Institute for Research & Executive Education

COMPUTER SCIENCE AND ENGINEERING

Time Series Forecasting with Chronos-2: An Evaluation on Wikipedia Pageviews Data

by

Bùi Hoàng Nhân

Student ID:21040009

Da Nang, 2019

Time Series Forecasting with Chronos-2: An Evaluation on Wikipedia Pageviews Data

Table of Contents

Abstract

This study investigates the performance of Chronos-2, a foundation model for time series forecasting, on Wikipedia pageview data for the Bitcoin article. Foundation models have demonstrated remarkable success across various machine learning domains, and their recent application to time series forecasting presents an opportunity to evaluate their effectiveness against traditional methods. This study compares zero-shot Chronos-2 against three strong baselines—Seasonal Naive, Exponential Smoothing (ETS), and Gradient Boosting with engineered lag features—using rolling-origin backtesting with five folds and rigorous statistical testing. Our results show that Gradient Boosting with engineered lag features significantly outperforms Chronos-2 (MASE 0.344 vs. 0.394, p=0.047), despite Chronos-2's zero-shot capability. Analysis reveals that explicit feature engineering capturing weekly seasonality provides substantial advantages for this domain-specific forecasting task. However, Chronos-2 demonstrates competitive performance without requiring domain expertise or feature engineering, suggesting its value for diverse time series portfolios where per-series optimization is impractical.

Keywords: Time series forecasting, Foundation models, Chronos-2, Wikipedia pageviews, Zero-shot learning, MASE, Rolling-origin backtesting

1. Introduction

Foundation models have revolutionized machine learning across multiple domains, from natural language processing to computer vision. Recently, this paradigm has been extended to time series forecasting, with models such as Chronos-2 offering the promise of zero-shot prediction capabilities across diverse temporal data. This study evaluates whether zero-shot foundation models can match or exceed traditional time series methods on a real-world forecasting task.

The research question addressed is whether Chronos-2, a pre-trained foundation model requiring no domain-specific training, can outperform carefully engineered traditional approaches on Wikipedia pageview forecasting. Wikipedia pageviews represent a compelling test case due to their public availability, clear temporal patterns, and practical relevance for content planning and resource allocation.

This study pursues four primary objectives. First, it compares zero-shot Chronos-2 against three strong baselines representing different forecasting paradigms. Second, it evaluates the calibration quality of Chronos-2's probabilistic forecasts. Third, it assesses the statistical significance of performance differences using appropriate non-parametric tests. Fourth, it analyzes error patterns and failure modes across models to inform practical deployment decisions.

The remainder of this report is organized as follows: Section 2 describes the dataset and preprocessing steps; Section 3 defines the forecasting problem setup; Section 4 details the methods and evaluation framework; Section 5 presents results; Section 6 discusses findings and implications; and Sections 7-9 address ethical considerations, reproducibility, and conclusions.

2. Dataset

2.1 Data Source and Collection

The dataset was obtained from the Wikimedia Pageviews API (https://wikimedia.org/api/rest_v1/), which provides access to daily pageview counts for all Wikipedia articles. The Bitcoin Wikipedia page was selected as the target series due to its sustained public interest and characteristic temporal patterns. Data

spans January 1, 2020 through December 31, 2024, comprising five years of daily observations. All Wikipedia content is licensed under CC BY-SA 3.0.

2.2 Preprocessing

The raw dataset contained 1,827 daily observations after cleaning. Missing value analysis confirmed complete data coverage with zero missing values detected. Outlier detection identified 38 extreme values (2.08% of data), which were addressed using winsorization at the 1st and 99th percentiles, establishing bounds of 3,563 to 36,912 pageviews. This approach preserves the temporal structure while mitigating the influence of extreme spikes. Frequency validation confirmed consistent daily observations with no gaps in the time series.

2.3 Data Split

Temporal ordering was strictly preserved to prevent data leakage. The training set comprises 60% of observations (1,096 days, January 1, 2020 through December 31, 2022). The validation set contains 20% (365 days, January 1, 2023 through December 31, 2023). The test set contains the remaining 20% (366 days, January 1, 2024 through December 31, 2024). This split ensures that all model development and hyperparameter decisions are made without access to test data.
[Insert Figure 1: train_val_test_split.png here]

---

3. Problem Setup

The forecasting task is defined as univariate time series prediction with a forecast horizon of H=30 days. No exogenous variables are incorporated, isolating the evaluation to each model's ability to capture patterns from historical pageview data alone.
Seasonal decomposition using STL confirmed strong weekly seasonality with period m=7, reflecting predictable day-of-week patterns in Wikipedia browsing behavior. This weekly cycle is clearly visible in the decomposition, with higher pageviews typically occurring on weekdays compared to weekends.
[Insert Figure 2: seasonality_decomposition.png here]
The Mean Absolute Scaled Error (MASE) serves as the primary evaluation metric. MASE is preferred over percentage-based metrics because it is scale-independent, symmetric, and well-defined even when actual values approach zero. A MASE value below 1.0 indicates that the forecast outperforms a naive seasonal baseline.

---

4. Methods

4.1 Models Evaluated

Seasonal Naive serves as the simplest baseline, forecasting each future value as the observed value from the same weekday in the previous week (seasonal period m=7). Despite its simplicity, this method provides a strong baseline for data exhibiting regular seasonal patterns.
Exponential Smoothing (ETS) implements the Holt-Winters method with additive trend and additive seasonality components. Parameters are optimized via Maximum Likelihood Estimation using the statsmodels library. The seasonal period is set to 7 to capture weekly patterns.
Gradient Boosting employs the LightGBM algorithm with engineered temporal features. The feature set includes lag values at 1, 7, 14, and 28 days; rolling mean statistics over 7 and 28-day windows; rolling

standard deviation over 7 and 28-day windows; and day-of-week encoding (0-6). Hyperparameters are fixed at 100 trees, maximum depth of 5, and learning rate of 0.05. Multi-step forecasting uses recursive (autoregressive) prediction.

Chronos-2 represents the foundation model approach. The amazon/chronos-t5-base variant was used, featuring a T5 encoder-decoder architecture with 200 million parameters. Critically, Chronos-2 operates in zero-shot mode with no fine-tuning on the Wikipedia pageviews data. The full training history is provided as context, and probabilistic forecasts are generated via 20 samples, from which quantiles at τ=0.1, 0.5, and 0.9 are extracted. Inference was performed on an NVIDIA RTX 3050 GPU (4GB VRAM) with CUDA 12.1, requiring approximately 3 seconds per fold.

4.2 Evaluation Framework

Rolling-origin backtesting with an expanding window was employed to simulate realistic forecasting conditions. Five folds were constructed, each producing 30-day-ahead forecasts. Fold 1 trains on days 1-1311 and forecasts days 1312-1341; Fold 2 trains on days 1-1341 and forecasts days 1342-1371; Fold 3 trains on days 1-1371 and forecasts days 1372-1401; Fold 4 trains on days 1-1401 and forecasts days 1402-1431; Fold 5 trains on days 1-1431 and forecasts days 1432-1461. Strict temporal ordering ensures no future information leaks into training data.

Point forecast metrics computed include MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), sMAPE (Symmetric Mean Absolute Percentage Error), and MASE. Probabilistic forecast metrics for Chronos-2 include pinball loss at quantiles τ=0.1, 0.5, and 0.9; interval coverage for the 80% nominal level; and mean interval width.

Statistical significance was assessed using the Wilcoxon Signed-Rank Test, a paired non-parametric test appropriate for comparing forecast errors without distributional assumptions. All models were compared against Gradient Boosting (the best baseline) at significance level α=0.05.

---

5. Results

5.1 Validation Performance

Table 1 presents cross-validation results across five folds. Gradient Boosting achieved the best validation performance (MASE=0.344±0.116), followed by Chronos-2 (MASE=0.394±0.079), Seasonal Naive (MASE=0.397±0.096), and ETS (MASE=0.508±0.036).

Table 1: Validation Performance (5-Fold Cross-Validation)

Model	MAE (mean ± std)	RMSE (mean ± std)	sMAPE (mean ± std)	MASE (mean ± std)
Gradient Boosting	**873 ± 293**	**1,150 ± 471**	**14.3% ± 3.9%**	**0.344 ± 0.116**
Chronos-2	999 ± 200	1,375 ± 428	16.7% ± 2.0%	0.394 ± 0.079
Seasonal Naive	1,007 ± 243	1,348 ± 385	16.6% ± 2.9%	0.397 ± 0.096
ETS	1,288 ± 92	1,662 ± 323	23.0% ± 3.3%	0.508 ± 0.036

[Insert Figure 5: mase_by_fold.png here]

5.2 Test Set Performance

Table 2 reports final holdout performance. All models show substantially higher errors on the test set compared to validation, with MASE values around 1.0-1.1 versus 0.34-0.40 during validation. This

indicates that the 2024 test period exhibited greater volatility than the 2023 validation period.

Table 2: Test Set Performance (Final Holdout)

Model	MAE	RMSE	sMAPE	MASE
Gradient Boosting	**2,739**	**4,231**	**35.8%**	**1.080**
Seasonal Naive	2,745	4,407	35.8%	1.082
Chronos-2	2,836	4,722	37.1%	1.118
ETS	6,984	8,453	149.5%	2.754

[Insert Figure 3: test_forecasts.png here]

5.3 Statistical Significance

Table 3 presents Wilcoxon Signed-Rank Test results comparing all models against Gradient Boosting. All comparisons yield statistically significant differences at $\alpha=0.05$.

Table 3: Statistical Significance Tests (Wilcoxon Signed-Rank)

Comparison	Test Statistic	p-value	Significant
Seasonal Naive vs. GB	4,527.0	0.033	Yes
ETS vs. GB	3,093.0	<0.001	Yes
Chronos-2 vs. GB	4,601.5	0.047	Yes

5.4 Probabilistic Forecast Quality

Table 4 summarizes Chronos-2's probabilistic calibration. The 80% prediction intervals achieve only 14.2% empirical coverage, indicating substantial under-coverage and overconfident predictions.

Table 4: Probabilistic Forecast Quality (Chronos-2)

Metric	Value	Expected
80% Interval Coverage	14.2%	80%
Mean Interval Width	769 pageviews	—
Pinball Loss (τ=0.1)	156	—
Pinball Loss (τ=0.5)	499	—
Pinball Loss (τ=0.9)	426	—

[Insert Figure 8: calibration_curve.png here]

---

6. Discussion

6.1 Interpretation of Main Findings

The superiority of Gradient Boosting is attributed to effective feature engineering that explicitly captures the strong weekly seasonality (m=7) present in Bitcoin Wikipedia pageviews, as evidenced by lag_7 contributing 35% of feature importance. The combination of short-term momentum (lag_1 at 22% importance), weekly patterns (lag_7 at 35%), and longer-term trends (rolling_mean_28 at 12%) provides a comprehensive representation of the data's temporal structure.

[Insert Figure 6: feature_importance.png here]

Chronos-2 demonstrates competitive performance despite operating in zero-shot mode without any training on Wikipedia data. The 14% MASE gap (0.344 vs. 0.394) represents the cost of generalization—Chronos-2's architecture is designed for broad applicability rather than optimal performance on any single

series. This trade-off may be acceptable in production settings where maintaining separate models for thousands of time series is impractical.

6.2 Error Analysis

Error analysis by forecast horizon reveals consistent degradation across all models, with MAE approximately doubling from h=1 (~600) to h=30 (~1,200). This pattern reflects the fundamental challenge of multi-step forecasting where uncertainty compounds over time.
[Insert Figure 4: error_by_horizon.png here]
Analysis by pageview level shows that all models struggle with high-traffic periods (MAE ~1,500-2,000 for >66th percentile), compared to low-traffic periods (MAE ~500-700 for <33rd percentile). Viral spikes and news events create inherently unpredictable volatility.
[Insert Figure 7: error_by_level.png here]
Chronos-2's prediction intervals showed significant under-coverage (14.2% vs. 80% nominal), likely due to the high volatility of the 2024 test period and the model's conservative zero-shot calibration. This finding highlights the importance of post-hoc calibration when deploying foundation models for uncertainty quantification.

6.3 Practical Implications

For single-series forecasting with clear domain patterns, carefully engineered features outperform zero-shot foundation models. Practitioners working with specific Wikipedia pages or similar data should invest in feature engineering. However, for diverse time series portfolios requiring rapid deployment, Chronos-2 offers substantial value through its zero-shot capability and probabilistic outputs.

6.4 Limitations

This study has several limitations. First, only the Bitcoin Wikipedia page was evaluated; results may not generalize to other pages or domains. Second, Chronos-2 was used with default settings without hyperparameter optimization, while Gradient Boosting used fixed rather than tuned hyperparameters. Third, only H=30 was tested; performance at other horizons remains unknown. Fourth, no exogenous features (news events, holidays) were incorporated. Fifth, the 2024 test period exhibited unusually high volatility. Sixth, Chronos-2 requires GPU resources (minimum 4GB VRAM), which may limit deployment options.

---

7. Ethical Considerations

This study uses publicly available Wikipedia pageview data, which contains no personally identifiable information. All data is aggregated at the page level and does not reveal individual user behavior. The dataset is licensed under CC BY-SA 3.0.
Potential biases exist due to topic selection; Bitcoin attracts particular demographic and interest-based audiences, and findings may not generalize to other Wikipedia topics. The forecasting models developed should not be used for market manipulation or to mislead investors about cryptocurrency trends.
Environmental impact warrants consideration. GPU inference for Chronos-2 contributes to carbon emissions; however, the computational cost (~3 seconds per fold on consumer hardware) is modest compared to training foundation models from scratch.

---

8. Reproducibility

All code, data, and results are publicly available at https://github.com/Vipproplayerone1/ts-chronos-gpu. The random seed is fixed at 42 for all stochastic operations. The computational environment includes Python 3.10, PyTorch 2.5.1, chronos-forecasting 2.2.0, and CUDA 12.1. Hardware requirements include an NVIDIA RTX 3050 with 4GB VRAM or equivalent. The full pipeline executes in approximately 20 minutes. Detailed instructions for reproducing all experiments are provided in the repository README.

---

9. Conclusion

This study evaluated Chronos-2, a foundation model for time series forecasting, against three traditional baselines on Bitcoin Wikipedia pageview data. Gradient Boosting with engineered lag features significantly outperformed all alternatives, achieving MASE of 0.344 compared to 0.394 for Chronos-2 ($p=0.047$). The performance advantage stems from explicit feature engineering that captures known weekly seasonality patterns.

For single-series forecasting with clear domain patterns, carefully engineered features outperform zero-shot foundation models. However, Chronos-2's ease of deployment and lack of required feature engineering make it valuable for diverse time series portfolios where per-series optimization is impractical.

Future work should extend this evaluation to multiple Wikipedia pages across different topics, incorporate exogenous features such as news events and holidays, explore fine-tuning Chronos-2 on Wikipedia-specific data, and investigate ensemble methods combining Gradient Boosting's accuracy with Chronos-2's probabilistic capabilities.

References

1. Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Mahoney, M. W., Torber, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). Chronos: Learning the Language of Time Series. *arXiv preprint arXiv:2403.07815*.
2. Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. https://otexts.com/fpp3/
3. Wikimedia Foundation. (n.d.). Pageviews API. Retrieved from https://wikimedia.org/api/rest_v1/
4. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30.
5. Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6(1), 3-73.
6. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688.