FINVERSE: TOWARD MORE REALISTIC EVALUATION FOR FINANCIAL TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) are increasingly evaluated on broad benchmarks that apply uniform conventional forecast evaluation metrics across heterogeneous time-series. However, in the financial domain, decision-relevant evaluation requires going beyond conventional metrics to assess forecasts in terms of directional accuracy, cross-sectional ranking, and portfolio outcomes. To this end, we introduce FinVerse, a financial time-series benchmark, collecting a broad collection of 116,897 series with 171.1M observations. FinVerse organizes financial data according to their economic semantics and evaluates target series across point-wise forecasting, cross-sectional ranking, and portfolio backtesting. Across 43 public TSFMs, model rankings on FinVerse show only moderate agreement with those on the existing benchmark, GIFT-Eval (r=0.43), suggesting that generic forecasting benchmarks do not fully capture the aspects of forecast quality that matter for financial decision-making. Furthermore, substantial variation in model performance across financial time-series categories and evaluation metrics provides a more informative basis for model selection, enabling practitioners to identify suitable models according to their target series and downstream objectives.

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 time-series foundation models (TSFMs) from many sources (Das et al., 2024; Liu et al., 2024; Woo et al., 2024; Goswami et al., 2024; Liu et al., 2025a; Taga et al., 2025; Ansari et al., 2025). By pretraining on large and heterogeneous collections of time-series, these models seek to capture transferable temporal patterns that generalize across domains.

As time-series foundation models advance, developing benchmarks for evaluating them becomes increasingly important. A central question in designing such benchmarks is whether strong benchmark performance translates into strong performance on the downstream decisions that rely on these forecasts. In other words, if a decision maker selects the best-performing model on a benchmark, should they expect it to perform well for the decisions they ultimately care about? Existing time-series benchmarks such as GIFT-Eval (Aksu et al., 2024), however, evaluate heterogeneous time-series using uniform conventional forecasting metrics, such as the Mean Absolute Scaled Error (MASE) and Continuous Ranked Probability Score (CRPS). Recent evidence suggests that strong performance on such benchmarks does not necessarily translate into strong performance in downstream financial decision-making (Saly-Kaufmann et al., 2026).

This discrepancy may arise for the following two reasons. First, conventional forecast accuracy metrics do not fully characterize the aspects of a forecast that matter for financial decision-making. For example, in stock forecasting, anticipating the direction or relative ranking of future returns may be more relevant to investment decisions than minimizing MASE or CRPS. Second, what constitutes a useful forecast varies substantially across financial time series, making a common evaluation criterion inadequate for capturing their distinct economic roles. For inflation indicators such as the Consumer Price Index (CPI), for instance, it may be important to predict not only the direction but also whether inflation is accelerating or decelerating.

To address these limitations, we introduce **FinVerse** (**Fin**ancial time-series benchmark uni**Verse**), a benchmark for economically informed evaluation of financial time-series forecasting. As shown in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

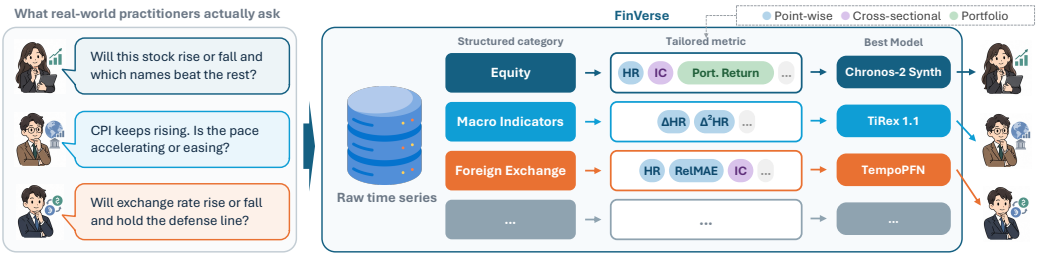

Figure 1: Overview of FinVerse. FinVerse tailors evaluation metrics to each time series, providing a more informative basis for model selection based on the economic characteristics of the target series.

Table 1: Coverage of the full FinVerse data by scope and category. D, BD, W, M, Q, and A denote daily, business-daily, weekly, monthly, quarterly, and annual frequencies, respectively. See Table 5 in the Appendix for detailed statistics at the detail-category level.

Scope	Category	Points	Frequencies
Inter-country	Foreign Exchange	294.4K	BD
	Commodity	261.9K	BD/W/M
	Crypto	5.0M	D
	Global Liquidity	4.5K	Q
	Global Risk	18.1K	D/BD/M
	Market Index	166.4K	BD
Country	Fixed Income	380.9K	BD/W/M
	Market Indicators	270.8K	D/BD/W/M
	Macro Indicators	165.2K	D/BD/W/M/Q/A
Individual	ETF	40.5M	BD
	Equity	121.7M	BD
	Fundamentals	2.2M	Q
	News Sentiment	75.0K	D

Figure 1, FinVerse is built on the principle that each time-series should be evaluated according to the criteria that its economic meaning makes relevant, rather than applying a common set of conventional forecast accuracy metrics across heterogeneous financial time-series. FinVerse organizes financial time-series according to their economic semantics and evaluates each target series using criteria aligned with its role in financial decision-making. This design extends evaluation beyond conventional forecast accuracy to capture complementary aspects of forecast quality through point-wise forecasting, cross-sectional ranking, and portfolio backtesting.

FinVerse contains 116,897 financial time-series with 171.1M observations. The data are organized into a three-level hierarchy consisting of *scope*, *category*, and *detail category*. At the highest level, time-series are classified into three scopes—*inter-country*, *country*, and *individual*—according to the geographic and economic scope. Within each scope, the series are further grouped into categories based on the type of financial time-series they represent, and each category is subsequently divided into more granular detail categories. Table 1 summarizes composition of FinVerse across categories. After establishing this hierarchy, we select 60,232 economically meaningful target series as the fixed evaluation set, comprising 13.6M observations collected from January 1, 2020 onward. FinVerse also provides data collected before January 1, 2020 for foundation model pre-training and fine-tuning.

In our experiments, we find that models that perform well under generic time-series benchmarks do not necessarily perform well on FinVerse, highlighting the need for financial-tailored benchmarks. Moreover, our experiments reveal that the relative performance of forecasting models changes substantially across financial time-series categories and evaluation criteria. This result suggests that FinVerse can support more informed model selection by helping practitioners choose models that are appropriate for both the characteristics of the target financial time series and the intended forecasting objective. The main contributions are as follows:

- We identify a key limitation of existing time-series benchmarks for financial forecasting: applying uniform conventional forecasting metrics across heterogeneous time-series may not adequately capture performance relevant to downstream financial decisions.
- We introduce FinVerse, a large-scale financial time-series benchmark comprising 116,897 series and 171.1M observations. FinVerse organizes financial series according to their economic semantics and evaluates them using criteria aligned with their roles in financial decision-making.
- We show that models performing well on generic time-series benchmarks do not necessarily perform well on FinVerse, providing empirical evidence for the need for financially tailored evaluation beyond conventional forecasting metrics.
- We demonstrate substantial variation in model performance across financial time-series categories and evaluation criteria. These results provide a more informative basis for model selection, helping practitioners identify models suited to their target time series and forecasting objectives.

2 Related Work

Time-Series Foundation Models (TSFMs). The foundation model paradigm has expanded from language to time series. Traditional forecasting models are often trained for a specific dataset, frequency, or forecasting horizon, which limits their ability to transfer across heterogeneous temporal data. In contrast, recent TSFMs aim to pretrain on large and diverse collections of time series and then support zero-shot or few-shot forecasting. Representative forecasting-oriented models include TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024; 2025), Moirai (Woo et al., 2024), Timer (Liu et al., 2024; 2025a), and TimePFN (Taga et al., 2025), which are pretrained on large collections of heterogeneous time series. The dominant architectural paradigm among recent foundation models is the decoder-only Transformer, which adapts autoregressive sequence modeling from language models to temporal prediction. Whether this generality transfers to finance is an open question: a recent study of foundation models on global equity returns reports that generic pretrained models underperform in zero-shot and fine-tuned settings, while models pretrained on financial data improve both forecast accuracy and economic utility (Rahimikia et al., 2025).

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 (Zhou et al., 2021). The Monash Time Series Forecasting Archive broadened the evaluation landscape by collecting diverse real-world forecasting datasets under a unified archive and metadata format (Godahewa et al., 2021). 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 (Woo et al., 2024). GIFT-Eval is a representative and widely used benchmark for general time-series forecasting model evaluation (Aksu et al., 2024). It provides a standardized evaluation suite over heterogeneous datasets and has become a practical reference point for comparing recent time-series foundation models.

Recent benchmark efforts have also begun to systematize financial time-series forecasting evaluation. FinBench and FinTSB provide dedicated benchmarks for financial time-series forecasting (Pederzoli et al., 2026; Hu et al., 2025), but their evaluation scopes are primarily centered on stock data which limits diversity in data coverage. However, FinVerse collects financial time series across multiple categories, including equity prices, exchange rates, commodities, crypto assets, fixed-income indicators, market indicators, macroeconomic indicators, ETFs, and fundamentals. It further supports multiple 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. 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 7–16.

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. Table 1 summarizes the category-level coverage, while Appendix Table 5 provides the detail-category breakdown. In total, FinVerse contains 3 scopes, 13 categories, and 101 detail categories with 116,897 financial series and 171.1M observations. FinVerse includes 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.

After organizing the hierarchy, we manually select 60,232 target series for evaluation based on their relevance to decision-making. These series contribute 13.6M observations over the evaluation period starting January 1, 2020. The resulting evaluation set spans 3 scopes, 10 categories, and 67 detail categories. The largest portion of it consists of business-daily observations from equity (8.0M) and ETF (4.5M) series, reflecting the large number of individual instruments represented in these categories, followed by quarterly company fundamentals (0.8M) and daily crypto series (140.5K). 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 complementary evaluation views: *point-wise forecasting*, *cross-sectional ranking*, and *portfolio performance*. Point-wise forecasting assesses individual predictions. In contrast, cross-sectional ranking evaluates whether a model correctly captures the relative ordering among a set of comparable assets forecast at the same point in time. Finally, portfolio performance translates model predictions into portfolio decisions under one fixed strategy and evaluates their realized performance through backtesting. The first two views are economic proxies: they measure properties of a forecast that financial decisions typically depend on, rather than the utility of a decision itself. Only the third view observes a realized outcome, and then only for the particular strategy defined below. In this section, we simply describe the metrics used for each evaluation view and their conceptual relevance to financial forecasting. Detailed mathematical definitions and implementation procedures are provided in Appendix C.

3.2.1 POINT-WISE FORECASTING

Hit Ratio (HR). HR measures directional correctness. 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. We also propose $\Delta\mathrm{HR}$ and $\Delta^2\mathrm{HR}$, which extend this directional criterion to first- and second-order changes, respectively. These variants are particularly useful for macroeconomic indicators such as CPI: $\Delta\mathrm{HR}$ evaluates whether the model correctly predicts whether the year-over-year value will increase or decrease, whereas $\Delta^2\mathrm{HR}$ evaluates whether it correctly predicts whether the rate of change will accelerate or decelerate relative to the previous period. These variants require the predicted and observed series to match multiple directional conditions simultaneously, as defined in Appendix C. They should therefore not be interpreted as simple binary directional-accuracy measures with a chance level of 0.5.

Relative Mean Absolute Error (RelMAE). FinVerse uses RelMAE, defined as the mean absolute error of the forecast divided by that of a naive forecast evaluated on the same cells. This differs from MASE, whose denominator is the in-sample seasonal-naive error computed from the history preceding the forecast. We adopt RelMAE to provide a direct comparison between model forecasts and simple naive forecasts under the same evaluation window, avoiding dependence on the length

and seasonal structure of the available history. For completeness, we also report results using the classical MASE definition in Appendix E.2. We include error-magnitude 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.

Three variants are considered. First, RelMAE_{rel} evaluates the accuracy of predicting relative changes (e.g., returns) against a no-change relative baseline. Second, RelMAE_{abs} evaluates the accuracy of predicting raw values against a no-change raw-value baseline defined over the same comparison window. Finally, RelMAE_{ori} evaluates raw-value forecasts against the conventional origin-value persistence baseline, where the most recently observed value is repeated across all future horizons. While RelMAE_{rel} and RelMAE_{abs} are evaluated only at financially meaningful forecasting horizons (e.g., one week, one month, one quarter, half a year, and one year ahead), RelMAE_{ori} measures forecasting error over the entire prediction horizon. Results using alternative baselines beyond persistence are also provided in Appendix E.3.

3.2.2 Cross-sectional Ranking

The information coefficient evaluates cross-sectional ranking quality. For a set of assets, FinVerse computes a Spearman rank correlation between predicted and realized future changes. IC is selected because portfolio and allocation decisions often depend on relative ordering across assets rather than independent point accuracy for each series. Therefore, it serves as a proxy for that ordering quality rather than as a measure of allocation outcomes. 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 Portfolio Performance

FinVerse also evaluates how forecasts behave once they are translated into portfolio decisions, converting forecasts into an equal-weight long-only strategy. At each rebalancing point, assets are ranked by their predicted signal, and the top 10% are selected to construct an equal-weight portfolio, with a minimum of ten assets. The selected portfolio is then held for the corresponding horizon before being rebalanced. We consider holding horizons of one week, one month, one quarter, half a year, and one year, restricted in each group to the horizons its sampling frequency admits. To reduce sensitivity to the choice of a calendar start date, FinVerse performs the backtest using every possible starting offset within each holding horizon and averages the resulting portfolio statistics.

For each portfolio, we measure annualized return, annualized volatility, Sharpe ratio, and maximum drawdown (MDD). These metrics capture complementary aspects of the realized investment outcomes. Annualized return measures portfolio profitability, while annualized volatility captures the variability of realized returns. The Sharpe ratio evaluates return relative to risk, and MDD measures the largest peak-to-trough portfolio loss. The portfolio aspect rewards lower risk as well as higher returns. We acknowledge that portfolio performance can depend on the choice of investment strategy. To start from a simple setting, we keep the portfolio construction assumptions minimal, following a setup similar to Pederzoli et al. (2026). Details are described in Appendix C.6.

3.3 Evaluation Periods and Forecast Horizons

We construct the evaluation test set using observations from the evaluation start date, `2020-01-01`, onward. 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.

Table 2: Prediction lengths and minimum lookback requirements per frequency.

Frequency	Pred. Length	Min. Lookback
Daily	365	365
Business-daily	260	260
Weekly	52	52
Monthly	12	12
Quarterly	4	4
Annual	1	4

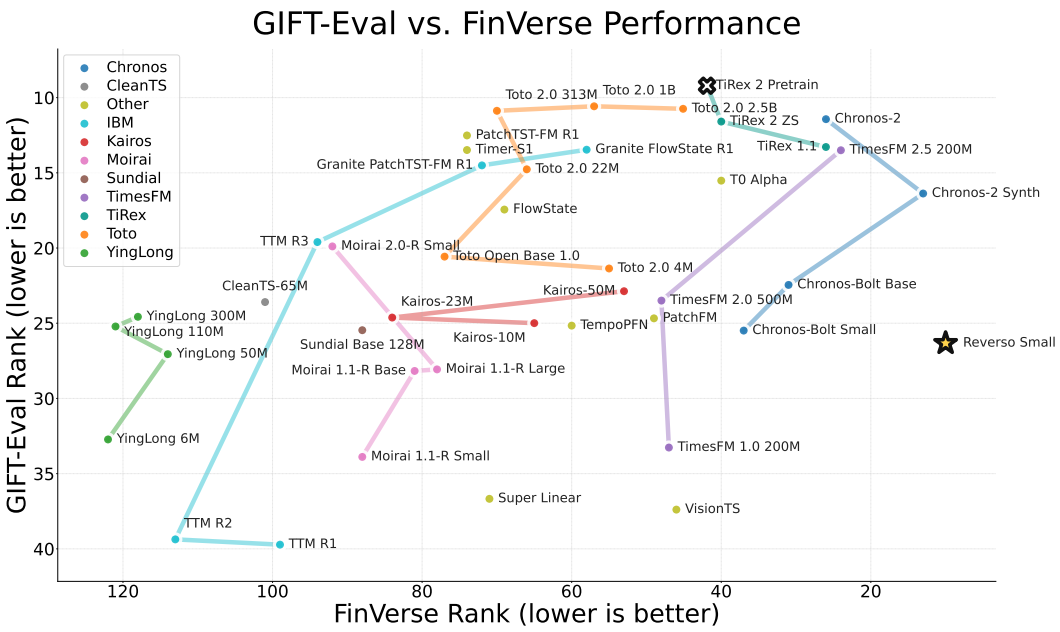

Figure 2: Comparison between FinVerse and GIFT-Eval rankings. The GIFT-Eval rank is the average of the public GIFT-Eval MASE rank and probabilistic forecast rank. The star marks the best-ranked model in FinVerse, and the X marker marks the best-ranked model in GIFT-Eval.

4 Experiments

4.1 Experimental Setup

Foundation Models. We evaluate FinVerse with 43 time-series foundation models, covering major model families such as TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024; 2025), Moirai (Woo et al., 2024), Timer (Liu et al., 2024; 2025a), Toto (Cohen et al., 2025), TiRex (Auer et al., 2025), and TTM (Ekambaram et al., 2024). The complete list of evaluated models is provided in Appendix B. All foundation models are evaluated in a zero-shot setting without any fine-tuning on FinVerse. However, given the release dates of the evaluated models and the fact that the FinVerse evaluation period begins in 2020, we cannot completely rule out the possibility that some evaluation-period data may have been included in their pretraining data. For all models, we use the minimum lookback window specified for each frequency in Table 2 as a context window.

Overall Rank Aggregation. FinVerse employs a diverse set of evaluation metrics, whose numerical magnitudes can vary substantially across financial time series. Directly averaging raw metric values would 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 (i.e., Pointwise Forecasting, Cross-sectional Ranking, and Portfolio Performance), 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 to contribute evenly to the final evaluation while preventing large categories from dominating the benchmark. As a result, smaller but economically important categories are appropriately reflected in the overall ranking.

4.2 Overall Benchmark Performance

Figure 2 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 agree only moder-

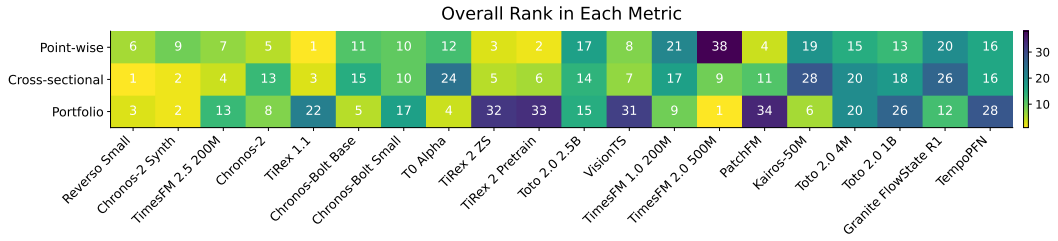

Figure 3: Top-20 models ranked separately under three aspects: point-wise forecasting, cross-sectional ranking, and portfolio performance aspects.

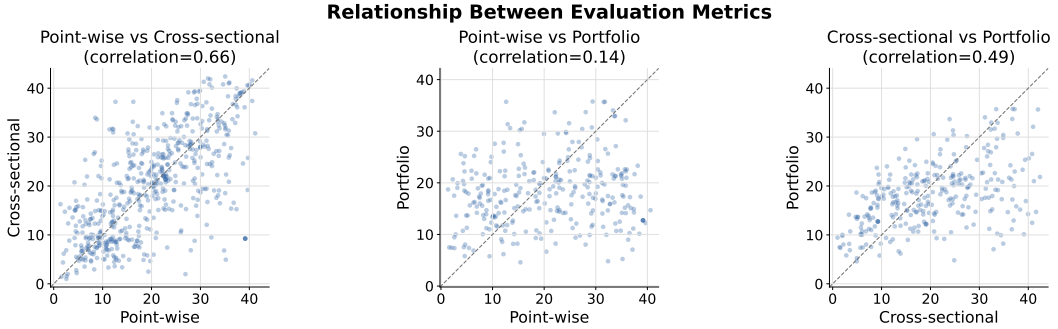

Figure 4: Pairwise relationship between point-wise, cross-sectional, and portfolio ranks.

ately (Pearson $r = 0.43$; Spearman $\rho = 0.41$; Kendall $\tau = 0.27$), indicating that they are related but far from interchangeable. For example, Reverso Small is the strongest FinVerse model despite a weaker GIFT-Eval rank. The result shows that strong performance on a generic benchmark does not necessarily translate into superior performance under our decision-relevant financial evaluation.

4.3 Performance Across Evaluation Groups

Figure 3 reports model rankings separately for the three evaluation aspects: point-wise forecasting, cross-sectional ranking, and portfolio performance. TiRex 1.1 performs best in the point-wise aspect, followed by TiRex 2 Pretrain. Cross-sectional ranking favors Reverso Small and Chronos-2 Synth. Portfolio performance also favors different models: TimesFM 2.0 500M ranks first, followed by Chronos-2 Synth and Reverso Small. Figure 4 further examines the relationships among the three evaluation aspects at the detail-category level, where each point is one model evaluated on one detail category–frequency unit and its coordinate is the geometric mean of its ranks under that aspect. Point-wise and cross-sectional ranking are the most closely related pair ($r = 0.66$; $\rho = 0.67$; $\tau = 0.49$). The point-wise–portfolio relationship is much weaker ($r = 0.14$, $\rho = 0.12$, $\tau = 0.08$), while the cross-sectional–portfolio relationship is moderately aligned ($r = 0.49$, $\rho = 0.48$, $\tau = 0.35$). Portfolio results therefore carry information that conventional per-series accuracy does not, and the three aspects are complementary rather than redundant, supporting a benchmark that evaluates forecasting models from three perspectives.

4.4 Performance Across Categories and Evaluation Metrics

Tables 3 and 4 further reveal substantial heterogeneity in model performance across both financial time-series categories and evaluation metrics. As shown in Table 3, at the category level, no single model consistently dominates: Chronos-2 Synth performs best for Commodity, Crypto, and Equity, TempoPFN performs best for Foreign Exchange and Market Index, and Toto 2.0 2.5B leads Market Indicators and ETF, while Reverso Small, TiRex 1.1, and VisionTS lead Fixed Income, Macro Indicators, and Fundamentals, respectively. A similar pattern emerges at the metric level, where different models achieve the strongest results for different forecasting objectives. Reverso Small and PatchFM jointly attain the highest hit ratio, while Reverso Small achieves the lowest RelMAE_{ori}.

Table 3: Category-wise rank aggregation. Lower values are better.

Model	Inter-country				Country			Individual		
	Commodity	Crypto	Foreign Exchange	Market Index	Fixed Income	Macro Indicators	Market Indicators	ETF	Equity	Fundamentals
Reverso Small	8.80	11.50	17.50	6.37	**8.33**	9.47	11.26	9.71	6.12	10.55
Chronos-2 Synth	**6.78**	**4.34**	8.08	10.10	11.08	12.54	13.47	6.34	**4.07**	13.85
TimesFM 2.5 200M	11.90	12.80	7.43	11.36	11.54	9.58	10.48	11.66	11.25	10.53
Chronos-2	11.46	11.55	13.75	12.70	13.98	8.41	11.29	10.11	13.46	9.35
TiRex 1.1	10.47	14.31	12.30	18.06	9.22	**5.50**	10.70	11.95	12.08	8.29
Chronos-Bolt Base	14.76	11.41	10.26	13.39	11.97	12.03	16.13	10.39	16.41	11.69
Chronos-Bolt Small	14.70	15.09	10.70	10.52	12.20	12.03	14.78	10.91	18.39	11.55
T0 Alpha	19.68	24.04	14.45	20.85	12.45	11.00	23.36	17.58	11.24	15.72
TiRex 2 ZS	12.26	13.09	15.61	19.43	10.41	6.85	15.95	20.66	22.89	8.13
TiRex 2 Pretrain	11.50	10.06	13.30	17.41	9.98	5.96	12.44	19.61	21.54	9.06
Toto 2.0 2.5B	7.31	13.50	10.37	10.19	9.56	16.82	**2.92**	**3.35**	8.62	21.73
VisionTS	22.91	23.60	9.95	16.30	12.27	10.69	26.10	20.95	24.57	**6.73**
TimesFM 1.0 200M	20.34	17.23	13.88	24.36	19.91	15.99	13.26	16.74	18.86	13.30
TimesFM 2.0 500M	21.88	14.80	16.07	27.83	29.84	30.33	32.91	12.37	11.95	15.11
PatchFM	12.34	14.56	17.38	11.26	8.51	7.40	11.15	13.75	11.33	19.99
Kairos-50M	12.97	20.87	23.92	11.33	14.25	15.19	11.13	8.69	16.89	22.72
Toto 2.0 4M	10.71	12.78	18.46	12.52	13.31	12.97	13.23	13.27	13.75	21.90
Toto 2.0 1B	6.93	12.35	11.49	11.98	10.62	15.24	5.45	5.04	7.37	21.52
Granite FlowState R1	12.82	22.54	20.15	15.47	13.71	15.13	20.68	19.57	21.18	23.44
TempoPFN	9.48	21.09	**4.39**	**4.66**	10.63	16.27	4.24	8.40	11.32	21.07
Kairos-10M	15.60	21.05	17.30	10.95	15.17	20.73	19.47	15.53	18.71	29.76
Toto 2.0 22M	11.29	9.63	15.60	11.34	14.05	15.57	7.50	13.05	15.13	19.55
FlowState	13.05	24.62	21.26	15.41	15.26	15.29	20.52	20.25	18.22	27.98
Toto 2.0 313M	8.42	11.59	14.28	12.51	12.80	15.03	5.62	8.22	10.12	19.15
Super Linear	22.02	21.74	14.39	22.31	16.43	18.99	18.09	27.86	21.11	9.86
Granite PatchTST-FM R1	27.73	20.26	22.26	24.09	31.80	29.56	29.03	28.57	20.63	19.57
PatchTST-FM R1	27.92	21.58	21.77	22.51	25.60	26.30	28.09	27.87	19.46	19.37
Timer-S1	20.64	19.48	19.82	26.10	18.87	18.51	23.76	28.82	27.65	12.24
Toto Open Base 1.0	15.94	15.74	34.67	12.78	14.97	20.40	22.85	17.68	17.74	16.77
Moirai 1.1-R Large	28.64	13.74	30.07	25.02	22.37	28.54	15.25	16.28	12.83	23.67
Moirai 1.1-R Base	27.51	18.48	24.52	28.04	20.11	27.86	26.41	20.10	18.98	21.15
Kairos-23M	22.03	23.89	14.87	22.44	23.31	19.38	23.53	18.05	23.07	24.41
Moirai 1.1-R Small	27.48	18.67	19.62	26.56	21.55	25.57	22.56	26.69	23.51	19.88
Sundial Base 128M	23.14	18.19	22.04	27.22	20.95	18.86	26.08	31.70	27.81	14.91
Moirai 2.0-R Small	25.99	20.48	25.88	21.52	23.27	21.78	25.70	23.10	17.61	22.68
TTM R3	26.50	25.65	27.38	16.74	27.53	28.44	26.77	26.87	28.10	17.36
TTM R1	24.00	24.81	21.84	29.99	27.58	22.91	25.30	30.03	23.43	16.58
CleanTS-65M	28.08	22.42	24.91	26.17	30.52	25.76	24.09	34.51	25.36	22.51
TTM R2	27.51	21.47	24.62	15.61	23.56	21.41	29.67	29.53	28.49	15.76
YingLong 50M	28.97	22.21	21.05	26.70	35.04	36.58	33.95	33.40	27.99	26.73
YingLong 300M	29.44	22.01	24.44	26.71	35.70	36.70	33.01	33.19	26.16	26.88
YingLong 110M	28.92	22.42	25.70	26.91	35.40	36.56	32.65	33.28	27.22	27.50
YingLong 6M	29.77	23.74	24.76	25.27	33.39	36.59	31.82	33.85	26.83	26.84

TiRex 1.1 leads $\Delta\mathrm{HR}$, $\Delta^2\mathrm{HR}$, and achieves the lowest RelMAE_{rel}, while TiRex 1.1 and TiRex 2 Pretrain jointly achieve the lowest RelMAE_{abs}. For the decision-oriented metrics, Chronos-2 Synth achieves the strongest IC, while TTM R1, Kairos-10M, TimesFM 2.0 500M, and Moirai 1.1-R Large attain the best return, Sharpe ratio, maximum drawdown, and volatility, respectively. Together, these results show that model superiority is highly dependent on both the type of financial time series being forecast and the criterion used to define a useful forecast. FinVerse therefore provides a more informative basis for model selection by allowing practitioners to identify models that are better suited to the characteristics of their target series and decision objectives.

5 Conclusion

We introduced FinVerse, a large-scale financial time-series benchmark designed to evaluate forecasting models according to the economic semantics. Moving beyond existing benchmarks that apply uniform conventional forecasting metrics across heterogeneous time series, FinVerse evaluates forecasts through three complementary aspects—point-wise forecasting, cross-sectional ranking, and portfolio backtesting—using evaluation criteria tailored to the economic roles of the target series. Our evaluation demonstrates that performance on generic time-series benchmarks does not neces-

Table 4: Metric-wise model performance across point-wise, cross-sectional, and portfolio-oriented metrics using averaged raw metric values. When we calculate the Sharpe ratio, we do not subtract the risk-free rate and instead use raw portfolio returns. Refer to Appendix C.5 for details on why the reported Sharpe ratio does not equal the ratio of the aggregated return and volatility values.

Model	HR↑	ΔHR↑	Δ^2HR↑	RelMAE_{rel} ↓	RelMAE_{abs} ↓	RelMAE_{ori} ↓	IC↑	Return↑	Sharpe↑	MDD↑	Vol.↓
Reverso Small	**0.664**	0.545	0.484	0.851	0.849	**0.981**	0.123	0.171	0.697	-0.349	0.470
Chronos-2 Synth	0.647	0.557	0.500	0.879	0.867	1.020	**0.147**	0.254	0.701	-0.349	0.705
TimesFM 2.5 200M	0.648	0.568	0.512	0.855	0.849	1.002	0.130	0.155	0.686	-0.373	0.478
Chronos-2	0.652	0.562	0.499	0.843	0.848	1.012	0.054	0.208	0.660	-0.365	0.650
TiRex 1.1	0.663	**0.576**	**0.520**	**0.835**	**0.833**	0.992	0.135	0.174	0.674	-0.369	0.596
Chronos-Bolt Base	0.638	0.560	0.494	0.879	0.878	1.049	0.073	0.165	0.706	-0.365	0.426
Chronos-Bolt Small	0.638	0.554	0.493	0.886	0.882	1.063	0.072	0.210	0.680	-0.382	0.623
T0 Alpha	0.624	0.571	0.511	0.934	0.930	1.207	0.119	0.251	0.669	-0.365	0.766
TiRex 2 ZS	0.649	0.571	0.517	0.838	0.837	1.007	0.133	0.166	0.640	-0.382	0.591
TiRex 2 Pretrain	0.651	0.573	0.519	0.836	**0.833**	1.001	0.134	0.180	0.631	-0.383	0.625
Toto 2.0 2.5B	0.651	0.546	0.488	0.961	0.936	1.139	0.098	0.205	0.679	-0.372	0.665
VisionTS	0.634	0.567	0.506	0.900	0.907	1.066	0.087	0.214	0.680	-0.401	0.669
TimesFM 1.0 200M	0.624	0.559	0.502	0.947	0.942	1.159	0.120	0.227	0.679	-0.373	0.637
TimesFM 2.0 500M	0.536	0.431	0.376	1.347	1.303	1.341	0.113	0.140	0.665	**-0.341**	0.424
PatchFM	**0.664**	0.565	0.503	0.871	0.859	1.022	0.116	0.168	0.646	-0.372	0.518
Kairos-50M	0.649	0.547	0.481	0.972	0.956	1.109	0.088	0.203	0.705	-0.358	0.650
Toto 2.0 4M	0.652	0.555	0.493	0.919	0.912	1.084	0.097	0.215	0.678	-0.367	0.678
Toto 2.0 1B	0.651	0.552	0.493	0.963	0.936	1.140	0.098	0.191	0.670	-0.372	0.566
Granite FlowState R1	0.634	0.565	0.490	1.022	0.990	1.240	0.042	0.214	0.702	-0.382	0.599
TempoPFN	0.657	0.513	0.443	0.978	0.943	1.066	0.056	0.198	0.642	-0.374	0.706
Kairos-10M	0.644	0.503	0.439	1.172	1.127	1.320	0.070	0.231	**0.708**	-0.374	0.710
Toto 2.0 22M	0.650	0.553	0.490	0.954	0.935	1.131	0.099	0.213	0.655	-0.367	0.564
FlowState	0.632	0.553	0.480	1.012	0.981	1.207	0.024	0.244	0.687	-0.379	0.739
Toto 2.0 313M	0.652	0.556	0.492	0.952	0.933	1.120	0.091	0.189	0.671	-0.370	0.596
Super Linear	0.599	0.548	0.486	0.952	0.941	1.116	0.109	0.255	0.670	-0.401	0.773
Granite PatchTST-FM R1	0.536	0.440	0.381	1.423	1.400	1.473	0.096	0.245	0.635	-0.385	0.732
PatchTST-FM R1	0.564	0.448	0.388	1.303	1.260	1.334	0.091	0.249	0.634	-0.387	0.755
Timer-S1	0.610	0.541	0.482	0.968	0.953	1.166	0.120	0.149	0.634	-0.397	0.608
Toto Open Base 1.0	0.596	0.504	0.442	0.995	0.975	1.117	0.079	0.232	0.638	-0.377	0.730
Moirai 1.1-R Large	0.531	0.447	0.394	1.247	1.180	1.406	0.096	0.160	0.626	-0.356	**0.373**
Moirai 1.1-R Base	0.536	0.450	0.393	1.207	1.166	1.348	0.087	0.144	0.624	-0.360	0.465
Kairos-23M	0.636	0.521	0.456	1.071	1.065	1.272	0.056	0.230	0.650	-0.374	0.749
Moirai 1.1-R Small	0.544	0.454	0.401	1.191	1.145	1.275	0.105	0.156	0.652	-0.380	0.461
Sundial Base 128M	0.607	0.546	0.487	0.986	0.972	1.200	0.110	0.225	0.638	-0.401	0.676
Moirai 2.0-R Small	0.573	0.496	0.444	1.137	1.099	1.369	0.081	0.189	0.621	-0.370	0.629
TTM R3	0.531	0.439	0.390	1.279	1.268	1.375	0.107	0.194	0.658	-0.400	0.648
TTM R1	0.590	0.449	0.382	2.167	1.962	1.658	0.068	**0.260**	0.664	-0.394	0.690
CleanTS-65M	0.567	0.486	0.420	1.253	1.237	1.415	0.090	0.242	0.644	-0.396	0.755
TTM R2	0.603	0.460	0.394	1.734	1.591	1.358	0.086	0.198	0.651	-0.407	0.643
YingLong 50M	0.421	0.272	0.230	1.718	1.644	1.782	0.054	0.234	0.645	-0.397	0.712
YingLong 300M	0.419	0.269	0.227	1.724	1.654	1.779	0.053	0.254	0.642	-0.398	0.755
YingLong 110M	0.419	0.269	0.229	1.718	1.647	1.774	0.046	0.249	0.640	-0.398	0.753
YingLong 6M	0.421	0.271	0.229	1.702	1.622	1.772	0.052	0.241	0.631	-0.402	0.717

sarily translate to performance under financially informed evaluation. Furthermore, by showing that relative model performance varies substantially across both target financial time series and evaluation metrics, FinVerse provides a more informative basis for model selection according to the economic characteristics of the target series and the intended forecasting objectives. For example, Reverso Small ranks first on FinVerse despite performing less strongly on the generic time-series benchmark, whereas Chronos-2 Synth ranks second on FinVerse while maintaining relatively strong performance across both financial and generic time-series evaluations. We hope that FinVerse provides a foundation for the development and evaluation of future time-series foundation models that better account for the distinct characteristics and decision-relevant properties of financial time series.

Limitation & Future Works. FinVerse has several limitations. First, our portfolio backtesting setup largely follows the protocol of Pederzoli et al. (2026) for a simple evaluation setting. However, this setup does not fully reflect practical considerations such as transaction costs and turnover. Second, ensuring strict point-in-time (PIT) consistency remains challenging. In particular, the timestamp associated with an observation may differ from its actual release time. Finally, the current benchmark has substantially greater coverage of the U.S. market than other regions. Addressing these limitations remains important future work. Additionally, we plan to extend FinVerse by incorporating multimodal financial information through methodologies such as Lee et al. (2026).

AI USE STATEMENT

In this work, we used generative AI tools to assist with many parts of the research process. Specifically, we used such tools for implementing methods, cleaning and reformatting data, developing theoretical models or conceptual frameworks, providing feedback on the evaluation methodology, assisting with translation, and drafting interpretations of results. We note that these tools were used only as an aid and only to produce drafts: every AI-assisted output was manually reviewed and verified by the authors before being used. We have not used generative AI tools to generate synthetic datasets, to formulate mathematical claims, to provide ingredients for or assist in the writing of proofs, or to propose or refine hypotheses, and supporting qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure, including editing code, drafting and editing parts of the manuscript, creating result figures and tables, and summarizing related literature. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

FinVerse is constructed primarily from publicly accessible financial and economic data sources, such as FRED and the U.S. EIA. During data collection, we made reasonable efforts to review and comply with the applicable licenses, terms of use, and redistribution conditions of the respective data sources. Based on our review, we believe that the data included in the public release can be appropriately used and distributed for research purposes. Nevertheless, given the scale and diversity of the collected data, if any licensing, redistribution, or other compliance issue is identified after release, we will promptly address it by removing, replacing, or otherwise modifying the affected data as appropriate. We are committed to maintaining the dataset in accordance with applicable data-use requirements and to addressing any concerns raised by data providers.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the relevant code in the Supplementary Material. Due to the large size of the dataset, we are unable to include the dataset itself in the submission materials. The dataset will be made publicly available upon acceptance.

REFERENCES

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. URL https://arxiv.org/abs/2410.10393.

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. URL https://arxiv.org/abs/2403.07815.

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. URL https://arxiv.org/abs/2510.15821.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning, 2025. URL https://arxiv.org/abs/2505.23719.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters, 2024. URL https://arxiv.org/abs/2408.17253.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models, 2025. URL https://arxiv.org/abs/2505.14766.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024. URL https://arxiv.org/abs/2310.10688.

EINK. Cleants-65m. https://huggingface.co/EINK/CleanTS-65M, 2026. Hugging Face model.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series, 2024. URL https://arxiv.org/abs/2401.03955.

Kun Feng, Shaocheng Lan, Yuchen Fang, Wenchao He, Sihan Lu, Shuqi Gu, Lintao Ma, Xingyu Lu, and Kan Ren. Kairos: Toward adaptive and parameter-efficient time series foundation models, 2025. URL https://arxiv.org/abs/2509.25826.

Xinghong Fu, Yanhong Li, Georgios Papaioannou, and Yoon Kim. Reverso: Efficient time series foundation models for zero-shot forecasting, 2026. URL https://arxiv.org/abs/2602.17634.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive, 2021. URL https://arxiv.org/abs/2105.06643.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models, 2024. URL https://arxiv.org/abs/2402.03885.

Lars Graf, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. FlowState: Sampling-rate-equivariant time-series forecasting, 2025. URL https://arxiv.org/abs/2508.05287. ICML 2026, PMLR 306.

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. URL https://arxiv.org/abs/2502.18834.

Jaehoon Lee, Suhwan Park, Taeyoon Lim, Seunghan Lee, Jun Seo, Dongwan Kang, Hwanil Choi, Minjae Kim, Sungdong Yoo, Soonyoung Lee, Yongjae Lee, and Wonbin Ahn. Fintexts: Financial text-paired time-series dataset via semantic-based and multi-level pairing, 2026. URL https://arxiv.org/abs/2603.02702.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models, 2024. URL https://arxiv.org/abs/2402.02368.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting, 2025a. URL https://arxiv.org/abs/2410.04803.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models, 2025b. URL https://arxiv.org/abs/2502.00816.

Vladyslav Moroshan, Julien Siems, Arber Zela, Timur Carstensen, and Frank Hutter. TempoPFN: Synthetic pre-training of linear rnns for zero-shot time series forecasting, 2025. URL https://arxiv.org/abs/2510.25502.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Liran Nochumsohn, Raz Marshanski, Hedi Zisling, and Omri Azencot. Super-linear: A lightweight pretrained mixture of linear experts for time series forecasting. *arXiv preprint arXiv:2509.15105*, 2025.

Sara Pederzoli, Marta Santacroce, Francesco Guerra, Marco Bergianti, and Francesco Del Buono. Finbench: A benchmarking framework for stock market prediction and portfolio allocation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 9661–9672, 2026. doi: 10.1145/3770855.3817595.

Eghbal Rahimikia, Hao Ni, and Weiguan Wang. Re(visiting) time series foundation models in finance, 2025. URL https://arxiv.org/abs/2511.18578.

Adir Saly-Kaufmann, Kieran Wood, Jan Peter-Calliess, and Stefan Zohren. Deep learning for financial time series: A large-scale benchmark of risk-adjusted performance, 2026. URL https://arxiv.org/abs/2603.01820.

Ege Onur Taga, M. Emrullah Ildiz, and Samet Oymak. TimePFN: Effective multivariate time series forecasting with synthetic data, 2025. URL https://arxiv.org/abs/2502.16294.

The Forecasting Company. t0: A time-series forecasting foundation model, 2026. URL https://huggingface.co/theforecastingcompany/t0-alpha.

Samy-Melwan Vilhes. Patchfm: A foundation model for univariate time series forecasting. https://huggingface.co/vilhess/PatchFM, 2026. Hugging Face model.

Xue Wang, Tian Zhou, Jinyang Gao, Bolin Ding, and Jingren Zhou. Output scaling: YingLong-delayed chain of thought in a large pretrained time series forecasting model, 2025. URL https://arxiv.org/abs/2506.11029.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

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. URL https://arxiv.org/abs/2012.07436.

A Detailed Data Coverage

This appendix provides the detailed coverage table for the released FinVerse data. Table 5 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.

Table 5: Coverage of the FinVerse data by detail category.

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	33,033
Inter-country	Commodity	Commodity ETF Broad	Business-daily	✓	2	10,433	3,174
Inter-country	Commodity	Commodity ETF Single	Business-daily	✓	5	25,154	7,935
Inter-country	Commodity	Commodity ETF Thematic	Business-daily	✓	2	8,149	3,174
Inter-country	Commodity	Future Front	Business-daily	✓	27	173,716	42,258
Inter-country	Commodity	Inventory Stock	Weekly	✓	7	10,985	2,303
Inter-country	Commodity	Spot	Business-daily	✓	3	28,307	4,741
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	140,483

Continued on next page

Scope	Category	Detail Category	Frequency	Eval.	Series	Points	Eval. Points
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	28,422
Inter-country	Market Index	Regional Index	Business-daily	✓	2	10,724	3,184
Country	Fixed Income	Bond Total Return Index	Business-daily	✓	10	7,830	7,750
Country	Fixed Income	Breakeven	Business-daily	✓	2	12,172	3,166
Country	Fixed Income	Breakeven	Monthly	✓	2	455	150
Country	Fixed Income	Corporate Yield	Business-daily	✓	2	21,823	3,160
Country	Fixed Income	Corporate Yield	Monthly	✓	2	2,574	150
Country	Fixed Income	Credit Spread	Business-daily	✓	10	7,830	7,749
Country	Fixed Income	Forward Inflation	Business-daily	✓	1	6,086	1,583
Country	Fixed Income	Inflation Expectation	Monthly	✓	7	3,949	530
Country	Fixed Income	Money Market	Business-daily	✓	8	20,958	12,317
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	6,328
Country	Fixed Income	Term Spread	Business-daily	✓	2	24,587	3,166
Country	Fixed Income	Treasury Bill	Business-daily	×	4	60,333	0
Country	Fixed Income	Treasury Yield	Business-daily	✓	11	154,357	17,402
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	12,879
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	14,310
Country	Macro Indicators	Activity	Monthly	✓	10	5,621	743
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	6
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	12
Country	Macro Indicators	Freight	Monthly	✓	1	313	73
Country	Macro Indicators	Gdp	Quarterly	✓	2	634	50
Country	Macro Indicators	Housing	Monthly	✓	4	2,367	231
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	673
Country	Macro Indicators	International Investment Position	Quarterly	✓	5	400	120
Country	Macro Indicators	International Investment Position	Annual	✓	1	50	6
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	1,041
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	374
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

Continued on next page

Scope	Category	Detail Category	Frequency	Eval.	Series	Points	Eval. Points
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	4,497,246
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	7,963,766
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	48,017
Individual	Fundamentals	Cash and Cash Equivalents	Quarterly	✓	4,642	206,110	81,457
Individual	Fundamentals	EPS Basic	Quarterly	✓	4,185	169,330	62,002
Individual	Fundamentals	EPS Diluted	Quarterly	✓	4,166	166,806	60,690
Individual	Fundamentals	Long-Term Debt	Quarterly	✓	3,249	100,936	40,488
Individual	Fundamentals	Net Income	Quarterly	✓	4,357	182,999	65,622
Individual	Fundamentals	Operating Cash Flow	Quarterly	✓	4,359	185,546	67,429
Individual	Fundamentals	Operating Income	Quarterly	✓	3,707	151,714	54,499
Individual	Fundamentals	R And D Expense	Quarterly	×	1,851	64,374	0
Individual	Fundamentals	Revenue	Quarterly	✓	3,677	99,475	49,241
Individual	Fundamentals	Shares Diluted	Quarterly	×	4,173	168,275	0
Individual	Fundamentals	Stockholders Equity	Quarterly	✓	4,697	210,564	90,538
Individual	Fundamentals	Total Assets	Quarterly	✓	4,699	196,491	90,137
Individual	Fundamentals	Total Liabilities	Quarterly	✓	4,238	160,730	77,290
Individual	News Sentiment	Sentiment Tone	Daily	×	11	37,488	0
Individual	News Sentiment	Sentiment Volume	Daily	×	11	37,488	0

B Foundation Model List

The evaluated models include TimesFM 1.0 200M, TimesFM 2.0 500M, and TimesFM 2.5 200M (Das et al., 2024); Chronos-2, Chronos-2 Synth, Chronos-Bolt Small, and Chronos-Bolt Base (Ansari et al., 2024; 2025); Moirai 1.1-R Small, Moirai 1.1-R Base, Moirai 1.1-R Large, and Moirai 2.0-R Small (Woo et al., 2024); Timer-S1 (Liu et al., 2024; 2025a); TempoPFN (Moroshan et al., 2025); 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 (Cohen et al., 2025); TiRex 1.1, TiRex 2 Pretrain, and TiRex 2 ZS (Auer et al., 2025); VisionTS (Chen et al., 2024); Sundial Base 128M (Liu et al., 2025b); YingLong 6M, YingLong 50M, YingLong 110M, and YingLong 300M (Wang et al., 2025); Kairos-10M, Kairos-23M, and Kairos-50M (Feng et al., 2025); Reverso Small (Fu et al., 2026); TTM R1, TTM R2, and TTM R3 (Ekambaram et al., 2024); PatchTST-FM R1 and Granite PatchTST-FM R1 (Nie et al., 2023); FlowState and Granite FlowState R1 (Graf et al., 2025); PatchFM (Vilhes, 2026); T0 Alpha (The Forecasting Company, 2026); Super Linear (Nochumsohn et al., 2025); and CleanTS-65M (EINK, 2026).

C Metric Definition

This section provides the detailed mathematical formulations of the evaluation metrics used in FinVerse. 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. Throughout, a forecast origin $o \in \mathcal{O}$ denotes the last time step observed by the model, so that the forecast window covers $o+1, \ldots, o+H$ and y_o is the most recent observation available when the forecast is made.

C.1 Hit Ratio

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

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

To capture higher-order directional dynamics, we further define

$$\Delta r_o^{(p,h)} = r_o^{(p,h)} - r_{o-1}^{(p,h)}, \qquad \Delta \hat{r}_o^{(p,h)} = \hat{r}_o^{(p,h)} - r_{o-1}^{(p,h)}, \tag{2}$$

$$\Delta^2 r_o^{(p,h)} = \Delta r_o^{(p,h)} - \Delta r_{o-1}^{(p,h)}, \qquad \Delta^2 \hat{r}_o^{(p,h)} = \Delta \hat{r}_o^{(p,h)} - \Delta r_{o-1}^{(p,h)}. \tag{3}$$

Let $\Delta^0 r_o^{(p,h)} = r_o^{(p,h)}$ and $\Delta^0 \hat{r}_o^{(p,h)} = \hat{r}_o^{(p,h)}$. We define the k-th order hit ratio as

$$\Delta^k \mathrm{HR}^{(p,h)} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \prod_{j=0}^{k} \mathbb{I}\Big[\mathrm{sign}\Big(\Delta^j \hat{r}_o^{(p,h)}\Big) = \mathrm{sign}\Big(\Delta^j r_o^{(p,h)}\Big)\Big], \qquad k \in \{0,1,2\}. \tag{4}$$

Here, $k = 0$, $k = 1$, and $k = 2$ correspond to HR, ΔHR, and Δ^2HR, respectively. 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, $p = 1\mathrm{y}$ computes the year-over-year change relative to the observation one year earlier, while $h = 1\mathrm{y}$ evaluates the prediction one year ahead. Different financial series use different (p, h) combinations according to their sampling frequency and economic interpretation. 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.

C.2 Relative Mean Absolute Error

FinVerse uses RelMAE, defined as the mean absolute error of the forecast divided by the mean absolute error of a naive forecast computed over the same evaluation cells. RelMAE_{rel} evaluates relative-change forecasts against a zero-change baseline. In other words, the baseline assumes no relative change over the forecast horizon:

$$\mathrm{RelMAE}_{rel}^{(p,h)} = \frac{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \left| \hat{r}_o^{(p,h)} - r_o^{(p,h)} \right|}{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \left| r_o^{(p,h)} \right|}. \tag{5}$$

RelMAE_{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}:

$$\mathrm{RelMAE}_{abs}^{(p,h)} = \frac{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |\hat{y}_{o+h} - y_{o+h}|}{\frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |y_{o+h} - y_{o+h-p}|}. \tag{6}$$

Finally, RelMAE_{ori} evaluates raw-value forecasts against the conventional origin-value persistence baseline, where the most recently observed value is repeated for all future horizons. This baseline is the persistence forecast from the origin, and the metric averages over every horizon $h = 1, \ldots, H$ rather than over a selected (p, h) grid:

$$\mathrm{RelMAE}_{ori} = \frac{\frac{1}{|\mathcal{O}|H} \sum_{o \in \mathcal{O}} \sum_{h=1}^{H} |\hat{y}_{o+h} - y_{o+h}|}{\frac{1}{|\mathcal{O}|H} \sum_{o \in \mathcal{O}} \sum_{h=1}^{H} |y_o - y_{o+h}|}. \tag{7}$$

C.3 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:

$$\mathrm{IC}_o^{(p,h)} = \rho_{\mathrm{Spearman}} \left(\left\{ \hat{r}_{i,o}^{(p,h)} \right\}_{i \in \mathcal{S}_o}, \left\{ r_{i,o}^{(p,h)} \right\}_{i \in \mathcal{S}_o} \right). \tag{8}$$

C.4 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 decile, subject to a floor of ten instruments: $|\mathcal{P}_o^{(h)}| = \min(|\mathcal{S}_o|, \max(10, \lceil 0.1\,|\mathcal{S}_o| \rceil))$. 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

$$R_o^{(h)} = \frac{1}{|\mathcal{P}_o^{(h)}|} \sum_{i \in \mathcal{P}_o^{(h)}} r_{i,o}^{(h,h)}. \tag{9}$$

Let $\mathcal{A}_h = \{0, \ldots, h-1\}$ denote the offset set, and let $o_{s,1}, \ldots, o_{s,K_s}$ 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

$$\mathrm{NAV}_{s,k}^{(h)} = \prod_{j=1}^{k} \left(1 + R_{o_{s,j}}^{(h)}\right). \tag{10}$$

Because $\mathrm{NAV}_{s,k}^{(h)}$ is observed only at rebalancing dates, it cannot reveal losses that occur and recover within a holding period. FinVerse therefore also tracks the portfolio value at every frequency step. Within the j-th holding period, the selected assets are bought equal-weight at $y_{i,o_{s,j}}$, the most recent observation available at that origin, and held, so the portfolio value after m steps is

$$V_{s,(j,m)}^{(h)} = \mathrm{NAV}_{s,j-1}^{(h)} \cdot \frac{1}{|\mathcal{P}_{o_{s,j}}^{(h)}|} \sum_{i \in \mathcal{P}_{o_{s,j}}^{(h)}} \frac{y_{i,\,o_{s,j}+m}}{y_{i,\,o_{s,j}}}, \qquad m = 1, \ldots, h, \tag{11}$$

with $\mathrm{NAV}_{s,0}^{(h)} = 1$. Because the holding is buy-and-hold, the end of each period coincides with the rebalance-level value, $V_{s,(j,h)}^{(h)} = \mathrm{NAV}_{s,j}^{(h)}$. Concatenating these values over $j = 1, \ldots, K_s$ yields the step-level value path $V_{s,1}^{(h)}, \ldots, V_{s,T_s}^{(h)}$ with $T_s = K_s h$, and the corresponding step returns are

$$R_{s,u}^{\mathrm{step}} = \frac{V_{s,u}^{(h)}}{V_{s,u-1}^{(h)}} - 1, \qquad V_{s,0}^{(h)} = 1. \tag{12}$$

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

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

Here, h denotes the holding horizon. Let P_f denote the number of frequency steps per year, with $P_f \in \{252, 365, 52, 12, 4, 1\}$ for business-daily, daily, weekly, monthly, quarterly, and annual series, respectively. The number of rebalancing trades per year is then $A_h = P_f/h$. For each offset s, the annualized return is computed from the rebalance-level returns,

$$\mathrm{AnnualReturn}_s^{(h)} = \left(\prod_{j=1}^{K_s} \left(1 + R_{o_{s,j}}^{(h)}\right) \right)^{\frac{A_h}{K_s}} - 1, \tag{14}$$

whereas the risk metrics are computed from the step-level path, so that losses occurring inside a holding period are reflected:

$$\mathrm{AnnualVolatility}_s^{(h)} = \mathrm{Std}\big(R_{s,u}^{\mathrm{step}}\big)\sqrt{P_f}, \qquad \mathrm{Sharpe}_s^{(h)} = \frac{\mathbb{E}\big[R_{s,u}^{\mathrm{step}}\big]}{\mathrm{Std}\big(R_{s,u}^{\mathrm{step}}\big)}\sqrt{P_f}, \tag{15}$$

$$\mathrm{MDD}_s^{(h)} = \min_u \left(\frac{V_{s,u}^{(h)}}{\max\left(1,\ \max_{v \le u} V_{s,v}^{(h)}\right)} - 1 \right). \tag{16}$$

The peak in the drawdown denominator includes the initial value of one, so a decline that starts before the first rebalancing date is also counted.

C.5 Return, Volatility, and Sharpe Ratio Aggregation

The Sharpe column is not the ratio of the Return and Vol. columns for three reasons. First, even within a single backtest, annualized return and the numerator of the Sharpe ratio are computed differently: annualized return is based on the compounded growth of the portfolio net asset value, whereas the Sharpe ratio uses the annualized arithmetic mean of one-step portfolio returns. Second, these statistics are computed separately for each holding-period offset and then averaged across offsets. As such, the average Sharpe ratio need not equal the ratio of the average return to the average volatility. Third, each column in the table is independently aggregated using a mean across all evaluated detail category–frequency units and holding horizons, so the ratio of the aggregated Return and Vol. columns need not equal the aggregated Sharpe ratio.

C.6 Assumptions of Portfolio Evaluation

1. **Trading frictions.** No transaction costs, bid–ask spreads, or market impact are charged. Every rebalancing date replaces the entire portfolio with the newly selected instruments, so turnover is effectively complete at each rebalance and is higher for the shorter holding horizons.
2. **Investable set, listing, and liquidity.** At each rebalancing date the portfolio is formed from the instruments that have an observed price at that date, and the availability of a price is taken as evidence that the instrument was listed and tradable then. An instrument whose price series ends simply leaves the cross-section from that date onward.
3. **Corporate actions.** Dividends and other cash distributions are not added back, so equity, ETF, and index returns are price returns rather than total returns, and the reported portfolio performance is correspondingly conservative for dividend-paying assets.
4. **Portfolio construction.** Positions are entered at the last observed price before the rebalancing origin, weighted equally at entry, and held without intra-period rebalancing, so weights drift with prices inside a holding period. The strategy is long-only, unlevered, and fully invested in the selected instruments. To avoid dependence on a single calendar start date, the backtest is repeated for every offset within the holding horizon and the resulting statistics are averaged.
5. **Trading calendars and the step-level path.** The step-level value path is evaluated on the union of the one-step date grids of the instruments in the group, and an instrument without an observation at a grid step carries its most recent price forward.
6. **Non-equity instruments.** The same rule is applied to every group for which a cross-sectional portfolio is meaningful, and each group is traded as its published price series. Foreign-exchange pairs use the quoted reference rate as published, so a position is a long exposure to that quoted rate and interest-rate differentials, i.e., carry, are not modeled. Commodity futures use front-month contract prices, so roll returns between contracts are not modeled. Commodity spot series are reference prices, which ignores storage and financing costs and makes them the least directly investable group. Country equity indices are price indices, so, as with equities and ETFs, dividends are excluded. Crypto pairs are daily closing prices on a seven-day calendar and are treated like any other price series. These groups therefore measure the value of the forecast signal for ranking instruments, not the return of an implementable trading strategy in each market.

D Metric Assignment by Data Category

This section shows the metric assignment used by FinVerse. Table 6 maps each compact variant ID to the corresponding (p, h) combinations. Tables 7–16 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.

Table 6: Mapping from compact variant IDs to (p, h) combinations. $1w$, $1m$, $1q$, $1h$, and $1y$ denote one week, one month, one quarter, one half-year, and one year, respectively.

(p, h) Variant	(p, h) combinations
D5	$(1w, 1w), (1m, 1m), (1q, 1q), (1h, 1h), (1y, 1y)$
D4	$(1m, 1m), (1q, 1q), (1h, 1h), (1y, 1y)$
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

Table 7: Metric assignments and variant IDs for inter-country Foreign Exchange categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
FX Pair	Business-daily	33,033	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5

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

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Future Front	Business-daily	42,258	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5
Spot	Business-daily	4,741	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Spot	Monthly	975	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D4
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D4
			Portfolio	return, vol, Sharpe, MDD	D4
Inventory Stock	Weekly	2,303	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Commodity ETF Single	Business-daily	7,935	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Commodity ETF Broad	Business-daily	3,174	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Commodity ETF Thematic	Business-daily	3,174	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A

Table 9: Metric assignments and variant IDs for inter-country Crypto categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Crypto Pair Close (Top 100)	Daily	140,483	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5

Table 10: Metric assignments and variant IDs for inter-country Market Index categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Country Index	Business-daily	28,422	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5
Regional Index	Business-daily	3,184	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A

Table 11: Metric assignments and variant IDs for country-level Fixed Income categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Treasury Yield	Business-daily	17,402	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D5
Credit Spread	Business-daily	7,749	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Bond Total Return Index	Business-daily	7,750	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Money Market	Business-daily	12,317	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
TIPS Yield	Business-daily	6,328	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Breakeven	Business-daily	3,166	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Corporate Yield	Business-daily	3,160	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Term Spread	Business-daily	3,166	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Forward Inflation	Business-daily	1,583	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
Inflation Expectation	Monthly	530	Point-wise	HR, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Breakeven	Monthly	150	Point-wise	HR, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Corporate Yield	Monthly	150	Point-wise	HR, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Mortgage Rate	Weekly	812	Point-wise	HR, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A

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

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Implied Volatility	Business-daily	12,879	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
US Equity Index	Business-daily	14,310	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A

Table 13: Metric assignments and variant IDs for country-level Macro Indicators categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Inflation	Monthly	673	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Labor	Monthly	1,041	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
GDP	Quarterly	50	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Housing Price	Monthly	747	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Productivity	Quarterly	72	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Money	Monthly	374	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Industrial Production by Industry	Monthly	2,100	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Labor by Industry	Monthly	1,200	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Activity	Monthly	743	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Capacity Utilization by Industry	Monthly	825	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Inventory	Monthly	518	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Spending	Monthly	450	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Energy Activity	Monthly	397	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Sentiment	Monthly	200	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Housing	Monthly	231	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Income	Monthly	225	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A

Continued on next page

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Credit	Monthly	148	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Regional Fed	Monthly	152	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Saving	Monthly	75	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Stress	Monthly	75	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Freight	Monthly	73	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Portfolio Flows	Quarterly	168	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Current Account	Quarterly	48	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
TIC Flows	Quarterly	95	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Financial Flows	Quarterly	24	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
External	Monthly	444	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
External	Quarterly	75	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Banking	Monthly	150	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Fiscal	Monthly	75	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	M10
			Point-wise	ΔHR, Δ^2HR	M4
			Point-wise	RelMAE_{ori}	N/A
Fiscal	Quarterly	24	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Fiscal	Annual	12	Point-wise	HR, ΔHR, Δ^2HR, RelMAE_{rel}, RelMAE_{abs}	A1
			Point-wise	RelMAE_{ori}	N/A
International Investment Position	Quarterly	120	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
International Investment Position	Annual	6	Point-wise	HR, ΔHR, Δ^2HR, RelMAE_{rel}, RelMAE_{abs}	A1
			Point-wise	RelMAE_{ori}	N/A

Continued on next page

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
FDI Flows	Quarterly	24	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
FDI Flows	Annual	6	Point-wise	HR, ΔHR, Δ^2HR, RelMAE_{rel}, RelMAE_{abs}	A1
			Point-wise	RelMAE_{ori}	N/A

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

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
ETF OHLCV Close	Business-daily	4,497,246	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5

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

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Equity OHLCV Close	Business-daily	7,963,766	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	D5
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	D5
			Portfolio	return, vol, Sharpe, MDD	D5

Table 16: Metric assignments and variant IDs for individual-series Fundamentals categories.

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Capital Expenditures	Quarterly	48,017	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Cash and Cash Equivalents	Quarterly	81,457	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
EPS Basic	Quarterly	62,002	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	Q6
EPS Diluted	Quarterly	60,690	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	Q6
Long-Term Debt	Quarterly	40,488	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Net Income	Quarterly	65,622	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	Q6

Continued on next page

Detail Category	Frequency	Eval. Points	Task	Metrics	(p, h) Variant
Operating Cash Flow	Quarterly	67,429	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Operating Income	Quarterly	54,499	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	Q6
Revenue	Quarterly	49,241	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
			Cross-sectional	IC	Q6
Stockholders Equity	Quarterly	90,538	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Total Assets	Quarterly	90,137	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A
Total Liabilities	Quarterly	77,290	Point-wise	HR, RelMAE_{rel}, RelMAE_{abs}	Q6
			Point-wise	ΔHR, Δ^2HR	Q3
			Point-wise	RelMAE_{ori}	N/A

E Additional Results

E.1 Model Size and Performance

Figure 5 examines whether overall FinVerse performance can be explained by model scale. Reverso Small achieves the best overall rank, followed by Chronos-2 Synth, TimesFM 2.5 200M, Chronos-2, and TiRex 1.1. Larger checkpoints do not consistently dominate, with several Toto, Timer, and YingLong variants ranking behind smaller or medium-sized models. These results suggest that financial decision-oriented performance cannot be explained by parameter count alone.

E.2 Disentangling the Effects of Evaluation Data and Metrics

Figure 2 compares model rankings between GIFT-Eval and FinVerse, where both the target time series and the evaluation metrics differ. To disentangle these two sources of variation, Figure 6 instead evaluates models on the same FinVerse target series using both the FinVerse evaluation metrics and MASE, thereby isolating the effect of the evaluation criteria. For every model and every detail category–frequency unit, Figure 6 plots the rank under MASE against the rank under each assigned metric group. For each metric group, models are first ranked separately for each metric within a detail category–frequency unit, and these metric-wise ranks are then aggregated using the geometric mean to obtain the group-level rank. MASE is nearly interchangeable with the RelMAE family (Pearson $r = 0.86$; Spearman $\rho = 0.86$; Kendall $\tau = 0.70$), which is expected since both divide the same absolute errors by a naive-forecast scale and differ only in the choice of naive forecast. Agreement drops to $r = 0.65$ ($\rho = 0.64$, $\tau = 0.47$) for the hit-ratio family, which evaluates the same forecasts in terms of direction rather than magnitude, to $r = 0.48$ ($\rho = 0.48$, $\tau = 0.34$) for cross-sectional IC, and to $r = 0.09$ ($\rho = 0.08$, $\tau = 0.06$) for the realized portfolio metrics, where the relationship all but disappears. A scaled point-forecast error is therefore a good proxy for another scaled point-forecast error but a poor proxy for decision-oriented evaluations. The ordering across the four panels also mirrors the aspect-level relationships in Figure 4, where the point-wise and portfolio views show the weakest relationship.

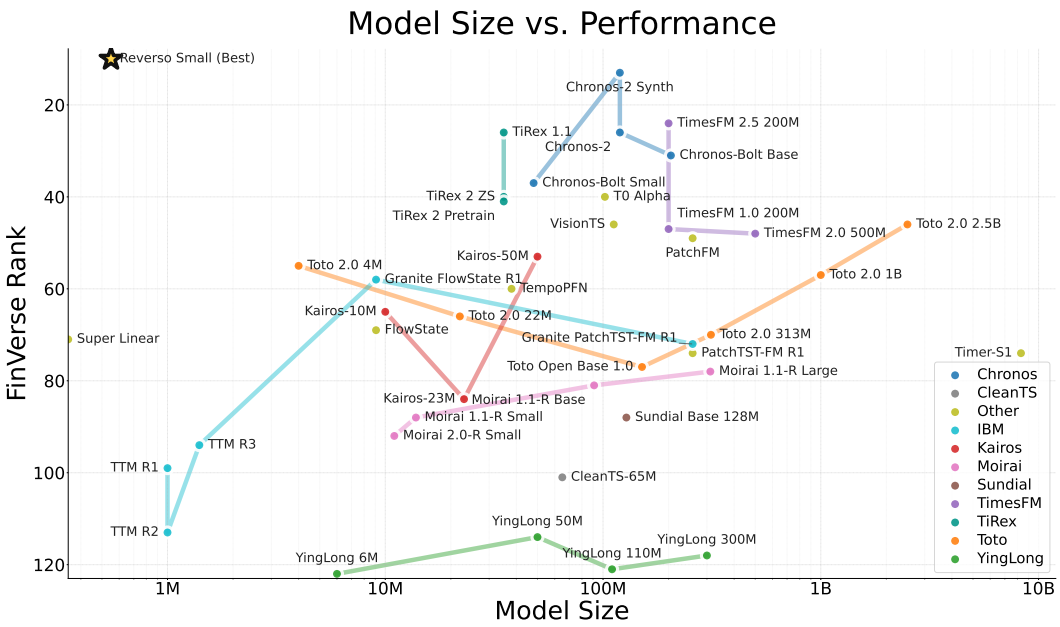

Figure 5: Relationship between model size and overall FinVerse rank. Lower rank indicates better performance.

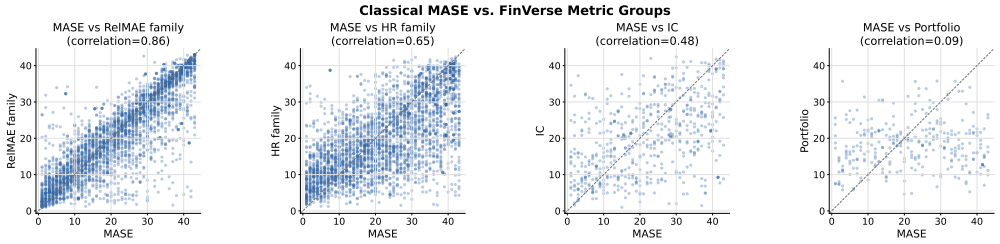

Figure 6: Rank under the classical $MASE$ versus metric group under FinVerse.

E.3 Alternative Naive Baselines

In this section, we evaluate five additional baselines alongside the persistence baseline used by RelMAE_{ori}. Let L denote the frequency-specific lookback window of Table 2. The seasonal baselines below use a calendar seasonal period m', namely $m' = 7$ for daily, 5 for business-daily, 52 for weekly, 12 for monthly, and 4 for quarterly data, with no seasonal cycle defined for annual data.

- **Persistence** (RelMAE_{ori}): repeats the last observation, $b_{o,h} = y_o$. It assumes no change over the horizon.
- **Seasonal naive** (RelMAE_{snaive}): repeats the most recent observation at the same seasonal position, $b_{o,h} = y_{o+h-m'\lceil h/m' \rceil}$. It assumes the last seasonal cycle repeats, and is not defined for annual series.
- **Seasonal mean** (RelMAE_{smean}): averages the observations at the same seasonal position over the last three years. It smooths the seasonal pattern instead of copying a single cycle, and like the seasonal naive it is not defined for annual series.
- **Random walk with drift** (RelMAE_{drift}): extrapolates the slope of the lookback window, $b_{o,h} = y_o + h\,(y_o - y_{o-L+1})/(L-1)$. It assumes the recent trend continues, i.e., a momentum view.
- **Lookback mean** (RelMAE_{mean}): predicts the mean of the last L observations for every horizon. It assumes immediate reversion to the recent average level.

- **AR(1)** (RelMAE_{ar1}): fits $y_t = c + \phi y_{t-1}$ by ordinary least squares on the last three years, or on the lookback window when that is longer, and iterates it forward from y_o, with ϕ clipped to $[-1, 1]$. It assumes gradual mean reversion at an estimated speed.

Table 17: Raw-value scaled error against alternative naive baselines. Avg. is the mean of the raw metric values, and K is the geometric mean of the model ranks for that metric.

Model	RelMAE_{ori}		RelMAE_{snaive}		RelMAE_{smean}		RelMAE_{drift}		RelMAE_{mean}		RelMAE_{ar1}	
	Avg.	K	Avg.	K	Avg.	K	Avg.	K	Avg.	K	Avg.	K
Reverso Small	**0.981**	**5.53**	**0.807**	**4.40**	**0.609**	4.71	0.805	**5.12**	**0.770**	**5.20**	**0.932**	**5.18**
Chronos-2 Synth	1.020	7.28	0.836	5.96	0.628	6.03	0.821	6.99	0.804	6.77	0.961	7.36
TimesFM 2.5 200M	1.002	8.26	0.820	6.36	0.620	6.27	0.826	7.86	0.780	7.39	0.953	7.99
Chronos-2	1.012	7.69	0.836	7.65	0.627	7.57	0.828	7.41	0.787	8.44	0.958	8.02
TiRex 1.1	0.992	6.35	0.823	4.85	0.617	**4.68**	**0.804**	6.30	0.778	5.81	0.937	6.03
Chronos-Bolt Base	1.049	11.77	0.863	11.71	0.653	11.88	0.863	11.69	0.815	12.18	0.998	12.29
Chronos-Bolt Small	1.063	11.38	0.865	11.19	0.654	11.22	0.874	11.85	0.824	12.00	1.005	12.42
T0 Alpha	1.207	19.17	0.962	18.10	0.713	17.81	1.009	19.13	0.907	18.31	1.135	19.71
TiRex 2 ZS	1.007	7.69	0.827	7.31	0.619	7.01	0.828	8.26	0.783	7.70	0.951	7.97
TiRex 2 Pretrain	1.001	7.03	0.824	6.23	0.616	5.88	0.823	7.15	0.781	6.85	0.946	6.99
Toto 2.0 2.5B	1.139	13.06	0.994	20.97	0.789	21.79	0.904	11.58	0.898	13.40	1.049	11.94
VisionTS	1.066	11.76	0.877	12.72	0.659	12.65	0.879	12.95	0.830	12.40	1.015	12.87
TimesFM 1.0 200M	1.159	19.79	0.917	18.10	0.686	17.41	0.964	19.36	0.875	18.22	1.088	19.73
TimesFM 2.0 500M	1.341	24.97	1.057	24.52	0.821	25.05	1.100	24.37	1.019	24.73	1.276	25.88
PatchFM	1.022	8.47	0.858	8.25	0.655	8.46	0.852	9.49	0.819	9.46	0.969	8.26
Kairos-50M	1.109	14.15	0.921	13.16	0.710	13.69	0.877	12.72	0.876	14.75	1.031	14.33
Toto 2.0 4M	1.084	13.61	0.919	15.02	0.704	15.50	0.877	13.28	0.853	13.79	1.007	12.17
Toto 2.0 1B	1.140	12.44	0.983	19.65	0.781	20.56	0.900	11.87	0.896	12.93	1.043	12.08
Granite FlowState R1	1.240	22.53	1.024	21.77	0.802	21.74	1.006	22.21	0.989	23.12	1.131	22.00
TempoPFN	1.066	11.16	0.863	9.86	0.661	9.97	0.871	10.97	0.839	11.32	1.016	11.21
Kairos-10M	1.320	23.67	1.054	22.10	0.840	22.29	1.046	22.52	1.048	23.80	1.235	23.15
Toto 2.0 22M	1.131	15.02	0.961	18.67	0.749	19.38	0.908	14.19	0.891	15.18	1.057	14.26
FlowState	1.207	19.33	0.998	18.92	0.780	18.76	0.982	19.78	0.969	19.78	1.088	19.09
Toto 2.0 313M	1.120	13.37	0.974	20.22	0.772	20.89	0.889	12.38	0.884	13.98	1.036	12.10
Super Linear	1.116	15.79	0.890	14.41	0.668	14.16	0.917	15.34	0.843	14.53	1.050	15.89
Granite PatchTST-FM R1	1.473	32.35	1.115	30.18	0.844	29.73	1.214	32.84	1.071	30.96	1.381	32.95
PatchTST-FM R1	1.334	27.33	1.040	25.16	0.776	24.99	1.081	27.60	1.001	26.53	1.248	27.51
Timer-S1	1.166	18.69	0.919	16.71	0.685	16.35	0.959	19.06	0.879	17.23	1.088	18.84
Toto Open Base 1.0	1.117	15.22	0.933	17.83	0.723	17.69	0.944	17.01	0.870	16.87	1.047	14.37
Moirai 1.1-R Large	1.406	25.51	1.096	25.82	0.819	25.72	1.175	26.51	1.047	26.39	1.334	26.06
Moirai 1.1-R Base	1.348	26.27	1.056	27.94	0.799	27.99	1.124	28.43	1.023	28.57	1.277	27.47
Kairos-23M	1.272	23.72	1.042	23.64	0.794	23.94	1.000	22.45	1.001	24.18	1.183	23.52
Moirai 1.1-R Small	1.275	23.47	1.003	24.62	0.749	24.57	1.059	25.14	0.962	24.89	1.199	25.54
Sundial Base 128M	1.200	21.63	0.943	19.19	0.703	18.86	0.991	21.82	0.899	19.38	1.120	21.63
Moirai 2.0-R Small	1.369	25.08	1.106	27.51	0.808	27.36	1.136	27.39	1.027	26.35	1.275	25.67
TTM R3	1.375	26.56	1.056	24.86	0.775	23.88	1.139	27.72	1.012	25.13	1.284	26.37
TTM R1	1.658	23.63	1.321	29.49	0.972	29.70	1.374	24.13	1.154	24.73	1.587	24.54
CleanTS-65M	1.415	30.46	1.081	28.63	0.797	28.20	1.168	30.96	1.045	29.80	1.308	30.81
TTM R2	1.358	21.14	1.087	22.99	0.829	23.11	1.086	20.48	0.984	21.29	1.269	21.73
YingLong 50M	1.782	39.51	1.312	38.21	0.970	37.91	1.469	39.72	1.283	38.91	1.655	39.75
YingLong 300M	1.779	39.18	1.309	37.86	0.968	37.56	1.467	39.38	1.281	38.58	1.652	39.41
YingLong 110M	1.774	38.74	1.303	37.39	0.964	37.05	1.462	38.94	1.278	38.09	1.647	38.93
YingLong 6M	1.772	38.64	1.304	37.31	0.966	37.10	1.462	38.83	1.278	38.10	1.648	38.86

Table 17 shows that model rankings, as measured by K, remain broadly stable across different naive baselines. Models that rank highly under RelMAE_{ori} generally maintain strong rankings when alternative baselines are used, while lower-ranked models similarly tend to remain near the bottom. This consistency suggests that the relative comparison of forecasting models is robust to the specific choice of baseline used to scale the raw-value forecasting error. In terms of absolute performance, among the 43 evaluated models, 2 outperform the persistence baseline, 25 outperform the seasonal naive baseline, 24 outperform the random-walk-with-drift baseline, 28 outperform the lookback-mean baseline, 9 outperform the AR(1) baseline, and all 43 outperform the seasonal-mean baseline.

Table 18: Overall FinVerse rank sum by forecast-origin year.

Model	Full	2020	2021	2022	2023	2024	2025
Reverso Small	**10**	28	16	20	**11**	**12**	**8**
Chronos-2 Synth	13	39	23	**14**	13	19	14
TimesFM 2.5 200M	24	33	52	30	20	49	28
Chronos-2	26	45	**11**	32	39	40	32
TiRex 1.1	26	31	40	**14**	27	28	33
Chronos-Bolt Base	31	23	78	32	53	59	66
Chronos-Bolt Small	37	**22**	54	45	51	73	69
T0 Alpha	40	53	46	32	44	53	21
TiRex 2 ZS	40	47	54	39	40	48	34
TiRex 2 Pretrain	41	41	55	35	48	46	26
Toto 2.0 2.5B	46	75	73	87	49	40	50
VisionTS	46	32	52	62	54	66	43
TimesFM 1.0 200M	47	50	27	44	64	65	77
TimesFM 2.0 500M	48	63	90	38	39	40	44
PatchFM	49	41	61	59	47	49	39
Kairos-50M	53	49	78	47	74	66	64
Toto 2.0 4M	55	84	54	84	55	35	81
Toto 2.0 1B	57	67	62	83	71	34	34
Granite FlowState R1	58	104	90	73	41	60	86
TempoPFN	60	91	24	93	73	44	49
Kairos-10M	65	75	73	85	69	60	59
Toto 2.0 22M	66	90	57	77	63	60	63
FlowState	69	108	74	87	53	56	87
Toto 2.0 313M	70	68	68	99	73	61	44
Super Linear	71	38	56	71	50	68	36
Granite PatchTST-FM R1	72	39	49	71	110	88	83
PatchTST-FM R1	74	40	54	80	111	87	90
Timer-S1	74	60	75	72	63	80	71
Toto Open Base 1.0	77	83	105	96	94	76	51
Moirai 1.1-R Large	78	90	56	63	85	69	57
Moirai 1.1-R Base	81	82	64	54	72	84	73
Kairos-23M	84	81	77	99	101	67	78
Moirai 1.1-R Small	88	95	90	60	61	78	67
Sundial Base 128M	88	57	67	66	74	82	89
Moirai 2.0-R Small	92	100	94	68	74	109	88
TTM R3	94	63	86	72	48	75	99
TTM R1	99	85	108	70	91	104	115
CleanTS-65M	101	69	81	85	74	80	116
TTM R2	113	69	105	116	83	82	113
YingLong 50M	114	111	90	97	120	103	112
YingLong 300M	118	105	98	101	124	122	114
YingLong 110M	121	101	80	84	124	116	128
YingLong 6M	122	111	91	102	108	105	107

E.4 Stability of the Model Comparison over Time

The results reported so far pool all forecast origins from 2020 onward. Because financial regimes change, a natural question is whether the model comparison itself is stable over time. We therefore repeat the evaluation separately for the forecast origins of each calendar year, using the same metric assignment and rank aggregation. We then compare the rankings across individual years as well as with the full-period ranking. The pairwise correlations between yearly rankings indicate moderate agreement across evaluation years, with Spearman correlations ranging from 0.42 to 0.80 (mean $\rho = 0.61$). These results suggest that, while model rankings exhibit some year-to-year variation,

the relative performance of models in FinVerse is reasonably robust across evaluation periods rather than being driven by a single year.

E.5 Detailed Model Rankings

Table 19 presents the overall FinVerse leaderboard, including each model's overall rank, aspect-level ranks, and corresponding K values. The overall rank is computed as the sum of the point-wise, IC, and portfolio placement ranks, while K denotes the geometric mean of metric-level ranks within each evaluation aspect. Tables 20 and 21 further break down model performance by financial data scope (Country, Individual, and Inter-country) and sampling frequency, respectively. Finally, Tables 22–30 provide a finer-grained breakdown by detail category, where each value represents the geometric mean of the corresponding metric-level model ranks and lower values indicate better performance. The detail-category results are split across multiple tables for readability. Repeated values across the four YingLong checkpoints are expected because their forecasts are almost constant over the horizon for the lower-frequency series regardless of checkpoint size.

Table 19: 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**	6	9.39	**1**	**7.95**	3	11.42
2	Chronos-2 Synth	13	9	11.38	2	8.25	2	10.58
3	TimesFM 2.5 200M	24	7	9.91	4	10.09	13	15.47
4	Chronos-2	26	5	9.18	13	15.32	8	14.55
5	TiRex 1.1	26	**1**	**6.73**	3	9.32	22	16.94
6	Chronos-Bolt Base	31	11	12.03	15	15.72	5	13.52
7	Chronos-Bolt Small	37	10	11.89	10	14.98	17	16.54
8	T0 Alpha	40	12	13.00	24	18.05	4	11.94
9	TiRex 2 ZS	40	3	8.03	5	11.22	32	19.76
10	TiRex 2 Pretrain	41	2	7.30	6	11.25	33	19.93
11	Toto 2.0 2.5B	46	17	13.89	14	15.53	15	16.12
12	VisionTS	46	8	10.96	7	12.27	31	19.37
13	TimesFM 1.0 200M	47	21	16.48	17	16.10	9	14.78
14	TimesFM 2.0 500M	48	38	27.28	9	13.30	**1**	**9.23**
15	PatchFM	49	4	9.14	11	15.06	34	20.10
16	Kairos-50M	53	19	15.86	28	19.04	6	13.62
17	Toto 2.0 4M	55	15	13.75	20	17.29	20	16.84
18	Toto 2.0 1B	57	13	13.34	18	16.63	26	17.45
19	Granite FlowState R1	58	20	16.38	26	18.25	12	15.34
20	TempoPFN	60	16	13.85	16	16.07	28	18.30
21	Kairos-10M	65	27	20.73	27	18.81	11	15.12
22	Toto 2.0 22M	66	18	14.85	25	18.13	23	17.11
23	FlowState	69	23	17.11	30	19.52	16	16.50
24	Toto 2.0 313M	70	14	13.64	29	19.44	27	18.06
25	Super Linear	71	22	17.08	12	15.14	37	20.66
26	Granite PatchTST-FM R1	72	39	28.57	19	17.07	14	15.82
27	PatchTST-FM R1	74	33	25.73	23	17.58	18	16.77
28	Timer-S1	74	24	17.99	8	13.12	42	22.83
29	Toto Open Base 1.0	77	25	18.68	31	20.42	21	16.85
30	Moirai 1.1-R Large	78	37	26.96	34	21.07	7	14.12
31	Moirai 1.1-R Base	81	34	26.18	37	23.41	10	14.96
32	Kairos-23M	84	29	21.09	36	23.32	19	16.77
33	Moirai 1.1-R Small	88	32	24.70	32	20.52	24	17.12
34	Sundial Base 128M	88	26	19.19	22	17.39	40	22.27
35	Moirai 2.0-R Small	92	30	22.71	33	21.05	29	19.05
36	TTM R3	94	35	26.32	21	17.29	38	20.83
37	TTM R1	99	31	22.87	43	26.81	25	17.40
38	CleanTS-65M	101	36	26.56	35	22.43	30	19.10
39	TTM R2	113	28	20.91	42	26.55	43	22.92
40	YingLong 50M	114	41	34.34	38	23.42	35	20.61
41	YingLong 300M	118	43	34.58	39	23.54	36	20.63
42	YingLong 110M	121	42	34.56	40	24.22	39	21.06
43	YingLong 6M	122	40	34.08	41	24.23	41	22.35

Table 20: Scope-wise rank aggregation. Lower values are better.

Model	Country	Individual	Inter-country
Reverso Small	9.32	9.91	9.45
Chronos-2 Synth	12.33	11.35	**7.06**
TimesFM 2.5 200M	9.88	10.71	11.15
Chronos-2	9.14	9.77	12.01
TiRex 1.1	**6.04**	8.92	12.44
Chronos-Bolt Base	12.11	11.95	13.28
Chronos-Bolt Small	12.11	12.03	13.22
T0 Alpha	11.41	15.38	19.60
TiRex 2 ZS	7.44	9.88	13.99
TiRex 2 Pretrain	6.55	10.66	12.49
Toto 2.0 2.5B	14.82	16.46	8.92
VisionTS	11.15	**8.57**	19.16
TimesFM 1.0 200M	16.46	14.09	19.51
TimesFM 2.0 500M	30.31	14.47	20.79
PatchFM	7.63	18.20	13.02
Kairos-50M	14.93	20.04	14.71
Toto 2.0 4M	13.03	19.89	12.21
Toto 2.0 1B	14.08	16.74	8.97
Granite FlowState R1	15.02	22.79	15.33
TempoPFN	14.78	18.07	8.29
Kairos-10M	19.75	26.64	15.41
Toto 2.0 22M	15.07	18.30	11.56
FlowState	15.40	25.96	15.77
Toto 2.0 313M	14.34	16.53	10.24
Super Linear	18.55	11.79	20.76
Granite PatchTST-FM R1	29.88	20.43	25.04
PatchTST-FM R1	26.23	20.10	24.94
Timer-S1	18.67	14.45	21.31
Toto Open Base 1.0	19.52	16.95	16.99
Moirai 1.1-R Large	27.11	21.46	25.34
Moirai 1.1-R Base	26.48	20.82	25.70
Kairos-23M	20.02	23.56	21.17
Moirai 1.1-R Small	24.85	20.82	24.67
Sundial Base 128M	19.31	17.10	22.94
Moirai 2.0-R Small	22.08	22.15	24.22
TTM R3	28.26	19.02	24.24
TTM R1	23.62	18.21	24.85
CleanTS-65M	26.39	23.77	26.39
TTM R2	21.89	17.80	23.43
YingLong 50M	36.28	27.45	26.27
YingLong 300M	36.45	27.38	27.02
YingLong 110M	36.29	28.00	27.07
YingLong 6M	35.97	27.47	27.23

Table 21: Frequency-wise rank aggregation. Lower values are better.

Model	Business-daily	Daily	Weekly	Monthly	Quarterly	Annual
Reverso Small	8.08	11.50	12.08	9.92	9.43	13.37
Chronos-2 Synth	7.38	**4.34**	8.08	11.57	15.41	18.90
TimesFM 2.5 200M	10.84	12.80	9.23	9.36	11.25	8.80
Chronos-2	13.05	11.55	10.17	8.73	8.83	14.46
TiRex 1.1	11.51	14.31	8.54	**4.66**	10.03	12.99
Chronos-Bolt Base	13.32	11.41	8.00	13.06	10.66	9.01
Chronos-Bolt Small	12.41	15.09	9.77	12.80	11.17	13.95
T0 Alpha	18.74	24.04	20.71	9.25	16.65	18.85
TiRex 2 ZS	15.94	13.09	10.98	6.03	9.53	10.52
TiRex 2 Pretrain	14.42	10.06	10.72	5.02	10.45	12.76
Toto 2.0 2.5B	**6.28**	13.50	7.35	16.42	19.82	33.00
VisionTS	19.70	23.60	28.40	10.65	**7.89**	9.56
TimesFM 1.0 200M	20.21	17.23	25.71	15.56	14.96	12.63
TimesFM 2.0 500M	21.91	14.80	25.64	33.53	17.19	15.13
PatchFM	11.04	14.56	10.67	6.72	16.11	15.37
Kairos-50M	12.81	20.87	15.80	17.85	14.60	20.12
Toto 2.0 4M	11.51	12.78	18.47	13.27	17.63	18.40
Toto 2.0 1B	7.38	12.35	**6.87**	14.62	19.42	34.40
Granite FlowState R1	16.12	22.54	14.12	15.16	18.32	23.46
TempoPFN	6.91	21.09	8.08	16.31	18.63	23.34
Kairos-10M	14.47	21.05	24.52	22.10	21.84	27.78
Toto 2.0 22M	11.20	9.63	13.94	16.09	17.02	33.99
FlowState	15.88	24.62	14.50	15.21	21.55	24.86
Toto 2.0 313M	9.31	11.59	8.31	14.46	18.45	28.98
Super Linear	21.58	21.74	23.58	18.52	12.68	11.63
Granite PatchTST-FM R1	28.20	20.26	32.31	29.43	22.99	25.11
PatchTST-FM R1	26.71	21.58	24.48	24.80	23.31	27.31
Timer-S1	25.02	19.48	10.94	17.24	15.75	16.87
Toto Open Base 1.0	18.56	15.74	16.13	18.72	18.75	20.68
Moirai 1.1-R Large	22.42	13.74	16.14	28.48	25.51	18.80
Moirai 1.1-R Base	23.53	18.48	17.35	27.30	23.79	27.85
Kairos-23M	21.55	23.89	19.51	21.14	20.00	13.39
Moirai 1.1-R Small	24.03	18.67	21.59	26.66	20.75	14.43
Sundial Base 128M	26.73	18.19	15.33	17.89	17.73	14.09
Moirai 2.0-R Small	23.33	20.48	25.35	23.28	20.71	12.74
TTM R3	25.06	25.65	34.40	28.46	21.63	15.91
TTM R1	26.24	24.81	33.02	29.12	13.69	**3.06**
CleanTS-65M	28.72	22.42	35.88	24.93	25.24	27.55
TTM R2	23.49	21.47	36.00	28.58	12.03	6.46
YingLong 50M	30.46	22.21	33.86	35.91	30.57	29.77
YingLong 300M	31.03	22.01	34.18	36.06	30.68	29.77
YingLong 110M	30.86	22.42	34.23	35.97	31.07	29.77
YingLong 6M	30.06	23.74	34.29	35.91	30.64	29.77

Table 22: Detail-category-wise rank aggregation (1/9).

Model	Activity	Banking	Bond Total Return Index	Breakeven	Capacity Utilization by Industry	Capital Expenditures	Cash and Cash Equivalents	Commodity ETF Broad
Reverso Small	8.12	17.04	6.55	9.45	5.84	29.05	14.48	10.94
Chronos-2 Synth	12.13	10.29	7.25	8.90	15.75	14.10	9.58	5.29
TimesFM 2.5 200M	6.17	11.52	15.76	7.81	7.93	19.28	10.17	10.05
Chronos-2	8.57	5.46	17.46	6.96	8.73	17.78	12.50	7.38
TiRex 1.1	3.54	7.02	11.99	6.94	3.82	10.59	3.76	12.11
Chronos-Bolt Base	11.05	12.18	7.08	13.03	9.28	14.89	11.85	18.08
Chronos-Bolt Small	14.99	18.28	**2.09**	18.32	11.79	19.50	12.36	13.64
T0 Alpha	10.23	19.04	19.42	8.23	8.65	14.67	10.69	27.63
TiRex 2 ZS	4.63	4.22	16.12	**6.28**	3.71	12.20	**3.03**	12.40
TiRex 2 Pretrain	**2.77**	**3.28**	11.34	6.73	**2.45**	12.70	3.81	11.56
Toto 2.0 2.5B	17.67	11.48	9.35	11.13	23.20	19.95	17.22	4.54
VisionTS	5.72	15.69	24.31	10.85	6.72	15.51	8.44	20.62
TimesFM 1.0 200M	13.30	18.61	27.68	19.39	11.70	14.51	10.92	25.13
TimesFM 2.0 500M	37.45	37.39	33.96	38.12	34.74	17.20	12.63	34.70
PatchFM	5.13	3.84	3.23	9.04	4.74	20.68	28.24	9.39
Kairos-50M	21.18	18.68	4.66	17.25	22.34	18.95	34.45	19.32
Toto 2.0 4M	15.45	10.44	12.47	14.53	9.97	20.48	25.09	5.81
Toto 2.0 1B	18.18	10.98	8.90	12.50	20.94	14.35	22.33	**2.89**
Granite FlowState R1	21.90	7.13	14.56	15.72	21.63	13.47	27.58	15.28
TempoPFN	13.15	16.59	7.45	17.20	19.12	30.44	28.93	6.52
Kairos-10M	22.74	19.53	23.55	16.09	29.87	29.92	40.11	11.55
Toto 2.0 22M	17.42	7.06	14.87	11.73	19.45	16.78	21.18	5.87
FlowState	15.66	9.81	11.28	18.18	18.29	22.72	33.47	12.71
Toto 2.0 313M	15.42	7.74	13.04	10.24	19.71	18.44	20.86	3.30
Super Linear	20.49	19.92	30.78	19.22	15.98	12.40	7.58	20.19
Granite PatchTST-FM R1	30.12	31.78	41.77	28.83	29.44	20.13	24.47	32.33
PatchTST-FM R1	21.25	22.67	38.00	16.58	24.16	15.20	24.97	27.10
Timer-S1	17.95	23.45	29.96	12.71	16.46	14.84	10.73	24.87
Toto Open Base 1.0	18.84	26.43	15.59	16.47	19.13	10.74	19.09	18.82
Moirai 1.1-R Large	34.69	35.60	26.26	26.07	29.38	15.91	28.38	37.05
Moirai 1.1-R Base	29.91	32.12	21.91	31.30	31.93	10.28	31.36	38.53
Kairos-23M	22.36	25.32	27.45	26.48	28.21	30.94	34.62	24.13
Moirai 1.1-R Small	32.13	34.03	22.36	29.03	27.24	20.19	27.07	36.63
Sundial Base 128M	18.32	19.46	31.71	18.42	15.92	13.99	12.43	27.61
Moirai 2.0-R Small	29.39	30.66	30.82	29.46	32.02	21.37	16.49	30.52
TTM R3	35.07	34.78	30.50	27.20	26.85	21.25	14.55	28.85
TTM R1	31.97	32.07	34.31	38.77	40.54	9.95	8.56	37.80
CleanTS-65M	26.89	29.44	39.90	27.47	23.82	15.11	26.89	34.93
TTM R2	26.67	30.21	6.54	33.83	41.37	**7.25**	17.31	29.92
YingLong 50M	39.94	39.10	35.65	35.99	36.26	21.05	31.94	36.07
YingLong 300M	39.93	39.10	36.76	36.04	36.30	21.05	31.94	37.09
YingLong 110M	39.64	39.10	37.74	36.50	36.30	21.01	32.00	35.07
YingLong 6M	39.88	39.10	34.71	34.81	36.30	21.01	31.94	35.82

Table 23: Detail-category-wise rank aggregation (2/9).

Model	Commodity ETF Single	Commodity ETF Thematic	Corporate Yield	Country Index	Credit	Credit Spread	Crypto Pair Close Top 100	Current Account
Reverso Small	5.90	**3.62**	12.13	**5.56**	10.75	9.39	11.50	8.29
Chronos-2 Synth	5.15	11.37	16.02	14.81	9.27	13.86	**4.34**	19.30
TimesFM 2.5 200M	12.31	8.37	11.73	15.24	15.44	6.28	12.80	5.96
Chronos-2	15.20	23.81	14.18	19.82	8.11	24.15	11.55	14.31
TiRex 1.1	7.64	7.58	12.40	18.06	**2.29**	9.19	14.31	11.32
Chronos-Bolt Base	15.93	7.23	13.64	13.80	15.65	15.24	11.41	10.16
Chronos-Bolt Small	16.29	4.17	15.28	10.26	16.27	18.23	15.09	9.92
T0 Alpha	25.96	28.86	11.47	18.39	11.65	24.93	24.04	15.09
TiRex 2 ZS	14.19	18.84	11.58	20.51	9.17	15.70	13.09	8.33
TiRex 2 Pretrain	15.05	17.18	11.61	20.58	7.20	15.74	10.06	9.99
Toto 2.0 2.5B	2.76	4.96	14.72	10.15	8.44	5.84	13.50	21.48
VisionTS	21.79	23.15	8.64	17.75	14.07	27.09	23.60	5.77
TimesFM 1.0 200M	24.50	10.27	12.92	25.60	18.63	23.74	17.23	17.03
TimesFM 2.0 500M	32.60	23.14	31.07	25.04	35.60	18.51	14.80	10.17
PatchFM	8.71	10.56	11.10	9.87	5.69	20.54	14.56	10.14
Kairos-50M	18.35	7.15	18.24	10.38	17.95	25.67	20.87	25.92
Toto 2.0 4M	9.57	12.81	18.11	11.44	7.84	**2.83**	12.78	20.82
Toto 2.0 1B	**1.50**	5.94	13.68	12.30	4.54	9.58	12.35	24.88
Granite FlowState R1	10.62	14.42	11.70	12.40	14.90	19.98	22.54	17.42
TempoPFN	6.84	7.57	9.78	6.66	15.33	9.38	21.09	27.25
Kairos-10M	13.79	14.94	14.05	9.05	20.13	17.45	21.05	36.50
Toto 2.0 22M	7.85	11.35	21.80	9.93	6.95	5.88	9.63	21.50
FlowState	14.09	18.50	17.63	12.17	13.36	5.65	24.62	21.50
Toto 2.0 313M	4.03	7.94	15.80	12.31	7.70	11.14	11.59	26.76
Super Linear	22.49	24.46	11.25	22.91	24.59	20.38	21.74	13.10
Granite PatchTST-FM R1	37.47	36.81	28.67	23.71	31.57	35.01	20.26	28.80
PatchTST-FM R1	34.87	33.67	27.59	22.26	24.35	33.38	21.58	27.21
Timer-S1	28.02	26.43	15.90	28.06	22.41	21.96	19.48	20.17
Toto Open Base 1.0	15.15	9.58	**7.00**	10.75	27.82	25.69	15.74	15.14
Moirai 1.1-R Large	30.85	29.73	22.83	21.41	34.82	26.59	13.74	31.46
Moirai 1.1-R Base	36.37	30.90	15.69	24.74	33.26	23.60	18.48	34.14
Kairos-23M	25.03	25.84	19.92	22.08	20.67	23.41	23.89	30.48
Moirai 1.1-R Small	33.72	33.63	17.87	23.29	34.94	12.40	18.67	30.55
Sundial Base 128M	29.42	27.24	13.71	28.12	24.84	16.15	18.19	15.59
Moirai 2.0-R Small	29.73	31.19	22.33	18.84	28.37	22.64	20.48	16.04
TTM R3	29.00	27.68	21.41	19.04	37.48	38.20	25.65	18.54
TTM R1	34.11	33.42	26.73	29.66	30.29	9.08	24.81	**4.14**
CleanTS-65M	35.44	36.61	30.03	23.56	27.18	21.53	22.42	23.11
TTM R2	29.22	23.90	19.13	21.30	26.49	39.19	21.47	5.64
YingLong 50M	36.90	36.88	35.14	26.30	39.05	28.57	22.21	30.67
YingLong 300M	37.57	38.20	36.74	26.15	39.05	30.82	22.01	30.67
YingLong 110M	36.88	36.90	36.75	26.28	39.05	31.59	22.42	30.67
YingLong 6M	36.37	37.02	33.85	24.69	39.05	22.26	23.74	30.67

Table 24: Detail-category-wise rank aggregation (3/9).

Model	EPS Basic	EPS Diluted	ETF OHLCV Close	Energy Activity	Equity OHLCV Close	External	FDI Flows	FX Pair
Reverso Small	10.25	9.94	9.71	10.45	6.12	5.24	16.02	17.50
Chronos-2 Synth	14.68	16.21	6.34	13.11	**4.07**	13.92	8.03	8.08
TimesFM 2.5 200M	10.56	10.55	11.66	13.78	11.25	9.99	14.61	7.43
Chronos-2	16.44	16.58	10.11	4.57	13.46	7.65	20.22	13.75
TiRex 1.1	10.74	11.58	11.95	**2.61**	12.08	**3.30**	10.31	12.30
Chronos-Bolt Base	17.81	18.17	10.39	12.02	16.41	13.81	13.78	10.26
Chronos-Bolt Small	21.08	21.18	10.91	6.91	18.39	14.92	20.38	10.70
T0 Alpha	20.97	21.21	17.58	11.66	11.24	11.84	14.45	14.45
TiRex 2 ZS	11.13	11.82	20.66	4.91	22.89	6.29	10.06	15.61
TiRex 2 Pretrain	11.47	11.79	19.61	4.28	21.54	5.01	9.21	13.30
Toto 2.0 2.5B	22.65	19.77	**3.35**	15.66	8.62	16.67	20.84	10.37
VisionTS	**7.11**	**7.31**	20.95	19.50	24.57	9.39	16.91	9.95
TimesFM 1.0 200M	15.02	15.02	16.74	15.29	18.86	19.63	9.00	13.88
TimesFM 2.0 500M	20.25	19.76	12.37	35.19	11.95	32.02	16.05	16.07
PatchFM	10.27	10.53	13.75	4.71	11.33	5.93	25.94	17.38
Kairos-50M	28.29	28.23	8.69	18.28	16.89	11.42	28.44	23.92
Toto 2.0 4M	19.41	17.13	13.27	11.27	13.75	14.60	21.37	18.46
Toto 2.0 1B	18.02	18.64	5.04	15.32	7.37	14.28	23.33	11.49
Granite FlowState R1	29.46	30.47	19.57	8.20	21.18	15.15	26.02	20.15
TempoPFN	22.38	22.37	8.40	17.40	11.32	16.49	26.62	**4.39**
Kairos-10M	33.43	34.46	15.53	22.22	18.71	20.79	30.54	17.30
Toto 2.0 22M	20.47	19.77	13.05	19.04	15.13	15.37	27.69	15.60
FlowState	32.49	33.71	20.25	8.77	18.22	13.10	34.12	21.26
Toto 2.0 313M	13.89	14.44	8.22	14.60	10.12	12.26	22.94	14.28
Super Linear	11.26	10.95	27.86	21.89	21.11	17.77	10.13	14.39
Granite PatchTST-FM R1	14.84	14.75	28.57	36.24	20.63	27.31	22.99	22.26
PatchTST-FM R1	14.27	14.13	27.87	33.23	19.46	22.85	20.86	21.77
Timer-S1	10.90	10.65	28.82	17.18	27.65	24.15	11.23	19.82
Toto Open Base 1.0	11.10	11.43	17.68	17.70	17.74	24.83	30.16	34.67
Moirai 1.1-R Large	21.82	22.43	16.28	30.23	12.83	37.19	**6.15**	30.07
Moirai 1.1-R Base	17.31	16.79	20.10	29.87	18.98	34.28	9.27	24.52
Kairos-23M	25.82	26.43	18.05	23.84	23.07	18.66	35.05	14.87
Moirai 1.1-R Small	12.48	12.05	26.69	31.90	23.51	28.44	8.13	19.62
Sundial Base 128M	14.62	15.44	31.70	21.75	27.81	25.51	10.06	22.04
Moirai 2.0-R Small	22.94	21.56	23.10	23.54	17.61	30.94	8.31	25.88
TTM R3	9.99	9.35	26.87	35.48	28.10	33.89	22.35	27.38
TTM R1	21.57	22.33	30.03	29.24	23.43	22.05	7.86	21.84
CleanTS-65M	21.67	20.88	34.51	29.24	25.36	28.71	24.04	24.91
TTM R2	20.25	20.53	29.53	32.73	28.49	15.31	9.16	24.62
YingLong 50M	20.21	19.49	33.40	39.46	27.99	38.83	32.36	21.05
YingLong 300M	20.89	20.26	33.19	40.87	26.16	38.83	32.36	24.44
YingLong 110M	22.91	23.35	33.28	40.28	27.22	38.83	32.36	25.70
YingLong 6M	20.42	20.21	33.85	40.02	26.83	38.83	32.36	24.76

Table 25: Detail-category-wise rank aggregation (4/9).

Model	Financial Flows	Fiscal	Forward Inflation	Freight	Future Front	GDP	Housing	Housing Price
Reverso Small	6.71	12.38	12.19	10.05	12.26	6.68	17.03	17.75
Chronos-2 Synth	20.61	15.24	11.47	7.69	**4.85**	24.92	8.44	11.96
TimesFM 2.5 200M	13.22	**7.94**	14.40	8.29	17.09	11.99	**1.91**	9.89
Chronos-2	4.56	10.78	13.60	7.93	10.78	12.41	9.74	8.70
TiRex 1.1	18.05	15.74	15.63	9.18	13.11	11.56	10.60	3.81
Chronos-Bolt Base	5.73	12.81	21.20	12.20	17.27	17.41	9.32	20.46
Chronos-Bolt Small	**3.92**	14.79	27.94	9.26	21.14	19.32	4.08	21.57
T0 Alpha	13.94	16.22	31.09	**2.09**	14.24	24.27	13.55	21.90
TiRex 2 ZS	11.02	13.65	10.71	7.12	18.09	14.98	5.81	9.58
TiRex 2 Pretrain	12.61	13.66	12.56	8.07	16.35	15.32	7.03	7.45
Toto 2.0 2.5B	20.29	18.48	4.50	18.18	9.49	4.30	26.21	5.32
VisionTS	8.43	15.97	8.51	9.12	24.95	14.74	6.53	16.43
TimesFM 1.0 200M	18.23	14.86	30.82	12.34	19.07	24.15	11.65	17.71
TimesFM 2.0 500M	21.09	29.18	39.15	32.98	18.26	29.45	35.64	36.14
PatchFM	7.89	11.76	14.66	9.61	17.07	7.51	10.76	8.31
Kairos-50M	26.29	14.64	25.65	18.07	8.87	**2.08**	25.43	16.96
Toto 2.0 4M	14.40	10.89	**4.21**	16.80	8.14	5.66	22.24	12.15
Toto 2.0 1B	20.35	17.88	7.58	19.83	10.72	6.41	27.21	3.08
Granite FlowState R1	7.08	21.75	18.09	15.36	13.94	13.28	19.88	6.55
TempoPFN	17.94	19.85	6.47	21.82	11.60	18.51	22.82	15.92
Kairos-10M	31.95	18.33	14.59	24.66	14.94	7.43	28.85	19.00
Toto 2.0 22M	13.67	14.43	5.01	20.50	12.01	7.53	26.36	8.48
FlowState	9.97	20.79	17.95	21.48	10.81	16.91	15.80	10.30
Toto 2.0 313M	18.44	17.66	10.12	18.92	10.71	7.64	23.36	**2.50**
Super Linear	18.31	25.88	6.96	10.47	22.01	24.12	11.82	25.59
Granite PatchTST-FM R1	30.88	14.79	32.33	30.43	20.44	26.55	32.30	33.58
PatchTST-FM R1	31.44	19.39	28.93	31.35	21.62	30.69	28.60	24.19
Timer-S1	23.37	12.10	12.01	12.52	24.01	30.27	7.96	22.95
Toto Open Base 1.0	17.65	20.10	25.71	15.36	21.30	25.49	17.84	29.10
Moirai 1.1-R Large	28.11	31.01	29.53	29.77	24.80	34.98	19.29	34.73
Moirai 1.1-R Base	36.87	27.22	34.33	22.15	16.53	32.89	22.87	33.69
Kairos-23M	24.95	19.22	14.72	27.72	20.90	15.24	21.97	16.04
Moirai 1.1-R Small	27.55	19.46	28.34	27.32	19.11	26.90	19.73	33.45
Sundial Base 128M	24.06	13.97	20.71	11.90	25.30	31.59	11.47	18.56
Moirai 2.0-R Small	14.48	20.04	21.23	31.35	25.64	24.78	18.09	30.45
TTM R3	28.63	22.93	20.75	21.02	23.95	36.89	25.27	37.03
TTM R1	10.06	17.90	33.90	40.74	17.80	19.26	30.30	28.98
CleanTS-65M	30.01	15.01	21.83	26.47	21.30	30.09	21.53	30.98
TTM R2	17.49	12.39	16.71	40.97	25.54	13.89	39.55	22.91
YingLong 50M	34.38	23.55	34.81	34.03	26.93	38.18	36.59	40.55
YingLong 300M	34.38	23.55	32.05	34.03	27.60	38.18	36.59	40.55
YingLong 110M	34.38	23.55	24.80	34.03	27.26	38.18	36.59	40.55
YingLong 6M	34.38	23.55	25.54	33.96	30.16	38.18	36.59	40.55

Table 26: Detail-category-wise rank aggregation (5/9).

Model	Implied Volatility	Income	Industrial Production by Industry	Inflation	Inflation Expectation	International Investment Position	Inventory	Inventory Stock
Reverso Small	8.83	10.17	5.72	12.01	17.02	12.45	7.42	14.02
Chronos-2 Synth	25.94	12.09	13.42	10.30	11.50	24.92	11.26	7.21
TimesFM 2.5 200M	9.94	14.78	9.38	9.37	12.05	20.60	6.33	5.04
Chronos-2	23.36	8.77	10.28	4.31	13.76	7.25	11.09	10.00
TiRex 1.1	9.17	**2.44**	2.92	**2.27**	7.60	19.84	2.27	9.66
Chronos-Bolt Base	28.22	12.78	11.28	12.31	15.01	7.95	12.96	8.91
Chronos-Bolt Small	27.23	13.97	12.19	16.66	11.34	8.59	16.69	10.07
T0 Alpha	18.73	9.18	8.88	11.83	5.74	18.45	11.63	20.87
TiRex 2 ZS	12.09	7.67	3.95	7.63	8.76	14.23	3.56	6.92
TiRex 2 Pretrain	8.31	5.98	**2.87**	6.36	7.69	16.63	**2.10**	7.45
Toto 2.0 2.5B	**5.01**	15.07	23.96	15.81	14.87	26.01	15.67	6.82
VisionTS	32.65	13.30	3.52	12.19	**3.58**	8.86	9.99	26.93
TimesFM 1.0 200M	7.90	12.48	10.73	17.69	12.92	23.23	20.18	21.02
TimesFM 2.0 500M	31.50	37.19	37.12	37.23	39.56	15.44	36.62	34.34
PatchFM	11.04	5.50	5.51	3.56	10.72	16.55	6.69	15.55
Kairos-50M	14.15	10.02	20.09	9.95	18.18	6.50	17.95	19.37
Toto 2.0 4M	16.77	10.33	13.86	15.60	15.08	14.66	10.87	20.60
Toto 2.0 1B	6.84	12.58	21.26	12.25	11.55	21.13	14.65	**4.83**
Granite FlowState R1	22.02	17.38	20.68	13.96	17.69	13.52	13.83	17.21
TempoPFN	7.76	10.91	18.21	14.86	19.85	17.27	14.94	7.08
Kairos-10M	23.62	16.89	23.33	16.56	20.40	18.88	19.60	32.95
Toto 2.0 22M	6.46	18.96	19.92	17.64	15.60	18.02	18.52	13.24
FlowState	21.30	16.07	20.43	18.46	18.20	14.62	12.38	15.39
Toto 2.0 313M	7.29	12.98	20.63	14.44	18.99	25.34	15.39	6.16
Super Linear	11.88	24.12	15.28	24.27	5.86	19.10	21.98	21.17
Granite PatchTST-FM R1	24.21	32.23	26.91	33.07	25.88	30.93	30.82	36.32
PatchTST-FM R1	24.47	23.55	26.30	27.85	18.75	31.87	24.93	31.20
Timer-S1	20.71	21.67	15.60	23.24	15.74	24.68	20.58	5.43
Toto Open Base 1.0	31.30	26.97	21.68	32.86	20.89	20.82	20.94	23.62
Moirai 1.1-R Large	6.11	34.84	30.93	31.94	26.33	33.62	35.13	24.67
Moirai 1.1-R Base	17.36	31.39	25.86	33.39	21.16	34.09	32.20	29.17
Kairos-23M	22.91	12.78	26.41	11.58	30.32	6.27	20.26	24.29
Moirai 1.1-R Small	14.13	32.21	28.60	31.09	23.84	26.98	33.45	30.29
Sundial Base 128M	22.25	19.61	16.25	21.83	18.70	23.73	23.89	9.79
Moirai 2.0-R Small	19.96	23.13	32.41	20.90	21.88	10.24	30.65	22.09
TTM R3	32.81	36.69	30.95	35.70	32.52	29.17	35.21	39.81
TTM R1	20.78	19.55	39.25	21.20	24.39	2.44	31.24	36.04
CleanTS-65M	15.71	32.50	25.36	28.42	25.31	29.63	28.22	34.11
TTM R2	38.86	20.67	35.88	20.24	25.54	**1.97**	26.31	39.87
YingLong 50M	33.66	40.27	37.45	40.16	34.07	36.83	39.74	31.32
YingLong 300M	31.82	40.27	37.49	40.27	34.21	36.83	39.51	31.32
YingLong 110M	30.80	40.27	37.49	40.05	34.72	36.83	39.02	31.32
YingLong 6M	30.45	40.27	37.49	40.11	34.32	36.83	39.15	31.32

Table 27: Detail-category-wise rank aggregation (6/9).

Model	Labor	Labor by Industry	Long Term Debt	Money	Money Market	Mortgage Rate	Net Income	Operating Cash Flow
Reverso Small	7.33	5.77	10.37	10.68	3.39	9.72	7.16	23.15
Chronos-2 Synth	10.12	10.69	17.91	11.47	12.29	9.55	7.61	13.45
TimesFM 2.5 200M	8.01	13.42	12.42	15.84	4.14	22.24	6.25	13.54
Chronos-2	7.69	7.21	**3.00**	8.23	15.08	10.43	19.76	21.14
TiRex 1.1	**2.35**	**2.28**	9.84	**2.68**	7.58	7.15	**4.77**	**8.55**
Chronos-Bolt Base	15.94	15.05	5.13	10.97	12.80	6.83	17.61	22.32
Chronos-Bolt Small	10.62	13.37	4.05	10.48	15.03	9.35	20.70	23.20
T0 Alpha	9.28	11.46	13.88	9.85	22.67	20.47	13.92	17.61
TiRex 2 ZS	3.60	5.30	6.15	5.61	3.28	21.51	5.91	15.58
TiRex 2 Pretrain	2.76	3.60	7.85	3.81	**3.25**	18.19	5.83	14.06
Toto 2.0 2.5B	19.41	14.77	20.51	14.71	4.64	8.19	23.56	15.67
VisionTS	9.90	13.08	3.53	11.43	20.50	30.67	9.85	19.98
TimesFM 1.0 200M	13.82	19.69	14.97	19.13	29.05	34.47	9.72	12.92
TimesFM 2.0 500M	35.05	33.24	11.76	30.87	32.95	16.76	16.27	17.94
PatchFM	4.81	4.81	24.02	3.62	5.03	**6.16**	18.70	20.40
Kairos-50M	16.22	18.88	30.92	18.85	15.66	11.76	31.42	22.16
Toto 2.0 4M	10.37	8.41	23.63	11.86	16.91	15.74	23.14	13.75
Toto 2.0 1B	19.27	9.50	21.25	15.36	4.82	11.46	24.24	15.45
Granite FlowState R1	21.47	16.43	13.48	12.96	22.61	10.59	38.41	11.77
TempoPFN	11.96	15.85	14.41	11.18	8.18	9.78	30.33	34.06
Kairos-10M	22.29	18.98	35.84	20.68	11.89	15.95	38.44	37.91
Toto 2.0 22M	13.55	7.86	18.62	10.72	13.07	15.02	24.11	10.29
FlowState	17.69	14.97	13.74	13.44	21.47	13.29	39.78	23.47
Toto 2.0 313M	15.27	7.89	16.15	14.24	7.46	12.88	27.17	12.25
Super Linear	21.14	21.57	12.39	20.95	32.49	27.59	9.19	10.43
Granite PatchTST-FM R1	35.31	33.44	29.19	31.80	40.32	27.25	12.60	17.24
PatchTST-FM R1	28.66	25.91	31.15	22.89	35.64	17.20	13.83	14.16
Timer-S1	18.95	25.26	20.88	20.33	26.55	30.28	6.91	9.88
Toto Open Base 1.0	24.89	29.67	24.52	19.55	17.13	9.26	26.33	14.85
Moirai 1.1-R Large	31.84	33.98	22.77	34.37	29.24	8.71	26.86	21.82
Moirai 1.1-R Base	32.71	34.97	27.23	32.44	27.32	8.15	28.00	12.00
Kairos-23M	19.64	20.43	24.96	21.19	24.71	14.18	29.74	34.42
Moirai 1.1-R Small	32.01	34.09	24.92	31.23	33.14	13.19	20.76	23.78
Sundial Base 128M	20.74	24.10	20.26	23.92	26.26	29.43	8.94	11.40
Moirai 2.0-R Small	28.38	30.01	19.05	29.43	25.60	30.97	30.84	23.87
TTM R3	34.15	36.21	19.76	37.25	34.00	27.82	10.76	23.01
TTM R1	36.99	36.27	23.03	33.44	33.83	29.08	23.50	12.98
CleanTS-65M	27.42	27.36	29.83	27.47	39.15	38.61	17.31	15.53
TTM R2	35.96	35.03	26.07	30.83	31.77	31.03	18.94	8.66
YingLong 50M	39.31	39.30	35.21	39.27	36.45	37.94	22.35	19.76
YingLong 300M	40.03	39.30	35.21	39.27	37.68	38.83	22.42	19.76
YingLong 110M	39.92	39.30	35.21	39.27	36.67	38.97	22.37	19.72
YingLong 6M	39.54	39.26	35.21	39.27	36.40	39.12	22.44	19.76

Table 28: Detail-category-wise rank aggregation (7/9).

Model	Operating Income	Portfolio Flows	Productivity	Regional Fed	Regional Index	Revenue	Saving	Sentiment
Reverso Small	7.28	6.84	9.14	11.71	9.01	7.42	6.10	11.21
Chronos-2 Synth	9.63	**5.22**	24.00	**4.58**	3.78	13.23	19.92	13.94
TimesFM 2.5 200M	5.76	7.43	16.52	7.72	5.36	8.01	6.31	14.45
Chronos-2	18.82	11.00	5.20	10.60	4.06	9.77	11.74	8.59
TiRex 1.1	**5.52**	6.50	19.29	7.87	18.06	8.04	11.22	3.97
Chronos-Bolt Base	15.81	10.79	6.69	14.33	12.39	11.04	14.59	9.15
Chronos-Bolt Small	18.79	9.40	8.34	20.34	11.21	12.65	10.04	8.32
T0 Alpha	13.93	14.07	25.01	7.50	28.76	17.80	5.86	5.06
TiRex 2 ZS	6.79	6.76	15.04	7.07	16.91	7.01	7.02	**3.29**
TiRex 2 Pretrain	6.03	6.65	17.61	6.46	11.33	8.07	**5.71**	4.68
Toto 2.0 2.5B	28.04	27.98	20.21	17.70	10.29	28.00	23.64	15.07
VisionTS	9.14	11.10	7.41	15.32	13.10	**6.10**	13.32	17.37
TimesFM 1.0 200M	9.08	6.04	21.19	18.55	21.44	11.56	15.94	18.26
TimesFM 2.0 500M	12.66	21.93	23.41	37.41	36.47	12.44	19.54	36.37
PatchFM	22.94	20.26	15.05	7.14	15.80	28.50	15.36	10.46
Kairos-50M	26.00	11.99	**2.43**	24.07	14.17	14.75	20.35	25.54
Toto 2.0 4M	27.19	19.87	11.54	10.95	15.76	24.92	23.11	18.47
Toto 2.0 1B	28.24	24.85	18.31	17.54	11.21	25.57	23.87	13.75
Granite FlowState R1	38.17	27.79	10.20	27.59	27.28	36.52	18.88	8.13
TempoPFN	28.61	12.02	8.88	15.38	**1.86**	15.58	15.59	16.92
Kairos-10M	33.50	21.29	4.26	28.26	17.87	19.99	34.70	28.07
Toto 2.0 22M	25.84	19.47	18.81	18.21	15.93	22.69	28.67	20.16
FlowState	39.99	36.61	11.13	28.08	28.17	38.93	12.55	5.91
Toto 2.0 313M	28.91	26.94	17.65	18.61	13.04	23.39	23.13	17.52
Super Linear	7.18	16.89	20.72	25.67	20.84	6.78	9.10	14.68
Granite PatchTST-FM R1	13.84	25.71	32.87	30.13	25.09	18.57	34.52	29.76
PatchTST-FM R1	13.02	26.73	33.66	35.14	23.16	18.60	31.41	30.59
Timer-S1	6.97	10.95	28.39	14.21	21.68	11.74	18.59	16.27
Toto Open Base 1.0	28.20	22.67	21.41	8.29	19.91	18.45	7.35	25.54
Moirai 1.1-R Large	24.29	28.85	34.71	18.95	37.29	24.29	11.66	28.99
Moirai 1.1-R Base	26.14	28.71	27.85	15.66	38.63	23.85	21.97	27.05
Kairos-23M	27.25	27.19	6.34	13.52	23.41	17.34	34.57	29.51
Moirai 1.1-R Small	20.32	22.12	23.08	17.04	37.16	20.08	20.31	27.07
Sundial Base 128M	10.35	13.69	29.70	17.43	25.04	13.30	13.68	16.22
Moirai 2.0-R Small	30.13	16.19	29.58	14.90	30.26	30.82	6.02	22.01
TTM R3	15.77	22.92	33.41	12.41	12.04	22.14	19.20	26.03
TTM R1	18.51	23.40	12.43	24.59	30.84	17.90	25.51	36.01
CleanTS-65M	17.09	29.50	31.96	26.22	34.28	20.55	25.72	20.57
TTM R2	16.38	12.37	12.45	31.71	7.04	17.26	33.12	38.58
YingLong 50M	22.62	33.67	38.95	37.06	27.76	29.56	36.73	35.66
YingLong 300M	22.60	33.67	38.95	37.32	28.18	29.29	37.83	35.66
YingLong 110M	22.53	33.67	38.95	37.30	28.60	29.59	35.82	35.57
YingLong 6M	22.56	33.67	38.95	37.23	26.80	29.54	36.69	35.66

Table 29: Detail-category-wise rank aggregation (8/9).

Model	Spending	Spot	Stockholders Equity	Stress	TIC Flows	TIPS Yield	Term Spread	Total Assets
Reverso Small	6.85	8.13	10.06	15.52	11.26	**2.32**	7.89	8.06
Chronos-2 Synth	13.55	8.82	24.03	16.01	25.04	10.16	4.67	19.17
TimesFM 2.5 200M	8.02	13.63	14.21	7.89	17.32	19.73	16.29	12.89
Chronos-2	12.08	10.47	**2.14**	7.09	4.44	19.78	22.41	**2.34**
TiRex 1.1	3.97	10.46	11.87	10.58	18.91	8.45	15.17	10.86
Chronos-Bolt Base	18.33	17.60	6.40	12.43	4.81	10.07	7.84	5.43
Chronos-Bolt Small	23.24	18.34	3.52	**7.04**	7.46	14.48	5.84	3.58
T0 Alpha	8.44	18.25	15.82	7.19	20.39	23.77	10.50	14.53
TiRex 2 ZS	7.19	8.80	10.12	8.42	13.14	20.19	23.57	8.39
TiRex 2 Pretrain	4.43	**7.92**	13.86	9.08	16.40	18.68	19.03	11.14
Toto 2.0 2.5B	16.87	10.96	22.53	27.30	20.31	5.20	10.28	22.40
VisionTS	5.07	21.22	3.65	16.33	7.63	27.86	20.77	3.04
TimesFM 1.0 200M	18.17	23.32	19.04	19.11	22.32	24.63	22.79	16.57
TimesFM 2.0 500M	37.98	16.29	15.23	31.74	21.72	18.33	24.04	13.52
PatchFM	**3.62**	11.24	16.95	11.06	10.78	5.14	9.13	25.18
Kairos-50M	16.72	14.87	13.45	13.98	**4.03**	13.93	**2.44**	16.36
Toto 2.0 4M	9.49	13.00	23.07	17.03	12.73	14.47	15.99	24.18
Toto 2.0 1B	14.12	12.49	23.89	25.63	16.77	9.62	13.20	25.08
Granite FlowState R1	17.40	10.50	13.38	19.58	5.89	5.74	23.07	19.99
TempoPFN	12.75	11.93	12.37	24.75	11.11	5.23	5.28	11.62
Kairos-10M	14.72	14.77	22.17	28.93	8.33	14.89	6.77	20.74
Toto 2.0 22M	10.50	14.15	14.25	24.67	13.00	15.62	16.63	21.46
FlowState	19.17	12.70	14.89	16.70	8.32	6.71	20.52	23.59
Toto 2.0 313M	13.67	13.45	19.53	27.53	19.53	13.51	11.14	19.94
Super Linear	24.42	22.03	12.65	24.53	19.84	27.93	28.26	11.46
Granite PatchTST-FM R1	31.98	25.75	31.05	33.92	34.54	41.75	35.06	28.92
PatchTST-FM R1	22.63	29.47	34.27	26.84	35.93	39.44	34.16	32.87
Timer-S1	22.99	22.10	21.20	18.03	29.64	33.45	31.61	19.14
Toto Open Base 1.0	27.65	12.51	17.97	10.03	18.38	20.79	23.19	14.47
Moirai 1.1-R Large	33.80	30.06	30.77	18.23	35.72	15.58	8.97	24.88
Moirai 1.1-R Base	32.52	32.52	21.98	14.65	29.70	12.36	15.48	26.49
Kairos-23M	16.13	19.72	15.99	15.59	9.41	24.53	18.55	16.30
Moirai 1.1-R Small	34.12	28.77	22.66	7.16	26.07	17.80	17.33	25.20
Sundial Base 128M	26.37	23.54	26.71	13.88	30.26	31.71	31.54	23.47
Moirai 2.0-R Small	24.00	23.72	19.11	10.57	23.97	23.29	22.82	20.94
TTM R3	37.85	23.52	30.93	11.92	34.93	30.28	36.76	26.88
TTM R1	22.68	18.61	14.44	25.54	15.13	19.14	22.61	16.52
CleanTS-65M	29.52	26.27	34.68	14.30	32.02	37.49	34.92	34.15
TTM R2	12.92	25.91	12.45	31.15	11.06	19.61	35.48	16.68
YingLong 50M	39.94	23.87	39.63	34.67	40.86	36.29	31.73	39.40
YingLong 300M	39.94	23.95	39.63	34.67	40.86	38.56	30.43	39.40
YingLong 110M	39.94	23.71	39.70	34.54	40.86	38.06	31.20	39.40
YingLong 6M	39.94	23.79	39.63	34.67	40.86	32.05	32.34	39.40

Table 30: Detail-category-wise rank aggregation (9/9).

Model	Total Liabilities	Treasury Yield	US Equity Index
Reverso Small	7.85	**2.46**	14.37
Chronos-2 Synth	17.40	18.00	6.99
TimesFM 2.5 200M	13.39	18.24	11.04
Chronos-2	3.42	29.36	5.45
TiRex 1.1	10.17	7.04	12.47
Chronos-Bolt Base	6.46	9.46	9.22
Chronos-Bolt Small	5.07	11.05	8.03
T0 Alpha	15.08	12.56	29.15
TiRex 2 ZS	7.39	9.01	21.04
TiRex 2 Pretrain	10.79	9.60	18.63
Toto 2.0 2.5B	21.19	6.38	**1.71**
VisionTS	**2.49**	20.89	20.87
TimesFM 1.0 200M	15.92	26.89	22.26
TimesFM 2.0 500M	13.77	20.04	34.38
PatchFM	31.19	5.89	11.27
Kairos-50M	17.83	15.23	8.76
Toto 2.0 4M	24.79	16.87	10.45
Toto 2.0 1B	25.08	7.82	4.34
Granite FlowState R1	22.90	5.90	19.43
TempoPFN	16.16	9.25	2.31
Kairos-10M	20.50	11.28	16.06
Toto 2.0 22M	22.09	17.63	8.71
FlowState	28.32	14.96	19.77
Toto 2.0 313M	20.52	12.52	4.34
Super Linear	9.83	32.41	27.53
Granite PatchTST-FM R1	27.97	39.56	34.81
PatchTST-FM R1	30.68	36.51	32.25
Timer-S1	18.25	22.63	27.27
Toto Open Base 1.0	13.74	14.81	16.67
Moirai 1.1-R Large	22.96	25.75	38.05
Moirai 1.1-R Base	24.61	16.24	40.18
Kairos-23M	19.02	27.70	24.18
Moirai 1.1-R Small	20.04	22.00	36.01
Sundial Base 128M	21.72	31.36	30.57
Moirai 2.0-R Small	16.55	10.20	33.09
TTM R3	26.17	22.75	21.84
TTM R1	14.60	29.44	30.81
CleanTS-65M	33.01	40.32	36.96
TTM R2	14.68	25.74	22.65
YingLong 50M	37.83	37.41	34.25
YingLong 300M	37.83	39.04	34.24
YingLong 110M	37.83	37.97	34.61
YingLong 6M	37.83	37.55	33.24