-min bars	2008-01 to 2026-09 (4,709 days)	Next-day RV forecast → vol-target sizing	Persistence, EWMA, GARCH/GJR, HAR, HARX, LightGBM, MLP	QLIKE; Sharpe of trend + vol-target book
S2 Intraday MR	7 mega-caps; 1-min bars; 20,700 NBBO snapshots	Validation 2022–23; lockbox 2024-01 to 2026-08	60-min vol/volume forecast → trade gate	Trailing gate (B), ToD-adjusted forecast gate (B2), learned gates	Net bps per trade; paired daily P&L
S3 Multi-sleeve	22 ETFs; 5 sleeves; 900 configs per arm	2012-01 to 2026-09; lockbox 2025-10+	Every forecast slot (returns, vol, uncertainty, regime, correlation)	HAR-based CLASSICAL arm; NONE; EWMA	OOS Sharpe, TIMESFM − CLASSICAL
S4 Regime-adaptive	25 ETFs; 74 TimesFM input series	2011-01 to 2026-09 (nested walk-forward)	Regime, vol, trend persistence, liquidity, uncertainty features	Same system without TimesFM features	OOS CAGR and Sharpe increment
S5 State allocator	4 sleeves + cash; 17 state series	2012-01 to 2026-09 (768 weekly decisions)	21-day state-variable forecasts → ridge/softmax allocator	Conventional features (B); AR(1) forecasts in identical format (C0)	ΔSharpe C − B
S6 Deployment	22 ETFs; LEAN cloud engine	2000-01 to 2026-09 (full period)	4-week ETF vol forecast → book vol target 18%, cap 1.5×	HARX vol model (A1)	ΔSharpe V10 − A1; parity in QuantConnect

Note. RV = realized variance; ToD = time of day; OOS = out of sample. All studies execute at the next open after a decision at the close (S2: one minute after the 5-min decision bar) and deduct measured bid–ask spreads plus slippage and commissions.

Figure 2

Estimation, Test and Post-Cutoff Windows by Study



Note. Light gray: training or development; mid gray: validation; dark gray: out-of-sample test; black: later-date evaluation window. For S1–S4, the black window begins at or immediately after the TimesFM 2.5 public release in September 2025. For S5 and S6, 2024+ is used as a stress-test window because the official model card dates some named pretraining sources through November 2023 (Google Research, 2025). This does not establish that every pretraining series ends in 2023. S6 has no holdout: its variants were selected on the full period.

Statistical Tests and the Metric Convention

Forecasts are compared by QLIKE, mean squared log error, or pinball loss with Diebold–Mariano tests, encompassing regressions, and a 90% model confidence set. Trading effects are paired Sharpe differences from a stationary bootstrap with a 20-day mean block. For the synthesis, every headline series is recomputed under one convention: daily net returns; Sharpe and Sortino ratios on returns in excess of DTB3; and CAGR, volatility, drawdown, and tail statistics on total returns. S2's intraday books hold no capital overnight and are treated as self-financing overlays. Every Sharpe ratio in this paper is therefore an excess-return Sharpe ratio. The Monte Carlo analysis jointly resamples S6 daily returns across series (10,000 paths, fixed seed); descriptive pooling uses DerSimonian–Laird random effects (DerSimonian & Laird, 1986). Where decision thresholds were not pre-registered, the 90% intervals are reported as descriptive uncertainty intervals rather than confirmatory tests.

Results

H1: Advantage Over Naive and GARCH Benchmarks Is Target-Dependent

H1 receives partial support (Table 3). In S1, zero-shot TimesFM's QLIKE for SPY is 0.286 against 0.428 for persistence, 0.393 for EWMA, and 0.342 for GARCH, with Diebold–Mariano t statistics from −5.0 to −7.5 across the three ETFs; its 80% quantile band covers 80.7% of outcomes against 73.3% for HAR. It also beats EWMA in S3 ($t = -2.23$) and the naive volatility forecast in S6. The advantage is not universal: TimesFM only ties EWMA in S4 (log R^2 .703 against .697) and loses to a time-of-day model on S2's 60-minute volume target (MSE 0.119 against 0.071), whose pre-close ramp is not represented in the three days of context used there. Figure 3 illustrates the typical behavior: TimesFM broadly tracks HAR, reacts somewhat more slowly to shocks, and widens its quantile band in crises.

Table 3

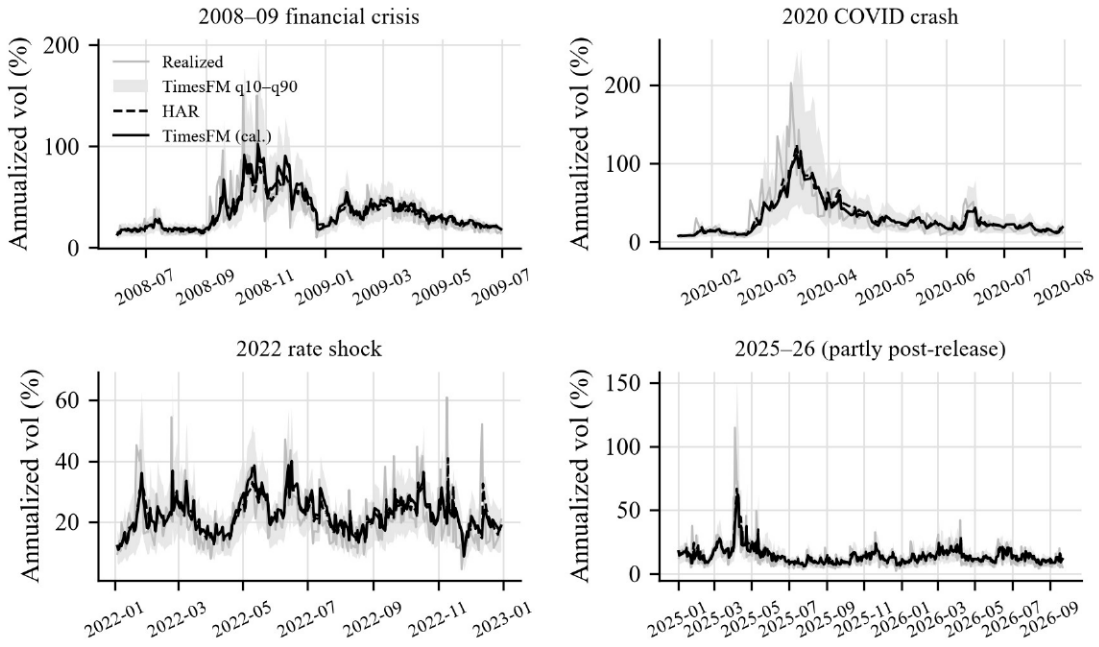
Forecast Accuracy of Zero-Shot TimesFM Against Conventional Volatility Forecasters (H1)

Study	Target	Loss	Persistence	EWMA	GARCH	HAR	HARX / best	TimesFM
S1 SPY	Next-day RV	QLIKE	0.428	0.393	0.342	0.289	0.252	0.286
S1 QQQ	Next-day RV	QLIKE	0.393	0.350	0.311	0.266	0.239	0.264
S1 IWM	Next-day RV	QLIKE	0.312	0.275	0.259	0.217	0.197	0.214
S2 (lockbox)	60-min RV	QLIKE	0.599	0.197	—	0.140	0.131[a]	0.154
S3 (2012–2025Q3)	20-day vol	QLIKE	—	0.459	—	0.478	—	0.423
S4 (2006–2025Q3)	20-day vol	QLIKE	2.214[b]	1.987	—	—	—	2.247
S6 equity ETFs	4-week vol	QLIKE	0.440	—	—	—	0.318	0.373
S6 non-equity ETFs	4-week vol	QLIKE	0.414	—	—	—	0.304	0.384

Note. Lower is better. QLIKE on variance forecasts; each study's own loss implementation and sample. [a]Walk-forward XGBoost, the best S2 model. [b]Trailing 20-day realized volatility. S1 sample: 2008-01 to 2026-09. S6: all weeks 2000–2026. Sources: timesfm-vol/reports/tables/daily_loss.csv, timesfm-mr/reports/g2_forecast_scores.csv, timesfm-sharpe/reports/03_forecast_eval.json, timesfm-cagr/reports/table_diag_vol.csv, timesfm-deploy/reports/forecast_skill.csv.

Figure 3

SPY Next-Day Volatility Forecasts Against Realized Values in Four Episodes (S1)



Note. Annualized volatility in percent. TimesFM (cal.) is the bias-calibrated zero-shot forecast; the shaded band is its 10th–90th percentile forecast.

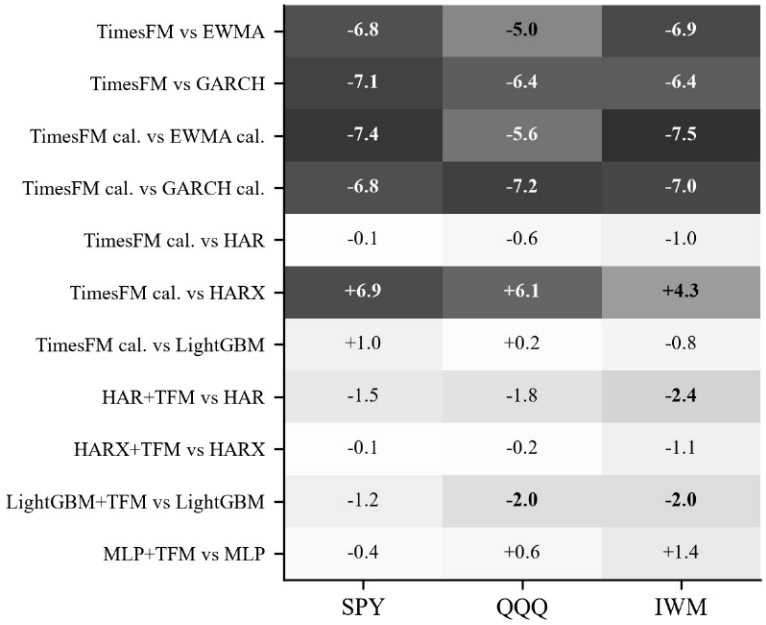
H2: Little Incremental Information

H2 is not supported (Figure 4). Against HAR, TimesFM is a statistical tie (Diebold–Mariano $t = -0.14, -0.57, -1.00$); against HARX it is clearly worse ($t = 6.88, 6.05, 4.30$).

Adding TimesFM to HARX changes little, and HARX is in the 90% model confidence set for all three ETFs while zero-shot TimesFM is in none. In a LightGBM feature-group ablation, the QLIKE contribution attributed to implied-volatility features is 0.0183 versus 0.0016 for TimesFM; in this specification, option-implied information therefore carries substantially more incremental signal. Even the TimesFM 3.0 research arm with VIX as a covariate trails HARX (t = 1.29). The other studies point in the same direction. In S2, walk-forward HAR and XGBoost beat TimesFM in every split (lockbox QLIKE 0.131 and 0.140 versus 0.154); its encompassing weight next to XGBoost is 0.16 and next to HAR on volume is −0.10. In S3, its QLIKE advantage over HAR is not significant (p = .226). In S5, it beats a format-matched AR(1) on some state variables, such as SPY–QQQ correlation (−9% pinball loss), but is worse on every return series. In S4, its return forecast has a cross-sectional rank IC of 0.013 against 0.057 for 12-1 momentum, and realized prices finish above its median forecast 58% of the time, consistent with a reversion bias in trending markets.

Figure 4

Diebold–Mariano Statistics for TimesFM Against Each Baseline (S1)



Note. QLIKE differential, first model minus second; negative values favor the first model. Bold: |t| > 1.96; shading proportional to |t|. n = 4,709 daily forecasts per ETF. cal. = bias-calibrated; TFM = TimesFM feature added.

H3: No Reliable Incremental Economic Value Against Strong Controls

H3 is not supported against the strongest same-slot conventional controls. One pre-registered S2 comparison—TimesFM versus a trailing gate—passes its original criterion, but the advantage largely disappears against the cheap time-of-day-adjusted forecast. Across the other designs, positive point estimates are small or unstable, while S4 is materially negative. Table 4 places every study's headline arms on one metric convention, and Table 5 reports each study's own test.

Table 4

Headline Arms of Every Study Under One Metric Convention

	Arm	CAGR %	Vol %	Sharpe	Sortino	Max DD %	Calmar	Skew	Ex. kurt.	VaR95 %	ES95 %	Worst yr %	Turnover ×/yr	Cost %/yr
S1	SPY buy-and-hold	11.4	19.2	0.59	0.82	−52.0	0.22	−0.38	11.7	1.81	2.93	−37.1	0.1	0.00
S1	Trend+VT, HAR	9.2	11.7	0.69	0.94	−18.9	0.49	−0.66	3.4	1.27	1.86	−13.9	15.4	0.22
S1	Trend+VT, HARX	9.3	11.6	0.70	0.96	−18.5	0.50	−0.67	3.2	1.25	1.84	−13.8	14.9	0.21
S1	Trend+VT, TimesFM	9.4	11.9	0.70	0.95	−18.1	0.52	−0.67	3.9	1.28	1.89	−14.1	12.6	0.18

	Arm	CAGR %	Vol %	Sharpe	Sortino	Max DD %	Calmar	Skew	Ex. kurt.	VaR95 %	ES95 %	Worst yr %	Turnover ×/yr	Cost %/yr
S2	Residual book, gate B (trailing)	−5.1	1.4	−3.69	−4.05	−13.1	−0.39	−1.60	11.8	0.18	0.28	−6.4	321.6	5.51
S2	Residual book, gate B2 (cheap forecast)	−3.1	1.3	−2.52	−2.84	−8.4	−0.37	−1.88	16.6	0.15	0.25	−4.5	232.6	3.99
S2	Residual book, gate C (TimesFM)	−2.6	1.1	−2.43	−2.79	−6.9	−0.37	−1.66	21.9	0.12	0.21	−3.4	186.2	3.16
S2	Raw VWAP book, gate C (TimesFM)	−5.0	3.7	−1.43	−1.59	−14.7	−0.34	−5.72	71.4	0.28	0.69	−7.0	272.5	5.07
S3	NONE (no forecasts)	4.1	6.3	0.40	0.53	−15.3	0.27	−1.22	9.1	0.60	1.00	−5.9	11.4	0.62
S3	CLASSICAL (HAR etc.)	4.8	5.9	0.54	0.73	−11.4	0.42	−0.87	5.1	0.57	0.93	−5.1	13.0	0.66
S3	TIMESFM	5.0	6.6	0.51	0.68	−15.5	0.32	−1.65	17.5	0.60	1.04	−6.6	11.6	0.62
S4	Without TimesFM (spliced)	6.0	8.9	0.53	0.72	−15.5	0.39	−0.70	4.8	0.87	1.40	−8.4	12.9	0.60
S4	With TimesFM (spliced)	3.2	7.5	0.25	0.33	−15.3	0.21	−0.90	6.1	0.75	1.20	−7.6	14.7	0.65
S5	Equal-weight sleeves	7.9	6.9	0.90	1.24	−11.3	0.69	−0.65	5.4	0.69	1.04	−9.2	5.2	0.11
S5	B conventional	8.6	8.0	0.85	1.17	−13.4	0.64	−0.82	6.4	0.78	1.27	−10.7	6.0	0.13
S5	C +TimesFM	8.7	7.8	0.89	1.23	−15.6	0.56	−0.68	4.7	0.76	1.24	−12.8	5.6	0.12
S5	C0 +AR(1) control	8.9	8.3	0.86	1.19	−16.0	0.55	−0.71	5.2	0.81	1.32	−12.6	5.5	0.12
S6	SPY buy-and-hold	8.3	18.6	0.41	0.57	−55.2	0.15	−0.08	10.2	1.79	2.82	−38.3	—	—
S6	A1 HARX, no TimesFM	10.0	14.5	0.58	0.80	−25.0	0.40	−0.63	3.6	1.51	2.22	−20.2	—	—
S6	A2 pure TimesFM	10.7	15.0	0.61	0.83	−25.3	0.42	−0.60	4.2	1.55	2.33	−20.6	—	—
S6	V10 deployed (HARX+TimesFM)	10.1	14.4	0.59	0.81	−25.7	0.39	−0.65	3.6	1.49	2.21	−20.7	—	—

Note. Daily net returns; Sharpe and Sortino on returns in excess of the 3-month T-bill (FRED DTB3, accrued by calendar days); CAGR, volatility, drawdown, skew, excess kurtosis, one-day historical VaR and expected shortfall (ES) at 95% on total returns. S2 is an intraday overlay with no overnight capital, so its daily P&L on $1 million is treated as an excess return. S2 covers only the 2024-01 to 2026-08 lockbox. S4 tracks are spliced from per-configuration return paths along the published walk-forward schedule; the projects' stitched simulation reports CAGR 3.1% and 5.9% (Sharpe 0.235 and 0.509). Turnover and cost drag come from each project's own accounting and are not strictly comparable across studies (S2 counts both legs and both sides). The largest gap between these Sharpe ratios and the projects' own figures is 0.017 (S4 splice); all others are within 0.003.

Table 5

Economic Value of TimesFM as Reported by Each Study (H3)

	Comparison	Unit	Estimate [90% interval]	p	Verdict
S1	Trend + vol target: TimesFM vs HAR sizing, SPY / QQQ / IWM	ΔSharpe	+0.008 / +0.018 / +0.044	.625 / .263 / .015	Not significant after BH
S1	HARX+TimesFM vs HARX sizing, SPY / QQQ / IWM	ΔSharpe	+0.002 / +0.007 / +0.021	.735 / .364 / .014	6 of 81 nominal, 0 after BH
S2	Residual book lockbox: C (TimesFM) − B (trailing)	Δ return %/yr	+2.7 [1.5, 3.9]	< .001	Pre-registered criterion met
S2	Residual book lockbox: C − B2 (cheap forecast)	Δ return %/yr	+0.5 [−0.4, 1.5]	.167	Not significant
S2	Raw book lockbox: C − B	Δ return %/yr	+0.3 [−1.1, 1.6]	.375	Not significant
S2	All gates, realistic costs, lockbox	Net bps/trade	−6.9 (resid. C); −3.9 (raw C)	—	Strategy discarded
S3	TIMESFM − CLASSICAL, 2012–2026 (pre-registered)	ΔSharpe	−0.030 [−0.256, 0.201]	.561	Not supported (p ≥ .10)
S3	TIMESFM − CLASSICAL, lockbox 2025-10+	ΔSharpe	−0.187 [−0.687, 0.230]	.774	—
S4	With − without TimesFM, full grid	Δ CAGR %/yr	−2.7 [−4.7, −0.8]	P(Δ > 0) = .008	Significantly negative increment
S5	C (+TimesFM) − B (conventional)	ΔSharpe	+0.041 [−0.105, 0.177][a]	.273	Not significant
S5	C0 (+AR(1), same format) − B	ΔSharpe	+0.007 [−0.153, 0.154][a]	.467	Control
S6	V10 (HARX+TimesFM) − A1 (HARX)	ΔSharpe	+0.007 [−0.007, 0.021]	.205	Not significant
S6	A2 (pure TimesFM) − A1 (HARX)	ΔSharpe	+0.025 [−0.010, 0.063]	.124	Not significant

Note. Paired block-bootstrap or project-specific tests, each study's own implementation and Sharpe convention (all excess of T-bills). p is one-sided for H3 (probability that the difference is ≤ 0) except S1, whose bt_incremental.csv reports a two-sided bootstrap p. [a]95% interval, as reported by S5. BH = Benjamini–Hochberg at a 10% false discovery rate over S1's 81 paired tests. The S2 residual C − B p is below the resolution of 3,000 bootstrap draws. A uniform re-estimate of every Sharpe difference is in Table 7 and Figure 13.

In S1, TimesFM sizing adds +0.008, +0.018 and +0.044 Sharpe over HAR sizing; of 81 paired comparisons 6 have $p < .05$ and 0 survive Benjamini–Hochberg control. The gain is smoothness (turnover 12.6 against 15.4 times a year in SPY), and TimesFM sizing does worse when VIX is at or above 25 (Sharpe 0.46 against 0.60).

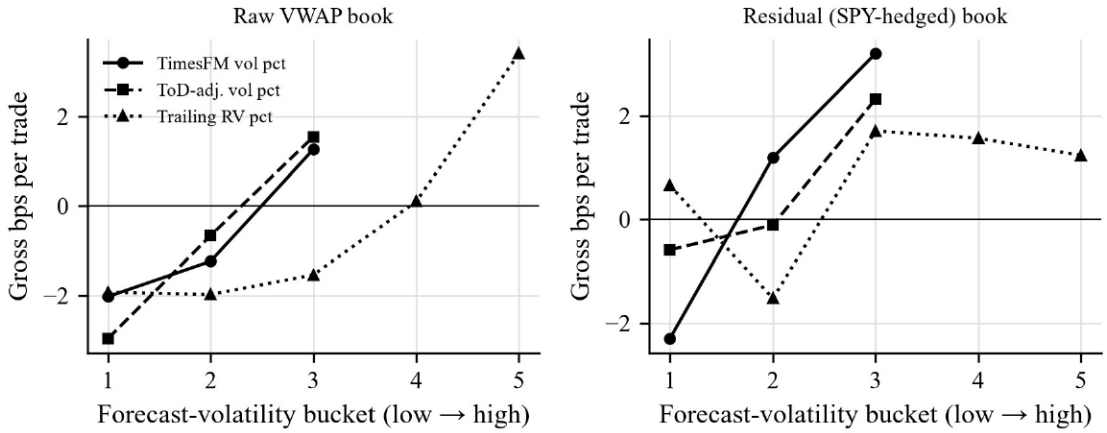
In S2 every book loses money after realistic costs in every period (all eight gates in both books; timesfm-mr/reports/tables); the ungated raw book's gross edge in training is −0.43 basis points per trade against 5.6 of cost. Gross expectancy rises with forecast volatility for every forecaster (Figure 5), so the volatility regime matters, but cheap forecasts find it as well as TimesFM. The TimesFM gate beats the trailing gate in the residual lockbox by 2.7 percentage points a year, meeting the pre-registered criterion, but beats the two-line time-of-day forecast gate by only 0.5 (90% interval −0.4 to 1.5).

In S3, the pre-registered endpoint failed: TIMESFM's Sharpe ratio is 0.515 against 0.546 for the classical arm (one-sided p = .561; Figure 6), and every active arm trails SPY (0.84), a 60/40 portfolio (0.83), and a monthly inverse-volatility basket (0.83). In S4, the TimesFM increment is materially negative: the walk-forward system compounds at 3.1% a year versus 5.9% without it (Sharpe 0.235 versus 0.509), and the increment is positive in only 0.8% of bootstrap resamples. Removing TimesFM's return forecasts alone recovers most of the gap (Figure 7), and neither version approaches buy-and-hold SPY (14.2%) or a plain top-three momentum rule (16.0%). In S5, the TimesFM allocator gains +0.041 Sharpe (95% interval [−0.105, +0.177]) with a maximum drawdown 2.1 points deeper; no feature group helps reliably

on its own (Figure 8), and equal-weighting the sleeves (0.898) is as good as any learned allocator. In S6, the HARX-plus-TimesFM model beats HARX alone by +0.007 Sharpe (90% interval [−0.007, +0.021]) and pure TimesFM by an insignificant +0.025. Cost stress tests do not change the rankings (Figure 9): TimesFM's S1 book degrades more slowly than HAR because it trades less, and it overtakes the S3 classical arm only at two to three times measured costs.

Figure 5

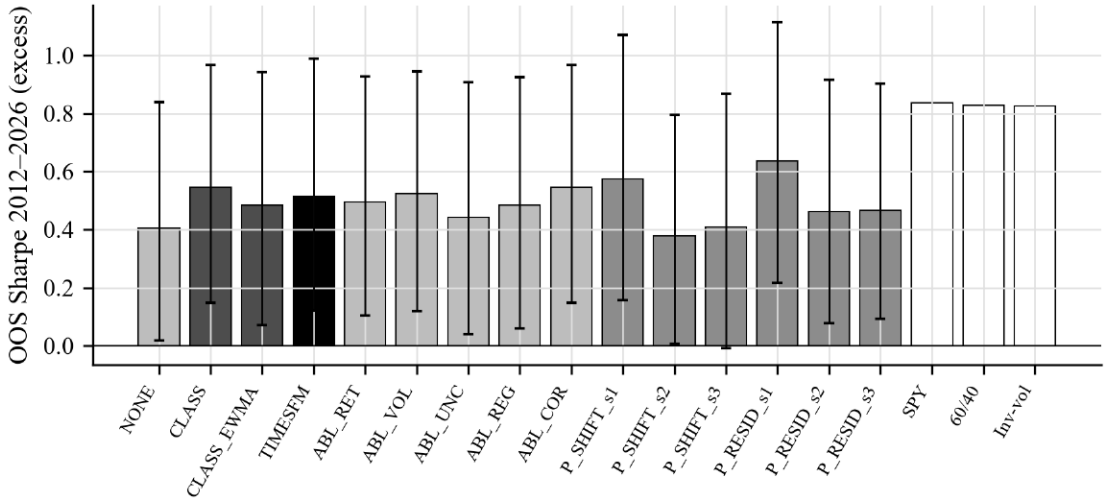
Gross Expectancy of Mean-Reversion Trades by Forecast-Volatility Bucket, Training Window (S2)



Note. Gross basis points per trade before costs, 2019-02 to 2021-12. Buckets are percentiles of each forecaster's 60-minute volatility forecast (terciles for the time-of-day-adjusted forecast). Average realistic costs are 5.6 bps (raw) and 11.1 bps (residual) per trade.

Figure 6

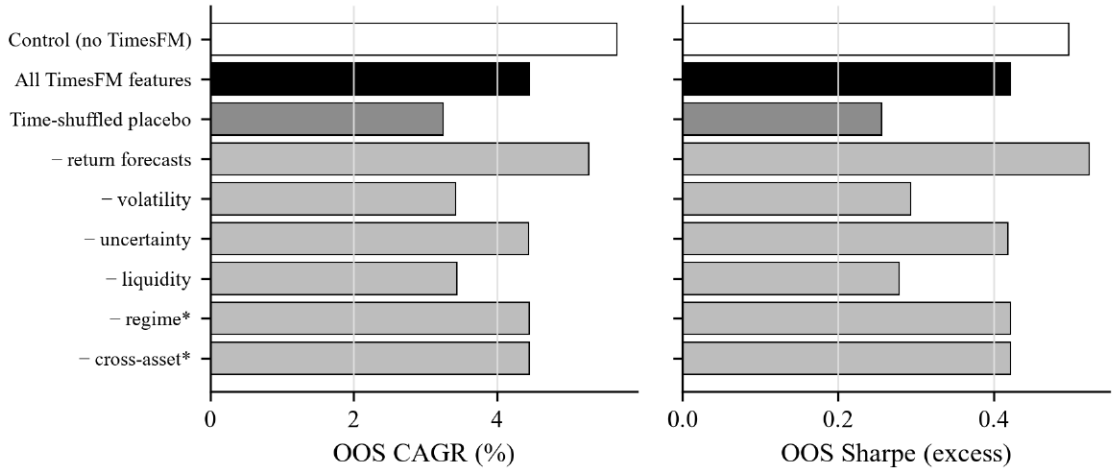
Out-of-Sample Sharpe Ratios of All S3 Arms and Three Passive Baselines, 2012–2026



Note. Error bars: 90% block-bootstrap intervals. Black: TIMESFM; dark gray: classical; light gray: classical with one TimesFM family swapped in; mid gray: placebos; white: passive baselines.

Figure 7

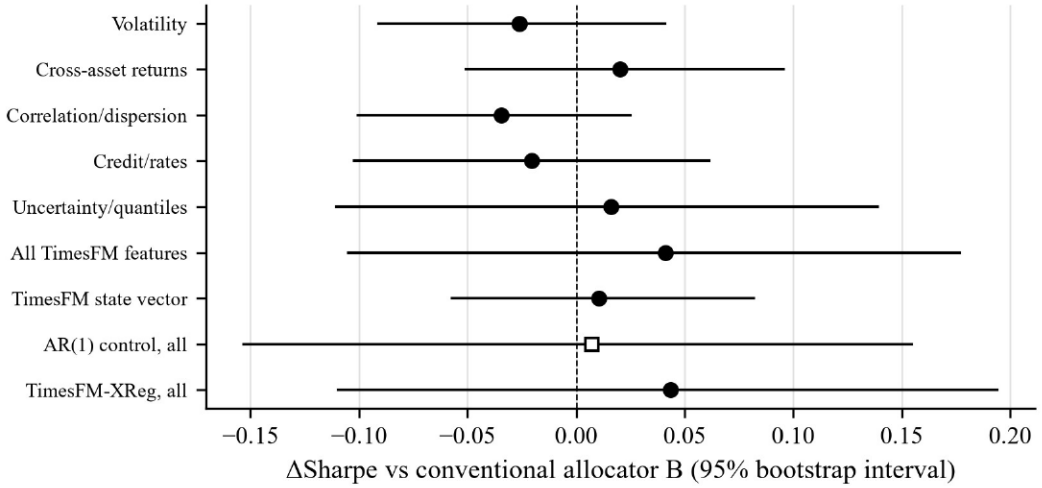
Out-of-Sample CAGR and Sharpe Ratio of the S4 Ablation Arms, 2011–2026



Note. All arms rerun the nested walk-forward on a common reduced grid. White: no TimesFM; black: all TimesFM features; mid gray: time-shuffled placebo; light gray: one group removed. *Uninformative by construction (identical selections).

Figure 8

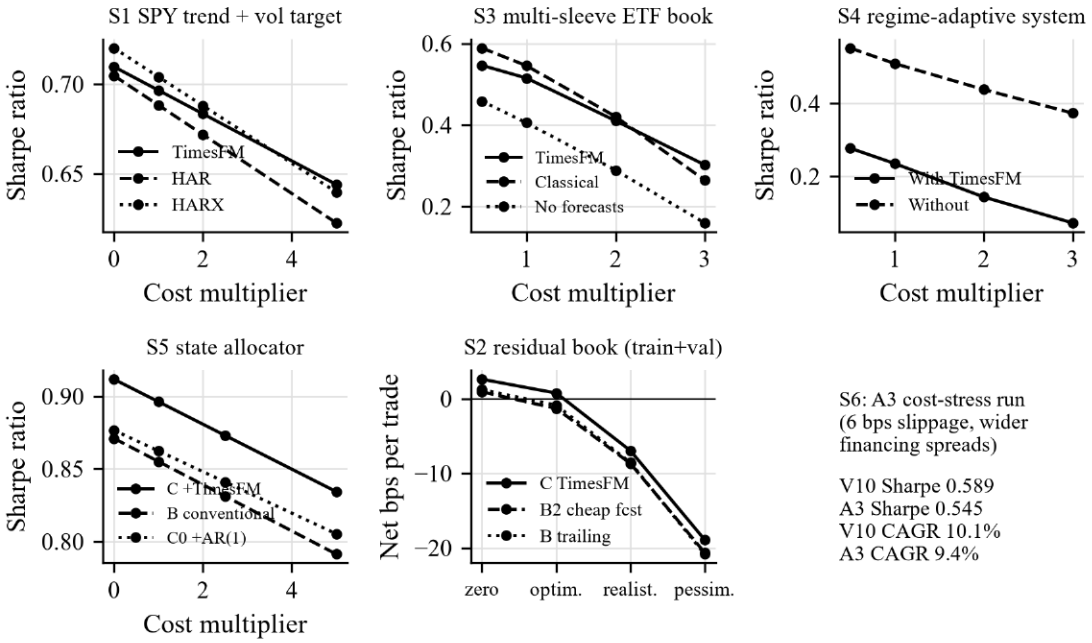
Sharpe Ratio Gains Over the Conventional Allocator by TimesFM Feature Group (S5)



Note. Paired block-bootstrap estimates with 95% intervals as reported by the project; open squares are AR(1) forecasts in the identical format. 768 weekly decisions, 2012–2026.

Figure 9

Sharpe Ratio and Net Expectancy Against Transaction-Cost Multipliers



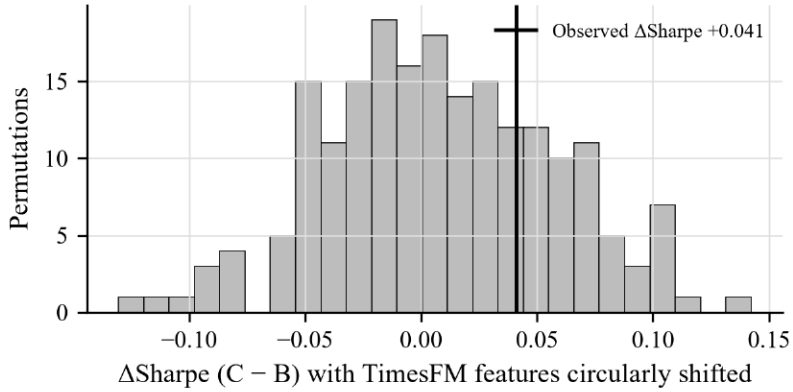
Note. Multiples of each study's measured cost model; the S2 panel shows train-plus-validation net basis points per trade under the four pre-registered cost scenarios.

H4: Placebos and Format-Matched Controls

H4 is not supported. In S3, a residual-shifted placebo scores a higher Sharpe ratio than TIMESFM (0.636 versus 0.515), so the pre-registered placebo condition fails. In S4, the point estimate favors real over time-shuffled features by +0.174 Sharpe, but the 90% interval [−0.043, +0.387] spans zero and both arms trail the no-TimesFM control. In S5, TimesFM beats a format-matched AR(1) by an insignificant +0.034 and exceeds 73% of 200 circular shifts of its own features (Figure 10). In S2, a one-session-stale TimesFM gate retains most of the live gate's effect (−7.5 versus −7.0 net basis points per trade), indicating that slow volatility regime information, rather than forecast freshness, explains most of the gating effect.

Figure 10

Permutation Distribution of the S5 Sharpe Gain With Circularly Shifted TimesFM Features



Note. Each draw shifts all TimesFM features by a random offset, destroying their timing but keeping their distribution; the vertical line is the observed gain.

H5: Forecast Skill in Later-Date Evaluation Windows

H5 is not supported. TimesFM's loss relative to its baseline worsens in 17 of 19 later-window comparisons (Table 6, Figure 11). In S1 it is slightly better than HAR before 2020 (−2.2% relative loss for SPY) and worse after release (+2.6%); in S5, its state-variable advantage over AR(1) (−5.6%) vanishes in 2024+ (+0.5%); and in S6, its equity-ETF gap to HARX

roughly doubles (Figure 12). S3's lockbox (ΔSharpe −0.187) and S4's post-release returns (−1.5% with TimesFM versus 16.0% without) point in the same direction. This pattern is compatible with pretraining overlap, but it is not diagnostic: later windows are short, financial regimes changed, and the public model documentation does not establish a single series-level pretraining cutoff.

Table 6

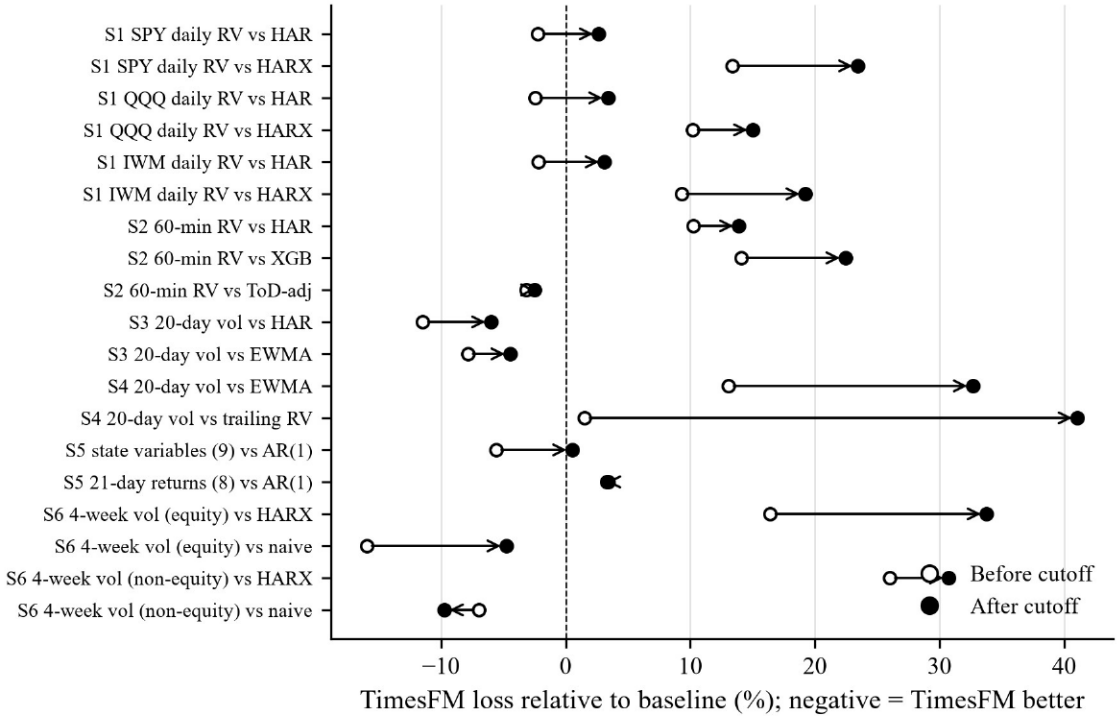
TimesFM Forecast Loss Relative to Baselines Across Earlier and Later Evaluation Windows (H5)

	Comparison	Loss	Before	Rel. loss %	After	Rel. loss %
S1	SPY daily RV vs HAR	QLIKE	2008–2019	−2.2	2025-09+	+2.6
S1	SPY daily RV vs HARX	QLIKE	2008–2019	+13.4	2025-09+	+23.5
S1	QQQ daily RV vs HAR	QLIKE	2008–2019	−2.4	2025-09+	+3.4
S1	QQQ daily RV vs HARX	QLIKE	2008–2019	+10.2	2025-09+	+15.0
S1	IWM daily RV vs HAR	QLIKE	2008–2019	−2.2	2025-09+	+3.1
S1	IWM daily RV vs HARX	QLIKE	2008–2019	+9.3	2025-09+	+19.2
S2	60-min RV vs HAR	QLIKE	2019–2025-09	+10.3	2025-10+	+13.9
S2	60-min RV vs XGB	QLIKE	2019–2025-09	+14.1	2025-10+	+22.4
S2	60-min RV vs ToD-adj	QLIKE	2019–2025-09	−3.2	2025-10+	−2.5
S3	20-day vol vs HAR	QLIKE	2012–2025Q3	−11.5	2025-10+	−6.0
S3	20-day vol vs EWMA	QLIKE	2012–2025Q3	−7.9	2025-10+	−4.4
S4	20-day vol vs EWMA	QLIKE	2006–2025Q3	+13.1	2025-09+	+32.7
S4	20-day vol vs trailing RV	QLIKE	2006–2025Q3	+1.5	2025-09+	+41.1
S5	state variables (9) vs AR(1)	Pinball	2008–2023	−5.6	2024+	+0.5
S5	21-day returns (8) vs AR(1)	Pinball	2008–2023	+3.4	2024+	+3.3
S6	4-week vol (equity) vs HARX	QLIKE	2000–2023	+16.4	2024+	+33.8
S6	4-week vol (equity) vs naive	QLIKE	2000–2023	−16.0	2024+	−4.8
S6	4-week vol (non-equity) vs HARX	QLIKE	2000–2023	+26.0	2024+	+30.7
S6	4-week vol (non-equity) vs naive	QLIKE	2000–2023	−7.0	2024+	−9.8

Note. Relative loss = TimesFM loss / baseline loss − 1; negative values favor TimesFM. For S1–S4, the later window begins at the model's public release (2025-09-15) or the first full month after it. For S5 and S6, 2024+ is used as a stress-test window because the official model card dates some named pretraining sources through November 2023 (Google Research, 2025). This is not evidence that every pretraining series ends in 2023. S2 values are recomputed from the forecast panel (post-release n = 9,632 symbol-hours). S5 rows average the per-series pinball-loss changes versus AR(1) reported in the project tables.

Figure 11

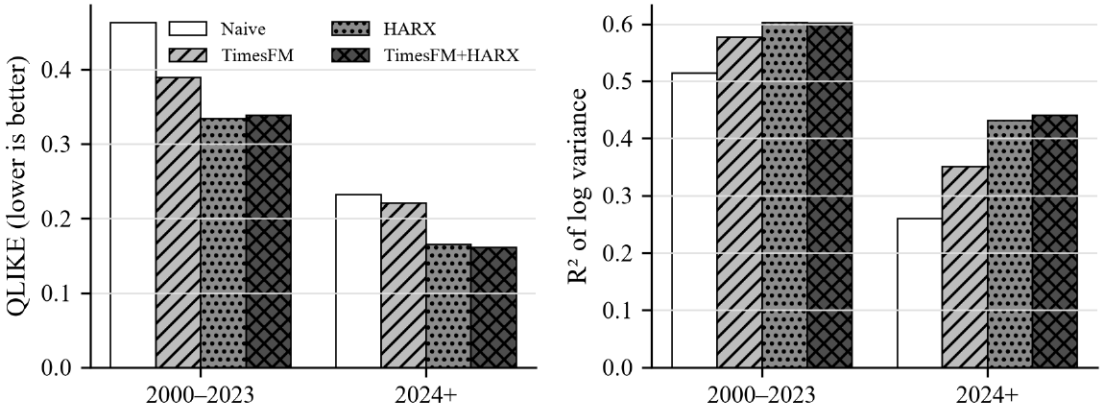
TimesFM Loss Relative to Each Baseline in Earlier (Open) and Later (Filled) Evaluation Windows



Note. Relative loss = TimesFM loss / baseline loss − 1, in percent; negative values favor TimesFM. Windows are listed in Table 6.

Figure 12

Four-Week ETF Volatility Forecast Skill Before and After 2024 (S6, Equity ETFs)



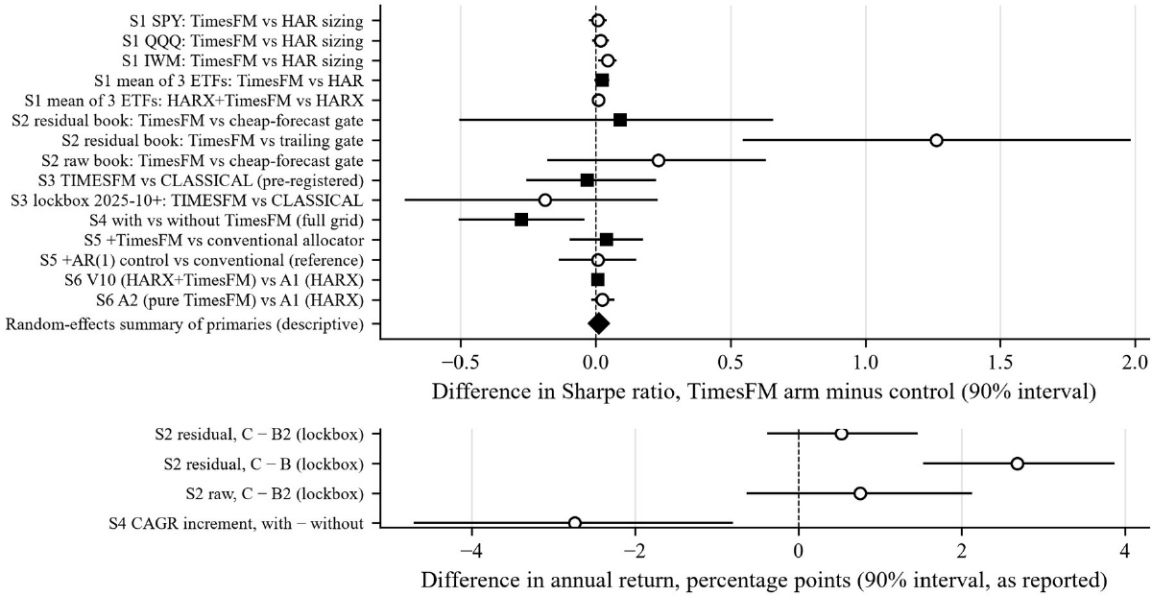
Note. QLIKE and R^2 of log variance. 2000–2023 n = 18,334 ETF-weeks; 2024+ n = 2,070.

Cross-Study Synthesis

Re-estimating every TimesFM-versus-control Sharpe difference with one bootstrap and one convention (Figure 13, Table 7) confirms the project-level pattern. S4 is significantly negative (−0.277). Three positive intervals exclude zero: S2's TimesFM gate against the trailing gate (+1.262), which shrinks to +0.092 with an interval spanning zero against the cheap-forecast gate; S1's IWM sizing gain (+0.044); and S1's mean HARX-plus-TimesFM gain of +0.010 (one-sided p = .048), which does not survive the Benjamini–Hochberg control applied to S1's own 81 tests. Pooling each study's primary comparison gives +0.012 Sharpe (90% interval [−0.004, +0.027]; I^2 = 9%). The studies share data, periods, implementation choices, and an author, so this pooled estimate is descriptive rather than an independent-replication meta-analysis; its interval includes zero.

Figure 13

Sharpe Ratio Effect of TimesFM in Every Study, Re-Estimated Under One Convention



Note. Upper panel: paired stationary-bootstrap differences (20-day mean block, 10,000 draws) with 90% intervals; squares are each study's primary comparison, pooled in the descriptive random-effects summary (diamond). Lower panel: effects reported in return units by S2 and S4.

Table 7

Uniform Re-Estimate of the TimesFM Sharpe Effect and Descriptive Pooling

	Primary comparison	ΔSharpe	90% interval	SE	FE weight %	RE weight %
S1	TimesFM vs HAR	+0.023	[−0.000, 0.047]	0.014	26.1	33.2
S2	TimesFM vs cheap-forecast gate	+0.092	[−0.501, 0.653]	0.353	0.0	0.1
S3	TIMESFM vs CLASSICAL (pre-registered)	−0.030	[−0.253, 0.221]	0.143	0.3	0.4
S4	with vs without TimesFM (full grid)	−0.277	[−0.503, −0.045]	0.139	0.3	0.5
S5	+TimesFM vs conventional allocator	+0.041	[−0.092, 0.172]	0.080	0.8	1.4
S6	V10 (HARX+TimesFM) vs A1 (HARX)	+0.007	[−0.006, 0.022]	0.009	72.5	64.5
	Fixed-effect pooled estimate	+0.011	[−0.001, 0.023]	—	100	—
	Random-effects (DerSimonian–Laird) estimate	+0.012	[−0.004, 0.027]	—	—	100

Note. Paired stationary bootstrap (Politis & Romano, 1994), mean block 20 days, 10,000 draws, on each study's daily net returns under the unified convention. Heterogeneity: $Q(5) = 5.49$, $p = .359$, $I^2 = 9\%$, $\tau^2 = 0.00006$. The studies share data, overlapping periods and one author, so the pooled estimate is descriptive and not a test. S1 enters as the mean of its three ETFs; S2 as the TimesFM gate against the cheap-forecast gate.

A program-wide sensitivity analysis further tempers the headline deployment Sharpe. The research program computed out-of-sample results for 16,440 configurations, arms, and design changes (Table 16, Appendix F). S6's reported deflated Sharpe ratio of .997 counts only its 10 core variants and uses their narrow cross-trial Sharpe dispersion (0.035). If the calculation instead uses the program-wide trial count together with the pooled trial-Sharpe dispersion (0.148), the deflated Sharpe ratio falls to .50; under that pooled dispersion it drops below .95 once roughly 18 trials are counted (Table 8). This is a deliberately conservative sensitivity analysis, not a claim that all 16,440 trials are independent or exchangeable.

Table 8

Deflated Sharpe Ratio of the Deployed V10 System Under Alternative Trial Counts and Trial-Sharpe Dispersions

Trials N	S6 V-variants (as in S6) (SD 0.035)	0.10 (SD 0.100)	S3 configs (SD 0.133)	Pooled S1+S3+S4+S6 (SD 0.148)
10	.997	.987	.976	.968
14	.997	.984	.969	.957
100	.995	.959	.905	.867
1,000	.993	.914	.792	.711
16,440	.990	.839	.623	.499
17,952	.990	.837	.617	.493

Note. Deflated Sharpe ratio (Bailey & López de Prado, 2014) = probability that V10's true Sharpe exceeds the expected maximum of N trials with the given cross-trial SD, adjusted for skewness and kurtosis of V10's 6,971 daily returns; V10 Sharpe = 0.589. Row N = 10 with the S6 SD reproduces the project's figure (.9972 vs reported .9974). With the pooled SD the DSR falls below .95 once N exceeds about 18.

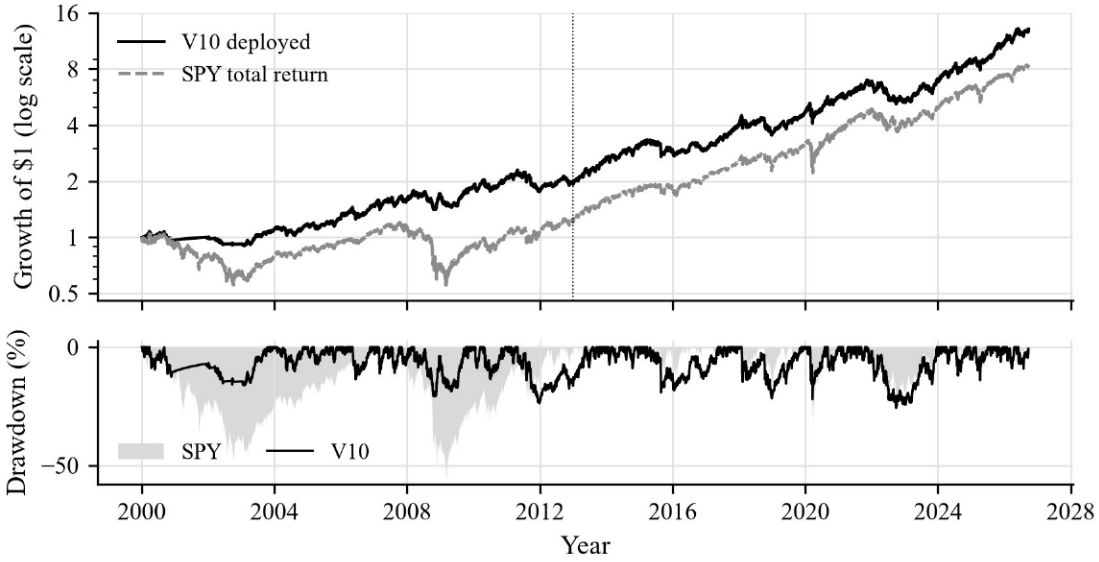
The Deployed System

V10 returned 10.1% a year from 2000 to 2026 with a Sharpe ratio of 0.589 and a maximum drawdown of −25.7%, against 8.3%, 0.406 and −55.2% for SPY, with a daily beta of 0.41 (Figure 14).

Its variant path (Table 9) shows where full-sample performance changes occurred: removing sector ETFs from the momentum universe is the largest improvement, while raising the volatility target increases CAGR without improving Sharpe. The system's full-sample lead over SPY is concentrated before 2013 (Table 10): 5.2% versus 1.5% a year in 2000–2012, but 14.8% versus 15.1% in 2013–2026, with similar Sharpe ratios and a shallower drawdown. V10 and A1 are nearly identical in every sub-period; their rolling one-year Sharpe difference exceeds 0.10 in absolute value in only 3.1% of windows, and V10 wins 12 of 27 calendar years (Figure 15). Target gross exposure averages 1.14 times net assets and reaches the cap in 24.3% of weeks, which by construction are low-forecast-volatility weeks.

Figure 14

Growth of One Dollar and Drawdowns of the Deployed System (V10) and SPY, 2000–2026



Note. LEAN cloud equity curves net of fees, slippage and financing; log scale. The dotted line marks 2013.

Table 9

Sequence of QuantConnect Cloud Runs in the Deployment Study (S6)

Run	Change	CAGR %	Vol %	Sharpe	Max DD %	Worst yr %	Orders
V01	Base: vol target 12%, IEF defensive	7.6	10.5	0.547	−18.0	−10.1	3,212
V02	Vol target 15%	8.6	13.0	0.533	−23.9	−11.2	3,339
V03	Vol target 18%	9.5	15.2	0.531	−28.5	−10.4	3,265
V04	3/6/12-month momentum blend	8.0	13.0	0.492	−29.9	−12.8	3,653
V05	Defensive = best of IEF/TLT/GLD	8.8	13.0	0.549	−23.8	−15.2	3,278
V06	SPY trend gate on momentum	8.0	12.4	0.512	−23.0	−14.3	3,142
V07	Permanent 30% defensive sleeve	8.6	12.9	0.536	−29.1	−8.5	3,035
V08	Trend sleeve only	7.6	11.6	0.511	−24.8	−17.9	1,472
V09	Momentum on 13 asset-class ETFs (no sectors)	9.1	12.3	0.594	−21.6	−17.1	2,571
V10	V09 with vol target 18% (deployed)	10.1	14.4	0.589	−25.7	−20.7	2,538
A1	V10 with HARX vol (no TimesFM)	10.0	14.5	0.582	−25.0	−20.2	2,600
A2	V10 with pure TimesFM vol	10.7	15.0	0.607	−25.3	−20.6	2,409
A3	V10 with 6 bps slippage, wider financing	9.4	14.4	0.545	−26.1	−21.2	2,539
V10-1999	V10 started 1999-01	10.3	14.4	0.591	−25.7	−20.7	2,584

Note. All runs 2000-01-03 to 2026-09-22 on the LEAN cloud engine, financed leverage cap 1.5×, weekly market-on-open rebalancing; Sharpe on excess of DTB3 computed from the exported daily equity curve. Runs were made sequentially and each choice was informed by earlier full-period results, so V10 is selected in sample. Source: timesfm-deploy/reports/variants.csv.

Table 10

Deployed System (V10), Its No-TimesFM Twin (A1) and SPY by Sub-Period

Period	CAGR V10	CAGR A1	CAGR SPY	Sharpe V10	Sharpe A1	Sharpe SPY	Max DD V10	Max DD SPY
2000–2009	6.5	6.2	−0.8	0.34	0.32	−0.05	−20.7	−55.2
2010–2019	9.7	9.9	13.4	0.65	0.66	0.88	−23.7	−19.3
2020–2026	16.4	16.3	15.5	0.79	0.79	0.67	−25.7	−33.7
2000–2012	5.2	5.0	1.5	0.29	0.27	0.08	−23.7	−55.2
2013–2026	14.8	14.8	15.1	0.82	0.82	0.80	−25.7	−33.7
Full 2000–2026	10.1	10.0	8.3	0.59	0.58	0.41	−25.7	−55.2

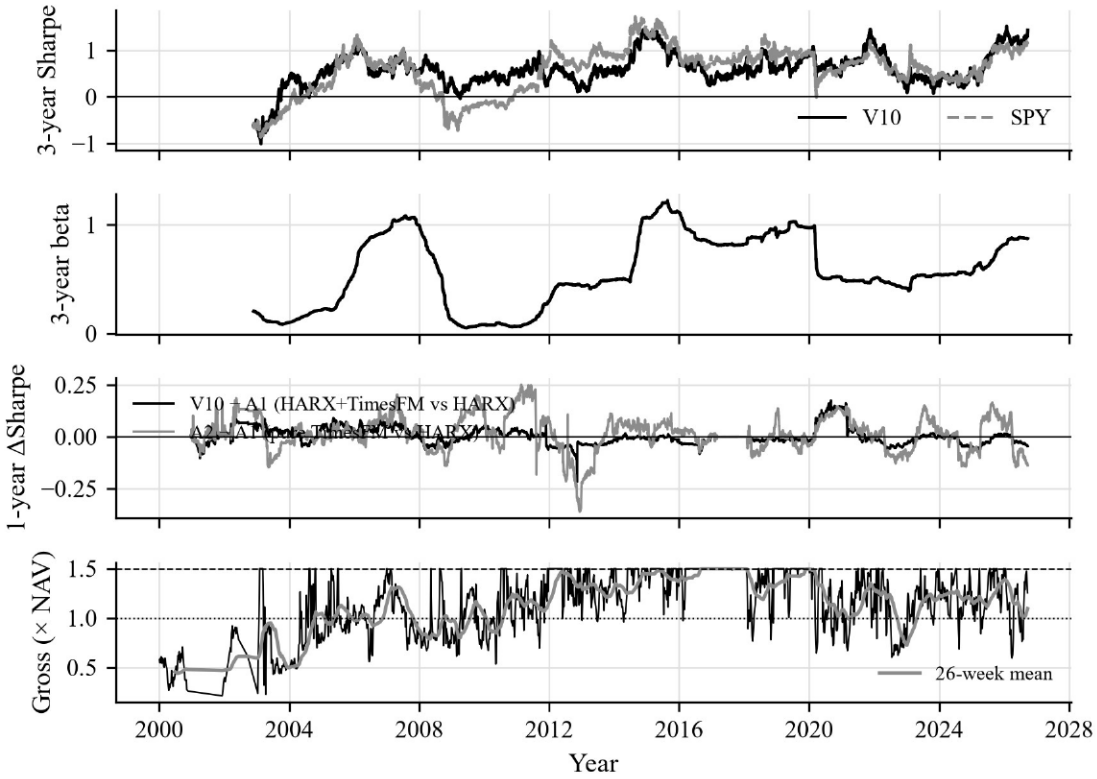
Note. CAGR and maximum drawdown in percent. Sub-period statistics follow the slicing convention of timesfm-deploy/scripts/04_final_stats.py and reproduce final_bundle.json to within 0.002 percentage points.

Daily returns are negatively skewed and fat-tailed, but less so than SPY's, and monthly returns are narrower than SPY's in both tails (Figure 16; Table 13, Appendix B); the calendar heatmap shows near-zero stretches in cash such as 2001 (Figure 17).

The block-bootstrap Monte Carlo (Table 11, Figure 18) puts V10's full-period Sharpe ratio between 0.30 and 0.88 (90% band) and its median maximum drawdown at −30.5%, deeper than the realized −25.7%. Over 10 years the chance of a drawdown worse than −30% is 22% for V10 against 67% for SPY, and V10 ends ahead of SPY in 65% of paths. The paired distribution of the TimesFM effect reproduces the project's bootstrap and is insensitive to block lengths of 5 and 60 days; S3's and S4's own Monte Carlo analyses give 5th-percentile Sharpe ratios of 0.17 and −0.16.

Figure 15

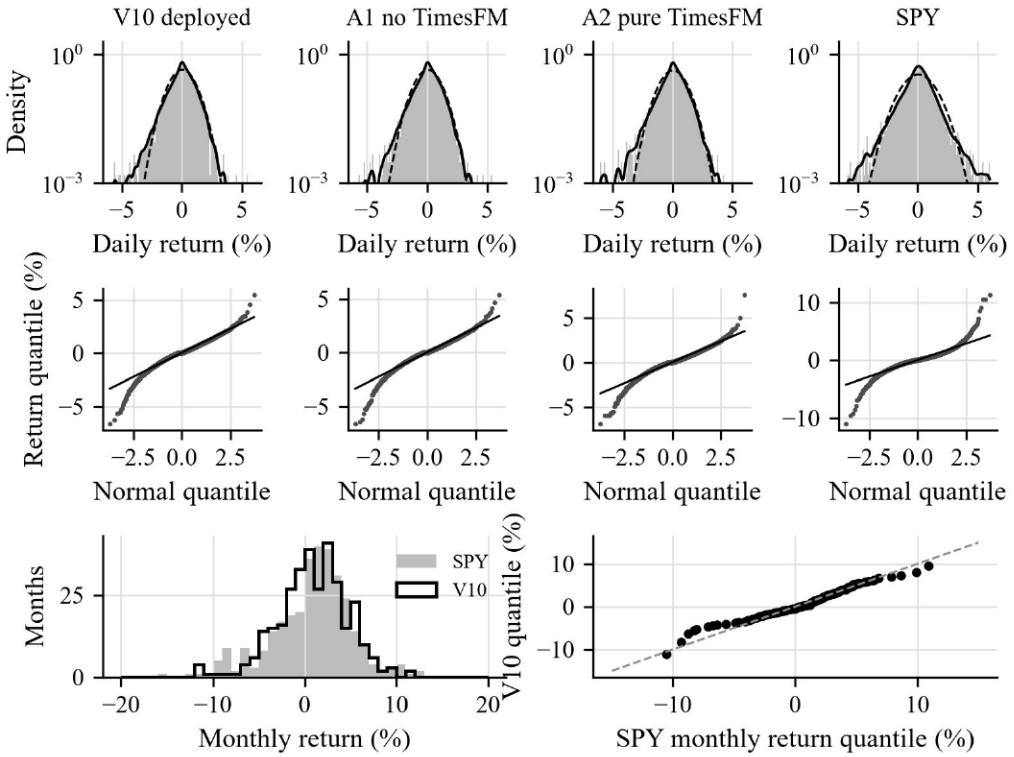
Rolling Risk, TimesFM Effect and Leverage of the Deployed System



Note. From top: 3-year Sharpe ratio of V10 and SPY; 3-year beta of V10 to SPY; 1-year Sharpe difference of V10 and of pure-TimesFM A2 against the HARX-only twin A1; weekly target gross exposure (dashed: 1.5 times cap; dotted: 1.0).

Figure 16

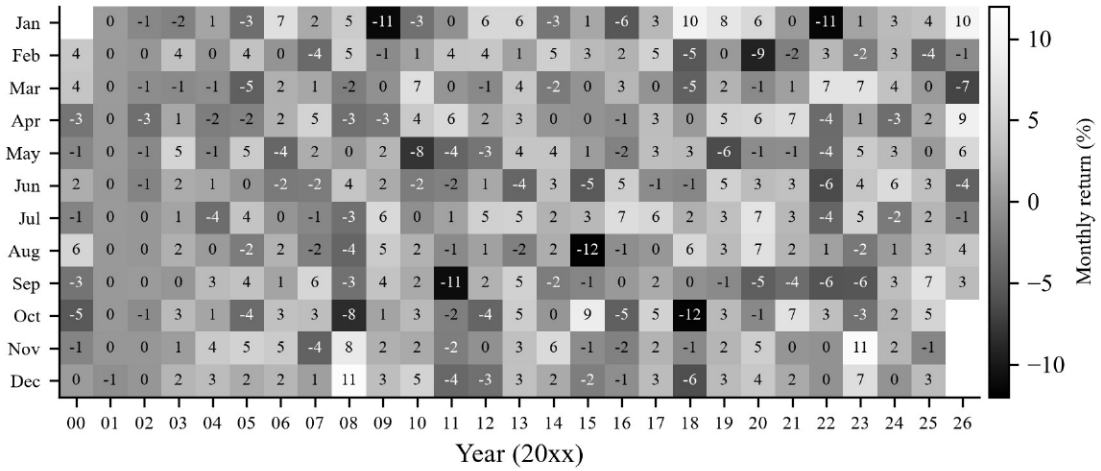
Daily and Monthly Return Distributions of the Deployed System, Its Ablations and SPY (S6)



Note. Top row: daily histogram, kernel density (solid) and normal density with equal mean and variance (dashed), log scale, exchange holidays excluded. Middle row: normal quantile–quantile plots. Bottom row: monthly return histograms of V10 and SPY, and V10 monthly quantiles against SPY's (dashed: 45-degree line).

Figure 17

Monthly Returns of the Deployed System (V10), 2000–2026



Note. Percent per month; darker cells are more negative.

Table 11

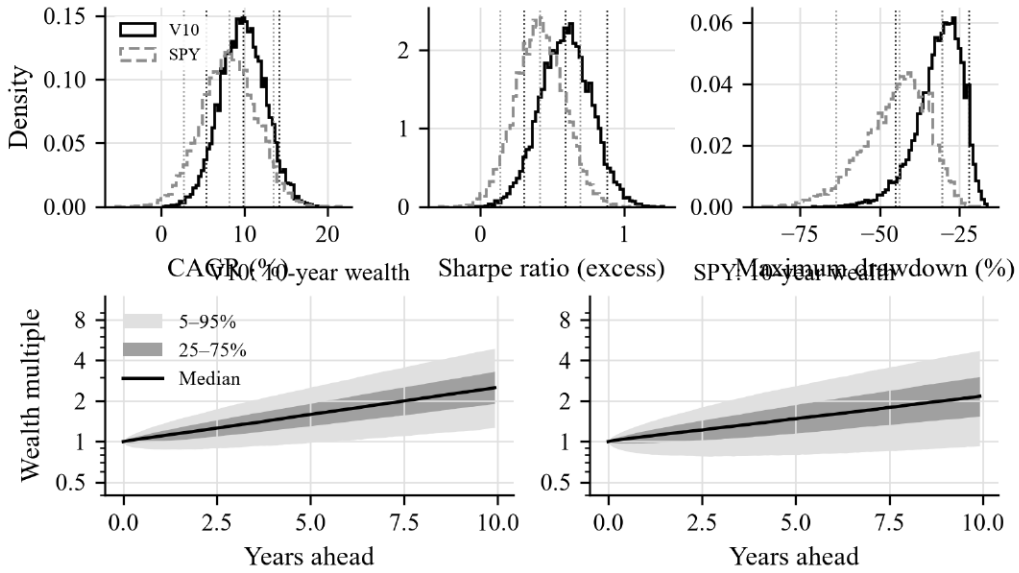
Stationary Block-Bootstrap Monte Carlo of the Deployment Study

Series	CAGR % median [5, 95]	Sharpe median [5, 95]	Max DD % median [5, 95]	P(DD < −20%) 10y	P(DD < −30%) 10y	P(DD < −40%) 10y	Median 10y wealth
V10	9.8 [5.4, 14.2]	0.59 [0.30, 0.88]	−30.5 [−45.1, −22.0]	77.9	22.3	4.0	2.51
A1	9.7 [5.3, 14.1]	0.59 [0.30, 0.87]	−30.9 [−45.7, −22.3]	78.8	23.6	4.3	2.49
A2	10.4 [5.7, 15.0]	0.61 [0.32, 0.90]	−31.4 [−46.1, −22.6]	81.2	25.6	4.9	2.64
SPY	8.1 [2.6, 13.5]	0.41 [0.14, 0.69]	−43.8 [−63.7, −30.6]	96.6	67.2	30.8	2.17

Note. 10,000 paths, mean block 20 trading days, seed 20260924, rows resampled jointly across series and the T-bill. Columns 2–4: full-length (6,971-day) paths; columns 5–8: 10-year (2,520-day) paths. P(V10 ends 10 years ahead of SPY) = .65; P(V10 ahead of A1) = .60. Paired ΔSharpe V10 − A1 median +0.007, 90% band [−0.006, 0.021] (block 5: [−0.008, 0.022]; block 60: [−0.007, 0.022]); the project's own bootstrap gives [−0.007, 0.021].

Figure 18

Monte Carlo Distributions of Full-Period Performance and Ten-Year Wealth Fan Charts, V10 and SPY



Note. Top: 10,000 stationary-bootstrap paths of 6,971 days (20-day mean block); dotted lines mark the 5th, 50th and 95th percentiles. Bottom: percentiles of wealth over jointly resampled 2,520-day paths, log scale.

H6: In-Platform Forecast Parity Demonstrated; Full Run Pending

H6 is partially supported. Forecast parity is demonstrated, but full-run completion was still pending at the manuscript build. The V2 project vendors the TimesFM 2.5 code into the algorithm and loads fp16 weight shards from QuantConnect's object store; against the reference package, 1,000 random forecasts differ by at most 0.00094 in log variance. The 2020 in-platform

smoke test compares 1,144 weekly forecasts with the offline pipeline (maximum difference 1.38×10^{-3}) and returns 22.1% versus 21.8% for V1. The full 2000–2026 in-platform run, created 2026-09-24 17:03:59 UTC, was 85% complete as of 2026-09-24 17:48 UTC and is therefore not used as evidence. Independent-engine reconciliations are directionally consistent elsewhere: S4's TimesFM track compounds at 3.11% in-house, 3.10% in Backtrader, and 2.85% in LEAN; S1's TimesFM book differs from its Backtrader replica by 0.02 points of CAGR and returns 9.26% a year in LEAN; and S5's LEAN replication trails Python by about 0.35% a year without changing the conclusion.

Discussion

The results suggest three reasons forecast improvements fail to become reliable trading improvements. First, the relevant conventional forecasts are already strong: volatility targeting depends primarily on getting the level of risk approximately right, and HAR augmented with implied volatility does so without a foundation model. TimesFM's smoother path can also be approximated with an EWMA or a no-trade band. Second, inexpensive forecasts identify many of the same regimes, as shown by the stale and time-of-day-adjusted gates in S2. Third, any residual information is small relative to strategy noise: S5's Sharpe difference has a standard error of about 0.08, so economically modest effects are difficult to distinguish even over roughly 15 years. Return-forecast use is least successful; in S4, TimesFM reduces exposure during trends that subsequently continue, consistent with the broader difficulty of forecasting excess returns (Welch & Goyal, 2008) and with evidence that heavy architectures need not beat simple baselines on noisy time-series targets (Zeng et al., 2023).

The deterioration in later-date windows raises a contamination concern but does not resolve it. The pattern is compatible with pretraining overlap, especially when evaluation series

are public, but it is also compatible with ordinary distribution shift, short post-release samples, and changing market microstructure. The defensible conclusion is therefore narrower: benchmark performance on potentially overlapping historical data should not be treated as equivalent to genuinely prospective evidence, and post-release evaluation deserves separate weight.

The deployment study reinforces that distinction between system performance and model contribution. Most of V10's full-sample relative performance is concentrated in the early period, when its trend and momentum rules reduce exposure during major bear markets; the TimesFM component adds only +0.007 Sharpe relative to the HARX-only twin. The deployment mandate is therefore met technically, but the model's incremental economic contribution is negligible in this specification. These results motivate five evaluation practices for financial foundation models: benchmark against the strongest same-slot conventional forecaster, include format-matched simple controls, report genuinely later-date evidence separately, disclose the full path of researcher choices and trials, and review model-weight licensing independently of source-code licensing.

Limitations and Threats to Validity

Pretraining overlap. Public documentation identifies several TimesFM 2.5 pretraining sources but does not disclose series-level membership for all data. Overlap with the market series used here therefore cannot be ruled out. Pre-2024 results should be interpreted as potentially exposed rather than as proven uncontaminated evidence; conversely, the short post-2023 and post-release windows may confound overlap with regime or distribution shift. **Hindsight in S6.** Its variants and universe were chosen using full-period results without a holdout, so V10's absolute performance is selected in sample. The V10-versus-A1 TimesFM ablation is cleaner

because those arms share all other design choices, but it is still embedded in a system selected through the same research path. **Dependence.** The studies share data vendors, periods, securities, implementation conventions, and one research program, so they are not independent replications. **Compute.** CPU-only inference constrained context length and forced S2 amendment A1, which dropped trade-count forecasts and TimesFM on holdout symbols (reported as not run). **Researcher degrees of freedom.** S3 ran one full-sample smoke test (amendment A0); S4's control-arm fixes followed its out-of-sample results; and decision rules for S1, S4, S5, and S6 were stated after the fact. The 432-configuration first S4 grid is counted in the program registry even though it was omitted from that project's own deflated Sharpe calculation. **Resampling and DSR sensitivity.** Stationary-bootstrap intervals and Monte Carlo paths inherit assumptions about dependence and resampling stationarity. The program-wide deflated-Sharpe calculation is intentionally conservative because its trials are not fully exchangeable; it should be read as sensitivity analysis rather than a formal multiplicity correction. **Engines and scope.** Cash handling differs across engines, the S4 synthesis uses spliced walk-forward tracks, and this paper does not test TimesFM 3.0 multivariate forecasting, fine-tuning, alternative covariate schemes, or passive execution in S2.

Conclusion

Across six designs and 16,440 logged trials, TimesFM 2.5 improves several volatility forecasts over naive and GARCH-family benchmarks, but not the strongest conventional baseline, and it does not deliver robust incremental net returns. Its weaker later-window performance warrants caution but does not prove contamination. Financial foundation models should be evaluated against strong same-slot baselines, on genuinely later data, with the full research path disclosed.

References

Aksu, T., Woo, G., Liu, J., Liu, X., Liu, C., Savarese, S., Xiong, C., & Sahoo, D. (2024). *GIFT-Eval: A benchmark for general time series forecasting model evaluation*. arXiv. https://doi.org/10.48550/arXiv.2410.10393

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*(2), 579–625. https://doi.org/10.1111/1468-0262.00418

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=gerNCVqqtR

Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, *40*(5), 94–107. https://doi.org/10.3905/jpm.2014.40.5.094

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, *61*(5), 458–471. https://doi.org/10.1090/noti1105

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

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

Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 235, pp. 10148–10167). PMLR. https://proceedings.mlr.press/v235/das24c.html

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

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

Google Research. (2025). TimesFM 2.5-200m [Model card]. Hugging Face. https://huggingface.co/google/timesfm-2.5-200m-pytorch

Google Research. (2026). TimesFM 3.0 (PyTorch) [Model card]. Hugging Face. https://huggingface.co/google/timesfm-3.0-pytorch

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. https://doi.org/10.3982/ECTA5771

Harvey, C. R., Hoyle, E., Korgaonkar, R., Rattray, S., Sargaison, M., & Van Hemert, O. (2018). The impact of volatility targeting. *The Journal of Portfolio Management*, *45*(1), 14–33. https://doi.org/10.3905/jpm.2018.45.1.014

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5–68. https://doi.org/10.1093/rfs/hhv059

Harvey, D. I., Leybourne, S. J., & Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics*, *16*(2), 254–259. https://doi.org/10.1080/07350015.1998.10524759

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.

Moreira, A., & Muir, T. (2017). Volatility-managed portfolios. *The Journal of Finance*, *72*(4), 1611–1644. https://doi.org/10.1111/jofi.12513

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. *Journal of Financial Economics*, *104*(2), 228–250. https://doi.org/10.1016/j.jfineco.2011.11.003

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870

Rasul, K., Ashok, A., Williams, A. R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Darvishi Bayazi, M. J., Adamopoulos, G., Riachi, R., Hassen, N., Biloš, M., Garg, S., Schneider,

A., Chapados, N., Drouin, A., Zantedeschi, V., Nevmyvaka, Y., & Rish, I. (2023). *Lag-Llama: Towards foundation models for probabilistic time series forecasting*. arXiv. https://doi.org/10.48550/arXiv.2310.08278

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10776–10787). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.722

Tan, M., Merrill, M. A., Gupta, V., Althoff, T., & Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, *37*, 60162–60191. https://doi.org/10.52202/079017-1922

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508. https://doi.org/10.1093/rfs/hhm014

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 235, pp. 53140–53164). PMLR. https://proceedings.mlr.press/v235/woo24a.html

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(9), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317

Appendix A

Data Validation

Defects were found by automated checks in each project's data pipeline: price jumps at symbol splices, bars dated before listing, returns inconsistent with published annual totals, and high–low ranges far outside the open–close body. After correction, S5's rebuilt total returns match published annual figures (SPY 2008 −36.8%, TLT 2022 −31.2%). Intraday bars are stamped at their end and daily rows contain only information available at the 16:00 close. In S2's 1-minute data all gaps above 20% are earnings events, at most 2% of minutes are missing, and zero-volume bars occur only in the March 2020 circuit breakers.

Table 12

Data Defects Found and Remedies Applied Across the Program

Defect	Example	Consequence if ignored	Remedy
Ticker reuse	Polygon HYG before 2007-04, EDV before 2007-12, SMH before 2011-12, META before 2022-06	Years of fictitious returns for another security	Hard-coded true inception dates; FB→META splice
Symbol change	QQQ has no bars 2004-12-01 to 2011-03-22 (traded as QQQQ)	1,588-day gap or a 42% jump at the join	Splice QQQQ
Corporate action	XLF 2016-09-19 XLRE spin-off: adjusted bars use factor 0.878 (true 0.812); splits endpoint lists a 1.139 split	Fake −6.9% or +14.6% day	Total return from unadjusted closes + distributions + integer splits only
Fat-finger wicks	SPY low 11.85 vs close 111.38 (2008-09-29); QQQ low 0.0 (2004-07-28)	Range-based vol estimators explode	Cap wick beyond body at 2.5× trailing median range (S3) or 6× session median (S1)
Dividends	adjusted=true covers splits only	Bond-ETF returns understated	Explicit dividend records
Clock mismatch	16:00 ETF close vs 15:00 cash-Treasury close	Spurious reversion (separate program; not a TimesFM study)	Align clocks with minute bars
Timestamp units	pandas 3 parquet index in µs vs ns integer keys	Silent empty joins, zero-trade backtests	Normalise to ns before integer comparisons

Note. Collected from the project data-validation code and notes; the clock-mismatch item comes from an adjacent research project and is listed because it affects any design mixing equity and rates closes. Details in Appendix A.

Appendix B

Additional Result Tables

Table 13

Moments of Daily Net Returns and Jarque–Bera Normality Tests

	Arm	n	Mean (bps)	SD %	Skew	Ex. kurt.	Min %	Max %	JB
S1	SPY buy-and-hold	4,708	5.02	1.21	−0.38	11.7	−12.9	11.5	26,892
S1	Trend+VT, HAR	4,708	3.76	0.74	−0.66	3.4	−5.5	4.0	2,572
S1	Trend+VT, HARX	4,708	3.81	0.73	−0.67	3.2	−4.9	3.6	2,330
S1	Trend+VT, TimesFM	4,708	3.85	0.75	−0.67	3.9	−6.2	4.8	3,296
S2	Residual book, gate B (trailing)	666	−2.10	0.09	−1.60	11.8	−0.6	0.5	4,055
S2	Residual book, gate B2 (cheap forecast)	666	−1.25	0.08	−1.88	16.6	−0.6	0.5	7,954
S2	Residual book, gate C (TimesFM)	666	−1.04	0.07	−1.66	21.9	−0.6	0.5	13,347
S2	Raw VWAP book, gate C (TimesFM)	650	−2.09	0.23	−5.72	71.4	−3.3	1.1	139,466
S3	NONE (no forecasts)	3,701	1.68	0.40	−1.22	9.1	−3.4	2.2	13,612
S3	CLASSICAL (HAR etc.)	3,701	1.95	0.37	−0.87	5.1	−2.8	1.7	4,495
S3	TIMESFM	3,701	2.01	0.42	−1.65	17.5	−5.4	3.2	48,612
S4	Without TimesFM (spliced)	3,953	2.48	0.56	−0.70	4.8	−3.9	2.7	4,056
S4	With TimesFM (spliced)	3,953	1.37	0.47	−0.90	6.1	−4.0	2.5	6,715
S5	Equal-weight sleeves	3,697	3.11	0.43	−0.65	5.4	−4.2	2.3	4,702
S5	B conventional	3,697	3.39	0.51	−0.82	6.4	−4.6	2.8	6,711
S5	C +TimesFM	3,697	3.44	0.49	−0.68	4.7	−3.2	2.8	3,649
S5	C0 +AR(1) control	3,697	3.51	0.52	−0.71	5.2	−4.2	3.1	4,492
S6	SPY buy-and-hold	6,971	3.75	1.17	−0.08	10.2	−10.9	11.3	30,470
S6	A1 HARX, no TimesFM	6,971	4.08	0.91	−0.63	3.6	−6.6	5.4	4,244
S6	A2 pure TimesFM	6,971	4.35	0.94	−0.60	4.2	−6.8	7.6	5,652
S6	V10 deployed (HARX+TimesFM)	6,971	4.10	0.90	−0.65	3.6	−6.6	5.5	4,235

Note. All series reject normality (Jarque–Bera $p < .001$; smallest JB = 2,330).

Appendix C

Amendment Logs

Table 14

Protocol Amendments and Researcher Degrees of Freedom

	Amendment	Data viewed before the change	Change
S2	A1	Before any forecast was evaluated	Compute cut: 16 series per origin instead of 29; trade-count forecasts and TimesFM on holdout symbols dropped
S2	A2	Before evaluation	Spread enters as a measured cost model; predicted spread dropped as a target
S2	A3	Before any gate was evaluated	Gate chosen by net bps per trade subject to ≥ 15% of ungated trades; common window from 2019-02
S2	Pre-lockbox	After validation, before lockbox	Choices frozen; strategy already discarded; lockbox opened once to test C > B
S3	A0	Full 2008–2026 (smoke test)	One full-sample smoke test printed before the development-window rule (disclosed)
S3	A1–A2	Full sample	Sleeve gross cap 4 → 2; walk-forward sleeve admission rule
S3	A3–A5	Development 2008–2011	SHY removed from trend/defensive sizing; momentum rank hysteresis; band and rebalance grids
S3	A6–A7	None (code review, tests)	Fat-finger wick cleaning; spread-fill bug fix
S3	A8	After OOS pass	Passive benchmarks bypass the 35% asset cap (benchmark bug); strategy arms unchanged
S4	v1 → v2	Control-arm OOS results	Momentum hysteresis and idle-capital fixes after seeing control results; v1 grid (432 configs) not in reported DSR
S5	α grid	Dry run on arms B and AR only	Ridge penalty grid widened to 1e5 before any TimesFM output existed
S6	V01–V10	Full-period cloud results	Sequential variant selection on 2000–2026 results (no holdout)

Note. Compiled from timesfm-mr/reports/AMENDMENTS.md, timesfm-sharpe/reports/AMENDMENTS.md, timesfm-allocator/REPORT.md §19 and the S4 and S6 project records. Every amendment is counted as an additional trial in the registry (Table 16).

Appendix D

Condensed Provenance Ledger

Every number in this paper is produced by a script from a file in one of the six project folders; provenance.csv lists each with its claim, source file, column or key, script and any discrepancy. The largest discrepancies are the S4 splice (CAGR within 0.15 points of the stitched simulation), S6's share of weeks at the leverage cap (24.3% with the project's 1.49 threshold, 23.7% at exactly 1.5) and S6's run count (14 rows in variants.csv against 15 in the study brief). None changes a conclusion. Table 15 lists the headline numbers.

Table 15

Condensed Provenance Ledger for Headline Numbers

Claim	Value	Source file	Script
S1 TimesFM vs HARX DM t, SPY	6.875	timesfm-vol/reports/tables/daily_dm.csv	s02_ingest.py
S1 paired tests surviving BH	0	timesfm-vol/reports/tables/bt_incremental.csv	s02_ingest.py
S2 lockbox resid. C − B2 %/yr	0.00526	timesfm-mr/reports/g7_paired_resid.json	s02_ingest.py
S2 registered trials	214	timesfm-mr/reports/TRIAL_REGISTRY.jsonl	s02_ingest.py
S3 TIMESFM − CLASSICAL ΔSharpe	-0.03043	timesfm-sharpe/reports/04_results.json	s02_ingest.py
S3 one-sided p	0.561	timesfm-sharpe/reports/04_results.json	s02_ingest.py
S4 CAGR increment	-0.02739	timesfm-cagr/reports/results.json	s02_ingest.py
S4 spliced TimesFM CAGR	0.03211	timesfm-cagr/data/results/arm_tfm.pkl	s01_series.py
S5 C − B ΔSharpe	0.04116	timesfm-allocator/results/tables/paired_tests.csv	s02_ingest.py
S5 AR(1) control ΔSharpe	0.007246	timesfm-allocator/results/tables/paired_tests.csv	s02_ingest.py
S6 V10 CAGR	0.101	timesfm-deploy/reports/final_bundle.json	s02_ingest.py
S6 V10 Sharpe	0.5891	timesfm-deploy/reports/final_bundle.json	s02_ingest.py
S6 V10 − A1 ΔSharpe	0.007335	timesfm-deploy/reports/ablation_boot.json	s02_ingest.py
S6 V10 2013–2026 CAGR	0.1483	out/series/S6.parquet	s05_rolling.py
S6 gross at cap	0.2435	timesfm-deploy/reports/final_bundle.json	s05_rolling.py
Uniform V10 − A1 ΔSharpe	0.007369	out/series/*.parquet	s06_synthesis.py
Random-effects pooled ΔSharpe	0.01173	out/tables/t_forest.csv	s06_synthesis.py
Program trials (core)	16,440	out/tables/t_trial_registry.csv	s06_synthesis.py
V10 DSR, pooled SD, core N	0.4993	out/tables/t_dsr_grid.csv	s06_synthesis.py
Post-cutoff comparisons worse	17	out/tables/t_prepost.csv	s06_synthesis.py
MC P(V10 beats SPY, 10y)	0.6488	scripts/s04_montecarlo.py	s04_montecarlo.py
V2 smoke max |Δ| log-var	0.00138	QC REST backtests/read a02c1b7588e6f3057b240df8dba0a0fb	s08_v2.py
V2 full-run status	In Progress...	QC REST backtests/read f6f3165e7fe65e0ff2d0d1a75bd2a3d4	s08_v2.py

Note. The full ledger (provenance.csv) has 3,262 rows with the column or key read and any discrepancy with other sources. Values are unformatted decimals (e.g., 0.1 = 10%).

Appendix E

Reproducibility

One command, `python scripts/build_all.py` in the timesfm-paper folder, regenerates every figure, table, number and both documents from the project outputs without running TimesFM or any cloud backtest (Python 3.11, pandas 3.0.6, NumPy, SciPy, matplotlib, reportlab, PyMuPDF; Node.js with docx; see README.md). Scripts s01–s09 assemble return series, extract reported numbers, compute the unified table and distributions, the Monte Carlo (10,000 paths, seed 20260924), the rolling analysis, the synthesis and deflated Sharpe ratios, the figures, the V2 status (read-only) and the tables; s10 writes the provenance ledger and build_paper.py the documents. Source projects: timesfm-vol (S1), timesfm-mr (S2), timesfm-sharpe (S3), timesfm-cagr (S4), timesfm-allocator (S5), and timesfm-deploy with the LEAN projects AIS TimesFM Deploy V1 and V2 (S6).

Appendix F

Program-Wide Trial Registry

Table 16

Program-Wide Trial Registry Used for the Deflated Sharpe Ratio

	Component	Trials	In core count	Source
S1	Paired with/without-TimesFM trading comparisons	81	Yes	timesfm-vol/reports/tables/bt_incremental.csv
S1	Vol-target sensitivity grid (504 configs x 3 ETFs)	1,512	No	timesfm-vol/reports/tables/bt_sensitivity.csv
S2	Registered gate/strategy configurations	214	Yes	timesfm-mr/reports/TRIAL_REGISTRY.jsonl
S2	Protocol amendments A1-A3	3	Yes	timesfm-mr/reports/AMENDMENTS.md
S3	Portfolio configs x arms (900 x 15)	13,500	Yes	timesfm-sharpe/reports/04_results.json
S3	Design amendments A0-A8	9	Yes	timesfm-sharpe/reports/AMENDMENTS.md
S4	Walk-forward grid, TimesFM + control arms	2,160	Yes	timesfm-cagr/reports/results.json
S4	v1 control grid run before v2 fixes (not in reported DSR)	432	Yes	timesfm-cagr/data/results/v1_arm_notfm.pkl
S5	Learned allocator trials	27	Yes	timesfm-allocator/results/tables/dsr.csv
S6	QuantConnect cloud runs (variants.csv)	14	Yes	timesfm-deploy/reports/variants.csv
	Total (core)	16,440		
	Total (including S1 sensitivity grid)	17,952		

Note. A trial is any configuration, arm or design change whose out-of-sample performance was computed and could have influenced what was reported. The brief for this paper listed 15 S6 cloud variants; variants.csv holds 14 rows, and 14 is used. Counting the list in the brief gives 15,997.