FUTURE-TS Empirical Benchmarking: A Real-Data, Temporally-Constrained Evaluation of Hosted Time-Series Foundation Models

James LePage
TSFM.ai · Parisi Labs
james.lepage@tsfm.ai

Paul Hultgren
TSFM.ai · Parisi Labs
paul.hultgren@tsfm.ai

April 2026; empirical run refreshed June 4, 2026

Abstract

This paper is the empirical companion to the FUTURE-TS design paper. We instantiate the benchmark on a task-native real-data suite, run the hosted TSFM.ai model catalog end to end, and report what the benchmark teaches once issue-time execution and strict scoring are enforced. The empirical suite, `future-ts-empirical-v2`, contains 15 tasks spanning public-development, blind-archive, and live tiers; 89 issue windows; 309 series-window requests; and 2382 realized targets from FRED, NOAA, USGS, Melbourne open data, CDC NSSP, MTA turnstiles, and NYC Open311.

The June 4, 2026 refresh surfaced 52 public hosted models after excluding internal evaluator-only entries, evaluated the public catalog across three context budgets, merged 51 submissions, and produced scored coverage for 47 models. The current winner is `Datadog/Toto-2.0-1B` with tier-weighted score 0.2369. The result is materially different from the earlier zero-shot-only run: `Datadog/Toto-2.0-1B` and `Datadog/Toto-2.0-313m` take the top two overall ranks, `NX-AI/TiRex` remains third, and `amazon/chronos-2` becomes the rank-aggregate consistency leader. The run also remains diagnostic rather than merely celebratory: five public catalog entries do not produce scored full coverage under the current common interface, exposing structural mismatches around fixed-channel models, minimum context requirements, maximum input lengths, and GPU memory pressure.

1 Empirical Suite

The empirical suite materialized for this paper is benchmarked under the identifier `future-ts-empirical-v2`. It contains 15 tasks, 89 issue windows, and 2382 realized targets. Fourteen tasks have six issue windows each; the hourly NOAA station-temperature task has five. Every issue window is constructed from task-native source histories that would have been available at issue time.

The suite covers 3 tiers, 6 tracks, and 9 domains. The tier split is 7 public-development tasks, 5 blind-archive tasks, and 3 live tasks. The track split is 5 forecasting-core tasks, 3 event tasks, 3 transfer tasks, 2 multivariate-relational tasks, 1 data-quality task, and 1 covariate-aware task.

2 Execution Protocol

All models in this paper are run through the hosted TSFM.ai catalog using the same multi-budget inference protocol. The runner queries the available model list from the TSFM.ai API,

Table 1: FUTURE-TS empirical v2 task suite. Fourteen tasks have 6 issue windows; NOAA hourly temperature has 5.

Task	Tier	Track	Metric	H	Series	Source
macro curve shift	public	event	event AUC	5	1	FRED DGS10 + DFF
macro unemployment	public	core	MAE	3	1	FRED UNRATE
retail sales	public	covariate-aware	sMAPE	3	1	FRED RSAFS
NOAA station temperature	public	core	MAE	24	3	NOAA global hourly
multisite entries	public	relational	wMASE	7	6	Melbourne pedestrian counts
turnstile entries	public	relational	wMASE	14	6	MTA turnstiles
streamflow daily	public	core	MAE	5	4	USGS NWIS
policy transfer	blind	transfer	sMAPE	4	1	FRED INDPRO
transfer sensor	blind	transfer	MAE	7	4	Melbourne pedestrian counts
MTA transfer station	blind	transfer	MAE	14	4	MTA turnstiles
respiratory ED share	blind	core	MAE	7	5	CDC NSSP
data gaps	blind	data quality	MAE	5	4	USGS NWIS
CO_2 weekly	live	core	MAE	4	1	NOAA GML Mauna Loa
Melbourne entries daily	live	relational	MAE	3	6	Melbourne pedestrian counts
Open311 spike event	live	event	event AUC	3	5	NYC Open311

materializes issue-time histories for every task window, and issues forecasting requests through `/v1/forecast/batch`. The execution path is case-first: each model receives all series belonging to a task window in one hosted request, model groups are chunked to a stable size, transient failures are retried, and only if a case-level request fails does the runner fall back to per-series recovery for the affected model-case pair. Models are scored only if they complete the full case set for the required budgets.

The June 4, 2026 refresh used `batch-size=4`, `concurrency=16`, `group-concurrency=4`, one retry, and a warmup pass across model classes. It evaluated three legal context budgets: `zero_shot` at 96 observations, `few_shot` at 192 observations, and `s16` at 288 observations. The monthly FRED tasks were widened to provide at least 32 observations of issue-time history, because the current Toto 2.0 endpoints require context lengths that are multiples of 32 and sufficient history to satisfy that minimum. The public run evaluated 52 requested models over all 89 issue windows for each budget.

This is a common-interface comparison, not a per-family ceiling-performance study. Every scored model receives the same issue-time histories and the same quantile request at levels 0.1, 0.5, and 0.9. The budgets change only the amount of issue-time context supplied to the hosted model; they do not add task-specific fine tuning, privileged covariates, or post-hoc calibration.

3 Scoring Geometry

The evaluator uses the FUTURE-TS scoring rules defined in the design paper. Each task produces a primary-metric raw score, a normalized task skill relative to a task anchor, uncertainty and efficiency scores where available, and a secondary-metrics dictionary that reports MASE, WAPE, sMAPE, and a pinball-sum proxy for CRPS when the submitted predictions support them.

The tier-weighted score is the headline scalar,

$$S = 0.25\, s_{\text{public}} + 0.35\, s_{\text{blind}} + 0.40\, s_{\text{live}},$$

with weights renormalised over tiers that actually produced a score. Adaptation AUC is computed from the three context-budget points rather than collapsed to a single zero-shot proxy. The paper therefore reports the scalar leaderboard together with tier winners, task winners, family means, and a rank-based aggregate.

Table 2: Top ten models in the June 4, 2026 FUTURE-TS empirical v2 run.

Rank	Model	Overall	F	U	A
1	Datadog/Toto-2.0-1B	0.237	0.233	0.696	0.229
2	Datadog/Toto-2.0-313m	0.221	0.248	0.703	0.236
3	NX-AI/TiRex	0.202	0.224	0.716	0.230
4	Salesforce/moirai-2.0-R-small	0.186	0.237	0.711	0.228
5	amazon/chronos-2	0.185	0.256	0.698	0.237
6	NX-AI/TiRex-1.1-gifteval	0.183	0.247	0.695	0.228
7	google/timesfm-2.5-200m-pytorch	0.177	0.179	0.473	0.202
8	google/timesfm-2.0-500m-pytorch	0.171	0.221	0.739	0.219
9	Datadog/Toto-2.0-22m	0.164	0.223	0.683	0.206
10	Datadog/Toto-2.0-4m	0.156	0.178	0.685	0.177

Table 3: Best model by benchmark tier, with the overall winner's position in that tier.

Tier	Tier leader	Mean tier skill	Winner position
blind archive	NX-AI/TiRex-1.1-gifteval	0.244	3rd
live	Datadog/Toto-2.0-1B	0.236	1st
public development	Datadog/Toto-2.0-313m	0.342	8th

4 Results

4.1 Leaderboard

The refreshed empirical run produces 47 scored models out of 52 surfaced public catalog entries. Thirty scored models have positive overall score. The winning model is `Datadog/Toto-2.0-1B` with overall score 0.2369.

The headline change is that multi-budget evaluation moves Toto 2.0 to the top of the table. `Datadog/Toto-2.0-1B` wins overall, `Datadog/Toto-2.0-313m` is second, and two smaller Toto 2.0 variants remain in the top ten. TiRex remains strong rather than disappearing: `NX-AI/TiRex` is third overall, and `NX-AI/TiRex-1.1-gifteval` is sixth. Chronos-2 remains top-five and becomes the most consistent model by mean rank.

4.2 Tier Structure

The tier breakdown makes the mechanism of the win clearer than the scalar leaderboard alone. Blind archive, live, and public development have different leaders.

This explains why `Datadog/Toto-2.0-1B` wins without sweeping the suite. It wins the live tier, ranks third in blind archive, and remains good enough in public development under the benchmark's tier weights. That profile is different from `Datadog/Toto-2.0-313m`, which wins public development; and `NX-AI/TiRex-1.1-gifteval`, which wins blind archive.

4.3 Rank Aggregate

The leaderboard also reports the mean of per-task ranks across the 15 tasks, with a 200-iteration bootstrap 95% confidence interval. Lower mean rank is better. The rank aggregate remains useful because it asks a different question from the tier-weighted scalar: which model is consistently strong across heterogeneous tasks.

Chronos-2 remains the most consistent model by mean rank, even though it is fifth by tier-weighted score. The overall winner ranks fifth on cross-task consistency, while Toto-2.0-313m ranks

Table 4: Top models by rank aggregate. Score order is the tier-weighted leaderboard rank.

Model	Score order	Rank order	$\bar{r}_s$	Rank CI
amazon/chronos-2	5	1	8.87	[4, 14]
Datadog/Toto-2.0-313m	2	2	9.00	[6, 12]
NX-AI/TiRex-1.1-gifteval	6	3	11.40	[7, 18]
Salesforce/moirai-2.0-R-small	4	4	11.67	[9, 17]
Datadog/Toto-2.0-1B	1	5	11.87	[9, 15]
google/timesfm-2.0-500m-pytorch	8	6	12.33	[7, 19]
amazon/chronos-t5-large	13	7	12.40	[8, 16]

Table 5: Best model by task in the June 4, 2026 empirical v2 suite.

Task	Winning model	Skill	Raw
blind_archive_melbourne_transfer_sensor	Maple728/TimeMoE-50M	0.526	1274.026
blind_archive_mta_transfer_station	NX-AI/TiRex-1.1-gifteval	0.728	1251.080
blind_archive_policy_transfer_fred	mldi-lab/Kairos_23m	0.183	0.336
blind_archive_respiratory_ed_pct	cisco-ai/cisco-time-series-model-1.0	0.108	0.840
blind_archive_usgs_sensor_gaps	Salesforce/moirai-1.0-R-small	0.035	709.169
live_melbourne_entries_daily	Datadog/Toto-2.0-313m	0.365	1722.530
live_noaa_co2_weekly	amazon/chronos-2	0.108	0.699
live_open311_incident_spike_public	AutonLab/MOMENT-1-small	0.375	0.688
public_dev_macro_curve_shift	Datadog/Toto-2.0-313m	0.659	0.830
public_dev_macro_unrate	mldi-lab/Kairos_50m	0.212	0.096
public_dev_melbourne_multisite_entries	google/timesfm-2.0-500m-pytorch	0.282	0.756
public_dev_noaa_temperature_station	amazon/chronos-2	0.663	1.317
public_dev_retail_sales	amazon/chronos-bolt-base	0.447	0.376
public_dev_transit_turnstile_entries	bytedance-research/Timer-S1	0.645	0.476
public_dev_usgs_streamflow_daily	Salesforce/moirai-1.0-R-large	0.044	668.472

second by mean rank. The overlap in rank CIs shows that the exact rank-aggregate ordering among the leaders is not load-bearing at 15 tasks; expanding the suite toward the full benchmark surface is the route to tighter intervals.

4.4 Task-Level Decomposition

Task-level winners turn over across families, which is the most important evidence that the suite is measuring multiple capabilities rather than a single generic forecasting axis.

No single family owns the task table. Chronos/Chronos-Bolt variants win three tasks, Toto, Kairos, and Moirai variants each win two, and TiRex, TimeMoE, Cisco, TimesFM, Timer-S1, and MOMENT each contribute task wins. The overall winner is therefore a tier-weighted portfolio result, not a universal task champion.

4.5 Family-Level Patterns

Because hosted model catalogs contain many related variants, family means help prevent over-reading a single model rank.

The family table reinforces the individual leaderboard without reducing the result to one model. TiRex has the strongest family mean, TimesFM remains broadly competitive, and Toto is no longer a tail observation: the five-model Toto family has positive mean overall score and four top-ten entries. Chronos remains strong, but its family mean is now below TiRex, TimesFM, and Toto.

Table 6: Family means in the June 4, 2026 empirical v2 run.

Family	Models	Mean overall
TiRex	2	0.193
TimesFM	2	0.174
Toto	5	0.167
Chronos	2	0.146
cisco-ai	1	0.144
Moirai	7	0.075
thuml	1	0.062
Chronos-Bolt	4	0.062
bytedance-research	1	0.037
mldi-lab	3	0.021
Kronos	1	0.021
TimeMoE	2	-0.002

4.6 Catalog Completeness and Failures

The refresh surfaced 52 public models and evaluated all 52. Fifty-one models produced at least one budget submission and 47 models were scored after merging the three budgets. Four models produced partial coverage and were excluded by the scoring gate: `Melady/TEMPO`, `NeoQuasar/Kronos-mini`, `NeoQuasar/Kronos-small`, and `thuml/timer-base-84m`. One additional catalog entry, `ibm-research/granite-timeseries-patchtst`, produced no scored coverage because the hosted model requires exactly seven input channels while the current runner sends benchmark-native univariate or case-level requests.

The failure modes are useful engineering signal. `thuml/timer-base-84m` requires at least 96 actual input points, which many short-history empirical windows cannot satisfy. `Melady/TEMPO` rejects NOAA hourly station-temperature windows whose input length exceeds its maximum of 336. `NeoQuasar/Kronos-mini` and `NeoQuasar/Kronos-small` hit GPU memory pressure on the long-context NOAA station-temperature task. The current scoring gate correctly prevents these partial submissions from influencing the leaderboard.

5 What the Benchmark Learns Already

The current empirical suite supports three claims. First, FUTURE-TS is operational as a real hosted-model benchmark: it materializes issue-time histories, invokes hosted TSFMs, writes benchmark-compliant submissions, validates them, and scores them from realized labels. Second, the benchmark is already rich enough to falsify simplistic "one best TSFM" narratives: task and tier winners turn over across model families. Third, catalog failures are themselves informative because they reveal fixed-channel assumptions, context minima, input-length ceilings, and runtime constraints that matter in real deployments.

6 What This Run Has Not Yet Established

The remaining gaps are run-scope gaps, not fundamental benchmark limitations. The present refresh intentionally reports the empirical v2 slice that was run end to end on the hosted catalog. FUTURE-TS already has the machinery to support a larger surface, multiple adaptation budgets,

repeated waves, and stricter manifest enforcement; those claims simply need to be exercised and reported in a future empirical wave.

First, scale is available but not fully used here. Fifteen tasks are enough to expose structure, but not enough to declare a universal strongest TSFM. The full `benchmarks/v1` surface is larger, and the rank-aggregate confidence intervals show why more tasks matter. The path forward is operational rather than conceptual: materialize the larger task surface, rerun the catalog, and report whether the top of the table is stable as the task count increases.

Second, budget coverage is now exercised for the empirical v2 surface but not yet stress-tested at full benchmark scale. The current refresh invokes `zero_shot`, `few_shot`, and `s16` as separate context budgets before merging the resulting submissions. Adaptation AUC is therefore a real three-point curve for this suite. The next evidence step is to repeat that same multi-budget protocol on the larger benchmark surface and compare whether the budget-response curves remain stable.

Third, repeated-wave stability is supported by the protocol but not measured in this single refresh. The present run is one materialized wave over 89 issue windows. Future empirical reports should schedule repeated waves, then report per-cutoff rank drift, winner frequency, and whether the tier-weighted winner remains stable under prequential evaluation.

Fourth, pretraining attestation is partly implemented but not externally established for this hosted-catalog run. The local runner produces valid submission manifests and honors issue-time constraints, and strict benchmark specs can require pretraining manifests. What this run does not yet have is model-author-side attestation for every hosted model. A sealed, manifest-required run on the full benchmark surface is therefore the next evidence step, not a missing piece of benchmark design.

7 Conclusion

The June 4, 2026 FUTURE-TS empirical v2 refresh changes the empirical answer. The paper is no longer a zero-shot-only TiRex result. It is a 52-model, three-budget public-catalog evaluation with 47 scored models, a Toto-2.0-1B overall winner, a strong TiRex and Chronos showing, and clear structural failures for five catalog entries under the current common interface.

The right scientific posture remains conditional. `Datadog/Toto-2.0-1B` is the winner of FUTURE-TS empirical v2 under a three-budget, single-wave, 15-task suite, scored with paired point and probabilistic metrics, a tier-weighted scalar, and a rank-based aggregate. That is meaningful. It is not the same as declaring a universally strongest TSFM. The platform can support the obvious next wave: expand the suite, repeat the multi-budget protocol, add repeated waves with prequential cutoffs, and require pretraining-data manifests so the overlap column carries real evidentiary weight.