# Evaluating Long-Range Temporal Structure in Foundation Model-Based Forecasts of Heartbeat Dynamics

**Anonymous Authors**[1]

## Abstract

We examine the long-range temporal structure of forecasts produced by Time-Series Foundation Models (TSFMs) on heartbeat dynamics using the MIT-BIH Normal Sinus Rhythm Database (NSRDB). Our findings indicate that these models do not adequately capture long-range dependencies, as reflected in growing errors in RR-interval predictions over longer forecast horizons. Code is available at (Anonymized for peer review).

## 1. Introduction

Time-Series Foundation Models (TSFMs) have demonstrated remarkable capabilities in forecasting diverse temporal data. By leveraging large-scale pre-training on various time-series datasets, TSFMs are capable of forecasting without task-specific training. This in-context learning capability offers a promising solution to regimes where there is a scarcity of available data, particularly in the healthcare domain, where high-quality longitudinal data is often difficult to acquire.

Although TSFMs demonstrate the potential to greatly advance physiological modeling, it remains unclear whether such models are actually capable of modeling the long-range structure of physiological data. We seek to ask the question *Is the model merely reproducing local beat-to-beat patterns, or does it capture physiologically plausible heart rate dynamics over longer time scales*?

We chose to evaluate the ability of TSFMs to model the long-range structure of heartbeat dynamics by forecasting Electrocardiogram (ECG) data. First, ECG is a very widespread modality given its importance in cardiac monitoring. Second, ECG data are characterized by a complex multiscale temporal hierarchy, where fine-grained morphological features are nested within broader physiological trends. The focus of our work goes beyond the local waveform morphology (milliseconds) and more to the long-range RR interval structure of ECG data (minutes to hours). Although ECG-specific foundation models have been proposed (Li et al., 2025a;b; Tang et al., 2025), the lack of publicly available weights and their predominant focus on classification rather than forecasting led us to focus our analysis on general-purpose time-series foundation models. Motivated by the above, we made the following contributions:

1. **Long-Range Physiological Signal Evaluation**: Our analysis focuses on evaluating the capacity of TSFMs to forecast RR intervals, representing long-range temporal structure in the minutes to hours range than simply local waveform-level (milliseconds)

2. **Physiological Distributional-Level Evaluation**: We evaluate TSFMs beyond the typical metrics used typically limited to MAE and RMSE. We also include Standard Deviation of Normal-to-Normal Intervals (SDNN) and Root Mean Square of Successive Differences (RMSSD) in order to evaluate if the TSFMs evaluate the fluctuations of the RR interval signal. Moreover, we also include a Kolmogorov-Smirnov (KS) goodness-of-fit metric as a metric to measure alignment with the underlying statistical generative process governing heartbeat dynamics (Subramanian & Ramsundar, 2025).

Our findings underscore the necessity of evaluating TSFMs through the lens of physiological validity before they can be reliably deployed for clinical forecasting.

## 2. Materials and Methods

### 2.1. Dataset

We use the MIT-BIH normal sinus rhythm dataset for benchmarking the TSFMs (The Beth Israel Deaconess Medical Center, 1990). This dataset contains 18 ECG recordings of subjects referred to the Arrhythmia Laboratory at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

*Table 1.* Characteristics of Time-Series Foundation Models Benchmarked

| | TimesFM-2.5 | Chronos-2 | Moirai 2.0 |
|---|---|---|---|
| Number of Parameters | 200M | 120M | 11.4M |
| Context Length | 16,000 | 8,192 | — |
| Maximum Pred. Length | — | 1,024 | — |

### 2.2. Time-Series Foundation Models

Although several foundation models have been developed for ECG and PPG data (Li et al., 2025a;b; Tang et al., 2025), most focus on classification or event detection rather than forecasting, and many lack publicly available weights, precluding inclusion in this study. While specialized time-series models exist (Nie et al., 2023), we instead evaluate the zero-shot forecasting capabilities of general-purpose Time-Series Foundation Models (TSFMs). This is motivated by the fact that modern TSFMs already incorporate patch-based transformer architectures (e.g., PatchTST), enabling modeling of long-range temporal dependencies.

We evaluate the following TSFMs (Table 1):

1. **TimesFM 2.5 (Google):** A decoder-only transformer using patching to efficiently capture long-range dependencies, pre-trained on large-scale synthetic and real-world time-series data (Das et al., 2024).

2. **Chronos-2 (Amazon):** Frames forecasting as a language modeling problem by quantizing continuous values into discrete tokens and predicting their distribution autoregressively (Ansari et al., 2025).

3. **Moirai 2.0 (Salesforce):** A lightweight decoder-only TSFM with single-patch input representation and quantile loss for probabilistic forecasting (Liu et al., 2026).

### 2.3. Evaluation Metrics

To assess the accuracy and distributional fidelity of the forecasted ECG signals, we employed several primary metrics. Let $y_i$ represent the ground truth signal and $\hat{y}_i$ represent the model's prediction for $n$ time points.

- **Mean Absolute Error (MAE):** Measures the average absolute magnitude of the forecast errors.

$$\text{MAE} = \frac{1}{H}\sum_{t=1}^{H}|y_t - \hat{y}_t| \tag{1}$$

- **Root Mean Squared Error (RMSE):** Provides a metric that disproportionately penalizes large outliers, which is critical for identifying significant deviations in cardiac cycle timing.

$$\text{RMSE} = \sqrt{\frac{1}{H}\sum_{t=1}^{H}(y_t - \hat{y}_t)^2} \tag{2}$$

- **Standard Deviation of NN intervals (SDNN):** Measures the overall variability of the RR intervals (normal-to-normal intervals), reflecting both short- and long-term components of heart rate variability.

$$\text{SDNN} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(RR_i - \bar{RR})^2} \tag{3}$$

where $RR_i$ denotes the $i$-th RR interval and $\bar{RR}$ is the mean RR interval. SDNN captures global variability and is sensitive to long-range temporal structure in the heartbeat dynamics.

- **Root Mean Square of Successive Differences (RMSSD):** Quantifies short-term variability by measuring the root mean square of differences between successive RR intervals.

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2} \tag{4}$$

RMSSD is primarily sensitive to high-frequency variations and reflects parasympathetic (vagal) activity, making it useful for assessing short-term temporal fidelity of the forecasted signals.

- **Kolmogorov-Smirnov (KS) Score:** We use the KS score as a Goodness-of-Fit (GOF) metric based on the Time Rescaling Theorem (Subramanian & Ramsundar, 2025). If a model correctly captures the heartbeat generative process, the transformed inter-event intervals $z_i$ should be independent and identically distributed uniform random variables on the interval $(0, 1)$.

For a model providing a conditional cumulative distribution $F(\tau|\mathcal{H}_t)$, we define the transformed variables:

$$z_i = F(\tau_i|\mathcal{H}_{i-1}) \tag{5}$$

The KS statistic is defined as the maximum deviation between the empirical cumulative distribution of the $z_i$ values and the CDF of a uniform distribution:

$$KS = \max_{1\le i\le N}\left|\frac{i-0.5}{N} - z_{(i)}\right| \tag{6}$$

where $z_{(i)}$ are the sorted transformed intervals. A well-fit model yields a KS score that falls within the 95% significance cutoff, defined as:

$$\text{Cutoff}_{95\%} \approx \frac{1.36}{\sqrt{N}} \tag{7}$$

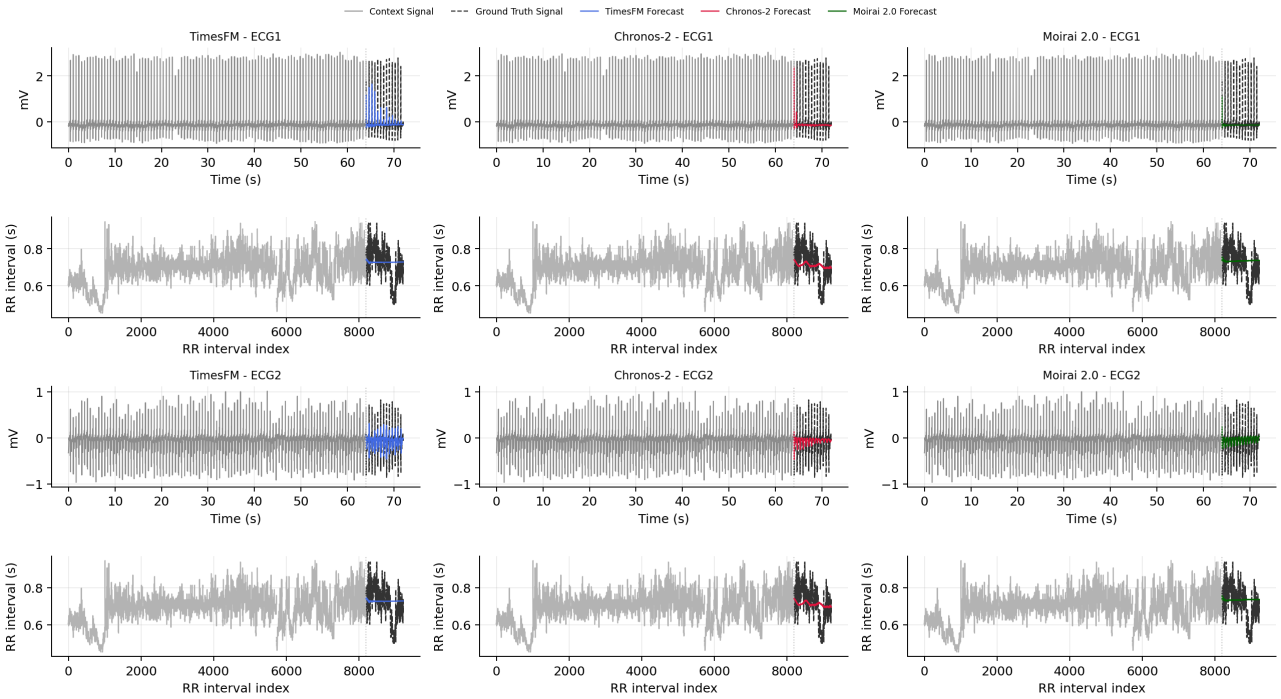


*Figure 1.* Time-Series Foundation Model (TSFM) forecasts of a single-subject's ECG data MIT-BIH NSRDB. Each panel compares the observed context window with the forecast horizon for ECG waveform and RR-interval prediction. TSFMs were given 8,192 samples of context and were tasked to forecast 1,024 future waveform or RR interval times.

## 3. Experimental Setup and Results

Figure 1 illustrates an example forecast for a particular patient. Qualitatively, we observe that the models cannot capture the long-range stochasticity that is reflected by an RR interval temporal signal, as opposed to the waveform where qualitatively, the TSFMs are capable of emulating.

For our first experiment, we evaluated RR interval forecasting performance across varying ECG context lengths and prediction horizons. This experiment tests whether longer historical context improves forecasting accuracy and whether errors accumulate as models forecast further into the future. We report subject-averaged MAE, RMSE, and KS statistics across the 18-subject cohort.

As shown in Table 2, forecasting performance generally worsened as the prediction horizon increased. This degradation was most apparent in RMSE and MAE, particularly for Chronos-2 and Moirai 2.0, suggesting that pointwise prediction errors accumulate over longer horizons. In contrast, increasing the context length from 2,048 to 4,096 or 8,192 samples did not consistently improve performance across models or metrics. These results suggest that simply providing additional historical ECG context is insufficient for reliably improving long-range RR interval forecasting.

For our next experiment, we compared two approaches for RR interval forecasting: direct prediction and extraction from model-generated ECG waveforms. Extracted RR intervals often achieved lower RMSE and MAE, indicating better pointwise accuracy. However, this advantage did not consistently extend to the KS metric. In several cases, direct prediction yielded lower KS despite higher RMSE/MAE, while in others the reverse held, suggesting a tradeoff between pointwise accuracy and distributional fidelity.

Lastly, we further analyzed error as a function of forecast beat index to characterize temporal degradation. As shown in Figure 2, all models exhibit increasing error, but with distinct dynamics. TimesFM shows gradual error growth, whereas Moirai 2.0 degrades rapidly and saturates early, and Chronos-2 exhibits delayed but sharp escalation at longer horizons. Critically, both TimesFM and Moirai 2.0 collapse in local variability (SDNN, RMSSD), producing near-zero RR variability relative to the ground truth. Moreover, all three models do not capture true physiological variability at all. This indicates that despite reasonable short-horizon accuracy, these models fail to preserve physiologically meaningful dynamics.

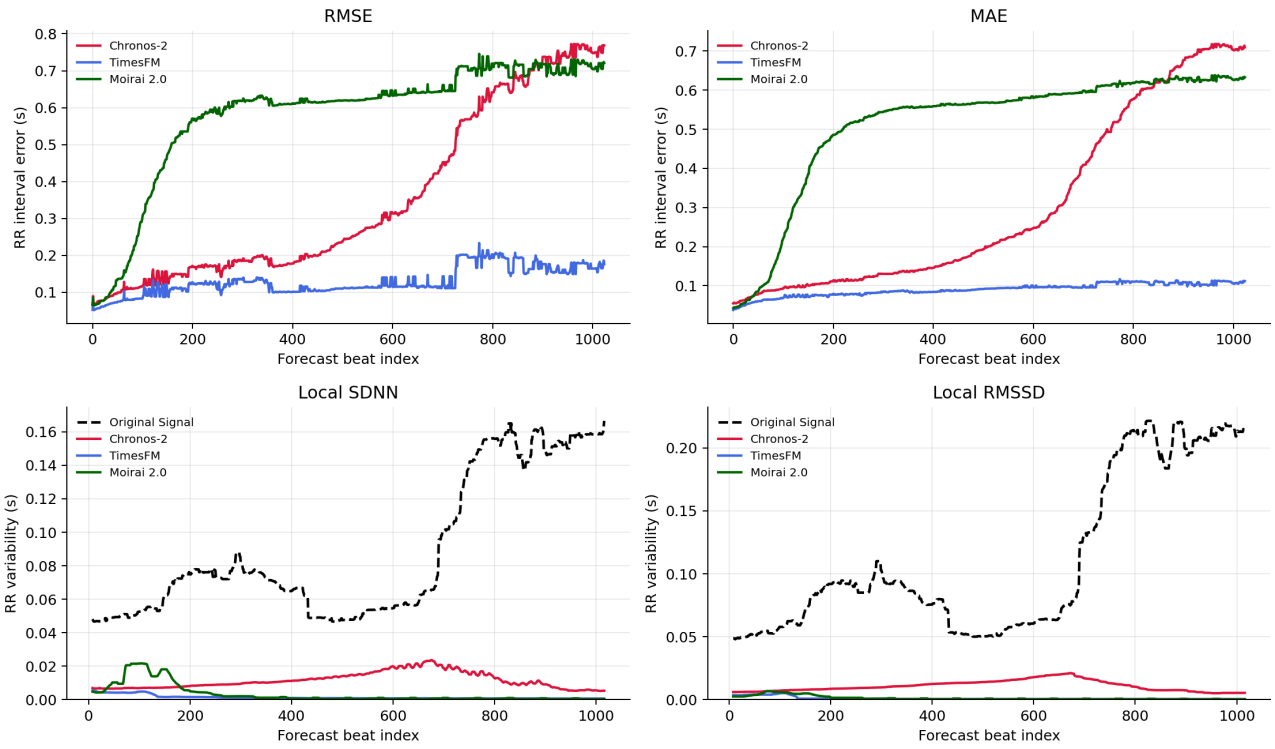


*Figure 2.* Per-beat evaluation of RR interval forecasting error and variability, averaged across the 18-subject cohort. The RMSE and MAE are plotted as a function of forecast beat index. Local variability metrics (SDNN and RMSSD) are compared to the ground truth RR interval signal (black dashed).

## 4. Discussion

The results highlight both the promise and limitations of Time-Series Foundation Models (TSFMs) for physiological data. While these models generate realistic short-term ECG waveforms, TSFM forecasting capabilities does not necessarily extend to long-range temporal scales. Analysis of RR interval dynamics reveals that performance degrades when modeling long-range structure.

A central question is whether TSFMs capture physiologically meaningful dynamics or primarily reproduce local statistical patterns. We observe that the models appear to rely on beat-level morphology rather than learning long-range cardiac behavior. This distinction is critical, as many clinically relevant processes such as autonomic regulation emerge over longer time scales. RR intervals provide a more direct probe of this limitation. Unlike waveform morphology, which can be matched locally, RR interval structure reflects integrated physiological control. Failure to model these dynamics indicates limited ability to capture meaningful temporal hierarchies.

These results have important implications for clinical deployment. Although TSFMs are attractive in low-data settings, caution is warranted for tasks requiring long-range physiological consistency.

Future work should prioritize evaluation frameworks that explicitly test long-range temporal modeling, such as RR interval forecasting. Improving performance may require architectural biases, extended context windows, and training objectives that enforce multi-scale temporal structure in physiological signals.

## References

Ansari, A. F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S. S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D. C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P. M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., and Bohlke-Schneider, M. Chronos-2: from univariate to universal forecasting, October 2025. URL http://arxiv.org/abs/2510.15821. arXiv:2510.15821.

Das, A., Kong, W., Sen, R., and Zhou, Y. A decoder-only foundation model for time-series forecasting,

April 2024. URL http://arxiv.org/abs/2310.10688. arXiv:2310.10688.

Li, H., Deng, B., Xu, C., Feng, Z., Schlegel, V., Huang, Y.-H., Sun, Y., Sun, J., Yang, K., Yu, Y., and Bian, J. Mira: medical time series foundation model for real-world health data. October 2025a. URL https://openreview.net/forum?id=Auy2DmlJBO.

Li, J., Aguirre, A. D., Junior, V. M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Brandon Westover, M., and Hong, S. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), June 2025b. ISSN 2836-9386. doi: 10.1056/AIoa2401033. URL https://ai.nejm.org/doi/10.1056/AIoa2401033.

Liu, C., Aksu, T., Liu, J., Liu, X., Yan, H., Pham, Q., Savarese, S., Sahoo, D., Xiong, C., and Li, J. Moirai 2. 0: when less is more for time series forecasting, February 2026. URL http://arxiv.org/abs/2511.11698. arXiv:2511.11698.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: long-term forecasting with transformers, March 2023. URL http://arxiv.org/abs/2211.14730. arXiv:2211.14730.

Subramanian, S. and Ramsundar, B. Density-based neural temporal point processes for heartbeat dynamics, November 2025. URL http://arxiv.org/abs/2511.22096. arXiv:2511.22096.

Tang, Z., Qi, J., Zheng, Y., and Huang, J. A comprehensive benchmark for electrocardiogram time-series. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6490–6499, Dublin Ireland, October 2025. ACM. ISBN 9798400720352. doi: 10.1145/3746027.3754729. URL https://dl.acm.org/doi/10.1145/3746027.3754729.

The Beth Israel Deaconess Medical Center, T. A. L. The mit-bih normal sinus rhythm database, 1990. URL https://physionet.org/content/nsrdb/.

# A. RR Forecast Context Length and Prediction Horizon Experiment

*Table 2.* Subject-averaged forecasting performance for RR interval prediction across the 18-subject cohort, context lengths, and prediction horizons. Direct denotes direct RR interval prediction; extracted denotes RR intervals derived from model-predicted waveforms via peak extraction.

| Context Length | Prediction Horizon | Output Type | Model | RMSE (s) ↓ | MAE (s) ↓ | KS ↓ |
|---|---|---|---|---|---|---|
| 2,048 | 256 | Direct | Chronos-2 | 0.24 | 0.14 | 0.43 |
| 2,048 | 256 | Direct | TimesFM | 0.20 | 0.11 | 0.49 |
| 2,048 | 256 | Direct | Moirai 2.0 | 0.49 | 0.37 | 0.48 |
| 2,048 | 256 | Extracted | Chronos-2 | 0.14 | 0.08 | 0.50 |
| 2,048 | 256 | Extracted | TimesFM | 0.16 | 0.10 | 0.51 |
| 2,048 | 256 | Extracted | Moirai 2.0 | 0.16 | 0.11 | 0.44 |
| 2,048 | 512 | Direct | Chronos-2 | 0.41 | 0.26 | 0.46 |
| 2,048 | 512 | Direct | TimesFM | 0.23 | 0.10 | 0.44 |
| 2,048 | 512 | Direct | Moirai 2.0 | 0.63 | 0.51 | 0.61 |
| 2,048 | 512 | Extracted | Chronos-2 | 0.20 | 0.10 | 0.46 |
| 2,048 | 512 | Extracted | TimesFM | 0.20 | 0.10 | 0.55 |
| 2,048 | 512 | Extracted | Moirai 2.0 | 0.24 | 0.12 | 0.49 |
| 2,048 | 1,024 | Direct | Chronos-2 | 0.66 | 0.42 | 0.51 |
| 2,048 | 1,024 | Direct | TimesFM | 0.34 | 0.11 | 0.40 |
| 2,048 | 1,024 | Direct | Moirai 2.0 | 0.68 | 0.44 | 0.67 |
| 2,048 | 1,024 | Extracted | Chronos-2 | 0.35 | 0.14 | 0.44 |
| 2,048 | 1,024 | Extracted | TimesFM | 0.25 | 0.10 | 0.56 |
| 2,048 | 1,024 | Extracted | Moirai 2.0 | 0.41 | 0.16 | 0.52 |
| 4,096 | 256 | Direct | Chronos-2 | 0.20 | 0.10 | 0.38 |
| 4,096 | 256 | Direct | TimesFM | 0.15 | 0.07 | 0.46 |
| 4,096 | 256 | Direct | Moirai 2.0 | 0.52 | 0.41 | 0.52 |
| 4,096 | 256 | Extracted | Chronos-2 | 0.14 | 0.08 | 0.49 |
| 4,096 | 256 | Extracted | TimesFM | 0.16 | 0.10 | 0.51 |
| 4,096 | 256 | Extracted | Moirai 2.0 | 0.15 | 0.11 | 0.44 |
| 4,096 | 512 | Direct | Chronos-2 | 0.36 | 0.21 | 0.42 |
| 4,096 | 512 | Direct | TimesFM | 0.19 | 0.08 | 0.42 |
| 4,096 | 512 | Direct | Moirai 2.0 | 0.64 | 0.52 | 0.63 |
| 4,096 | 512 | Extracted | Chronos-2 | 0.23 | 0.10 | 0.45 |
| 4,096 | 512 | Extracted | TimesFM | 0.20 | 0.10 | 0.55 |
| 4,096 | 512 | Extracted | Moirai 2.0 | 0.23 | 0.12 | 0.49 |
| 4,096 | 1,024 | Direct | Chronos-2 | 0.66 | 0.41 | 0.50 |
| 4,096 | 1,024 | Direct | TimesFM | 0.33 | 0.10 | 0.39 |
| 4,096 | 1,024 | Direct | Moirai 2.0 | 0.69 | 0.44 | 0.67 |
| 4,096 | 1,024 | Extracted | Chronos-2 | 0.39 | 0.16 | 0.44 |
| 4,096 | 1,024 | Extracted | TimesFM | 0.24 | 0.10 | 0.55 |
| 4,096 | 1,024 | Extracted | Moirai 2.0 | 0.34 | 0.13 | 0.51 |
| 8,192 | 256 | Direct | Chronos-2 | 0.17 | 0.08 | 0.33 |
| 8,192 | 256 | Direct | TimesFM | 0.15 | 0.07 | 0.40 |
| 8,192 | 256 | Direct | Moirai 2.0 | 0.55 | 0.44 | 0.56 |
| 8,192 | 256 | Extracted | Chronos-2 | 0.19 | 0.09 | 0.47 |
| 8,192 | 256 | Extracted | TimesFM | 0.21 | 0.10 | 0.49 |
| 8,192 | 256 | Extracted | Moirai 2.0 | 0.22 | 0.13 | 0.44 |
| 8,192 | 512 | Direct | Chronos-2 | 0.24 | 0.12 | 0.38 |
| 8,192 | 512 | Direct | TimesFM | 0.18 | 0.08 | 0.39 |
| 8,192 | 512 | Direct | Moirai 2.0 | 0.66 | 0.55 | 0.69 |
| 8,192 | 512 | Extracted | Chronos-2 | 0.23 | 0.11 | 0.45 |
| 8,192 | 512 | Extracted | TimesFM | 0.20 | 0.10 | 0.54 |
| 8,192 | 512 | Extracted | Moirai 2.0 | 0.18 | 0.10 | 0.47 |

*Continued on next page*

*Table 2.* Subject-averaged forecasting performance for RR interval prediction across the 18-subject cohort, context lengths, and prediction horizons. Direct denotes direct RR interval prediction; extracted denotes RR intervals derived from model-predicted waveforms via peak extraction. *Continued.*

| Context Length | Prediction Horizon | Output Type | Model | RMSE (s) ↓ | MAE (s) ↓ | KS ↓ |
|---|---|---|---|---|---|---|
| 8,192 | 1,024 | Direct | Chronos-2 | 0.59 | 0.31 | 0.47 |
| 8,192 | 1,024 | Direct | TimesFM | 0.34 | 0.10 | 0.37 |
| 8,192 | 1,024 | Direct | Moirai 2.0 | 0.76 | 0.52 | 0.74 |
| 8,192 | 1,024 | Extracted | Chronos-2 | 0.41 | 0.16 | 0.44 |
| 8,192 | 1,024 | Extracted | TimesFM | 0.25 | 0.11 | 0.54 |
| 8,192 | 1,024 | Extracted | Moirai 2.0 | 0.29 | 0.12 | 0.49 |