

Financial Forecasting with a Small Time-Series Foundation Model

A Local Evaluation of Timestral-28M

Philipp D. Dubach

Zurich, Switzerland

phdubach@pm.me

22 October 2026

Abstract

We evaluate Timestral-28M on daily returns, volume and range volatility using rolling annual adaptation and a locally reserved retrospective 2026 evaluation whose forecast origins precede model release. Thirty-six full-parameter trajectories require 101.6 minutes of training and validation on a 16 GB Apple M2 CPU. Among the evaluated combinations, the original model achieves 4.70% equal-asset relative-MAE skill for fourteen stocks’ fifth-session log-volume against conditional OLS (descriptive 95% interval: 2.08–6.98%). The adapted ensemble’s direct skill against the original is 0.14% [−1.15, 1.48]. Returns show no consistent benchmark advantage, and volatility results differ between stocks and equity funds. Unknown pretraining overlap limits generalization. The volume result supports a research baseline, with adaptation benefit and trading profitability unestablished.

Keywords: time-series foundation models, financial forecasting, domain adaptation, probabilistic prediction.

1 Motivation and related work

Chronos [1] and TimesFM [2] transfer forecasting patterns across numerical series. Timestral provides multivariate, covariate-aware forecasts at 21 quantile levels [3]. Mistral’s reported family-level benchmark rankings do not establish financial performance for the 28M checkpoint.

Recent financial evaluations find sparse improvements over a random-walk benchmark despite favorable rankings of pretrained models [4]. We ask whether trading activity is more forecastable than returns and whether adaptation improves pretrained forecasts. This case study tests one model and recipe; execution value requires a separate decision-based evaluation.

2 Experimental design

Data and targets. Daily Yahoo data from 2010 onward cover fourteen U.S. stocks (Appendix A) and three equity ETFs (SPY, QQQ, IWM). TLT, GLD and USO are excluded from adaptation and selection and tested for transfer. For adjusted close P , volume V , high H and

low L , the targets are

$$\begin{aligned} r_t &= \log(P_t/P_{t-1}), & v_t &= \log(1 + V_t), \\ s_t &= \log\left(\max\left\{\frac{\log(H_t/L_t)}{\sqrt{4\log 2}}, 10^{-8}\right\}\right). \end{aligned} \quad (1)$$

Forecasts condition on 512 completed daily bars; intraday availability and point-in-time provider vintages are unestablished. Offsets $h \in \{1, 5\}$ denote the daily observation at $t + h$, including for returns. The log Parkinson range proxy [5] omits overnight jumps and is not intraday realized volatility.

Chronology and adaptation. For each 2023–2025 test year, training labels span the five years ending two years earlier, followed by one validation year. The 2026 fold trains on 2020–2024 and validates on 2025. Evaluation retains 150 historical origins and forty 2026 origins, spaced five sessions apart. The latter run from 2 January to 13 October, with the last fifth-session outcome on 20 October.

Each task, year and seed resets the same 27,821,184 pretrained parameters for 500 AdamW updates at peak learning rate 10^{-5} . Validation selects steps 0, 100, 250 or 500 by mean asset/offset MAE ratios. Selection baselines are zero returns, persistence for $h = 1$ volume, the trailing 20-session median for $h = 5$ volume, and HAR-style range forecasts. OLS is an evaluation supplement, not a selection baseline. With $y \in \{r, v, s\}$ and Q the checkpoint’s 21 quantile levels, define $\rho_\tau(u) = u(\tau - \mathbf{1}_{u < 0})$. Training averages

$$\frac{1}{a_{i,t}} \left[\frac{1}{21} \sum_{\tau \in Q} \rho_\tau(y_{i,t+h} - \hat{q}_\tau) + \frac{1}{2} |y_{i,t+h} - \hat{q}_{0.5}| \right], \quad (2)$$

over sampled rows and both offsets. Past-context scales $a_{i,t}$ are RMS returns, RMS lag-five log-volume differences, and RMS lag-one log-range differences, floored at $10^{-4}, 10^{-3}, 10^{-3}$ respectively. Every fourth update uses related-series groups; inference is univariate. Three selected quantile arrays form an elementwise-median ensemble. The $3 \times 4 \times 3 = 36$ runs vary training order, not market histories. Step zero wins three of 27 historical selections.

Table 1: Illustrative 2026 comparisons at offset five, chosen after inspecting the results. Values are equal-asset skill percentages, Eq. (3). “Adapted” denotes the task-specific selected ensemble; HAR is the specified log-proxy regression.

Forecast / comparator	Skill	95% interval
Stocks: log volume, base / OLS	4.70	[2.08, 6.98]
Stocks: log volume, adapted / OLS	4.94	[2.76, 7.11]
Stocks: log volume, adapted / base	0.14	[−1.15, 1.48]
Stocks: return, adapted / median	−0.38	[−0.62, −0.14]
Stocks: log range, adapted / HAR	2.26	[0.69, 3.34]
ETFs: log range, adapted / HAR	−5.08	[−9.22, −1.64]

Comparators and uncertainty. Return comparators are zero, past means and medians. Volume comparators include persistence, rolling medians and OLS on the preceding 512 observations (Appendix A). The direct HAR-style OLS [6] uses current log-range and trailing 5/22-session means, with a fitted-residual median adjustment; it is not a high-frequency HAR-RV comparison.

For A assets, report equal-asset relative MAE skill,

$$S = \frac{100}{A} \sum_{i=1}^A \left(1 - \frac{\text{MAE}_{i,\text{model}}}{\text{MAE}_{i,\text{comparator}}} \right). \quad (3)$$

Positive skill means lower error. Descriptive 95% percentile intervals use 1,000 paired circular resamples with fixed four-origin blocks (about twenty sessions), sharing dates across assets. They condition on fitted models and selection, with no multiplicity correction or repeated training. Historical blocks cross fold boundaries and wrap endpoints; year counts are not fixed. Block-length sensitivity was not tested.

3 Results

Figure 1 summarizes 32 equity comparisons with named comparators; Table 1 illustrates the fifth-session results. The adapted ensemble’s historical stock log-volume skill against OLS is 2.68% [0.64, 4.92] at $h = 1$ and 2.97% [1.06, 4.95] at $h = 5$. Its historical $h = 1$ skill against the original is 1.08% [0.49, 1.69], without a clear 2026 continuation. In 2026, adapted stock $h = 1$, equity-ETF and external-fund volume intervals against OLS all span zero. Adapted stock $h = 5$ skill against rolling20 is 3.36% [−0.84, 7.31], so OLS superiority does not establish superiority to every cheap rule.

Selected historical return-MAE intervals against the original span zero. In 2026, adapted stock $h = 5$ returns lose to the past median by 0.38%. Adapted stock range improves on HAR by 2.26%, while equity ETFs lose 5.08%. The original already has approximately the adapted stock-range point skill. The adapted return ensemble’s nominal 90% stock coverage is 83.75% at $h = 1$ and 84.11% at $h = 5$.

All runs use CPU FP32 and two threads on a 16 GB M2. Training and validation sum to 101.6 minutes. The

observed interval from first training start to queue completion is 118.1 minutes, including preparation, gaps and overlapping analyses; initial setup, earlier studies and final audits are excluded. Separate verification code re-plays chronology and scores after relocation with original-workspace reads blocked.

4 Scope and implications

The 2023–2025 outcomes had already been examined. The OLS supplement froze before own-task adapted-volume and 2026 scores, but its historical comparisons remain adaptive. The 2026 outcomes were locally reserved before evaluation; all origins nevertheless precede Timestral’s release. This is a retrospective local hold-out. Unknown pretraining overlap, including overlap through correlated series, prevents interpreting chronological fine-tuning splits as proof of an uncontaminated foundation-model test [7].

A fixed surviving panel, retrospective adjustments, forty new origins and one recipe limit generalization. All twelve 2026 MAE-skill intervals comparing task-adapted forecasts with the original span zero across three tasks, two offsets and two equity panels. This leaves adaptation benefit unestablished. Daily log-volume accuracy alone establishes no intraday execution or capacity-allocation benefit. No strategy, costs, fills or profits were evaluated. A next test should freeze choices before collecting post-release observations and evaluate volume forecasts in a specified decision.

References

- [1] Abdul Fatir Ansari et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, (2024). URL <https://arxiv.org/abs/2403.07815>. arXiv:2403.07815.
- [2] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167, 2024. URL <https://proceedings.mlr.press/v235/das24c.html>.
- [3] Mistral AI. Introducing Timestral. Official model release, 15 October, 2026. URL <https://mistral.ai/news/timestral/>.
- [4] Miquel Noguer I Alonso and Rodolfo Pereira Franklin. Pre-trained time-series foundation models for financial return forecasting, 2026. URL <https://arxiv.org/abs/2606.27100>. Preprint, arXiv:2606.27100.
- [5] Michael Parkinson. The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, 53(1):61–65, 1980. doi: 10.1086/296071. URL <https://doi.org/10.1086/296071>.
- [6] Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2): 174–196, 2009. doi: 10.1093/jffinec/nbp001. URL <https://doi.org/10.1093/jffinec/nbp001>.
- [7] Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Rethinking evaluation in the era of time series foundation models: (Un)known information leakage challenges, 2026. URL <https://arxiv.org/abs/2510.13654>. Preprint, arXiv:2510.13654v3, revised 25 February.

A Comparisons and reproducibility

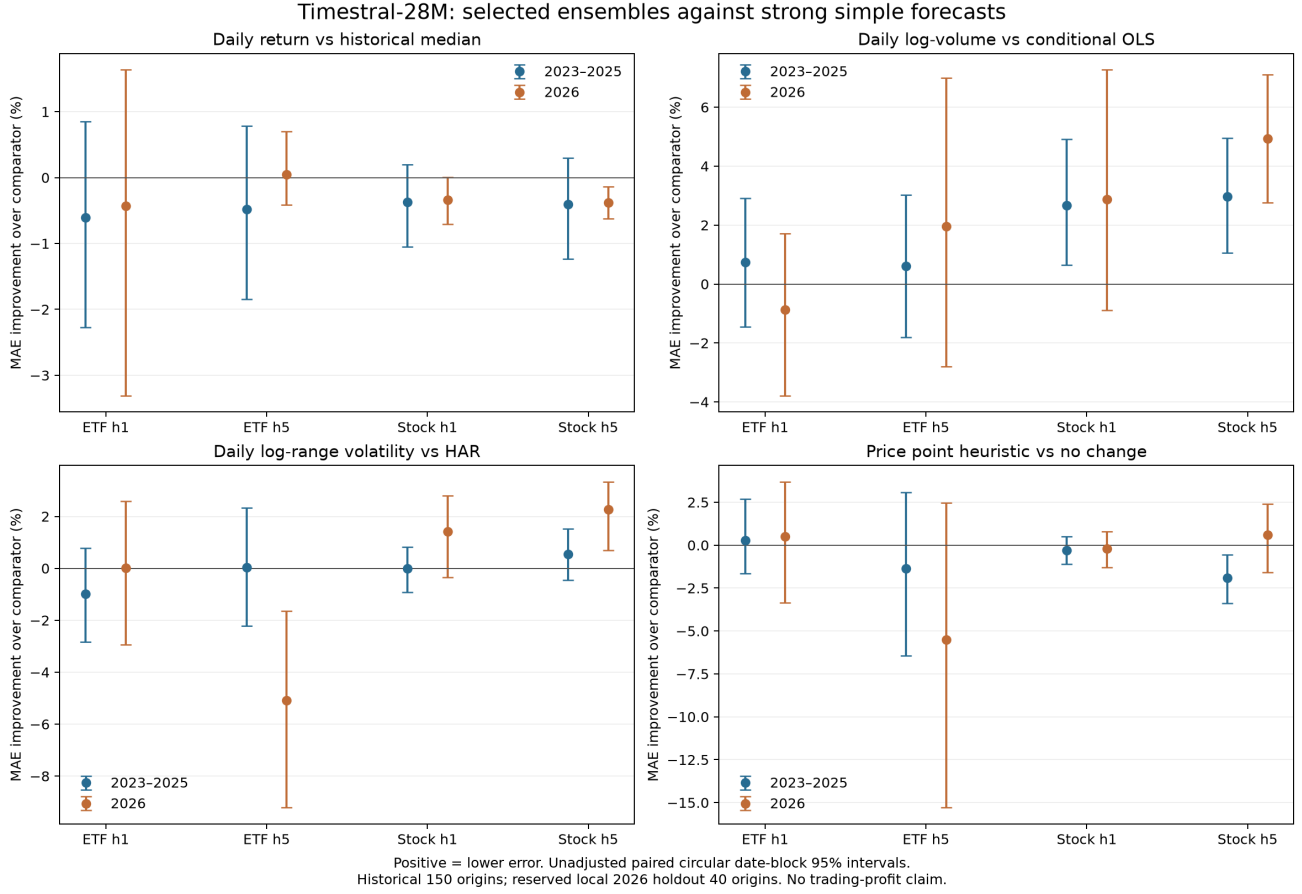


Figure 1: Thirty-two selected-ensemble equity comparisons with descriptive, unadjusted paired 95% date-block intervals. ETF denotes SPY/QQQ/IWM; the separately archived TLT/GLD/USO transfer tests are not shown. Blue: 2023–2025 (150 origins); orange: retrospective local 2026 (forty origins). Offset-five daily returns are not cumulative returns. The price heuristic $\exp(\sum_{j=1}^h \hat{q}_{0.5,j})$, normalized by origin price, is not a cumulative-distribution median. Marginal quantiles are not summed into price bands. Figure metadata preserves source-cell and file hashes.

Causal conditional volume comparator. At origin t , the direct regression for $h \in \{1, 5\}$ fits only fully observed historical pairs inside the 512-session context:

$$v_{u+h} = \beta_0 + \beta_1 v_u + \beta_2 \bar{v}_{u,5} + \beta_3 \bar{v}_{u,22} + \mathbf{1}_{h=1} \beta_4 v_{u+h-5} + \epsilon_{u+h}, \quad u + h \leq t, \quad (4)$$

where $\bar{v}_{u,k} = k^{-1} \sum_{j=0}^{k-1} v_{u-j}$. At $h = 5$ the last value equals v_u and is omitted. Point forecasts add the median fitted residual to the OLS prediction; residual quantiles provide an approximate distribution, without guaranteed conditional calibration.

Availability and research record. Archive: `timestral-market-forecasting`, experiment commit `cf5a68f`. Stocks: AAPL, MSFT, AMZN, GOOGL, CSCO, INTC, JPM, BAC, XOM, CVX, JNJ, PG, KO, MCD. Frozen recipes and model revisions are in `experiment/walkforward/protocol.json` and `experiment/models/manifest.json`. The repository preserves data, weights, dependencies, forecasts and audits, including earlier studies; run `make verify`. The paper’s `evidence.json` records source rows for Table 1 and all twelve direct adaptation comparisons. Redistribution follows original terms; no public deposit is claimed.

Authorship and declarations. P.D.D.: conceptualization and research supervision. AI assistance: implementation, verification, literature search and drafting. Ethics: public market data, no human participants. Funding and competing interests: not specified in the research record.