

THE MITHAI SHOP CHALLENGE

When Moving Festivals Meet Time-Series Foundation Models
Independent research brief | 23N AI Labs | 27 October 2026

Research status. Controlled synthetic simulation; not a peer-reviewed paper. This brief summarizes the public 23N Labs research release (21 October 2026) for technical discussion. No real retailer transactions or realized savings are claimed.

Abstract

We examine whether pretrained time-series foundation models (TSFMs) can forecast festival-driven demand when the calendar date of a major event shifts between years. A reproducible synthetic Ahmedabad retailer models three perishable sweet products, corporate gifting, wedding-season effects, capacity-limited sales, and changing festival dates. Thirteen forecasting methods and configurations are evaluated for forecast accuracy, peak timing, probabilistic reliability, and downstream production decisions. History-only models sometimes concentrate predicted demand around prior-year festival dates. A shared external calendar correction improves accuracy across tested model families. In the simulated decision environment, accounting for stockouts and calibrating forecasts can matter as much as selecting the foundation model.

1. Research question and design

The motivation was to test TSFMs outside commonly benchmarked domains using a culturally specific but technically generalizable demand pattern. Diwali moved from 20 October 2025 to 8 November 2026, a 19-day shift. We asked: (i) can history-only methods place the demand peak correctly; (ii) does explicit knowledge of future festival dates help; and (iii) do improvements in forecasts translate into better inventory decisions?

The study simulates daily kilograms demanded and sold from 2019 through 2026 for kaju katli, motichoor ladoo and gulab jamun. Up to 30 independently randomized shops were used in the base setting, with additional stress scenarios. The generator includes weekday and seasonal patterns, festival uplift, corporate orders booked in advance, weddings, bulk orders, noise, and sales censoring when inventory is insufficient.

2. Experimental methods

Three pretrained models—Birla AI Labs Yug-1, Google TimesFM 2.5, and Amazon Chronos-Bolt—were run **zero-shot** with a maximum historical context of 2,048 daily observations and a 56-day forecast horizon. Traditional and operational comparisons included seasonal-naïve (52 weeks), ETS, calendar regression, censoring-aware calendar regression, a shopkeeper 1.10 multiplier and calibrated shopkeeper. An external calendar correction was also applied to Yug-1, TimesFM, Chronos-Bolt and ETS, producing 13 evaluated configurations.

The calendar correction was learned from past observed sales around known festival dates, then applied to future dates. It was an **application-layer adjustment**, not a native covariate feature of Yug-1. The calendar regression used ridge regression on log-transformed sales with time trend, weekdays, annual Fourier seasonality, wedding-season indicators and festival-relative date bins. Censoring-aware regression additionally addresses the distinction between observed sales and latent demand.

3. Evaluation metrics

Metric	Definition / interpretation
WAPE	Sum of absolute errors divided by total latent demand; lower is better.
MAE	Mean absolute difference between predicted and latent demand, in kg.
Weighted quantile loss	Quantile forecast error, accounting for asymmetric over/under prediction.

Metric	Definition / interpretation
80% interval coverage	Fraction of actual demand inside the model's 10th–90th percentile interval.
Peak timing error	Predicted minus actual date of maximum seven-day demand total; negative means early. Only scored for distinct peaks.
Bias-adjusted phantom surge	Excess predicted weight around previous festival dates after accounting for overall forecast bias.
Business regret	Feasible perfect-hindsight contribution minus contribution from the forecast-driven production policy; lower is better.

4. Calendar-shift findings

Method	Distinct peak	Median peak error	Phantom surge
Yug-1	48%	-17 days	+29%
TimesFM 2.5	33%	-16 days	+15%
Chronos-Bolt	23%	-19 days	+17%
Seasonal naive	100%	-20 days	+45%
ETS	0%	Not defined	+10%
Calendar-informed group	88–96%	-1 to 0 days	+4 to +11%

These are the specified 19-day shift scenarios. A model without a distinct peak does not receive a forced timing error; the distinct-peak share must therefore accompany timing comparisons.

5. Effect of explicit calendar information

WAPE improvement (percentage points)	Base	R1	R2	R3
Yug-1 + calendar	10.3	5.6	9.4	4.7
TimesFM + calendar	6.2	2.9	6.0	2.6
Chronos-Bolt + calendar	8.0	4.2	7.3	3.7
ETS + calendar	10.4	6.7	9.6	5.9

R1 varies festival shape by year; R2 introduces festival-independent local shocks; R3 combines both. These are **reductions in error**, not final WAPE values. The gains show the value of explicit event information but do not demonstrate superiority of one architecture.

6. Production-decision results

Every method was evaluated under common simulated business rules: a 56-day plan, product shelf lives, capacity limits, cost and price assumptions, and an alternative weekly replanning policy. The table reports festival-season regret per 56-day product plan after training-only bias correction.

Method	Simulated regret (INR)
Censoring-aware calendar regression	47,602
Standard calendar regression	50,855
TimesFM 2.5 + calendar	52,847
Yug-1 + calendar	52,882

Method	Simulated regret (INR)
Chronos-Bolt + calendar	54,102
Yug-1 history only	87,051
TimesFM history only	82,656
Chronos-Bolt history only	91,129

The Yug-1 + calendar versus TimesFM + calendar difference was only INR 35 per plan, with a 95% paired bootstrap interval spanning -1,914 to +1,603: **inconclusive**. Censoring-aware calendar regression had lower regret than Yug-1 + calendar by INR 5,280 (95% interval -7,302 to -3,208). These are synthetic scenario-specific results, not realized profits.

7. Sampling sensitivity and validation

Yug-1 generated 100 sampled paths per forecast. Across five seeds (67067, 1, 2, 3 and 4), nominal 80% festival-period coverage ranged from 68% to 80%; festival WAPE ranged from 41.1% to 42.6%. The original seed was favorable to Yug-1 in 15 of 16 business comparisons, motivating the final multi-seed analysis. This suggests a useful investigation of Monte Carlo variability and per-series random sampling; it does not by itself prove an implementation defect.

The public verification record reports 42 archived synthetic forecast files and 241,920 forecast rows, with no nonfinite or crossed-quantile rows and 18/18 independently recomputed metric matches. The publication did not rerun inference. Approximately 367 MB of raw synthetic forecast archives remain outside the public repository; code, notebooks, pinned model revisions, result JSON, checksums and audit scripts are public.

8. Technology and reproducibility

The documented workflow uses Python 3.11, NumPy, pandas, SciPy, statsmodels, Jupyter notebooks, local CPU computation and a free Kaggle Tesla T4 GPU. Forecast archives use compressed NPZ files; summarized metrics and evidence are recorded in JSON/Markdown. Festival dates are a versioned factual compilation from Office Holidays with a Drik Panchang cross-check for Diwali, rather than a dynamically refreshed prediction-time API. Model revisions and source histories were pinned and hash-checked.

9. Limitations and next experiments

The dataset is synthetic, so effect sizes depend on simulator assumptions. Most robustness findings were developed post hoc; bootstrap intervals describe variability across simulated shops and model seeds rather than real-world population uncertainty. Native covariate capabilities were not evaluated. The next useful experiments would compare native versus external calendar inputs where supported, validate on consented real retailer data, evaluate uncertainty calibration across seeds and series, and vary the business decision policy.

10. 23N AI Labs and research collaboration

23N AI Labs is a founder-led independent applied AI research and engineering initiative. Its work combines controlled model evaluation, reproducible experimentation, agent-assisted research workflows and the engineering of specialized AI applications. The lab's role is not to replace foundational-model researchers, but to investigate model behavior in challenging settings and help translate results into reliable systems. A practical follow-up to this study would be a jointly defined, scoped evaluation question with agreed inputs, success metrics, publication rights and review responsibilities.

References and provenance

23N Labs Research Team (2026). *The Mithai Shop Challenge*, public release, 21 October 2026.

<https://github.com/23nlabs/mithai-shop-challenge>

Primary technical source: <https://github.com/23nlabs/mithai-shop-challenge/blob/main/final/technical-appendix.md>

Reproduction instructions: <https://github.com/23nlabs/mithai-shop-challenge/blob/main/REPRODUCIBILITY.md>

This discussion paper is an edited summary of the public research release, not a new peer-reviewed study. Independent of and not endorsed by the model publishers.

Terminology, Model Attribution and References

Supplement to *The Mithai Shop Challenge: When Moving Festivals Meet Time-Series Foundation Models*. Definitions are specific to the published evaluation where indicated.

A. Terminology and notation

Term	Definition
TSFM	Time-Series Foundation Model; a broadly pretrained model for forecasting time-indexed numeric observations.
Zero-shot inference	Applying a pretrained model to new series without task-specific fine-tuning.
Context length / horizon	Number of historical observations supplied / future time steps predicted. This study used up to 2,048 days and a 56-day horizon.
Covariate	Additional explanatory feature. Known future covariates may include festival dates, but model-specific interfaces differ.
Seasonal naïve	Baseline that repeats a value from a corresponding earlier season; this study used a 52-week offset.
ETS	Error–Trend–Seasonality family of statistical forecasting models; the implemented history-only baseline used weekly additive seasonality.
Calendar-aware regression	Ridge regression on $\log(1 + \text{sales})$ using trend, weekday, annual Fourier terms, wedding indicators and festival-relative bins.
Censoring / latent demand	Recorded sales may be capped by stock availability, concealing customers' true demand. Latent demand is the simulator's uncapped target.
WAPE / MAE	Weighted Absolute Percentage Error = total absolute error / total demand; Mean Absolute Error = mean absolute error in kilograms.
wQL / quantiles	Weighted Quantile Loss evaluates probabilistic forecasts; quantiles (e.g., P10, P50, P90) describe forecast distribution levels.
Coverage	Observed fraction of outcomes inside a prediction interval; nominal 80% coverage used P10–P90.
Peak timing error	Predicted minus actual date of a distinct maximum seven-day demand window; negative indicates an early prediction.
Phantom surge	Extra forecast weight around previous festival dates; the bias-adjusted version removes overall level bias.
Regret	Feasible perfect-hindsight simulated contribution minus forecast-policy simulated contribution; lower is better.
Bootstrap interval	Uncertainty interval estimated by repeated resampling; the study resampled simulated shops and Yug-1 seeds.
Ablation / information parity	Ablation changes one input or component; information parity means comparisons use equivalent available information.

B. Model attribution

Birla AI Labs — Yug-1. A 271.8-million-parameter zero-shot probabilistic time-series foundation model. The official citation credits Aaditya Jain, Debdeep Sanyal, Aaryan Nagpal, Dhruv Kumar, Murari Mandal and Saurabh Deshpande (2026). The model card also acknowledges Varya Srivastava and additional contributors. The study used a pinned Yug-1 checkpoint and package version 0.1.1.

Google Research — TimesFM 2.5. A later checkpoint in the TimesFM family. The foundational paper, *A Decoder-Only Foundation Model for Time-Series Forecasting*, is by Abhimanyu Das, Weihao Kong, Rajat Sen and

Yichen Zhou (ICML 2024). The experiment used TimesFM 2.5, not the original TimesFM checkpoint.

Amazon — Chronos-Bolt. The experiment used *Chronos-Bolt base*, from the Amazon Chronos model family. The associated paper, *Chronos: Learning the Language of Time Series*, is by Abdul Fatir Ansari and coauthors (Transactions on Machine Learning Research, 2024). It should not be called “Cortova” or conflated with Chronos-2.

C. References

- [1] Jain, A., Sanyal, D., Nagpal, A., Kumar, D., Mandal, M., & Deshpande, S. (2026). *Yug: A foundation model for zero-shot probabilistic time-series forecasting*. Birla AI Labs. <https://github.com/birla-ai-labs/yug>
- [2] Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). *A decoder-only foundation model for time-series forecasting*. Proceedings of ICML. <https://openreview.net/forum?id=jn2iTJas6h>
- [3] Ansari, A. F., Stella, L., Turkmen, C., et al. (2024). *Chronos: Learning the language of time series*. Transactions on Machine Learning Research. <https://www.amazon.science/publications/chronos-learning-the-language-of-time-series>
- [4] Google Research. *TimesFM 2.5 checkpoint and documentation*. <https://huggingface.co/google/timesfm-2.5-200m-transformers>
- [5] Amazon. *Chronos-Bolt base model card*. <https://huggingface.co/amazon/chronos-bolt-base>
- [6] 23N Labs Research Team. (2026). *The Mithai Shop Challenge* (synthetic controlled study; code, methodology and results). <https://github.com/23nlabs/mithai-shop-challenge>
- [7] 23N Labs Research Team. (2026). *Technical Appendix, final v3*. <https://github.com/23nlabs/mithai-shop-challenge/blob/main/final/technical-appendix.md>

D. Acknowledgments and scope

We acknowledge Birla AI Labs for making Yug-1 available with documented model versions; Google Research for TimesFM; and Amazon for Chronos and Chronos-Bolt. This is an independent evaluation and is not affiliated with or endorsed by any model publisher. Model names and paper citations identify the underlying third-party research; all numerical retailer outcomes in this report are simulated.

Scientific caution: The study did not test native future-covariate interfaces across all model families. Calendar-aware variants used a separate application-layer correction. Most robustness results are post hoc, and the raw forecast archives are not included in the public repository.