

FinVerse: Financial Time-Series Benchmark

Toward A More Realistic Evaluation for Financial Time-series Forecasting

Jaehoon Lee, Jun Seo, Seunghan Lee, Tae Yoon Lim, Dongwan Kang, Hwanil Choi, Minjae Kim, Sungdong Yoo, Junhyeok Kang, Sangjun Han, Soonyoung Lee, Wonbin Ahn*

LG AI Research

As time-series foundation models have emerged, the need for benchmarks that can evaluate their forecasting ability in meaningful ways has become increasingly important. Existing time-series forecasting benchmarks provide useful standardized comparisons, but they often evaluate heterogeneous series with uniform error-based metrics. Strong performance under such metrics does not necessarily imply that a model’s forecasts will support the best real-world decisions across domains. For example, in stock forecasting, correctly predicting whether a price will rise or fall can be more directly relevant to realized returns than minimizing point-wise forecast error alone. To this end, we introduce FinVerse, a finance-domain time-series forecasting benchmark that takes a first step toward more realistic evaluation. The released FinVerse data artifact contains 116,897 financial time series with 171.1M observations, of which 60,232 series with 17.4M observations are selected as evaluated targets based on their economic relevance to financial decisions. Unlike generic forecasting benchmarks that primarily emphasize uniform point-forecast or probabilistic accuracy, FinVerse defines 11 metric families comprising 78 evaluation metrics and assigns the most appropriate evaluation metrics to each individual time series based on its underlying economic meaning. Our analysis of 43 public time-series forecasting foundation models shows that strong performance under generic forecasting criteria does not necessarily translate into useful financial forecasts. This finding highlights the need for domain-aware benchmarks that evaluate models under objectives closer to real-world decision making.

Leaderboard: <https://huggingface.co/spaces/LG-AI-Research/FinVerse>

1 Introduction

The recent success of large language models has renewed interest in foundation models across a broad range of domains. This trend has naturally extended to time-series modeling, where a growing body of work aims to build general-purpose models that can forecast, represent, or reason over temporal observations from many sources [2, 6, 10, 19, 20, 24, 27]. As these models become larger and more general, the field increasingly needs evaluation protocols that can distinguish models that are merely accurate under generic statistical criteria from models that are useful for realistic downstream decisions.

Well-defined natural language benchmarks have played this role in the development of language models. [11, 17, 25, 26] In natural language processing, benchmark suites provide shared tasks, standardized data splits, and interpretable metrics that make it possible to compare models and track progress over time. More importantly, strong performance on a well-designed benchmark is expected to correlate with practical capability on real tasks. A benchmark therefore acts not only as a leaderboard, but also as a compact definition of what the community currently considers meaningful progress.

Time-series forecasting does not yet have an equally satisfactory benchmark standard. Existing benchmarks, including widely used suites such as GIFT-Eval [1], have provided valuable common ground for comparing forecasting models. However, they typically evaluate heterogeneous time series with a largely uniform set of error-based forecast accuracy metrics. This design raises a practical question: if a decision maker selects the best-performing model on such a

*Corresponding author.

Table 1: Coverage of the full FinVerse data artifact by scope and category. D, BD, W, M, Q, and A denote daily, business-daily, weekly, monthly, quarterly, and annual frequencies, respectively.

Scope	Category	Series	Points	Frequencies
Inter-country	Foreign Exchange	25	294.4K	BD
	Commodity	59	261.9K	BD/W/M
	Crypto	4,700	5.0M	D
	Global Liquidity	39	4.5K	Q
	Global Risk	6	18.1K	D/BD/M
	Market Index	22	166.4K	BD
Country	Fixed Income	75	380.9K	BD/W/M
	Market Indicators	32	270.8K	D/BD/W/M
	Macro Indicators	199	165.2K	D/BD/W/M/Q/A
Individual	ETF	25,531	40.5M	BD
	Equity	30,745	121.7M	BD
	Fundamentals	55,442	2.2M	Q
	News Sentiment	22	75.0K	D

benchmark, should they expect reasonable performance on the decisions that actually depend on the forecast? We argue that the answer is often no, because the meaning of a good forecast changes across domains and target series. For example, in stock forecasting, correctly anticipating the direction or relative ranking of future returns can be more relevant than minimizing price error. For macroeconomic indicators such as CPI, both the direction of change and the acceleration or deceleration relative to the previous period may matter. A benchmark that applies the same MASE- and CRPS-centered evaluation to all series cannot fully capture these domain-specific requirements.

To address this gap, we propose **TimeVerse** (**Time**-series benchmark **uniVerse**), an initiative toward more realistic evaluation of time-series forecasting. In this paper, as its first domain-specific benchmark, we introduce **FinVerse**, which focuses on financial time-series forecasting.¹ FinVerse is designed around the principle that each time series should be evaluated according to the decision context implied by its economic meaning. Rather than treating financial forecasting as a homogeneous error minimization problem, the benchmark organizes financial time series around three evaluation aspects: point-wise forecast accuracy, cross-sectional ranking quality, and portfolio-level backtesting. Across these three evaluation aspects, we define 11 metric families comprising 78 evaluation metrics and assign the most appropriate metric to each individual time series based on its underlying economic meaning. This design aims to provide a more realistic and practically informative view of model performance in financial forecasting.

FinVerse collects financial time series across a broad hierarchy of scopes and categories. The benchmark covers inter-country series such as foreign exchange rates, commodities, crypto assets, and global market indices; country-level series such as fixed income, macroeconomic indicators, and market indicators; and individual-instrument series such as equities, ETFs, and firm fundamentals. This structure allows the benchmark to evaluate models on diverse financial time series that require different notions of predictive usefulness.

The released FinVerse data artifact contains 116,897 financial time series with 171.1M observations. Among them, we select 60,232 economically meaningful target series with 17.4M observations as the fixed evaluation set, manually choosing series whose semantics make them important for financial forecasting and decision making. From the remaining historical data before the 2020-01-01 evaluation start date, FinVerse constructs a training set with 66,134 time series and 100.3M observations for foundation model pretraining and fine-tuning. Using 2020-01-01 as the evaluation start date, the benchmark evaluates selected financial time series. Table 1 summarizes the coverage of the FinVerse data artifact. Through this benchmark, we observe that models that perform well under generic time-series benchmarks do not necessarily produce useful results for financial forecasting tasks. These findings highlight the need for financial-domain benchmarks that evaluate models under decision-oriented settings rather than relying solely on generic error-based forecasting accuracy.

The main contributions of this report are as follows:

¹We plan to extend this benchmark to additional domains, such as medicine and energy, where realistic evaluation likewise requires domain-specific metrics.

- We identify a key limitation of existing time-series benchmarks: their reliance on uniform error-based metrics can obscure whether a model is useful for realistic domain decisions.
- We introduce FinVerse, a large-scale finance benchmark consisting of 116,897 financial time series (171.1M observations). FinVerse addresses the limitations of existing time-series benchmarks by replacing one-size-fits-all error-based evaluation with decision-oriented metrics tailored to the underlying economic meaning of each financial time series.
- We provide empirical evidence that strong performance on existing generic error-based time-series benchmarks does not necessarily imply reasonable performance on practical financial forecasting objectives.

2 Related Work

2.1 Time-Series Foundation Models

The foundation model paradigm has recently expanded from language and vision to time-series modeling. Traditional forecasting models are often trained for a specific dataset, domain, frequency, or forecasting horizon, which limits their ability to transfer across heterogeneous temporal data. In contrast, recent time-series foundation models aim to pretrain on large and diverse collections of time series and then support zero-shot or few-shot forecasting on unseen datasets. Representative forecasting-oriented models include TimesFM [6], Chronos [2, 3], Moirai [27], Timer [19, 20], and TimePFN [24], which are pretrained on large collections of heterogeneous time series.

The dominant architectural paradigm among recent time-series foundation models is the decoder-only Transformer, which adapts autoregressive sequence modeling from language models to temporal prediction. Given a historical sequence, these models generate future observations in an autoregressive manner, enabling flexible forecasting across diverse domains without task-specific architectural modifications. Representative examples include TimesFM, Chronos, and Timer. Beyond forecasting-oriented models, general-purpose representation models such as MOMENT [10] learn transferable time-series representations that can be adapted to a broad range of downstream tasks, including forecasting, classification, anomaly detection, imputation, and representation learning. Together, these advances have substantially improved the generalization ability of time-series models and accelerated the development of general-purpose time-series foundation models.

2.2 Time-Series Benchmarks

Several benchmark efforts have shaped modern time-series forecasting evaluation. TSLib-style long-sequence forecasting protocols popularized standardized comparisons across canonical datasets [30]. The Monash Time Series Forecasting Archive broadened the evaluation landscape by collecting diverse real-world forecasting datasets under a unified archive and metadata format [9]. More recently, LOTSA, introduced together with Moirai, scaled pretraining and evaluation toward a large collection of heterogeneous time-series datasets, supporting the development of universal forecasting models [27].

GIFT-Eval is a representative and widely used benchmark for general time-series forecasting model evaluation [1]. It provides a standardized evaluation suite over heterogeneous datasets and has become a practical reference point for comparing recent time-series foundation models. However, its general-purpose design necessarily abstracts away many domain-specific objectives. In financial forecasting, the practical value of a forecast often depends on direction, ranking, risk, or portfolio-level outcomes rather than only generic error-based point or probabilistic forecast accuracy. FinVerse is designed to complement such general benchmarks by evaluating models under finance-specific metrics and decision contexts.

2.3 Financial Forecasting

Forecasting is a central task in finance because many downstream decisions depend on expectations about future market behavior, risk, and economic conditions. Recent work has studied financial forecasting from several complementary angles, including spatial-temporal modeling for stock time series, adaptation under distribution shift for stock trend forecasting, and volatility prediction from options-derived signals [23, 28, 29]. These studies aim to improve forecasting performance on financial time series, whose non-stationary dynamics make accurate prediction particularly challenging.

Recent benchmark efforts have also begun to systematize financial time-series forecasting evaluation. FinTSB provides a dedicated benchmark for financial time-series forecasting [13], but its evaluation scope is primarily centered on stock data, and its prediction target focuses on next-day prices, which limits diversity in both data coverage and evaluation

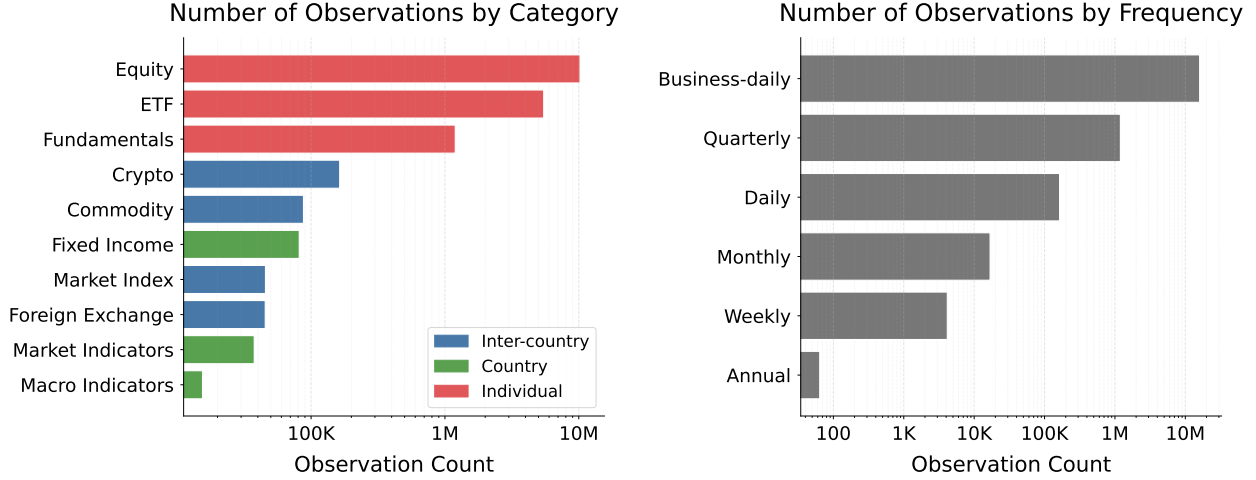


Figure 1: Observation coverage of the selected FinVerse evaluation targets by category and frequency.

periods. In contrast, FinVerse collects financial time series across multiple data categories, including equity prices, exchange rates, commodities, crypto assets, fixed-income indicators, market indicators, macroeconomic indicators, ETFs, and fundamentals. It further supports multiple data frequencies and evaluation horizons, aiming to provide a broader and more practically useful evaluation setting for financial forecasting models.

3 FinVerse

FinVerse is built as a finance-domain benchmark universe rather than as a simple collection of heterogeneous financial time series. The benchmark first organizes diverse financial signals into a hierarchical taxonomy defined by Scope, Category, and Detail Category. For each Detail Category–Frequency combination, FinVerse assigns evaluation metrics that reflect the semantic meaning and economic role of the corresponding time series. The metric assignments for each Detail Category–Frequency combination are summarized in Tables 6–15. Section 3.1 describes how the benchmark is structured under this hierarchy and reports evaluation data coverage statistics. Section 3.2 then explains the domain-aware metrics used to evaluate financial time series, and Section 3.3 describes the evaluation periods and forecast horizons used to construct forecasting tasks.

3.1 Benchmark Universe Construction

FinVerse organizes financial time series along a scope–category–detail category hierarchy. The benchmark contains three scopes. The inter-country scope covers cross-country global series, including Foreign Exchange, Commodity, Crypto, Global Liquidity, Global Risk, and Market Index categories. The country scope covers country-level economic and market conditions, including Fixed Income, Macro Indicators, and Market Indicators. The individual scope covers individual tradable instruments, firm-level targets, and sector-level signals, including ETF, Equity, Fundamentals, and News Sentiment categories. The detailed category-level coverage for each category is reported in Appendix Table 4. In total, the released full data artifact contains 3 scopes, 13 categories, and 101 detail categories with 116,897 financial series and 171.1M observations. The released data include daily, business-daily, weekly, monthly, quarterly, and annual frequencies. Note that the country and individual scopes are centered on the U.S. market, the world’s largest financial market, whereas the inter-country scope includes cross-country and global financial series.

Among all collected financial time series, we manually select 60,232 target series (17.4M observations) to form the fixed evaluation set based on their underlying economic meaning and relevance to financial forecasting and decision-making. The resulting evaluation set spans 3 scopes, 10 evaluated categories, 67 evaluated detail categories, and 75 evaluated detail-category–frequency units. Figure 1 summarizes the observation-level coverage of the selected evaluation targets. The largest portion of the evaluation set consists of business-daily observations from equity (10.2M observations) and ETF (5.5M observations) series, reflecting the large number of individual instruments represented in these categories. Although country- and inter-country-level series, which are primarily available at monthly, quarterly, and annual frequencies, contain fewer observations, they remain essential components of the benchmark because they capture economically important macroeconomic and market conditions.

3.2 Benchmark Evaluation Metrics

FinVerse uses metric families that correspond to three evaluation views: point-wise forecast quality, cross-sectional ranking quality, and realized portfolio performance. For a single series, let y_t denote the realized value at time t , let $\hat{y}_{t+h}$ denote the forecast for horizon h , and let $\mathcal{O}$ be the set of valid forecast origins.

3.2.1 Point-wise Forecasting

Hit Ratio. Hit ratio measures directional correctness. For a comparison window p and forecast horizon h , define the realized and predicted changes as

$$r_o^{(p,h)} = \frac{y_{o+h} - y_{o+h-p}}{y_{o+h-p}} - 1, \quad \hat{r}_o^{(p,h)} = \frac{\hat{y}_{o+h} - \hat{y}_{o+h-p}}{y_{o+h-p}} - 1. \quad (1)$$

The hit ratio is

$$\text{HR}^{(p,h)} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathbb{I} \left[\text{sign}(\hat{r}_o^{(p,h)}) = \text{sign}(r_o^{(p,h)}) \right]. \quad (2)$$

For first- and second-order directional changes, define

$$\Delta r_o^{(p,h)} = r_o^{(p,h)} - r_{o-1}^{(p,h)}, \quad \Delta \hat{r}_o^{(p,h)} = \hat{r}_o^{(p,h)} - r_{o-1}^{(p,h)}, \quad (3)$$

and

$$\Delta^2 r_o^{(p,h)} = r_o^{(p,h)} - 2r_{o-1}^{(p,h)} + r_{o-2}^{(p,h)}, \quad \Delta^2 \hat{r}_o^{(p,h)} = \hat{r}_o^{(p,h)} - 2r_{o-1}^{(p,h)} + r_{o-2}^{(p,h)}. \quad (4)$$

The derivative hit ratios are then

$$\Delta \text{HR}^{(p,h)} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathbb{I} \left[\text{sign}(\hat{r}_o^{(p,h)}) = \text{sign}(r_o^{(p,h)}) \right] \mathbb{I} \left[\text{sign}(\Delta \hat{r}_o^{(p,h)}) = \text{sign}(\Delta r_o^{(p,h)}) \right], \quad (5)$$

$$\Delta^2 \text{HR}^{(p,h)} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathbb{I} \left[\text{sign}(\hat{r}_o^{(p,h)}) = \text{sign}(r_o^{(p,h)}) \right] \mathbb{I} \left[\text{sign}(\Delta \hat{r}_o^{(p,h)}) = \text{sign}(\Delta r_o^{(p,h)}) \right] \mathbb{I} \left[\text{sign}(\Delta^2 \hat{r}_o^{(p,h)}) = \text{sign}(\Delta^2 r_o^{(p,h)}) \right]. \quad (6)$$

The comparison window p specifies the reference interval used to compute the realized and predicted changes (e.g., week-over-week, month-over-month, quarter-over-quarter, half-year-over-half-year, or year-over-year), whereas the forecast horizon h specifies how far into the future the prediction is evaluated. For example, $p = 1y$ computes the year-over-year change relative to the observation one year earlier, while $h = 1y$ evaluates the prediction one year ahead. Different financial series use different (p, h) combinations according to their sampling frequency and economic interpretation. The complete (p, h) assignments for each financial series and evaluation metric are provided in Tables 6–15.

HR is motivated by the fact that many financial decisions depend more on predicting the direction of movement than on reconstructing the exact future value. For example, in asset-price forecasting, correctly predicting whether the price will rise or fall after a given horizon is often more important than minimizing the level error. The proposed ΔHR and $\Delta^2 \text{HR}$ extend this directional criterion to first- and second-order changes, respectively. These variants are particularly useful for macroeconomic indicators such as CPI, where long-term trends make it important to predict not only whether the year-over-year value will increase or decrease, but also whether the rate of change is accelerating or decelerating relative to the previous period.

Mean Absolute Scaled Error. MASE is included because scaled error metrics have been widely adopted in time-series forecasting benchmarks and provide a scale-independent measure that is comparable across heterogeneous series. Unlike the original MASE, which scales forecast errors using the in-sample error of a seasonal or naive forecast, FinVerse adapts the scaling baseline to better reflect financial forecasting objectives, where the magnitude of future values is often as important as their direction. While the proposed metrics share the same intuition as the original MASE—measuring forecast error relative to a simple baseline predictor—the baseline is defined directly for the prediction target under evaluation. This design is motivated by financial time series (e.g., price), which often exhibit weak or unstable seasonality compared with many benchmark datasets, making persistence-based baselines more appropriate than the seasonal-naive scaling used in the original MASE.

Specifically, three variants are considered. MASE_{rel} measures forecast error on relative changes against a no-change relative baseline. Since $r_o^{(p,h)}$ denotes the relative change from y_{o+h-p} to y_{o+h} , this baseline corresponds to predicting zero relative change (i.e., a gross return of one). FinVerse computes

$$\text{MASE}_{rel}^{(p,h)} = \frac{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |\hat{r}_o^{(p,h)} - r_o^{(p,h)}|}{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |r_o^{(p,h)}|}. \quad (7)$$

MASE_{abs} evaluates raw-value forecasts against a no-change raw-value baseline defined over the same comparison window. In other words, the baseline predicts the future value by repeating the observation at y_{o+h-p} :

$$\text{MASE}_{abs}^{(p,h)} = \frac{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |\hat{y}_{o+h} - y_{o+h}|}{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |y_{o+h} - y_{o+h-p}|}. \quad (8)$$

Finally, MASE_{ori} evaluates raw-value forecasts against the conventional origin-value persistence baseline, where the most recently observed value is repeated for all future horizons. Since the raw-value metric always uses $p = h$, the baseline reduces to persistence from the forecast origin:

$$\text{MASE}_{ori} = \frac{\frac{1}{|\mathcal{O}|H} \sum_{o \in \mathcal{O}} \sum_{h=1}^H |\hat{y}_{o+h} - y_{o+h}|}{\frac{1}{|\mathcal{O}|H} \sum_{o \in \mathcal{O}} \sum_{h=1}^H |y_o - y_{o+h}|}. \quad (9)$$

We include MASE-based metrics because, although directional correctness is often the primary objective in financial forecasting, predictions with unrealistic magnitudes are of limited practical value. For example, in foreign exchange forecasting, where price movements are typically small, a model that consistently predicts excessively large changes may achieve reasonable directional accuracy while remaining unsuitable for real-world use. Accordingly, point-wise evaluation should consider both directional behavior and magnitude accuracy. These three MASE variants capture complementary aspects of financial forecasting. MASE_{rel} evaluates the accuracy of predicting relative changes (e.g., returns), whereas MASE_{abs} measures the accuracy of predicting the raw values themselves. Furthermore, MASE_{rel} and MASE_{abs} are evaluated only at financially meaningful forecasting horizons (e.g., one week, one month, one quarter, half a year, and one year ahead), while MASE_{ori} measures forecasting error over the entire prediction horizon.

3.2.2 Cross-sectional Ranking

Information Coefficient. The information coefficient evaluates cross-sectional ranking quality. For a set of assets $\mathcal{S}_o$ available at origin o , FinVerse computes a Spearman rank correlation between predicted and realized future changes:

$$\text{IC}_o^{(p,h)} = \rho_{\text{Spearman}} \left(\left\{ \hat{r}_{i,o}^{(p,h)} \right\}_{i \in \mathcal{S}_o}, \left\{ r_{i,o}^{(p,h)} \right\}_{i \in \mathcal{S}_o} \right). \quad (10)$$

IC is selected because portfolio and allocation decisions often depend on relative ordering across assets rather than independent point accuracy for each series. This is used for groups where relative ranking is meaningful, such as currencies, commodities, crypto assets, global market indices, ETFs, equities, selected fixed-income groups, and selected fundamentals.

3.2.3 Real-world Portfolio Backtesting

Portfolio evaluation converts forecasts into an equal-weight long-only strategy. At each rebalance origin, the benchmark ranks assets by the horizon- h predicted signal and selects the top 10% fraction $\mathcal{P}_o^{(h)} \subset \mathcal{S}_o$. For a holding horizon h , the selected set $\mathcal{P}_o^{(h)}$ is reconstructed at rebalance origins spaced h frequency steps apart and held for the next h steps. To avoid depending on a single calendar start date, FinVerse runs this backtest for every possible offset within the h -step window and averages the resulting portfolio statistics. Thus, for a one-month horizon, starts such as January 1 and January 2 define separate monthly-rebalanced paths rather than discarding all but one start date. The realized portfolio return for a rebalance origin is

$$R_o^{(h)} = \frac{1}{|\mathcal{P}_o^{(h)}|} \sum_{i \in \mathcal{P}_o^{(h)}} r_{i,o}^{(h,h)}. \quad (11)$$

Let $\mathcal{A}_h = \{0, \dots, h-1\}$ denote the offset set, and let $o_{s,1}, \dots, o_{s,K_s}$ denote the rebalance origins in chronological order for offset s , where K_s is the number of valid rebalancing periods in that offset path. The cumulative net asset value for offset s is

$$\text{NAV}_{s,k}^{(h)} = \prod_{j=1}^k \left(1 + R_{o_{s,j}}^{(h)} \right). \quad (12)$$

The resulting return sequence is summarized with annualized return, annualized volatility, Sharpe ratio, and maximum drawdown for each offset, and the reported metric is the average across offsets:

$$\text{Metric}^{(h)} = \frac{1}{|\mathcal{A}_h|} \sum_{s \in \mathcal{A}_h} \text{Metric}_s^{(h)}. \quad (13)$$

Table 2: Frequency-specific prediction lengths and minimum lookback requirements.

Frequency	Prediction Length	Minimum Lookback
Daily	365	365
Business-daily	260	260
Weekly	52	52
Monthly	12	12
Quarterly	4	4
Annual	1	4

Here, h denotes the holding horizon. The annualization factor is selected directly from the portfolio holding horizon:

$$A_h \in \{52, 12, 4, 2, 1\} \quad (14)$$

for one-week, one-month, one-quarter, half-year, and one-year horizons, respectively. For each offset s , the annualized return, annualized volatility, Sharpe ratio, and maximum drawdown are computed as

$$\text{AnnualReturn}_s^{(h)} = \left(\prod_{j=1}^{K_s} (1 + R_{o_{s,j}}^{(h)}) \right)^{\frac{A_h}{K_s}} - 1, \quad \text{AnnualVolatility}_s^{(h)} = \text{Std}(R_{o_{s,j}}^{(h)}) \sqrt{A_h}, \quad (15)$$

$$\text{Sharpe}_s^{(h)} = \frac{\mathbb{E}[R_{o_{s,j}}^{(h)}]}{\text{Std}(R_{o_{s,j}}^{(h)})} \sqrt{A_h}, \quad \text{MDD}_s^{(h)} = \min_k \left(\frac{\text{NAV}_{s,k}^{(h)}}{\max_{\ell \leq k} \text{NAV}_{s,\ell}^{(h)}} - 1 \right). \quad (16)$$

These metrics are included because financial forecasts are ultimately useful only when they can support stable downstream decisions under realistic return and risk tradeoffs. We acknowledge that many investment strategies are possible and that benchmark outcomes can vary with the chosen strategy; FinVerse therefore starts with a simple portfolio rule and leaves richer strategy families for future extensions.

Overall, the benchmark defines six metric families for point-wise forecasting, one metric family for cross-sectional ranking, and four metric families for realized portfolio evaluation. Combined with the predefined (p, h) or (h) combinations, these yield a total of 78 evaluation metrics.

3.3 Evaluation Periods and Forecast Horizons

The default training artifact contains observations before the 2020-01-01 evaluation start date. As summarized in Table 2, we define the prediction length and the minimum lookback window size for each sampling frequency. The prediction lengths are chosen to correspond to financially meaningful forecasting horizons (i.e., one week, one month, one quarter, half a year, and one year ahead), while the minimum lookback window sizes are selected to provide sufficient historical context for making meaningful forecasts at the corresponding horizons.

4 Evaluation with FinVerse

4.1 Foundation Models

We evaluate FinVerse with a broad set of recent time-series foundation models and forecasting checkpoints. The evaluated models include TimesFM 1.0 200M, TimesFM 2.0 500M, and TimesFM 2.5 200M [6]; Chronos-2, Chronos-2 Synth, Chronos-Bolt Small, and Chronos-Bolt Base [2, 3]; Moirai 1.1-R Small, Moirai 1.1-R Base, Moirai 1.1-R Large, and Moirai 2.0-R Small [27]; Timer-S1 [19, 20]; TempoPFN [12]; Toto Open Base 1.0, Toto 2.0 4M, Toto 2.0 22M, Toto 2.0 313M, Toto 2.0 1B, and Toto 2.0 2.5B [5]; TiRex 1.1, TiRex 2 Pretrain, and TiRex 2 ZS [7]; VisionTS [4]; Sundial Base 128M [18]; YingLong 6M, YingLong 50M, YingLong 110M, and YingLong 300M [16]; Kairos-10M, Kairos-23M, and Kairos-50M [15]; Reverso Small [22]; TTM R1, TTM R2, and TTM R3 [8]; PatchTST-FM R1 and Granite PatchTST-FM R1 [21]; FlowState and Granite FlowState R1 [14]; PatchFM; T0 Alpha; Super Linear; and CleanTS-65M.

4.2 Overall Rank Aggregation

FinVerse contains heterogeneous evaluation metrics with different numerical characteristics, and the numbers of time series vary substantially across detail category–frequency combinations. Directly averaging raw metric values would

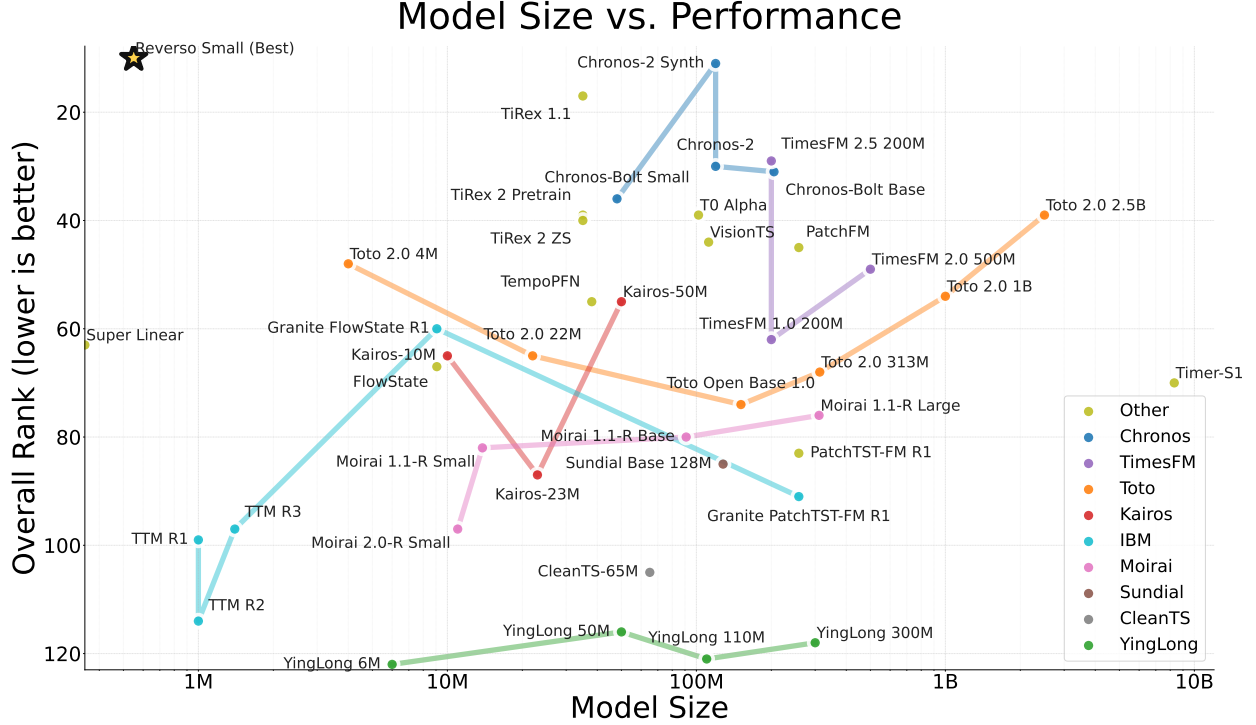


Figure 2: Relationship between model size and overall FinVerse rank. Lower rank indicates better performance.

therefore overemphasize certain metrics and categories. To address this issue, FinVerse adopts a rank-based aggregation strategy rather than directly averaging raw metric values.

Specifically, we first compute the average score of each evaluation metric within every detail category–frequency combination. Models are then ranked separately for each metric within each detail category–frequency combination. Next, for each of the three evaluation groups—Point-wise Forecasting, Cross-sectional Ranking, and Realized Portfolio Backtesting—we compute the geometric mean of the corresponding metric ranks, denoted by K . The models are then ranked according to the K value of each evaluation group. Finally, the overall rank is obtained by summing the three group-level ranks.

This aggregation strategy enables heterogeneous metrics with different scales and properties to contribute more evenly to the final evaluation while preventing categories containing many time series from dominating the benchmark. As a result, smaller but economically important categories are appropriately reflected in the overall ranking.

4.3 Performance in FinVerse

4.3.1 Overall Benchmark Performance

Figure 2 compares model size with the overall FinVerse rank computed from the three aspect placement ranks. Reverso Small achieves the best overall rank, followed by Chronos-2 Synth, TiRex 1.1, TimesFM 2.5 200M, and Chronos-2. In contrast, larger checkpoints do not consistently dominate: several Toto, Timer, and YingLong variants remain behind smaller or medium-sized models in the overall ranking. These results suggest that financial decision-oriented evaluation cannot be explained by parameter count alone. Instead, model pretraining data, inductive biases, horizon handling, and robustness to financial distribution shifts appear to contribute at least as much as model scale.

Figure 3 compares the overall FinVerse rank with the corresponding GIFT-Eval ranking for models evaluated on both benchmarks. Across the 43 matched models, the two rankings have only a moderate correlation (Pearson $r = 0.40$), indicating that they are related but far from interchangeable. For example, Chronos-2, TiRex 1.1, and TimesFM 2.5 200M rank strongly on both benchmarks, while Reverso Small is the strongest FinVerse model despite a weaker GIFT-Eval rank. These differences indicate that strong performance on a generic time-series benchmark does not necessarily translate into superior performance under FinVerse’s decision-oriented financial evaluation.

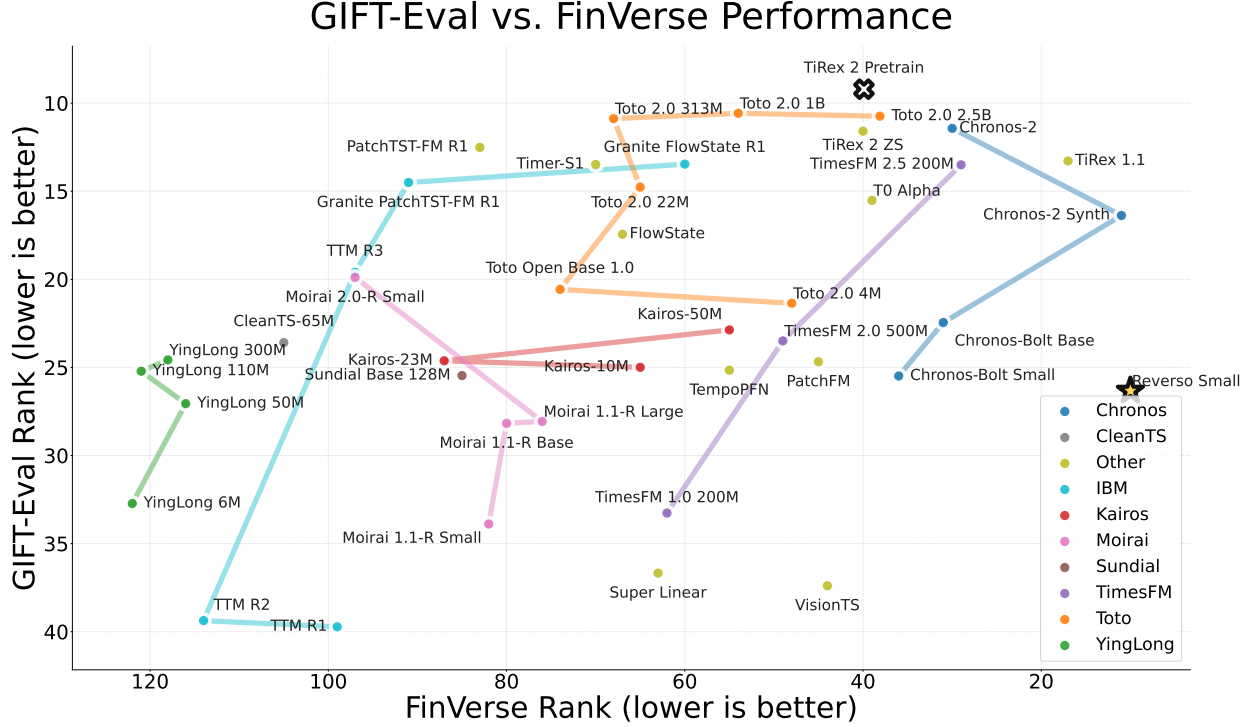


Figure 3: Comparison between FinVerse and GIFT-Eval rankings for matched models. The GIFT-Eval rank is the average of the public GIFT-Eval MASE rank and probabilistic forecast rank. The star marks the best-ranked model in FinVerse, and the X marker marks the best-ranked model in GIFT-Eval.

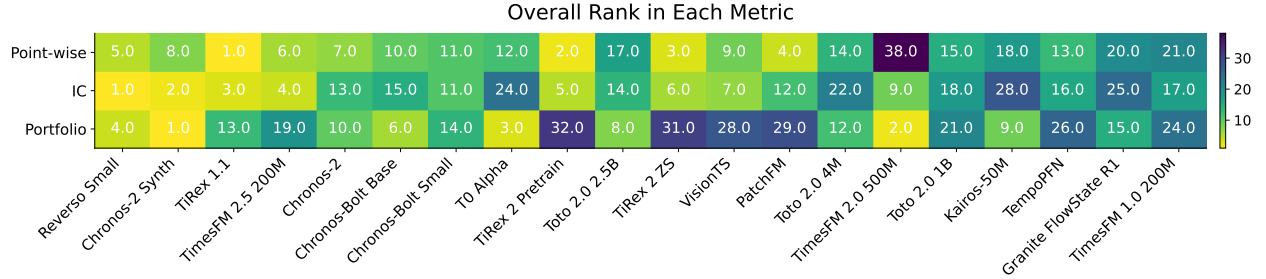


Figure 4: Top-20 models ranked separately under point-wise, cross-sectional, and portfolio-oriented evaluation aspects.

4.3.2 Performance Across Evaluation Aspects

Figure 4 reports model rankings separately for the three evaluation aspects: point-wise forecasting, cross-sectional ranking, and portfolio backtesting. TiRex 1.1 performs best in the point-wise aspect, followed by TiRex 2 Pretrain, TiRex 2 ZS, PatchFM, and Reverso Small. Cross-sectional ranking favors Reverso Small, Chronos-2 Synth, TiRex 1.1, TimesFM 2.5 200M, and TiRex 2 Pretrain. Portfolio evaluation again favors a different set of models: Chronos-2 Synth ranks first, followed by TimesFM 2.0 500M, T0 Alpha, Reverso Small, and Moirai 1.1-R Large. These results demonstrate that strong point-wise forecasting performance does not necessarily imply superior cross-sectional ranking or portfolio performance.

Figure 5 further examines the relationships among the three evaluation aspects at the model-detail-category-frequency level. Point-wise and cross-sectional rankings are strongly related in this view ($r = 0.72$). The point-wise-portfolio relationship is weaker ($r = 0.37$), while the cross-sectional-portfolio relationship is moderately aligned ($r = 0.60$). Portfolio-level outcomes therefore capture information beyond conventional per-series forecasting accuracy and cross-sectional rank correlation. Overall, the three evaluation aspects are complementary rather than redundant, supporting the need for a benchmark that evaluates forecasting models from multiple financial perspectives.

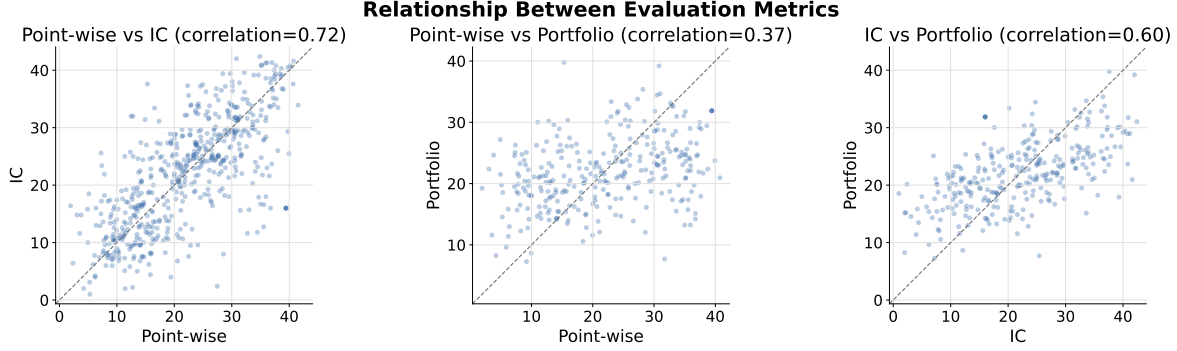


Figure 5: Pairwise relationship between point-wise, cross-sectional, and portfolio ranks at the model–detail-category–frequency level.

4.3.3 Metric-wise Analysis

Table 3 further decomposes model performance by metric family using averaged raw metric values. Even within the point-wise evaluation aspect, the best-performing model differs across metrics. Reverso Small achieves the strongest HR value, TiRex 1.1 leads ΔHR and $\Delta^2\text{HR}$, PatchFM leads MASE_{rel} , Toto 2.0 2.5B leads MASE_{abs} , and Chronos-2 leads MASE_{ori} . For cross-sectional and portfolio-oriented metrics, Chronos-2 Synth achieves the strongest IC, Return, and MDD values, Moirai 1.1-R Small leads Sharpe, and TimesFM 2.0 500M achieves the lowest Volatility value. These results further demonstrate that models preferred by point-wise metrics are not necessarily those preferred by ranking-based or portfolio-oriented evaluation. Overall, the metric-wise analysis reinforces the central motivation of FinVerse. Point-wise forecasting accuracy, directional correctness, cross-sectional ranking quality, and portfolio performance each capture distinct aspects of financial forecasting utility, and no single model consistently dominates across all evaluation criteria.

5 Known Limitations and Future Work

FinVerse is designed as an extensible benchmark rather than a closed final artifact. This release prioritizes a broad, reproducible evaluation interface, but several limitations remain in temporal alignment and domain coverage.

5.1 Calendar and Release-Date Alignment

FinVerse currently normalizes weekly, monthly, quarterly, and annual observations to standard calendar period ends. This convention provides a consistent origin grid across heterogeneous series, but it does not fully preserve series-specific reporting calendars. For example, U.S. fiscal-year series may end in September rather than December, USDA crop-year series may follow agricultural calendars ending in June, and weekly series may be released on weekdays other than the normalized weekly anchor.

A related limitation is point-in-time availability. Some economic indicators are assigned to the period they measure, even though the value is released only later. For example, a January CPI observation is economically associated with January, but it is typically released in early to mid-February. If the observation is timestamped at January 31 without an explicit release-date field, the benchmark may treat information as available earlier than it would have been in a real forecasting setting.

5.2 Future Extensions

Several extensions are planned. First, FinVerse should improve point-in-time temporal alignment by adding calendar-aware preprocessing and release-date metadata for fiscal, crop-year, macroeconomic, and filing-date-sensitive series. Second, FinVerse should broaden coverage for index constituents, sector indices, survey-based indicators, fixed-income instruments, and country-specific datasets beyond the United States. Third, future releases should expand the set of decision-oriented evaluation tasks and portfolio construction protocols so that the benchmark can cover a wider range of practical financial use cases.

Table 3: Metric-wise model performance across point-wise, cross-sectional, and portfolio-oriented metrics using averaged raw metric values. Best performance in each metric column is bolded.

Model	HR	Δ HR	Δ^2 HR	MASE _{rel}	MASE _{abs}	MASE _{ori}	IC	Return	Sharpe	MDD	Vol.
Reverso Small	0.673	0.542	0.481	11.548	11.652	1.573	0.146	0.352	0.638	-0.282	1.173
Chronos-2 Synth	0.646	0.550	0.495	17.969	7.936	1.612	0.153	0.946	0.632	-0.227	2.786
TiRex 1.1	0.666	0.577	0.518	13.713	7.162	1.548	0.152	0.390	0.603	-0.297	1.845
TimesFM 2.5 200M	0.653	0.568	0.509	12.242	10.373	1.596	0.147	0.317	0.605	-0.298	1.222
Chronos-2	0.659	0.560	0.495	21.008	7.474	1.505	0.072	0.657	0.586	-0.286	2.622
Chronos-Bolt Base	0.645	0.559	0.493	19.900	6.783	1.576	0.100	0.353	0.637	-0.303	1.172
Chronos-Bolt Small	0.644	0.551	0.491	19.643	6.898	1.561	0.099	0.592	0.970	-0.307	2.415
T0 Alpha	0.621	0.566	0.507	15.088	10.151	1.838	0.129	0.800	0.588	-0.285	2.976
TiRex 2 Pretrain	0.652	0.574	0.515	16.253	6.454	1.547	0.151	0.488	0.553	-0.300	2.346
Toto 2.0 2.5B	0.661	0.547	0.483	15.239	1.524	1.809	0.106	0.539	0.638	-0.296	2.166
TiRex 2 ZS	0.650	0.572	0.513	17.120	6.481	1.533	0.150	0.441	0.567	-0.306	2.079
VisionTS	0.639	0.568	0.506	19.033	6.459	1.562	0.108	0.619	0.646	-0.321	2.543
PatchFM	0.672	0.565	0.500	4.105	11.809	1.653	0.127	0.423	0.606	-0.303	1.837
Toto 2.0 4M	0.662	0.556	0.493	14.976	4.521	1.661	0.100	0.602	0.619	-0.285	2.465
TimesFM 2.0 500M	0.534	0.417	0.365	211.602	163.602	205.471	0.131	0.297	0.698	-0.266	1.037
Toto 2.0 1B	0.661	0.553	0.489	12.339	3.117	1.819	0.104	0.471	0.609	-0.298	2.167
Kairos-50M	0.659	0.548	0.481	28.325	20.003	1.730	0.104	0.560	0.605	-0.289	2.309
TempoPFN	0.668	0.514	0.446	35.542	15.494	1.648	0.080	0.573	0.575	-0.302	2.757
Granite FlowState R1	0.642	0.563	0.486	55.796	4.328	2.218	0.048	0.611	0.645	-0.305	2.686
TimesFM 1.0 200M	0.622	0.556	0.499	19.547	10.575	7.237	0.134	0.634	0.917	-0.299	2.122
Super Linear	0.592	0.534	0.474	15.109	6.970	1.674	0.119	0.931	0.585	-0.318	3.042
Kairos-10M	0.655	0.506	0.441	61.964	23.403	2.114	0.077	0.840	0.624	-0.295	2.695
Toto 2.0 22M	0.658	0.554	0.488	11.000	3.499	1.676	0.104	0.553	0.594	-0.288	2.340
FlowState	0.641	0.550	0.476	47.539	5.101	2.080	0.025	0.748	0.619	-0.302	2.864
Toto 2.0 313M	0.660	0.557	0.490	15.043	5.001	1.856	0.096	0.490	0.597	-0.297	2.066
Timer-S1	0.606	0.533	0.477	16.586	10.489	1.785	0.132	0.513	2.633	-0.314	2.021
Toto Open Base 1.0	0.596	0.495	0.434	15.642	5.919	1.845	0.091	0.728	0.570	-0.293	2.909
Moirai 1.1-R Large	0.532	0.438	0.388	48.699	10.707	2.238	0.108	0.355	0.597	-0.278	1.127
Moirai 1.1-R Base	0.538	0.440	0.385	24.502	6.860	1.957	0.101	0.366	2.667	-0.288	1.255
Moirai 1.1-R Small	0.548	0.445	0.393	34.274	11.824	1.840	0.117	0.357	2.681	-0.300	1.208
PatchTST-FM R1	0.565	0.440	0.383	20.755	12.921	2.630	0.107	0.900	2.637	-0.304	2.997
Sundial Base 128M	0.603	0.540	0.483	16.902	9.571	1.832	0.123	0.734	0.561	-0.318	2.507
Kairos-23M	0.642	0.522	0.457	39.604	8.560	1.857	0.079	0.810	0.570	-0.305	2.878
Granite PatchTST-FM R1	0.539	0.433	0.377	19.615	12.962	2.463	0.109	0.756	0.537	-0.305	2.987
Moirai 2.0-R Small	0.569	0.494	0.442	9.469	6.866	2.043	0.093	0.568	0.544	-0.294	2.583
TTM R3	0.531	0.427	0.380	5.792	7.628	2.048	0.121	0.539	0.579	-0.322	2.277
TTM R1	0.586	0.450	0.384	31.108	8.051	7.918	0.085	0.790	0.582	-0.313	2.749
CleanTS-65M	0.567	0.477	0.413	19.269	11.818	2.159	0.098	0.763	0.543	-0.313	2.944
TTM R2	0.611	0.459	0.393	30.052	7.214	2.054	0.103	0.583	2.662	-0.329	2.069
YingLong 50M	0.406	0.276	0.240	20.777	12.918	2.627	0.054	0.691	0.560	-0.313	2.589
YingLong 300M	0.404	0.273	0.237	20.621	12.927	2.624	0.052	0.786	0.549	-0.310	2.860
YingLong 110M	0.405	0.274	0.239	21.437	12.920	2.618	0.047	0.945	0.546	-0.313	2.874
YingLong 6M	0.406	0.275	0.239	20.852	12.897	2.617	0.051	0.670	0.550	-0.315	2.643

Beyond finance, FinVerse is intended to be the first domain-specific member of a broader time-series benchmark universe. The same principle can be applied to domains such as medicine and energy, where realistic evaluation also requires metrics aligned with domain-specific decisions rather than a single generic point-forecast score.

6 Conclusion

This report introduced FinVerse, a finance-domain benchmark in the broader Time-Series Benchmark Universe. The benchmark is motivated by a gap between generic error-based time-series forecasting accuracy and the decision-oriented objectives that often determine usefulness in financial settings. Instead of evaluating all series through a single point-forecast metric family, FinVerse organizes financial forecasting around three complementary aspects: per-series accuracy, cross-sectional ranking, and portfolio backtesting.

FinVerse is built from released data artifacts that share a common atomic time-series schema. The full data artifact contains 116,897 financial time series with 171.1M observations, including 60,232 evaluated target series with

17.4M observations that are selected according to their economic meaning and relevance to financial decisions. Using 2020-01-01 as the evaluation start date, the benchmark evaluates 75 detail-category–frequency units using 11 metric families expanded into 78 metric columns. For model development, the pre-evaluation portion before 2020-01-01 provides 66,134 training time series with 100.3M observations. This structure allows heterogeneous financial data to be represented through a common atomic universe while assigning metrics according to the economic meaning of each series.

Our empirical results over 43 public forecasting models show that model quality changes substantially depending on the financial objective being measured. Reverso Small achieves the best overall FinVerse rank, Chronos-2 Synth leads the portfolio aspect, and TiRex 1.1 leads the point-wise aspect. Point-wise accuracy, cross-sectional ranking quality, and portfolio outcomes are correlated but not fully aligned, and no single model dominates all metric families. These findings support the central premise of FinVerse: strong error-based forecasting performance alone does not necessarily imply usefulness for financial decisions involving direction, relative ranking, risk, or realized portfolio returns.

References

- [1] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation, 2024.
- [2] Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025.
- [3] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series, 2024.
- [4] Xinyu Chen, Jialin Li, Yang Liu, Yong Liu, Zhen Wang, and Mingsheng Long. VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters, 2024.
- [5] Luke Nicholas Darlow, Samuel Cohen, Sjoerd van Steenkiste, Artem Zhohus, Soumyasundar Pal Sharma, Tim Beckers, Rahul Ramesh, Qingsong Wen, Artur Dubrawski, and Anoop Nair. Toto: Time series optimized transformer for observability, 2025.
- [6] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024.
- [7] Samuel Dooley, Gurnoor Singh Khurana, Siddhartha Mohapatra, Kiarash Shaloudegi Esfahani, Viraj Marathe, Chao Zhang, Chih-Chun Hsu, Julian Hunt, Adina Crainiceanu, Dan Suci, Gregory Wornell, Jae-Gil Lee, Greg Ver Steeg, Rahul Puri, Artur Dubrawski, and Willie Neiswanger. Tirez: Zero-shot forecasting across long and short horizons with enhanced in-context learning, 2025.
- [8] Vijay Ekambaram, Arindam Jati, Nam H. Nguyen, Pranay Dayama, Chandra Reddy, Jayant Kalagnanam, Zhen Liu, Yan He, and Sinno Jialin Pan. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series, 2024.
- [9] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive, 2021.
- [10] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models, 2024.
- [11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.
- [12] Noah Hollmann, Samuel Müller, Katharina Eggenberger, and Frank Hutter. TempoPFN: Synthetic pre-training of linear rnns for zero-shot time series forecasting, 2025.

- [13] Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu-tao Xia, Dawei Cheng, and Changjun Jiang. FinTSB: A comprehensive and practical benchmark for financial time series forecasting, 2025.
- [14] IBM Research. Flowstate: A small state space model for time series forecasting, 2025.
- [15] Kairos Contributors. Kairos: A time series foundation model, 2025.
- [16] Yuxuan Li, Haoran Zhang, Jianmin Wang, and Mingsheng Long. Yinglong: A unified foundation model for time series forecasting, 2025.
- [17] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekogul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
- [18] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models, 2025.
- [19] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting, 2025.
- [20] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models, 2024.
- [21] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023.
- [22] Reverso Contributors. Reverso: Time series forecasting checkpoints, 2025.
- [23] Emi Soroka and Artem Arzyn. Data-efficient realized volatility forecasting with vision transformers, 2025.
- [24] Ege Onur Taga, M. Emrullah Ildiz, and Samet Oymak. TimePFN: Effective multivariate time series forecasting with synthetic data, 2025.
- [25] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 2019.
- [26] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pages 353–355, 2018.
- [27] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024.
- [28] Wenbo Yan and Ying Tan. Double-path adaptive-correlation spatial-temporal inverted transformer for stock time series forecasting, 2024.
- [29] Lifan Zhao, Shuming Kong, and Yanyan Shen. DoubleAdapt: A meta-learning approach to incremental learning for stock trend forecasting, 2023.
- [30] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting, 2021.

A Detailed Data Coverage

This appendix provides the detailed coverage table for the released FinVerse data. Table 4 reports, for each scope, category, detail category, and frequency, whether the group contains evaluated target series, together with the number of series, observed data points in the full data artifact, and valid evaluation points after forecast-origin filtering.

Table 4: Detailed full-data coverage by scope, category, detail category, frequency, evaluation-target status, observed points, and valid evaluation points.

Scope	Category	Detail Category	Frequency	Eval.	Series	Points	Eval. Points
Inter-country	Foreign Exchange	FX Index	Business-daily	×	4	30,316	0
Inter-country	Foreign Exchange	FX Pair	Business-daily	✓	21	264,091	34,503
Inter-country	Commodity	Commodity ETF Broad	Business-daily	✓	2	10,433	3,298
Inter-country	Commodity	Commodity ETF Single	Business-daily	✓	5	25,154	8,245
Inter-country	Commodity	Commodity ETF Thematic	Business-daily	✓	2	8,149	3,298
Inter-country	Commodity	Future Front	Business-daily	✓	27	173,716	43,601
Inter-country	Commodity	Inventory Stock	Weekly	✓	7	10,985	2,303
Inter-country	Commodity	Spot	Business-daily	✓	3	28,307	4,932
Inter-country	Commodity	Spot	Monthly	✓	13	5,123	975
Inter-country	Crypto	Crypto Pair Close (Other)	Daily	×	831	794,412	0
Inter-country	Crypto	Crypto Pair Close (Top 100)	Daily	✓	100	179,443	120,963
Inter-country	Crypto	Crypto Pair High	Daily	×	932	959,788	0
Inter-country	Crypto	Crypto Pair Low	Daily	×	931	959,424	0
Inter-country	Crypto	Crypto Pair Open	Daily	×	931	968,233	0
Inter-country	Crypto	Crypto Pair TVL	Daily	×	24	33,254	0
Inter-country	Crypto	Crypto Pair Volume	Daily	×	951	1,082,920	0
Inter-country	Global Liquidity	FX Year Over Year Change	Quarterly	×	6	801	0
Inter-country	Global Liquidity	Stock	Quarterly	×	15	1,692	0
Inter-country	Global Liquidity	Year Over Year Change	Quarterly	×	18	1,965	0
Inter-country	Global Risk	Credit Spread	Business-daily	×	1	781	0
Inter-country	Global Risk	Leading Indicator	Monthly	×	2	1,489	0
Inter-country	Global Risk	Uncertainty	Daily	×	1	15,092	0
Inter-country	Global Risk	Uncertainty	Monthly	×	2	694	0
Inter-country	Market Index	Country Index	Business-daily	✓	20	155,667	31,391
Inter-country	Market Index	Regional Index	Business-daily	✓	2	10,724	3,304
Country	Fixed Income	Bond Total Return Index	Business-daily	✓	10	7,830	5,230
Country	Fixed Income	Breakeven	Business-daily	✓	2	12,172	3,304
Country	Fixed Income	Breakeven	Monthly	✓	2	455	150
Country	Fixed Income	Corporate Yield	Business-daily	✓	2	21,823	3,302
Country	Fixed Income	Corporate Yield	Monthly	✓	2	2,574	150
Country	Fixed Income	Credit Spread	Business-daily	✓	10	7,830	5,230
Country	Fixed Income	Forward Inflation	Business-daily	✓	1	6,086	1,652
Country	Fixed Income	Inflation Expectation	Monthly	✓	7	3,949	530
Country	Fixed Income	Money Market	Business-daily	✓	8	20,958	12,543
Country	Fixed Income	Mortgage Rate	Weekly	✓	3	5,617	812
Country	Fixed Income	Swap Rate	Business-daily	×	7	29,820	0
Country	Fixed Income	TIPS Yield	Business-daily	✓	4	22,478	6,604
Country	Fixed Income	Term Spread	Business-daily	✓	2	24,587	3,304
Country	Fixed Income	Treasury Bill	Business-daily	×	4	60,333	0
Country	Fixed Income	Treasury Yield	Business-daily	✓	11	154,357	18,161
Country	Market Indicators	Fin Conditions	Weekly	×	6	17,316	0
Country	Market Indicators	Fin Stress	Weekly	×	1	1,687	0
Country	Market Indicators	Fin Stress	Monthly	×	1	434	0
Country	Market Indicators	Implied Volatility	Business-daily	✓	9	55,495	13,361
Country	Market Indicators	Leading Index	Monthly	×	1	794	0
Country	Market Indicators	Money Spread	Business-daily	×	1	9,407	0
Country	Market Indicators	Recession Flag	Daily	×	1	62,606	0
Country	Market Indicators	Recession Flag	Monthly	×	1	2,056	0
Country	Market Indicators	Recession Prob	Monthly	×	1	705	0
Country	Market Indicators	Sentiment	Business-daily	×	1	9,478	0
Country	Market Indicators	US Equity Index	Business-daily	✓	9	110,839	14,868
Country	Macro Indicators	Activity	Monthly	✓	10	5,621	744
Country	Macro Indicators	Banking	Weekly	×	2	5,562	0
Country	Macro Indicators	Banking	Monthly	✓	2	1,902	150
Country	Macro Indicators	Capacity Utilization By Industry	Monthly	✓	11	8,385	825
Country	Macro Indicators	Credit	Monthly	✓	2	1,696	148
Country	Macro Indicators	Current Account	Quarterly	✓	4	650	48
Country	Macro Indicators	Energy Activity	Monthly	✓	6	2,557	397
Country	Macro Indicators	Energy Refinery	Weekly	×	5	7,302	0
Country	Macro Indicators	External	Monthly	✓	6	2,460	444
Country	Macro Indicators	External	Quarterly	✓	3	951	75
Country	Macro Indicators	FDI Flows	Quarterly	✓	1	317	24
Country	Macro Indicators	FDI Flows	Annual	✓	1	80	4

Continued on next page

Scope	Category	Detail Category	Frequency	Eval.	Series	Points	Eval. Points
Country	Macro Indicators	Financial Flows	Quarterly	✓	1	316	24
Country	Macro Indicators	Fiscal	Monthly	✓	1	546	75
Country	Macro Indicators	Fiscal	Quarterly	✓	1	240	24
Country	Macro Indicators	Fiscal	Annual	✓	2	222	8
Country	Macro Indicators	Freight	Monthly	✓	1	313	73
Country	Macro Indicators	Gdp	Quarterly	✓	2	634	50
Country	Macro Indicators	Housing	Monthly	✓	4	2,367	219
Country	Macro Indicators	Housing Price	Monthly	✓	10	3,099	747
Country	Macro Indicators	Housing Price	Quarterly	×	1	204	0
Country	Macro Indicators	Income	Monthly	✓	3	2,421	225
Country	Macro Indicators	Industrial Production By Industry	Monthly	✓	28	19,776	2,100
Country	Macro Indicators	Inflation	Monthly	✓	10	6,793	675
Country	Macro Indicators	International Investment Position	Quarterly	✓	5	400	120
Country	Macro Indicators	International Investment Position	Annual	✓	1	50	4
Country	Macro Indicators	Inventory	Monthly	✓	7	2,870	518
Country	Macro Indicators	Labor	Weekly	×	2	6,189	0
Country	Macro Indicators	Labor	Monthly	✓	14	7,872	1,046
Country	Macro Indicators	Labor By Industry	Monthly	✓	16	15,744	1,200
Country	Macro Indicators	Money	Weekly	×	1	1,220	0
Country	Macro Indicators	Money	Monthly	✓	5	4,035	375
Country	Macro Indicators	Policy Rate	Business-daily	×	1	18,453	0
Country	Macro Indicators	Policy Rate	Daily	×	3	22,267	0
Country	Macro Indicators	Portfolio Flows	Quarterly	✓	7	2,231	168
Country	Macro Indicators	Productivity	Quarterly	✓	3	948	72
Country	Macro Indicators	Regional Fed	Monthly	✓	2	994	152
Country	Macro Indicators	Saving	Monthly	✓	1	807	75
Country	Macro Indicators	Sentiment	Monthly	✓	3	1,763	200
Country	Macro Indicators	Spending	Monthly	✓	6	3,690	450
Country	Macro Indicators	Stress	Monthly	✓	1	434	75
Country	Macro Indicators	TIC Flows	Quarterly	✓	4	851	95
Individual	ETF	ETF OHLCV Close	Business-daily	✓	5,107	8,100,873	3,919,924
Individual	ETF	ETF OHLCV High	Business-daily	×	5,107	8,098,720	0
Individual	ETF	ETF OHLCV Low	Business-daily	×	5,107	8,095,707	0
Individual	ETF	ETF OHLCV Open	Business-daily	×	5,107	8,108,763	0
Individual	ETF	ETF OHLCV Volume	Business-daily	×	5,103	8,143,493	0
Individual	Equity	Equity OHLCV Close	Business-daily	✓	6,143	24,193,386	7,743,280
Individual	Equity	Equity OHLCV High	Business-daily	×	6,144	24,189,072	0
Individual	Equity	Equity OHLCV Low	Business-daily	×	6,144	24,106,760	0
Individual	Equity	Equity OHLCV Open	Business-daily	×	6,145	24,253,740	0
Individual	Equity	Equity OHLCV Volume	Business-daily	×	6,169	24,986,230	0
Individual	Fundamentals	Capital Expenditures	Quarterly	✓	3,442	140,619	60,843
Individual	Fundamentals	Cash and Cash Equivalents	Quarterly	✓	4,642	206,110	84,757
Individual	Fundamentals	EPS Basic	Quarterly	✓	4,185	169,330	75,581
Individual	Fundamentals	EPS Diluted	Quarterly	✓	4,166	166,806	74,206
Individual	Fundamentals	Long-Term Debt	Quarterly	✓	3,249	100,936	43,229
Individual	Fundamentals	Net Income	Quarterly	✓	4,357	182,999	80,909
Individual	Fundamentals	Operating Cash Flow	Quarterly	✓	4,359	185,546	83,997
Individual	Fundamentals	Operating Income	Quarterly	✓	3,707	151,714	67,678
Individual	Fundamentals	R And D Expense	Quarterly	×	1,851	64,374	0
Individual	Fundamentals	Revenue	Quarterly	✓	3,677	99,475	60,974
Individual	Fundamentals	Shares Diluted	Quarterly	×	4,173	168,275	0
Individual	Fundamentals	Stockholders Equity	Quarterly	✓	4,697	210,564	93,125
Individual	Fundamentals	Total Assets	Quarterly	✓	4,699	196,491	91,002
Individual	Fundamentals	Total Liabilities	Quarterly	✓	4,238	160,730	77,746
Individual	News Sentiment	Sentiment Tone	Daily	×	11	37,488	0
Individual	News Sentiment	Sentiment Volume	Daily	×	11	37,488	0

B Metric Assignment by Data Category

Appendix B expands the metric assignment used by FinVerse. Table 5 maps each compact variant ID to the corresponding (p, h) combinations. Tables 6–15 then provide the detailed metric assignment for each evaluated scope-category group. Each table lists the detail category, frequency, number of evaluation points, evaluation aspect, metric families, and (p, h) variant used for that task. Together, these tables specify how FinVerse connects each financial data group to the metrics that are economically meaningful for that group.

Interpretation of (p, h) variants. The notation (p, h) follows the target definition in Section 3.2. For a forecast origin o , h is the forecast horizon, so the forecast is evaluated at the future endpoint $o + h$. The parameter p is the comparison window ending at that endpoint: relative-change metrics use y_{o+h-p} as the reference value and evaluate $r_o^{(p, h)} = y_{o+h} / y_{o+h-p} - 1$. Thus, p determines the economic movement being measured, while h determines how far

Table 5: Mapping from compact variant IDs to (p, h) combinations.

(p, h) Variant	(p, h) combinations
D5	$(1w, 1w), (1m, 1m), (1q, 1q), (1h, 1h), (1y, 1y)$
M10	$(1m, 1m), (1q, 1m), (1q, 1q), (1h, 1m), (1h, 1q), (1h, 1h), (1y, 1m), (1y, 1q), (1y, 1h), (1y, 1y)$
M4	$(1m, 1m), (1q, 1m), (1h, 1m), (1y, 1m)$
Q6	$(1q, 1q), (1h, 1q), (1h, 1h), (1y, 1q), (1y, 1h), (1y, 1y)$
Q3	$(1q, 1q), (1h, 1q), (1y, 1q)$
A1	$(1y, 1y)$
N/A	N/A

ahead that movement is evaluated. Diagonal variants such as $(1m, 1m)$ or $(1y, 1y)$ measure the change from the forecast origin to the target endpoint, which is the natural setting for asset-like returns, cross-sectional ranking, and portfolio horizons. Non-diagonal variants such as $(1y, 1m)$ measure a longer-window quantity at a shorter forecast horizon; for example, they evaluate the year-over-year value that will be observed one month after the forecast origin. In the compact IDs below, $1w$, $1m$, $1q$, $1h$, and $1y$ denote one week, one month, one quarter, one half-year, and one year, respectively. The variant IDs are used only to keep the metric-assignment tables compact; the corresponding (p, h) combinations are listed explicitly in Table 5.

Table 6: Metric assignments and variant IDs for inter-country Foreign Exchange evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
FX Pair	Business-daily	34,503	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5

Table 7: Metric assignments and variant IDs for inter-country Commodity evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Future Front	Business-daily	43,601	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5
Spot	Business-daily	4,932	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5
Spot	Monthly	975	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5
Inventory Stock	Weekly	2,303	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Commodity ETF Single	Business-daily	8,245	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Commodity ETF Broad	Business-daily	3,298	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Commodity ETF Thematic	Business-daily	3,298	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A

Table 8: Metric assignments and variant IDs for inter-country Crypto evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Crypto Pair Close (Top 100)	Daily	120,963	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5

Table 9: Metric assignments and variant IDs for inter-country Market Index evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Country Index	Business-daily	31,391	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5
Regional Index	Business-daily	3,304	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A

Table 10: Metric assignments and variant IDs for country-level Fixed Income evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Treasury Yield	Business-daily	18,161	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
Credit Spread	Business-daily	5,230	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	D5
Bond Total Return Index	Business-daily	5,230	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Money Market	Business-daily	12,543	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
TIPS Yield	Business-daily	6,604	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Breakeven	Business-daily	3,304	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Corporate Yield	Business-daily	3,302	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Term Spread	Business-daily	3,304	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Forward Inflation	Business-daily	1,652	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A
Inflation Expectation	Monthly	530	Point-wise	HR, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Breakeven	Monthly	150	Point-wise	HR, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Corporate Yield	Monthly	150	Point-wise	HR, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Mortgage Rate	Weekly	812	Point-wise	HR, $MASE_{abs}$	D5
			Point-wise	$MASE_{ori}$	N/A

Table 11: Metric assignments and variant IDs for country-level Market Indicators evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(<i>p, h</i>) Variant
Implied Volatility	Business-daily	13,361	Point-wise	HR, MASE _{rel} , MASE _{abs}	D5
			Point-wise	MASE _{ori}	N/A
US Equity Index	Business-daily	14,868	Point-wise	HR, MASE _{rel} , MASE _{abs}	D5
			Point-wise	MASE _{ori}	N/A

Table 12: Metric assignments and variant IDs for country-level Macro Indicators evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(<i>p, h</i>) Variant
Inflation	Monthly	675	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Labor	Monthly	1,046	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
GDP	Quarterly	50	Point-wise	HR, MASE _{rel} , MASE _{abs}	Q6
			Point-wise	Δ HR, Δ^2 HR	Q3
			Point-wise	MASE _{ori}	N/A
Housing Price	Monthly	747	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Productivity	Quarterly	72	Point-wise	HR, MASE _{rel} , MASE _{abs}	Q6
			Point-wise	Δ HR, Δ^2 HR	Q3
			Point-wise	MASE _{ori}	N/A
Money	Monthly	375	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Industrial Production by Industry	Monthly	2,100	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Labor by Industry	Monthly	1,200	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Activity	Monthly	744	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Capacity Utilization by Industry	Monthly	825	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Inventory	Monthly	518	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Spending	Monthly	450	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A
Energy Activity	Monthly	397	Point-wise	HR, MASE _{rel} , MASE _{abs}	M10
			Point-wise	Δ HR, Δ^2 HR	M4
			Point-wise	MASE _{ori}	N/A

Continued on next page

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Sentiment	Monthly	200	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Housing	Monthly	219	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Income	Monthly	225	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Credit	Monthly	148	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Regional Fed	Monthly	152	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Saving	Monthly	75	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Stress	Monthly	75	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Freight	Monthly	73	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Portfolio Flows	Quarterly	168	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
Current Account	Quarterly	48	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
TIC Flows	Quarterly	95	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
Financial Flows	Quarterly	24	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
External	Monthly	444	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
External	Quarterly	75	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
Banking	Monthly	150	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Fiscal	Monthly	75	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	M10
			Point-wise	ΔHR , $\Delta^2 HR$	M4
			Point-wise	$MASE_{ori}$	N/A
Fiscal	Quarterly	24	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3

Continued on next page

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
			Point-wise	MASE _{ori}	N/A
Fiscal	Annual	8	Point-wise Point-wise	HR, Δ HR, Δ^2 HR, MASE _{rel} , MASE _{abs} MASE _{ori}	A1 N/A
International Investment Position	Quarterly	120	Point-wise Point-wise Point-wise	HR, MASE _{rel} , MASE _{abs} Δ HR, Δ^2 HR MASE _{ori}	Q6 Q3 N/A
International Investment Position	Annual	4	Point-wise Point-wise	HR, Δ HR, Δ^2 HR, MASE _{rel} , MASE _{abs} MASE _{ori}	A1 N/A
FDI Flows	Quarterly	24	Point-wise Point-wise Point-wise	HR, MASE _{rel} , MASE _{abs} Δ HR, Δ^2 HR MASE _{ori}	Q6 Q3 N/A
FDI Flows	Annual	4	Point-wise Point-wise	HR, Δ HR, Δ^2 HR, MASE _{rel} , MASE _{abs} MASE _{ori}	A1 N/A

Table 13: Metric assignments and variant IDs for individual-series ETF evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
ETF OHLCV Close	Business-daily	3,919,924	Point-wise Point-wise Cross-sectional Portfolio	HR, MASE _{rel} , MASE _{abs} MASE _{ori} IC return, vol, Sharpe, MDD	D5 N/A D5 D5

Table 14: Metric assignments and variant IDs for individual-series Equity evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Equity OHLCV Close	Business-daily	7,743,280	Point-wise Point-wise Cross-sectional Portfolio	HR, MASE _{rel} , MASE _{abs} MASE _{ori} IC return, vol, Sharpe, MDD	D5 N/A D5 D5

Table 15: Metric assignments and variant IDs for individual-series Fundamentals evaluation categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Capital Expenditures	Quarterly	60,843	Point-wise Point-wise Point-wise	HR, MASE _{rel} , MASE _{abs} Δ HR, Δ^2 HR MASE _{ori}	Q6 Q3 N/A
Cash and Cash Equivalents	Quarterly	84,757	Point-wise Point-wise Point-wise	HR, MASE _{rel} , MASE _{abs} Δ HR, Δ^2 HR MASE _{ori}	Q6 Q3 N/A
EPS Basic	Quarterly	75,581	Point-wise Point-wise Point-wise Cross-sectional	HR, MASE _{rel} , MASE _{abs} Δ HR, Δ^2 HR MASE _{ori} IC	Q6 Q3 N/A Q6
EPS Diluted	Quarterly	74,206	Point-wise Point-wise Point-wise Cross-sectional	HR, MASE _{rel} , MASE _{abs} Δ HR, Δ^2 HR MASE _{ori} IC	Q6 Q3 N/A Q6

Continued on next page

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Long-Term Debt	Quarterly	43,229	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
Net Income	Quarterly	80,909	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	Q6
Operating Cash Flow	Quarterly	83,997	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
Operating Income	Quarterly	67,678	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	Q6
Revenue	Quarterly	60,974	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
			Cross-sectional	IC	Q6
Stockholders Equity	Quarterly	93,125	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
Total Assets	Quarterly	91,002	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A
Total Liabilities	Quarterly	77,746	Point-wise	HR, $MASE_{rel}$, $MASE_{abs}$	Q6
			Point-wise	ΔHR , $\Delta^2 HR$	Q3
			Point-wise	$MASE_{ori}$	N/A

C Additional Result Tables

This appendix provides three complementary views of model performance. Table 16 reports the overall leaderboard used in the main analysis. It shows each model’s overall FinVerse rank, computed as the sum of the point-wise, IC, and portfolio placement ranks, together with the aspect-level ranks and the corresponding K values. The K values are geometric means of metric-level ranks within each evaluation aspect. Table 17 reports scope-wise performance. It aggregates metric-level ranks separately within the Country, Individual, and Inter-country scopes, showing whether a model performs consistently across different financial data scopes or is relatively stronger in a particular scope. Table 18 reports category-wise performance. It further breaks down rank aggregation by financial category, such as Foreign Exchange, Commodity, Crypto, Fixed Income, Macro Indicators, ETF, Equity, and Fundamentals. This table highlights which models are preferred for specific financial data categories and shows that strong overall performance does not necessarily imply uniformly strong performance across all categories.

Table 16: Overall and aspect-wise model rankings. Overall is the sum of the point-wise, IC, and portfolio placement ranks. K columns report the geometric mean of metric ranks within each aspect.

Rank	Model	Overall Rank	Point		IC		Portfolio	
			Rank	K	Rank	K	Rank	K
1	Reverso Small	10	5	9.02	1	8.16	4	12.10
2	Chronos-2 Synth	11	8	11.02	2	8.17	1	9.18
3	TiRex 1.1	17	1	6.42	3	9.36	13	14.91
4	TimesFM 2.5 200M	29	6	9.45	4	10.24	19	16.44
5	Chronos-2	30	7	9.50	13	15.39	10	14.55
6	Chronos-Bolt Base	31	10	12.19	15	15.67	6	13.81
7	Chronos-Bolt Small	36	11	12.20	11	14.98	14	15.20
8	T0 Alpha	39	12	12.83	24	18.10	3	10.60
9	TiRex 2 Pretrain	39	2	7.02	5	11.14	32	19.53
10	Toto 2.0 2.5B	39	17	14.97	14	15.52	8	14.23
11	TiRex 2 ZS	40	3	7.94	6	11.19	31	19.47
12	VisionTS	44	9	11.47	7	12.50	28	18.62
13	PatchFM	45	4	8.46	12	15.21	29	18.66
14	Toto 2.0 4M	48	14	13.70	22	17.39	12	14.90
15	TimesFM 2.0 500M	49	38	28.43	9	13.25	2	10.33
16	Toto 2.0 1B	54	15	14.22	18	16.69	21	16.89
17	Kairos-50M	55	18	15.14	28	19.09	9	14.46
18	TempoPFN	55	13	13.31	16	16.06	26	17.99
19	Granite FlowState R1	60	20	16.23	25	18.14	15	15.56
20	TimesFM 1.0 200M	62	21	16.32	17	16.07	24	17.89
21	Super Linear	63	23	17.40	10	14.77	30	19.15
22	Kairos-10M	65	27	19.96	27	18.90	11	14.78
23	Toto 2.0 22M	65	19	15.17	26	18.18	20	16.60
24	FlowState	67	22	16.72	29	19.36	16	15.66
25	Toto 2.0 313M	68	16	14.59	30	19.57	22	17.22
26	Timer-S1	70	24	17.75	8	13.16	38	22.27
27	Toto Open Base 1.0	74	26	19.19	31	20.38	17	16.21
28	Moirai 1.1-R Large	76	37	27.03	34	21.09	5	12.21
29	Moirai 1.1-R Base	80	36	26.64	37	23.38	7	13.98
30	Moirai 1.1-R Small	82	32	24.88	32	20.55	18	16.33
31	PatchTST-FM R1	83	33	25.45	23	17.52	27	18.54
32	Sundial Base 128M	85	25	19.02	20	17.34	40	22.75
33	Kairos-23M	87	28	20.69	36	23.31	23	17.80
34	Granite PatchTST-FM R1	91	39	28.55	19	16.94	33	19.56
35	Moirai 2.0-R Small	97	30	23.08	33	20.77	34	20.56
36	TTM R3	97	34	26.32	21	17.37	42	23.64
37	TTM R1	99	31	24.00	43	26.57	25	17.93
38	CleanTS-65M	105	35	26.36	35	22.21	35	20.69
39	TTM R2	114	29	21.32	42	26.51	43	24.51
40	YingLong 50M	116	41	34.31	38	23.58	37	21.77
41	YingLong 300M	118	43	34.61	39	23.70	36	21.21
42	YingLong 110M	121	42	34.53	40	23.93	39	22.58
43	YingLong 6M	122	40	33.99	41	24.32	41	23.08

Table 17: Scope-wise rank aggregation. Lower values are better.

Model	Country	Individual	Inter-country
Reverso Small	9.21	9.03	9.12
Chronos-2 Synth	12.25	9.27	6.80
TiRex 1.1	5.96	8.21	11.43
TimesFM 2.5 200M	9.75	9.48	10.66
Chronos-2	9.08	13.40	11.33
Chronos-Bolt Base	12.02	14.18	12.56
Chronos-Bolt Small	11.99	15.72	11.92
T0 Alpha	11.38	15.90	18.32
TiRex 2 Pretrain	6.47	11.09	11.60
Toto 2.0 2.5B	15.73	15.78	11.02
TiRex 2 ZS	7.39	10.92	13.22
VisionTS	11.08	10.76	18.33
PatchFM	7.52	17.39	11.86
Toto 2.0 4M	13.28	19.24	12.41
TimesFM 2.0 500M	30.24	14.62	21.80
Toto 2.0 1B	14.82	16.30	11.43
Kairos-50M	14.67	18.81	14.67
TempoPFN	14.59	17.61	7.76
Granite FlowState R1	14.95	25.95	15.11
TimesFM 1.0 200M	16.40	12.99	20.85
Super Linear	18.47	11.79	19.42
Kairos-10M	19.61	24.96	14.96
Toto 2.0 22M	15.41	18.67	12.45
FlowState	15.29	28.62	15.00
Toto 2.0 313M	14.96	17.85	12.22
Timer-S1	18.53	12.78	20.72
Toto Open Base 1.0	19.41	20.09	16.32
Moirai 1.1-R Large	26.93	20.80	23.61
Moirai 1.1-R Base	26.45	21.43	24.92
Moirai 1.1-R Small	24.97	19.98	24.08
PatchTST-FM R1	26.15	17.60	25.66
Sundial Base 128M	19.20	16.09	22.81
Kairos-23M	19.94	23.51	20.88
Granite PatchTST-FM R1	29.88	17.93	27.28
Moirai 2.0-R Small	22.06	24.82	24.66
TTM R3	28.33	16.70	25.25
TTM R1	23.88	20.86	25.44
CleanTS-65M	26.34	21.88	27.44
TTM R2	21.91	18.75	23.87
YingLong 50M	36.21	24.99	27.17
YingLong 300M	36.40	25.00	27.61
YingLong 110M	36.21	25.69	28.06
YingLong 6M	35.89	24.89	27.83

Table 18: Frequency-wise rank aggregation. Lower values are better.

Model	Business-daily	Daily	Weekly	Monthly	Quarterly	Annual
Reverso Small	7.33	11.85	11.86	9.98	8.90	16.53
Chronos-2 Synth	7.09	4.30	8.20	11.50	15.17	14.42
TiRex 1.1	10.24	14.85	8.36	4.60	10.11	12.45
TimesFM 2.5 200M	9.91	16.45	9.34	9.28	10.64	8.19
Chronos-2	12.33	12.36	10.20	8.66	10.45	18.61
Chronos-Bolt Base	12.59	12.50	7.93	12.94	11.42	13.36
Chronos-Bolt Small	11.57	14.08	9.53	12.70	12.92	13.28
T0 Alpha	18.12	21.50	20.42	9.15	17.35	22.63
TiRex 2 Pretrain	13.24	9.67	10.89	4.99	10.73	11.18
Toto 2.0 2.5B	9.33	12.90	7.63	16.37	19.58	33.25
TiRex 2 ZS	14.79	13.28	10.46	6.01	10.13	10.96
VisionTS	18.96	23.78	28.45	10.54	8.90	9.46
PatchFM	9.89	15.21	10.66	6.63	15.34	17.73
Toto 2.0 4M	12.59	12.96	18.29	13.19	16.57	22.87
TimesFM 2.0 500M	21.97	13.15	25.56	34.18	18.06	14.07
Toto 2.0 1B	10.92	12.81	6.58	14.59	18.99	35.90
Kairos-50M	12.03	20.66	15.61	17.92	13.02	15.20
TempoPFN	6.40	21.72	7.97	16.24	18.87	19.57
Granite FlowState R1	15.07	22.86	14.19	15.41	18.87	23.77
TimesFM 1.0 200M	20.47	19.90	26.10	15.53	14.53	11.31
Super Linear	20.61	21.80	24.40	18.26	13.00	13.54
Kairos-10M	13.77	19.51	24.40	22.18	20.17	26.28
Toto 2.0 22M	13.14	10.04	13.58	16.04	16.57	33.38
FlowState	14.63	24.36	14.30	15.41	21.93	20.51
Toto 2.0 313M	12.64	11.57	8.35	14.41	19.32	31.23
Timer-S1	23.74	21.87	11.05	17.16	15.04	13.89
Toto Open Base 1.0	17.92	13.24	16.23	18.81	21.39	21.30
Moirai 1.1-R Large	21.67	11.44	16.15	28.50	26.13	12.48
Moirai 1.1-R Base	23.26	17.71	17.39	27.20	25.13	28.30
Moirai 1.1-R Small	23.76	18.44	21.73	26.65	20.14	21.77
PatchTST-FM R1	27.78	21.49	24.39	24.66	22.07	25.02
Sundial Base 128M	25.88	18.68	16.51	17.87	17.28	12.67
Kairos-23M	21.41	23.28	19.42	21.09	18.82	15.77
Granite PatchTST-FM R1	29.63	21.82	32.35	29.65	21.63	23.38
Moirai 2.0-R Small	23.40	20.47	25.18	23.33	21.80	11.92
TTM R3	25.16	27.01	34.44	28.75	20.55	19.58
TTM R1	26.85	22.99	33.17	29.33	13.76	4.01
CleanTS-65M	29.28	22.78	35.95	24.96	24.33	24.97
TTM R2	23.61	26.08	36.03	28.45	11.17	6.05
YingLong 50M	29.83	22.11	34.05	36.50	29.89	27.14
YingLong 300M	30.70	18.62	34.37	36.64	30.10	27.14
YingLong 110M	30.49	21.04	34.42	36.55	30.43	27.14
YingLong 6M	29.31	22.38	34.41	36.49	29.92	27.14