

A Pre-Registered Zero-Shot Benchmark of Time-Series Foundation Models against a Deployed Probabilistic Forecaster

Indivd Analytics

October 1, 2026

Abstract

We report a pre-registered benchmark of three 2026-class zero-shot time-series foundation models (TSFMs) — Chronos-Bolt-small, TimesFM-2.x-small and Moirai-small — against the probabilistic footfall forecaster our platform deploys in production (the *served stack*). The entire evaluation specification, including a falsifiable decision rule, was frozen and made public *before* any estimation ran: it carries sha256 `5ac1f9b97a232ddd6e6dc7b2a5ca9f4e4807fd5b288944be3f4eafcc45957e9b` and is available at its registered URL [5]. Every model — and two tuning-free naive baselines — was scored on exactly the same evaluation points: **20091** same-day band forecasts (issued for the day they forecast, “horizon-0”) over **133** benchmark-eligible stores spanning **3** ISO weeks, with the data cut at 2026-10-01T17:24:32Z. Accuracy (SMAPE), signed bias, prediction-interval coverage at each model’s own nominal level, and pre-defined failure modes were computed with identical formulas for every model. Two of the three challengers beat the served stack materially on accuracy: the served median SMAPE was **44.228** versus **34.109** for Chronos-Bolt-small and **32.358** for TimesFM-2.x-small (deltas **+10.12** and **+11.87** SMAPE points), with the win holding in the low-volume tercile; Moirai-small is a clear outlier (**124.927**; delta **-80.70**). Critically, however, the pre-registered decision rule also requires *failure-mode non-regression* — a challenger may not introduce new failure modes such as flat, low-variety forecasts on stores that the served stack handles fine — and both winning models fail it. The recorded verdict, quoted verbatim from the frozen rule evaluation, is: **“accuracy-only win, disqualified by chronos (failed cond3_failure_nonregression); timesfm (failed cond3_failure_nonregression)”**. In other words the honest outcome is an *accuracy-only* result that is *disqualified* from promotion: no TSFM is promoted to a material zero-shot win, and the served stack is not declared the winner either. A secondary, non-gating comparison on a longer-history upstream zone surface — where no served baseline exists — finds that a seasonal-naive baseline is the most accurate model overall (median SMAPE **23.66** versus **41.10** for the best TSFM), i.e. the zero-shot challengers transfer *poorly* to the long-history, highly intermittent target. All numbers in this paper are generated by code from frozen result artifacts; none is hand-typed. The benchmark decides nothing about serving; adoption is a separate governance decision.

1 Problem and Motivation

Foundation models pre-trained on large time-series corpora now promise competent forecasting *without* per-series training — a zero-shot transfer regime [1, 3, 8], with dedicated benchmark suites emerging for exactly this evaluation [6]. If a general model can match or beat a fleet-tuned production stack on our own data, with no per-store fitting, that has direct operational value: it would collapse the cost of onboarding new stores and would offer a credible challenger for the populations our own modelling refuses to train on (short, sparse or degenerate series). This question must be answered honestly, because the temptation to overtake a single headline metric is exactly what a benchmark is supposed to check.

We therefore pose a narrow, falsifiable question: *does a 2026-class TSFM, used zero-shot on identical held-out windows and identical pairs, materially beat the deployed served forecaster on the fleet’s hourly visitation series — including on the known hard classes (low-volume stores, weekend and holiday structure)?* “The served stack wins” is an explicitly valid outcome, and so is “no material difference”.

1.1 Pre-registration before estimation

The methodological core of this study is a discipline rather than a result: the entire specification — models, window, pair-set construction, metric formulas and a falsifiable decision rule — was frozen and pub-

lished *before* any estimation ran [5]. The frozen specification carries sha256 5ac1f9b97a232ddd6e6dc7b2a5ca9f4e4807fd5b288944be3f4eafcc45957e9b (over the document excluding its own hash line) and is archived at

<https://analytics.indivd.com/papers/tsfm-prereg-phase1-2026-10-01-5ac1f9b9.md>

which serves as the registration record. Nothing in that specification may be edited or reinterpreted here; where this paper paraphrases it, the specification governs. The value of the pre-registration is that the decision rule, including the clause that ultimately disqualifies the accuracy win, was fixed in advance and could not be tuned after seeing the numbers. The registration record is published on this platform’s own Papers page, and its integrity rests on the recorded hash matching the served document; we disclose that no independent third-party registry timestamp exists for it. Disciplined, pre-specified comparison design of this kind is standard practice in large empirical forecasting evaluations such as the M5 competition [7], where strong baselines and fixed evaluation protocols were fixed before the fact.

1.2 Scope, terminology and anonymity

All stores are referred to only by stable anonymised labels $L_1, \dots, L_n$ (the 1-indexed order of the internal location identifier among active locations, ascending); no client, brand, venue or city is named. All paper-facing quantities are relative or scale-free where a raw magnitude could be linking, and the raw identifiers appear only in machine-readable files marked sensitive. The evaluated surface is the fleet’s hourly store-visitation series, ‘location_visits’, and a secondary upstream zone-visitation series.

2 Method

2.1 The challenger: strict zero-shot

The binding challenger definition is *zero-shot*: a TSFM receives the univariate historical series only and is used with *no* per-store training, fine-tuning or calibration on our data — no weather, no calendar features, no hyperparameter search. The served stack, by contrast, consumes an exogenous feature surface (including weather and calendar structure) and is tuned on our data. The comparison is therefore “as-deployed served stack versus zero-config transfer models” and is reported as such; the structural disadvantage of the challenger is disclosed up front and is not “fixed” by any per-store tuning, because doing so would break the zero-shot definition.

2.2 Pair-set construction (identical across all models)

The primary (gated; cohort gates defined in §3.2) surface is the served-pair set: public forecast band rows (prediction interval, confidence level 0.9, prediction type location, metric ‘location_visits’) joined to the canonical observed visitation series on exact timestamp, restricted to same-day horizon-0 rows (**generated_at date = timestamp date**) with the latest vintage taken per (store, timestamp). This yield is the *evaluated pair set*; every model — including the served baseline and both naive baselines — is scored on the identical pairs.

For each pair (ℓ, t) every challenger receives the *identical* continuous hourly grid of the store’s history up to 00:00 UTC of t ’s UTC date, with every absent hour (closed or missing) filled with 0.0; the model forecasts the full target day in one call and the step at t ’s hour-of-day is extracted. The fixed context window is **2048** hours; in practice every store’s history exceeded the window and was truncated to it (all **2093** forecast tasks), so the effective context is the most recent **2048** hours for every store-day. Context strictly precedes the target day, matching the served stack’s horizon-0 information-time semantics. Two tuning-free naive baselines are computed on the same grid: seasonal-naive ($t - 168$ h) and a 28-day hour-of-day climatology. Prediction-interval lower bounds are floored at 0.0 (visitation counts are non-negative); point forecasts are deliberately *not* clipped, so that gross over-prediction is measured rather than masked.

2.3 Metrics (frozen formulas, identical for every model)

Per-pair SMAPE is $100 \cdot |F - A| / ((|A| + |F|) / 2)$ with the convention $\text{SMAPE}(A = 0, F = 0) = 0$; we report the pooled mean and median, a *pooled-typical* median (the pooled set excluding pairs belonging to stores the scored model classifies into a failure class, per model — so a pooled-typical comparison between

two models is not on strictly identical pairs, unlike the pooled median), signed percentage bias on pairs with $A > 0$, and the empirical prediction-interval coverage at each model’s *own* recorded nominal level — never a forced common nominal. Coverage is reported with 95% Clopper–Pearson exact intervals [2]. A secondary daily-aggregate SMAPE aggregates pairs to (store, UTC day) before scoring. Widths and calibration are compared per model at recorded nominals; a model whose coverage falls below $0.6\times$ its nominal would be flagged “poorly calibrated” (a reported defect, not a verdict gate). SMAPE is used with its known caveats in mind [4].

2.4 Failure modes and the decision rule

Two failure classes are evaluated per model, on its own outputs. A store enters the *gross-over-prediction* class if on more than 5% of its pairs $F > 10\times$ the causal maximum observed value strictly before the target day. A store enters the *flat-serving* class if its pairs contain at most five distinct point values rounded to two decimals — a reproducibility/interpretability guard, not an accuracy metric.

The pre-registered decision rule (section 6 of the frozen specification; “prereg” hereafter) is evaluated mechanically on the gated pair set. For each TSFM m let $\Delta_m = \text{served median SMAPE} - m\text{'s median SMAPE}$ (positive = the challenger is more accurate). A *material zero-shot win* by m^* requires *all* of: (1) $\Delta_{m^*} \geq 2.0$ SMAPE points; (2) m^* ’s pooled-typical median beats served by at least 5% relative; (3) *failure-mode non-regression*: no store enters m^* ’s gross-over class that is not already in the served stack’s, and m^* ’s flat-serving store count does not exceed the served stack’s count by more than two; and (4) the condition-1 win also holds on the low-volume tercile. *Served-stack-wins* is recorded iff served beats *every* TSFM by at least 2.0 points; *parity* is the residual case (no material difference, with parsimony and interpretability favouring the served stack). Any challenger that beats served by at least 2.0 points but fails (2), (3) or (4) is recorded as an *accuracy-only win*, *disqualified by* the stated failed condition — never silently promoted. A separate deep-subset read carries *no go/no-go* rule.

3 Data

3.1 One byte-frozen pull; two surfaces

The study makes one read-only pull per surface and records the sha256 of every file in the results file; there is no re-pull and no refresh. The primary surface is the served band surface described in §2.2: it yields, before cohort gating, **204** pairable locations, and the evaluation window is all horizon-0 pairs with target timestamp in $[2026-09-15\ 00:00Z, 2026-10-01T17:24:32Z)$, a same-day rolling-forecast surface (essentially all evaluable pairs are horizon-0; no zone-grain served surface exists). This surface is only about 17 days deep, so absolute series depth cannot be required of it; validity comes from the fact that *every* model sees exactly the same pairs, not from depth.

The secondary, non-gating surface is a long-history upstream zone-visitation series for the deep cohort (stores whose first upstream signal is at or before mid-2025), covering multiple seasons. A TSFM-versus-naive comparison runs there over two disjoint held-out holiday windows spanning two regimes (a winter christmas/new-year window and a spring easter window); no served baseline exists on that surface, so it is descriptive only.

3.2 Cohort gates and census

Gates are applied to the pairable location set and are computed, not assumed: at least 512 hourly rows in the store’s continuous grid, trading-day continuity at least 0.85, at least 80% non-zero open hours, and at least one in-window pair. The gates yield a benchmark cohort of **133** stores, excluding **71** pairable locations (48 excluded by the non-zero-hours gate, 14 by the continuity gate, 9 by both). The evaluated set is **20091** same-day pairs across **3** distinct ISO weeks (2026-W38, 2026-W39, 2026-W40), so the window is not flagged insufficient. Volume terciles (store-level, by total observed visits over the 28 days preceding the store’s first in-window target) contain 5847 (low), 6290 (mid) and 7954 (high) pairs; the tercile cut-points are 2152 and 6011. Every metric is reported per stratum: ISO week, weekend vs weekday, day-of-week, and volume tercile.

4 Results

4.1 Primary pooled results

Table 2 reports the pooled gated-surface metrics for all six models. The zero-shot TSFMs Chronos-Bolt-small and TimesFM-2.x-small beat the served stack on the headline accuracy metric by a wide margin: served median SMAPE **44.228** versus **34.109** (Chronos-Bolt-small) and **32.358** (TimesFM-2.x-small). The signed deltas Δ_m are **+10.12** and **+11.87** points respectively; both also win on the pooled-typical median (the model-specific failure-class exclusion) by **25.98%** and **29.71%** relative. Moirai-small is the clear outlier, at **124.927** (delta **-80.70**), systematically *under*-predicting (its median signed bias is **-71.46%**). Figure 1 shows the pooled medians and fig. 2 the deltas against the pre-registered materiality bar.

Two caveats about the served stack itself deserve honest emphasis. First, its *median* signed bias is near zero (**+2.72%**) but its *mean* bias is very large (**+208.56%**) — a heavy over-prediction tail. Chronos-Bolt-small is the best-centred model (**-0.30%**). Second, the served band at its nominal 0.90 level under-covers overall (coverage **0.706**), yet its *store-median* coverage is **0.915**: a small set of stores drives the overall miss. All models clear the “not poorly calibrated” floor, and coverage is compared at each model’s own recorded nominal (served 0.90; the TSFMs 0.80 due to decile-only quantile heads), never a forced common level (fig. 4); because the levels differ, cross-model coverage numbers are indicative, not like-for-like.

4.2 Decision rule: the accuracy win is disqualified

Table 3 evaluates the four conditions of the pre-registered rule for each TSFM. Chronos-Bolt-small and TimesFM-2.x-small each clear conditions (1), (2) and (4) — including the low-volume (sparse) tercile, where the accuracy deltas are **+3.19** and **+5.03** points — but both *fail* condition (3), the failure-mode non-regression clause: each classifies **3** stores into the flat-serving class versus the served stack’s **0** (the rule permits no more than the served count plus two). Moirai-small fails conditions (1), (2) and (4) outright, and also fails condition (3) (**8** flat-serving stores, above the served **+2** allowance). The gross-over-prediction class is empty for every model, so condition (3) turns entirely on the flat-serving guard.

No model is therefore promoted to a material zero-shot win; served-stack-wins is not triggered either (the best delta is positive and far from -2.0); and the outcome is not parity (the best delta is well beyond the materiality bar). The pre-registered residual case is recorded exactly, and the recorded verdict (prereg §6, verbatim) is:

“accuracy-only win, disqualified by chronos (failed cond3_failure_nonregression);
timesfm (failed cond3_failure_nonregression)”

The verdict code recorded in the frozen results file is **ACCURACY-ONLY WIN (DISQUALIFIED)**. This is the central honesty result of the study: on accuracy alone the zero-shot challengers are ahead, but the pre-registered non-regression clause — fixed before the numbers were seen — blocks promotion.

4.3 Stratified structure

Table 4 reports median SMAPE per stratum. The zero-shot TSFMs beat the served stack by a wide margin in the high volume tercile and by a smaller margin in the mid tercile (served **40.720** versus **36.330** for Chronos-Bolt-small and **34.799** for TimesFM-2.x-small); seasonal-naive is remarkably strong in the high-volume tercile. The served stack is weakest on the weekend stratum (median SMAPE **49.202** versus **43.242** on weekdays); this reflects its weekend under-coverage (**0.533** at nominal 0.90 versus **0.741** on weekdays) and a heavy mean over-prediction tail on both strata (weekend mean bias **+156.69%**, weekday **+218.13%**), while its weekend median bias is slightly negative. These patterns are descriptive and carry no additional verdict weight beyond the rule already evaluated.

4.4 Secondary deep-subset read (descriptive; no go/no-go rule)

Table 5 and fig. 5 report the non-gating deep-subset read on the multi-season upstream zone surface (**4527** pairs; TSFM and naive baselines only — no served baseline exists there). The honest finding is that *neither naive baseline is beaten by any zero-shot TSFM*: seasonal-naive is the most accurate model overall (median SMAPE **23.66**) and in both holiday strata, while the best TSFM (TimesFM-2.x-small,

41.10) trails it by roughly seventeen SMAPE points. Chronos-Bolt-small degrades sharply across the two regimes, and Moirai-small is far off. This strongly intermittent, open-hours-only target (closed hours are 0 by the identical grid convention) is exactly the regime where zero-shot transfer struggles; the read is reported descriptively and feeds the hard-class discussion below, carrying no verdict weight.

5 Validation and Quality

5.1 Derive-from-artifact discipline

Every number, table and figure input in this paper is produced by a single generator, `phase3/code/gen_paper.py`, reading the frozen results files (`phase2/results/output_tsfm.json` and, for the secondary deep read, `phase2/results/output_tsfm_deep.json`) plus the frozen per-store, per-stratum and cohort CSVs; no value is hand-typed. The generator writes the LaTeX body, the figure PDFs and the table fragments, and it fails loudly on any unresolved template token. Regenerating the manuscript from the frozen artifacts reproduces the LaTeX sources byte-identically; the build identity hash for this build is `c1472d22bde7`. The benchmark evaluation itself is regeneration-deterministic by construction: pinned package versions, recorded checkpoint repository identifiers, seed 42 and deterministic inference flags are recorded in the results-file metadata, and re-running the evaluation over the frozen pull reproduces the results file byte-identically (a check performed before this note was written).

5.2 Anonymity and the compiled-artifact scan

Anonymity is enforced from the first artifact: prose, tables and figure axes/annotations carry only L_k labels and scale-free quantities. Because figure text lands in the compiled PDF’s text layer, the leak scan runs over the *compiled PDF*, not only the LaTeX sources: it searches for any raw cohort identifier in identifier context and for client/venue names, and it asserts the sources contain no raw identifiers outside the sensitive machine files. Raw identifiers live only in the sensitive machine-readable artifacts (including the frozen results files, whose failure-mode store lists carry raw identifiers) and are never read into the manuscript.

5.3 Scope control

A machine-checked study-scope manifest (`study-scope.yaml`) fixes the data window, store set, join unit and as-of date, each with provenance, and a verifier proves the shipped panel obeys it. The manifest is amended only by dated entries, never by silent edits.

6 Discussion

The study delivers a clean, if uncomfortable, answer. On the gated same-day surface, two of three zero-shot TSFMs beat the deployed served stack materially on accuracy — and by a margin that is large by the standards of this estate (roughly ten to twelve SMAPE points) — yet the pre-registered rule declines to promote them because they regress on a reproducibility guard (flat-serving store count). Three readings follow.

First, the result is a genuine *accuracy* signal, not noise: the win holds on the pooled-typical median and in the low-volume tercile, precisely the sparse-store hard class the benchmark was motivated by. Second, it is *not* an adoption recommendation: the non-regression clause exists because a forecaster whose output collapses to a handful of distinct values on some stores is operationally fragile regardless of its average error, and a benchmark that ignored that would be over-claiming. Third, the contrast between the two surfaces is instructive: the same zero-shot models that lead on the short, dense, same-day location surface *trail* a trivial seasonal-naïve baseline on the long, highly intermittent, multi-season zone surface. Transfer strength is therefore surface-dependent, and a single headline number would have hidden that.

The broader lesson is the value of the pre-registration itself. The disqualifying clause was written before the numbers, so the study reports an accuracy win *and* its disqualification together, without an opportunity to move the goalposts. That, rather than any single metric, is the durable product of this exercise. One asymmetry deserves explicit acknowledgement: the pre-registered failure classes (gross over-prediction, flat-serving) bind only the challengers’ promotion prospects, while the served stack’s

own pathologies — a heavy mean over-prediction tail (mean signed bias **+208.56%**) and overall under-coverage at its nominal (**0.706** versus 0.90) — sit outside the frozen failure classes and could not have disqualified it. The benchmark decides nothing about serving; any adoption of a challenger is a separate governance decision that would have to weigh the flat-serving regression against the accuracy gain.

7 Limitations

We restate, without softening, the five frozen disclosures of the pre-registration (prereg §8):

- *Gated-cohort scope.* The gated verdict covers autumn-regime, horizon-0, location-grain, open-hours pairs only; holiday spikes and multi-regime transfer are tested only on the secondary deep-subset read, where no served baseline exists.
- *Univariate challenger versus exogenous served stack.* The challenger is history-only, whereas the served stack is exogenous and fleet-tuned; this structural disadvantage is disclosed up front and no per-store tuning is permitted to remove it.
- *Continuous-grid 0.0 fill.* The TSFM context is a continuous hourly grid with closed or absent hours filled as 0.0, while the served stack consumes its own feature surface; the evaluated pair set and metric functions are identical across models, but the input-construction difference is part of “zero-config transfer”.
- *Horizon-0 only.* The served band surface is a same-day rolling forecast, so no multi-day horizon comparison is possible; the few horizon-positive rows are excluded by a structural rule, not by the results.
- *Per-model nominals.* Coverage is compared at each model’s own recorded nominal (served 0.90; the TSFMs 0.80, their decile-only quantile heads); coarser-quantile models have wider native bands by construction, so cross-model coverage statements are made at each model’s own level, never a forced common one.

Beyond these, the flat-serving guard is a reproducibility and interpretability consideration, not an accuracy defect; it is precisely why the accuracy wins above are not promoted. Finally, this benchmark makes no statement about the served production system and implies no change to it.

References

- [1] Abdul Fatir Ansari, Lorenzo Stella, Caner Türkmen, 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 Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.
- [2] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.
- [3] Abhimanyu Das, Weihao Kong, Rajat Subhra Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.
- [4] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
- [5] Indivd Analytics. Phase-1 pre-registration spec: zero-shot time-series foundation model benchmark. <https://analytics.indivd.com/papers/tsfm-prereg-phase1-2026-10-01-5ac1f9b9.md>, 2026. Registered 2026-10-01; sha256 5ac1f9b97a232ddd6e6dc7b2a5ca9f4e4807fd5b288944be3f4eafcc45957e9b.
- [6] Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, and Bin Yang. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 5595–5606, 2025.

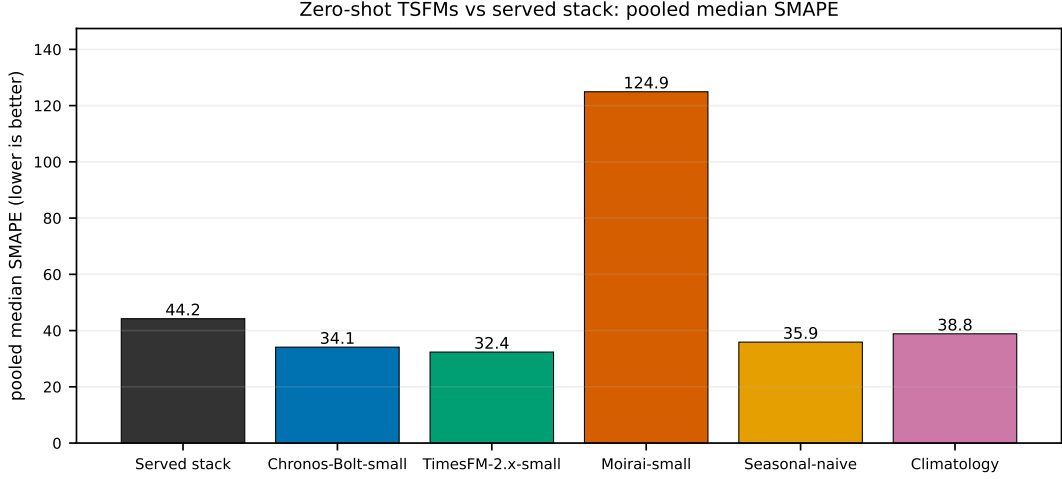


Figure 1: Pooled median SMAPE over the gated pair set, by model (lower is better). The served stack is highlighted; the two leading zero-shot TSFMs sit materially below it and Moirai-small far above. Values injected from `phase2/results/output_tsfm.json` by `phase3/code/gen_paper.py`; `tab:primary`.

model	package / version	checkpoint	point	PI nominal
Served stack	as-deployed	as-deployed	predicted value	0.90
Chronos-Bolt-small	chronos-forecasting 2.3.2	amazon/chronos-bolt-small	median ($q_{0.50}$)	0.80
TimesFM-2.x-small	timesfm 3.0.2	google/timesfm-2.5-200m-pytorch	$q_{0.50}$	0.80
Moirai-small	uni2ts 2.0.0	Salesforce/moirai-1.0-R-small	median of 100 samples	0.80
Seasonal-naive	n/a (tuning-free)	n/a	n/a	n/a
Climatology	n/a (tuning-free)	n/a	n/a	n/a

Table 1: Model shortlist and realized prediction-interval (PI) nominals. Package versions, checkpoint repository identifiers and quantile heads are recorded in the frozen results file `meta`; generated by `phase3/code/gen_paper.py`.

- [7] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4):1325–1336, 2021.
- [8] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

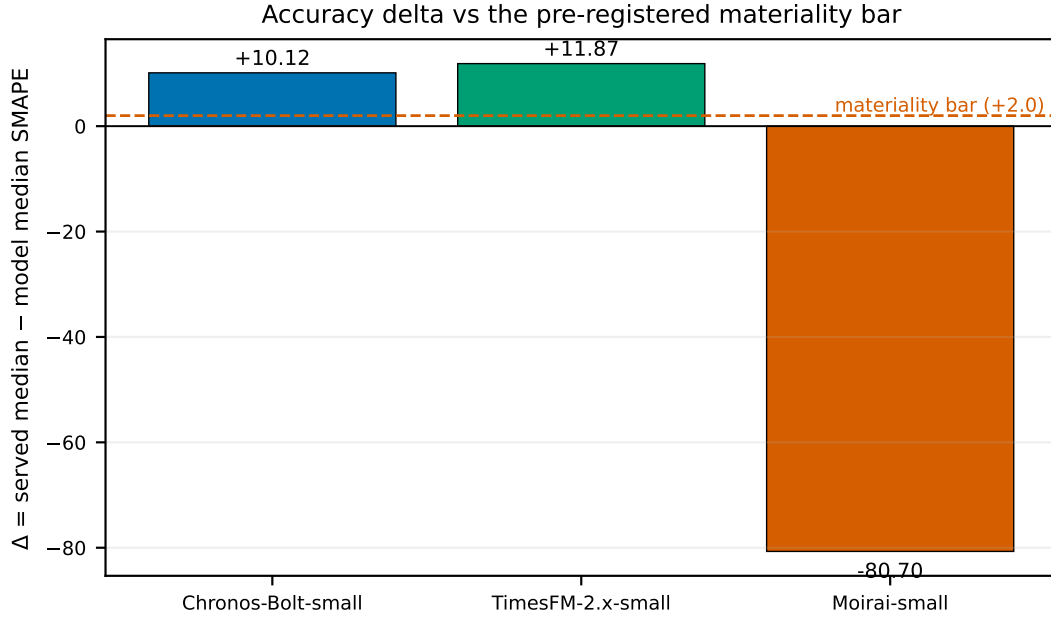


Figure 2: Accuracy delta $\Delta_m = \text{served median SMAPE} - \text{model median SMAPE}$ (positive = challenger more accurate), with the pre-registered materiality bar of two SMAPE points. The winning challengers clear the bar; the served stack does not beat every model by the bar, so *served-stack-wins* is not triggered either. Values injected from `phase2/results/output_tsfm.json`; `tab:decision`.

model	SMAPE median	SMAPE mean	pooled-typ. median	bias med	bias mean	PI cov. @nominal	daily-agg SMAPE med
Served stack	44.228	61.675	44.228	+2.72	+208.56	0.706 @0.90	24.860
Chronos-Bolt-small	34.109	54.918	32.739	-0.30	+16.27	0.808 @0.80	17.569
TimesFM-2.x-small	32.358	51.845	31.087	-4.22	+19.14	0.748 @0.80	16.597
Moirai-small	124.927	116.280	115.007	-71.46	-47.04	0.727 @0.80	109.782
Seasonal-naive	35.897	52.998	35.897	-2.14	+22.91	n/a	17.544
Climatology	38.849	53.224	38.849	+2.81	+47.64	n/a	25.221

Table 2: Primary pooled results over the gated pair set (prereg §5). SMAPE mean/median, pooled-typical median, signed % bias, prediction-interval coverage at each model’s own nominal, and daily-aggregate SMAPE. Generated from `phase2/results/output_tsfm.json`.

pre-registered condition (prereg §6)	Chronos-Bolt	TimesFM-2.x	Moirai
(1) $\Delta \geq 2.0$ pts	✓ +10.12	✓ +11.87	× -80.70
(2) pooled-typical $\geq 5\%$ rel	✓ 26.0%	✓ 29.7%	× -160.0%
(3) failure non-regression	× (flat 3)	× (flat 3)	× (flat 8)
(4) low-tercile $\Delta \geq 2.0$	✓ +3.19	✓ +5.03	× -140.93
material zero-shot win (1/2/3/4)	×	×	×

Table 3: Pre-registered decision rule evaluated mechanically (prereg §6). Conditions (1)–(4) and the material-win conjunction per TSFM; Δ is served median SMAPE minus model median SMAPE. The served stack’s flat-serving count is **0** (with zero gross-over-prediction stores for every model), so condition (3) permits at most served +2 flat-serving stores. Generated from `phase2/results/output_tsfm.json`.

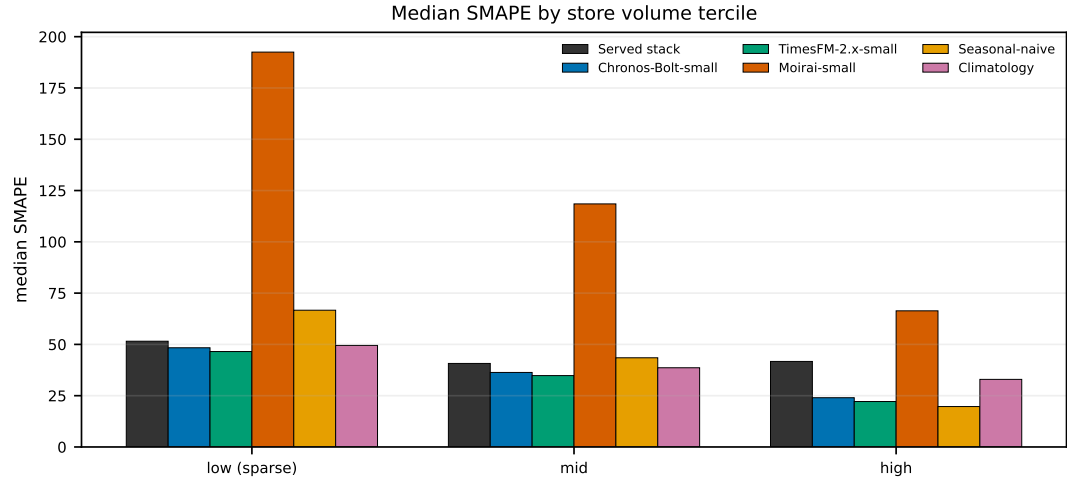


Figure 3: Median SMAPE by store volume tercile (low/mid/high) for the challengers, the served stack and the naive baselines. The zero-shot TSFMs lead in the mid and high terciles and win the low (sparse) tercile modestly. Values injected from the frozen per-stratum CSVs via `phase2/results/output_tsfm.json; tab:strata`.

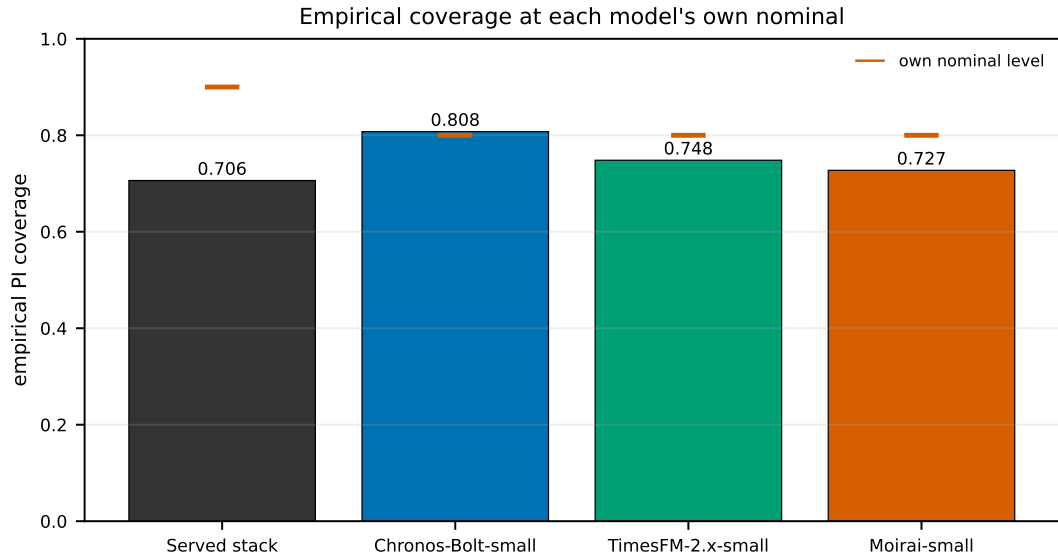


Figure 4: Empirical prediction-interval coverage by model, with each model's own recorded nominal level shown as a marker. Coverage is compared at each model's own nominal, never a forced common level. Values injected from `phase2/results/output_tsfm.json`.

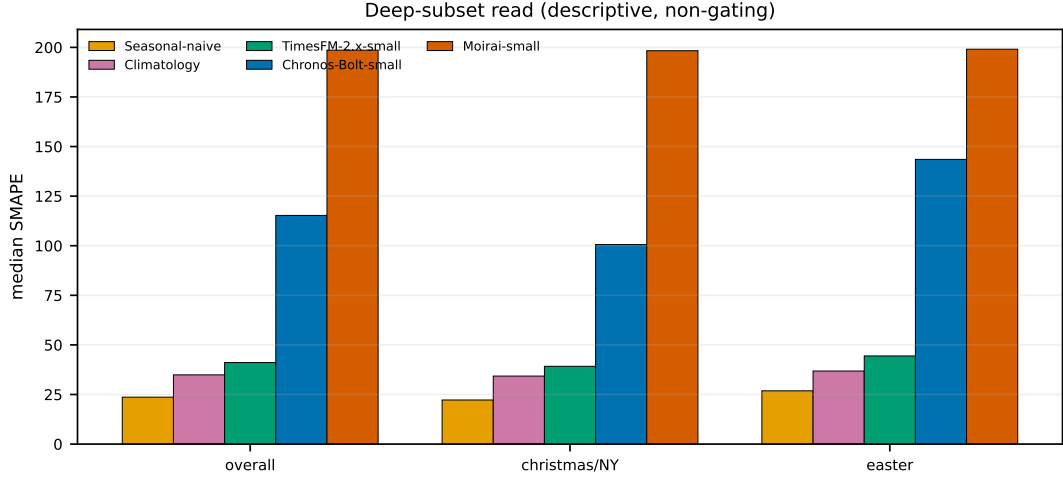


Figure 5: Secondary, non-gating deep-subset read: median SMAPE by model on the multi-season upstream zone surface, overall and in the two held-out holiday windows. No zero-shot TSFM beats seasonal-naive. No served baseline exists on this surface; no go/no-go rule applies. Values injected from `phase2/results/output_tsfm_deep.json`; `tab:deep`.

stratum	Served stack	Chronos-Bolt-small	TimesFM-2.x-small	Moirai-small	Seasonal-naive	Climatology
2026-W38	48.010	34.497	32.032	114.321	34.783	39.389
2026-W39	44.709	35.539	34.096	130.887	36.364	40.827
2026-W40	39.770	31.228	30.051	129.249	38.095	35.644
weekday	43.242	33.095	31.518	121.440	37.736	36.501
weekend	49.202	41.275	37.274	143.376	28.722	56.209
Fri	42.329	34.055	33.132	122.017	35.294	34.211
Mon	45.046	36.055	31.713	160.611	40.000	37.876
Sat	51.483	42.983	38.445	156.290	34.783	62.159
Sun	44.728	36.708	34.795	104.372	20.690	44.290
Thu	39.829	32.606	31.160	102.407	40.000	36.620
Tue	46.481	32.010	30.110	121.380	37.500	36.715
Wed	43.135	31.892	31.538	112.002	35.484	36.866
low	51.553	48.361	46.522	192.487	66.667	49.505
mid	40.720	36.330	34.799	118.487	43.478	38.615
high	41.698	24.011	22.160	66.367	19.708	32.953

Table 4: Median SMAPE per stratum (ISO week, weekend/weekday, day-of-week, volume tercile) by model. All four stratum families of the frozen prereg §4 are reported; every value is generated from `phase2/results/output_tsfm.json`.

model	n	SMAPE median	SMAPE mean	christmas/NY median	easter median
Seasonal-naive	4527	23.66	55.68	22.22	26.85
Climatology	4527	34.89	62.20	34.28	36.83
TimesFM-2.x-small	4527	41.10	72.75	39.21	44.43
Chronos-Bolt-small	4527	115.27	116.90	100.60	143.53
Moirai-small	4527	198.60	166.24	198.31	199.08

Table 5: Secondary non-gating deep-subset read (no go/no-go rule): median and mean SMAPE overall and in the two held-out holiday windows, TSFM and naive baselines only. Generated from `phase2/results/output_tsfm_deep.json`.