What Does an Observability Foundation Model Know?

Dhyey Dharmendrakumar Mavani
Independent
ddmavani2003@gmail.com

Rian Atri
Keiji AI
rianatri@keiji.ai

Tairan Ji
Independent
tairanjiryan@gmail.com

Abstract

A linear probe can show that a label is recoverable from a model's hidden states, but not whether that goes beyond what the input already reveals, or whether the model uses it. We audit Toto, an observability forecasting foundation model, on the Benchmark of Observability Metrics (BOOM) across five series-disjoint resplits, comparing linear probes on its frozen residual stream with models that read the raw input window and with Toto's architecture stripped of its trained configuration. Short-vs-medium cadence and metric type are more linearly recoverable from Toto's residuals than from the strongest raw-window model in every resplit (macro-F_1 0.766 vs. 0.633 and 0.545 vs. 0.498). Domain is nearly tied, and series cardinality is recovered far better from the raw window. MOMENT-base shows related cadence, metric-type, and domain readouts. Recoverability is not use: exchanging Toto's residuals with those of high-burst donors moves a future-burstiness readout as intended but does not make forecasts consistently burstier than a randomized donor. A BOOM-trained coordination probe has negative zero-shot R^2 on the tested external benchmarks. We report each label against its strongest baseline.

1 Introduction

Time-series foundation models (TSFMs) serve increasingly as zero-shot forecasters, and probing their hidden states is a natural way to ask what they have learned. Consider two telemetry series: request latency sampled every few seconds and disk usage sampled every few minutes, and suppose a linear probe on a forecaster's hidden states tells them apart. The model may have learned sampling cadence in pretraining. Or the two series may simply look different in the raw window, for instance in their step-to-step variation, so that any classifier on the input would separate them just as well. And even if the hidden states separate them better than the input does, the forecaster need not use that distinction when it predicts the next values. A probe score alone cannot tell these cases apart.

We study this question for Toto [Cohen et al., 2024, 2025], an observability forecasting foundation model, on the Benchmark of Observability Metrics (BOOM) [Datadog, 2025], following probing work that separates decodability from use [Tenney et al., 2019a,b, Conneau et al., 2018, Rogers et al., 2020, Belinkov, 2022, Hewitt and Liang, 2019, Meng et al., 2022, Turner et al., 2024, Geiger et al., 2024, 2025, Olsson et al., 2022]. We read "know" operationally: a label is known to the extent that a linear probe on Toto's frozen residual stream predicts it on held-out series. This is a claim about linear recoverability from one representation under one probe family, not about semantic understanding or all the information the model contains. For each label we ask three questions. Is it more linearly recoverable from Toto's residuals than from the input window itself? Does that depend on Toto's trained weights and their order? And does moving the readout move the forecast? *Input baselines*—a probe on six summary statistics of the input and four models trained on the raw window—answer the first. *Backbone baselines*, Toto with random weights or with its trained blocks in shuffled order, answer the second. A donor exchange tests the third on future burstiness, and zero-shot probe transfer asks whether a probe fitted on BOOM carries over unchanged to external datasets. Figure 1 summarizes the first two questions for each label.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

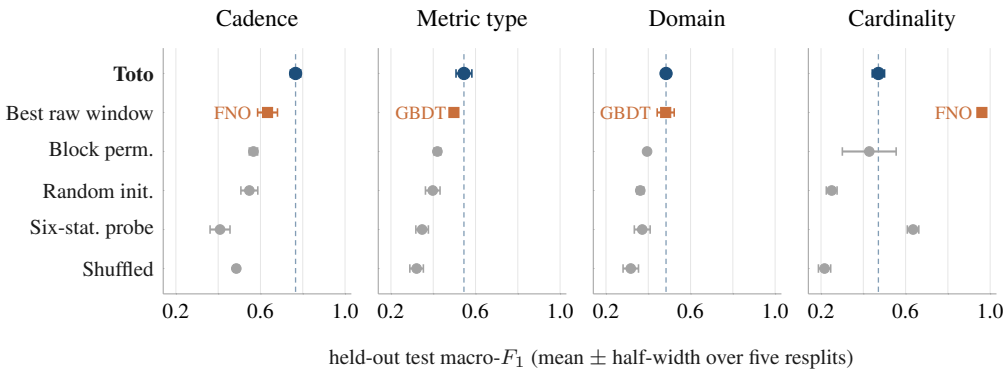

Figure 1: Held-out test macro-F_1 of Toto, its baselines, and the shuffled-label floor for each taxonomy label (means ± half-widths over five resplits; §4). The dashed line marks Toto's mean; the orange square is the strongest raw-window model for that label.

Contributions

First, we give a simple audit design for forecasting representations: every held-out probe score is paired, on the same windows, with input and backbone baselines, and a separate donor exchange tests forecast response. Second, we report a label-by-label result on BOOM across five series-disjoint resplits: cadence and metric type are more linearly recoverable from Toto's residuals than from every tested raw-window model in all five resplits, and their gaps over the six-statistic probe and random initialization persist within the tested strata of the other labels; domain is nearly tied with its strongest raw-window model, and cardinality is recovered far better from the raw window. Third, we bound how far probe readouts can be interpreted. MOMENT-base shows related cadence, metric-type, and domain readouts on the same resplits, with narrower baselines. In both models, exchanging residuals moves the targeted future-burstiness readout in every resplit without consistently moving the forecast endpoint—forecast burstiness for Toto and future mean absolute error (MAE) for MOMENT—and a BOOM-trained coordination probe has negative zero-shot R^2 on the tested external benchmarks. We never treat a moved readout as evidence that the model uses it to forecast.

2 Related Work

Forecasting foundation models and observability

TSFMs differ in pretraining objective and use: Chronos and TimesFM cast broad-domain forecasting as autoregressive prediction, whereas MOMENT is a general-purpose encoder trained with masked reconstruction [Ansari et al., 2024, Das et al., 2024, Goswami et al., 2024]. Toto adapts the forecasting-foundation-model paradigm to multivariate telemetry, with the accompanying BOOM benchmark supplying real observability series and forecast tasks [Cohen et al., 2024, 2025, Datadog, 2025]. An interpretability audit asks instead whether frozen hidden states make series attributes accessible to a simple readout.

Probing, selectivity, and controls

Linear probes have long asked which properties are recoverable from intermediate representations [Alain and Bengio, 2016, Conneau et al., 2018, Tenney et al., 2019a]; probe capacity and control tasks matter because a flexible probe can expose structure available without the representation under study [Hewitt and Liang, 2019, Belinkov, 2022]. Recent TSFM work makes this question concrete: Wiliński et al. analyze layer structure and probe concepts such as trend and periodicity, then steer models with concept-derived interventions; Pandey et al. probe synthetic temporal factors and their compositions [Wiliński et al., 2025, Pandey et al., 2025]. We ask the same question for telemetry taxonomy labels on real BOOM series, with raw-window neural and tree models over the same context as comparators.

Decodability, interventions, and model use

A model need not use a readable direction to compute its output. Control-task probing makes this explicit, while causal tracing and editing, activation steering, and causal abstraction test how internal changes relate to behavior [Hewitt and Liang, 2019, Meng et al., 2022, Turner et al., 2024, Geiger et al., 2024, 2025]. These methods come largely from language-model settings; we adopt their reading-vs-intervening distinction. Related TSFM work has already used concept-based interventions to alter temporal patterns and forecasts [Wiliński et al., 2025]. We use a donor exchange to test whether moving a readout moves a forecast.

Forecast evaluation and transfer

Forecast Evaluation (FEV) emphasizes realistic task coverage, covariates, reproducible comparisons, and aggregation across tasks [Shchur et al., 2025]. Public long-horizon suites likewise enable cross-dataset forecast evaluation [Zhou et al., 2021, THUML, 2024]. Our external test is narrower than either forecast transfer or representation transfer: it applies a BOOM-trained probe, unchanged, to external FEV configurations and long-sequence datasets.

3 Setup

Models

Toto is a decoder-only transformer over patchified multivariate inputs [Cohen et al., 2024, Nie et al., 2023, Dosovitskiy et al., 2021]; we load the public `Datadog/Toto-Open-Base-1.0` checkpoint without fine-tuning. We divide each 1024-observation context into 16 non-overlapping 64-observation patches and read activations at the patch embedding and layers 0–11. The MOMENT-base encoder [Goswami et al., 2024] uses masked-reconstruction pretraining; we load it in embedding mode and read activations at layers 3, 6, 9, and 11. MOMENT uses a 512-observation context with 8-observation patches, so its six-statistic probe pools a different final patch than Toto's in Table 1. Further protocol details are in App. A.

Data

BOOM has $2{,}807$ observability series and approximately 350 million observations [Cohen et al., 2025, Datadog, 2025]. The fixed context-and-target length excludes 489 series (17.4%), including daily and weekly cadences, so the cadence label here is a Short-vs-Medium bucket. All results use five seeded, series-disjoint resplits (seeds 42–46): a 70/15/15 split capped at 500 training series gives 500/421/422 train/validation/test series per resplit, leaving the rest unused. Test sets differ across resplits but can overlap. Series too short for the fixed context contribute no windows; up to four windows are sampled per remaining series (App. A). Zero-shot probe transfer uses 11 FEV configurations [Shchur et al., 2025] and four long-sequence time-series forecasting (LSTF) datasets (ETTh1, ETTh2, weather, electricity) [THUML, 2024, Zhou et al., 2021]; the FEV configurations may share underlying sources (IDs in App. A).

Pretraining/evaluation overlap

Toto was pretrained on an internal observability corpus; overlap with BOOM is unknown. The series-disjoint resplits prevent within-study train/test leakage.

Labels

We audit four BOOM-metadata taxonomy labels: cadence (sampling-frequency metadata, restricted to the Short and Medium buckets that fit the fixed context-and-target length), metric type (gauge, rate, distribution), domain (Application Usage, Infrastructure, Database), and series cardinality (variate count in bins 1, 2–8, 9–32, above 32). Other and multilabel metric/domain entries are outside these singleton-label probes.

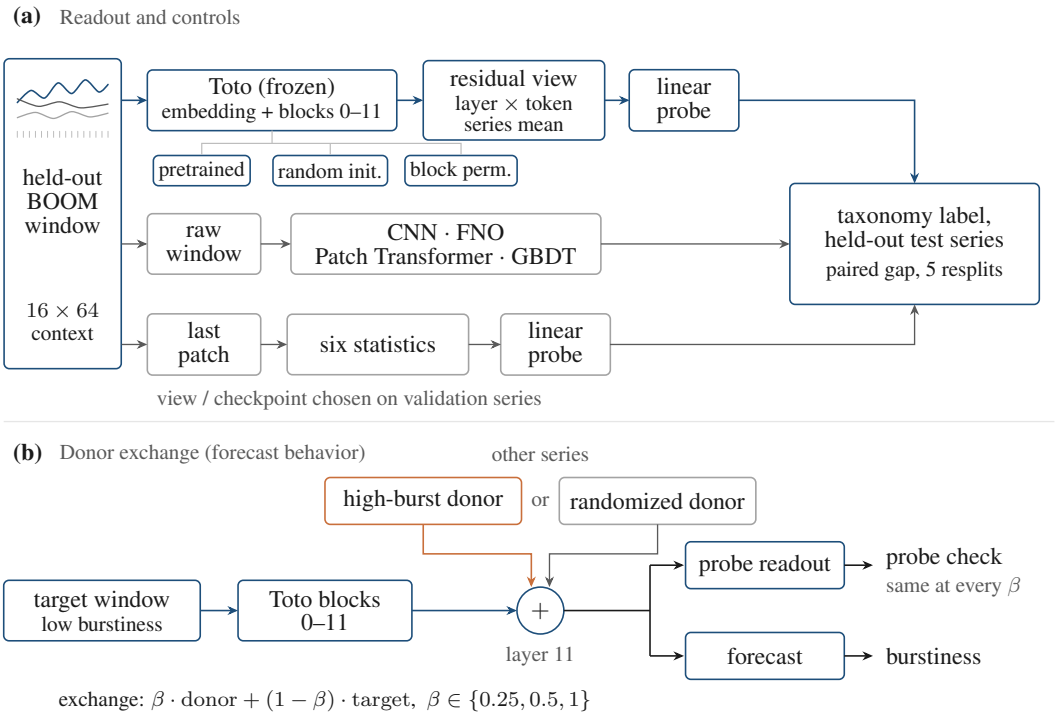

Figure 2: Method. (a) Linear probes on frozen Toto residual views (pretrained backbone, or the randomly initialized and block-permuted backbone baselines) versus the input baselines (raw-window models and a six-statistic probe), compared as paired gaps on held-out series. (b) Layer-11 donor exchange, read by the probe and the forecast.

Dynamic labels are engineered statistics of the context and the next patch. For each variate i, the robust scale s_i is 1.4826 times the median absolute deviation (MAD) of its context, clamped away from zero. Burstiness is the largest absolute deviation from the context median, divided by s_i, over the final context patch (current) or the true next patch (future). Sparsity is the fraction of exact zeros in the same two patches. Shift risk averages $|\bar{x}_i^{\text{next}} - \bar{x}_i^{\text{last}}|/s_i$ over all variates of the series.

For coordination, let d_i be the shift of variate i's mean from the final context patch to the next patch. A variate is active if $|d_i| > 0.5\, s_i$; coordination is the larger of the fractions of active variates that move up and that move down, and zero if no variate is active.

4 Method

Held-out readouts

Figure 2 sketches the design. We fit logistic-regression probes for categorical labels and ridge probes for continuous labels, and report macro-F_1 and R^2. The candidate views are 13 layers by 3 token positions with series-mean pooling (39 views). We select the view by validation accuracy (macro-F_1 for cardinality) and keep probe regularization fixed (App. A).

Baselines

Input baselines ask whether a label is already recoverable from the input window. The six-statistic probe is a linear probe on the last 64-observation patch's mean, standard deviation, sparsity, mean absolute value, maximum absolute value, and cross-variate spread.

Table 1: Held-out test macro-F_1 on BOOM (means $\pm$ half-widths over five resplits). A: Toto against the input baselines. The strongest raw-window model is chosen by mean validation macro-F_1; the paired gap is Toto minus that model, and wins count the resplits in which it is positive. B: backbone baselines and the shuffled-label floor. All raw-window models are in App. D.

A. Input baselines

Label	Toto	Six-stat. probe	Strongest raw-window model		Paired gap	Wins
Cadence	0.766 ± 0.024	0.409 ± 0.047	FNO	0.633 ± 0.047	$+0.133 \pm 0.063$	5/5
Metric type	0.545 ± 0.037	0.347 ± 0.030	GBDT	0.498 ± 0.007	$+0.047 \pm 0.036$	5/5
Domain	0.484 ± 0.011	0.371 ± 0.037	GBDT	0.482 ± 0.040	$+0.002 \pm 0.040$	2/5
Cardinality	0.471 ± 0.029	0.635 ± 0.027	FNO	0.961 ± 0.016	-0.490 ± 0.044	0/5

B. Backbone baselines and shuffled-label floor

Label	Random init.	Block perm.	Shuffled labels
Cadence	0.547 ± 0.040	0.567 ± 0.020	0.486 ± 0.011
Metric type	0.398 ± 0.035	0.420 ± 0.016	0.321 ± 0.032
Domain	0.362 ± 0.017	0.394 ± 0.011	0.317 ± 0.037
Cardinality	0.250 ± 0.025	0.428 ± 0.127	0.217 ± 0.029

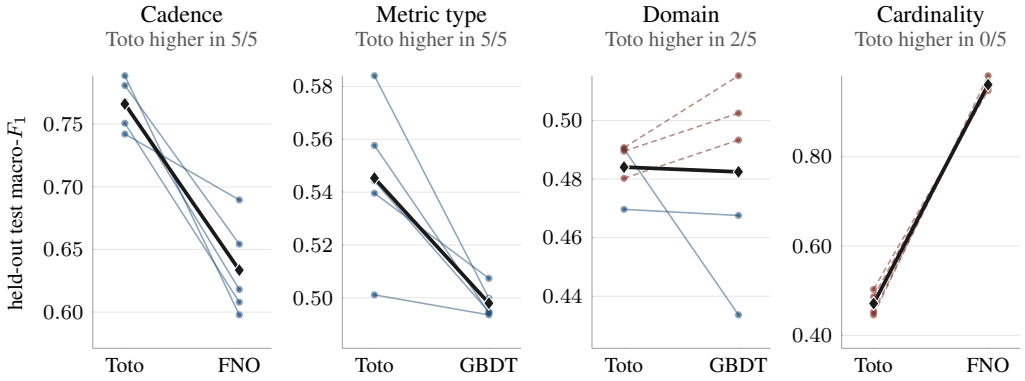

Figure 3: Held-out test macro-F_1 in each resplit for Toto and the strongest raw-window model (thin lines: solid where Toto is higher, dashed where the raw-window model is higher) and their five-resplit means (bold). Each panel has its own vertical scale.

Four raw-window models are trained from scratch on the same windows: a convolutional network (CNN), a Fourier neural operator (FNO), and a patch Transformer, which receive the masked full context, with variates as input channels zero-padded to a common count, plus a coverage channel (the share of unpadded variates); and a gradient-boosted tree (GBDT) on 15 engineered features, including the six last-patch statistics and full-context spectral and autocorrelation summaries.

Backbone baselines ask whether a readout depends on Toto's trained configuration: Toto with randomly initialized weights, and a block permutation that runs the 12 pretrained transformer blocks in a non-identity order (its view search also admits per-variate pooling, 78 views, which can only favor it). Both keep Toto's 64-observation patching, its parameter-free input scaling, and the same windows, so these shared components do not explain gaps over them. A shuffled-label probe on Toto's residuals gives a chance-level reference, the *shuffled-label floor*.

The raw-window models are not evaluators of the same kind as the probe: they are nonlinear and trained from scratch on the 500 training series of each resplit, whereas the probe is linear on a representation pretrained elsewhere. A gap in Toto's favor therefore means that a label is more linearly recoverable from Toto's residuals than these models, with this data and training budget, learn to recover it from the window; the window itself may still carry the information.

For label y, baseline c, and resplit s, the paired macro-F_1 gap is $\Delta_{y,c,s} = F_1(y, \text{pretrained}, s) - F_1(y, c, s)$. Each label's strongest raw-window model is the one with the highest mean validation macro-F_1 across the five resplits; the paired gap is computed against it in every resplit. Choosing by mean test macro-F_1 would select the same models, and for cadence and metric type the gap is positive against all four raw-window models in all five resplits.

Common-support probes

Some taxonomy labels are correlated, so a readout of one could partly reflect another. Three prespecified common-support probes therefore retest the gaps over the six-statistic probe and random initialization within one fixed stratum of the other two labels. Each averages windows within series, balances target classes by series, selects the view on validation data, and measures test performance once. Raw-window models were not refit within strata.

Donor exchange

Our donor exchange tests forecast behavior by blending residuals from real donor series into a target series' state. In each resplit we fit a future-burstiness probe on training series at the fixed layer-11, all-context, series-mean view, and evaluate 40 triples at residual blends $\beta \in \{0.25, 0.5, 1.0\}$. Each triple pairs a held-out target window from the lowest future-burstiness quartile with a high-burst donor (a top-quartile window from another series) and a randomized donor (any eligible test window from another series, not matched on covariates or taxonomy). Because the probe is linear, the difference between the two exchanges' probe scores equals β times the difference between the two donors' probe scores. The *probe check* is the fraction of targets for which the high-burst exchange raises the readout more than the randomized exchange. It is therefore the same at every blend. The check confirms that the exchange moves the readout in the intended direction; only the forecast endpoints vary with β. The primary forecast endpoint is whether the forecast is burstier under the high-burst donor than under the randomized donor; the secondary endpoint is the ratio of their weighted absolute percentage errors (WAPE).

Reporting conventions

The resplit is the unit of analysis: each resplit draws disjoint train, validation, and test series from the same BOOM corpus. Unless stated otherwise, $\pm$ is a Student-$t(4)$ 95% half-width across the five resplit-level values; it measures variability across resplits of one corpus and is descriptive rather than a significance test. A *win* is a resplit with a positive paired gap, so 5/5 reports consistency of direction within BOOM.

Zero-shot probe transfer

We apply the BOOM-trained coordination probe, with its BOOM weights and standardization and without refitting, to Toto activations on 11 FEV configurations and four LSTF datasets. The target is computed from each external context and its true next patch. The test asks only whether this fitted readout carries over unchanged; it does not measure forecast accuracy, and whether Toto's representations of external series encode coordination is a separate question for a probe refit on those series.

5 Results

5.1 Cadence and metric type are more recoverable from Toto than from input baselines

Cadence and metric type are the two labels on which Toto's readout beats every baseline (Table 1). Against the strongest raw-window model, the paired gap is positive in all five resplits for both labels (Figure 3); it is also positive against the other three raw-window models in every resplit, and both labels beat random initialization and block permutation in 5/5. The cadence margin is the larger one, and it holds even though the FNO itself is well above the shuffled-label floor and the six-statistic probe; the metric-type margin is smaller but consistent across resplits.

Domain is nearly tied. Toto beats both backbone baselines, but its macro-F_1 is level with the GBDT's and its accuracy is lower (0.534 vs. 0.573, 0/5 resplits), so a raw-window model recovers domain about as well.

Cardinality reverses the pattern. The CNN, FNO, and patch Transformer all reach macro-F_1 near 0.95, far above Toto (App. D). These neural models see the variate count directly, through the zero-padded variate channels and the coverage channel. The two input baselines that see neither, the six-statistic probe and the GBDT, are much weaker than the neural models but still above Toto (0.635 and 0.663 vs. 0.471), so the reversal does not rest on the coverage channel alone.

5.2 Common-support probes preserve two gaps in tested strata

Domain and metric type are moderately associated (series-level Cramér's $V = 0.402 \pm 0.011$), whereas cadence is nearly independent of both (0.055 ± 0.012 with domain and 0.050 ± 0.018 with metric type). Within fixed strata of the other two labels (Table 2), cadence and metric type keep positive gaps over the six-statistic probe and random initialization in every resplit, and their means stay well above the six-statistic probe. Domain's gaps are smaller and less consistent across resplits; it is also the label most associated with metric type, so composition and direct signal are hardest to separate for it. The strata are a targeted robustness check on small, balanced test support that varies by resplit. Rotated held-out-combination tests within BOOM are in App. B.

5.3 MOMENT shows related taxonomy readouts on BOOM

On the same resplits, MOMENT-base's cadence, metric-type, and domain readouts beat both its random initialization and the six-statistic probe in all five resplits (Table 3), although the margins over random initialization are modest and smallest for domain. Its cardinality readout matches random initialization and, like Toto's, trails the six-statistic probe. MOMENT's baselines are narrower than Toto's, since they exclude the raw-window models and block permutation, and its dynamic readouts are comparable to the six-statistic probe (App. C). Cross-model recurrence on one benchmark does not rank architectures.

5.4 Exchanges move the readout without consistently moving the forecast

The donor exchange asks whether moving the readout moves the forecast. In Toto, the probe check is above 0.5 in all five resplits, so the exchange moves the future-burstiness readout as intended (Figure 4, left). The forecasts do not follow: the fraction of targets whose forecast is burstier under the high-burst than under the randomized donor is below one half on average at every blend and above it in at most one resplit (Table 6). The secondary endpoint favors the high-burst donor at the point estimate, with median WAPE ratios below one, but its half-widths span one.

MOMENT's matched interchange patches its validation-selected layer-3, all-context view and draws its null donors from the nearest context-matched, non-high-burst windows of other series, with taxonomy penalties (App. A). It shows the same dissociation: the probe check is above parity in every resplit, while the fraction of targets with lower future MAE averages below parity at all three blends (Figure 4, right; Table 9). Because MOMENT's null donors are matched to its high-burst donors on variate count, last-patch statistics, and taxonomy, its contrast holds these donor properties approximately fixed. Toto's randomized donor is not matched, so Toto's contrast may also reflect donor differences other than burstiness.

Table 2: Common-support probes: held-out test macro-F_1 within one fixed stratum of the other two labels (means $\pm$ half-widths over five resplits; classes balanced by series). Test support is 11–19 (cadence), 17–26 (metric type), and 11–18 (domain) series per class. Wins are against random initialization and against the six-statistic probe.

Target	Fixed stratum	Toto	Random init.	Six-stat.	Wins
Cadence	Infrastructure/gauge	0.697 ± 0.060	0.586 ± 0.121	0.485 ± 0.137	5/5; 5/5
Metric type	Application Usage/Short	0.502 ± 0.053	0.395 ± 0.048	0.371 ± 0.034	5/5; 5/5
Domain	Short/gauge	0.406 ± 0.073	0.350 ± 0.091	0.376 ± 0.087	4/5; 3/5

Table 3: MOMENT-base held-out test macro-F_1 on the same five resplits (means $\pm$ half-widths; views chosen on validation series). The last column is the paired gap to the six-statistic probe and the number of resplits in which it is positive.

Label	MOMENT	Random init.	Six-stat.	Gap to six-stat. (wins)
Cadence	0.632 ± 0.113	0.556 ± 0.105	0.401 ± 0.099	$+0.231 \pm 0.044$ (5/5)
Metric type	0.499 ± 0.022	0.452 ± 0.029	0.351 ± 0.042	$+0.148 \pm 0.030$ (5/5)
Domain	0.437 ± 0.014	0.409 ± 0.020	0.337 ± 0.034	$+0.100 \pm 0.023$ (5/5)
Cardinality	0.259 ± 0.010	0.257 ± 0.018	0.668 ± 0.013	-0.409 ± 0.012 (0/5)

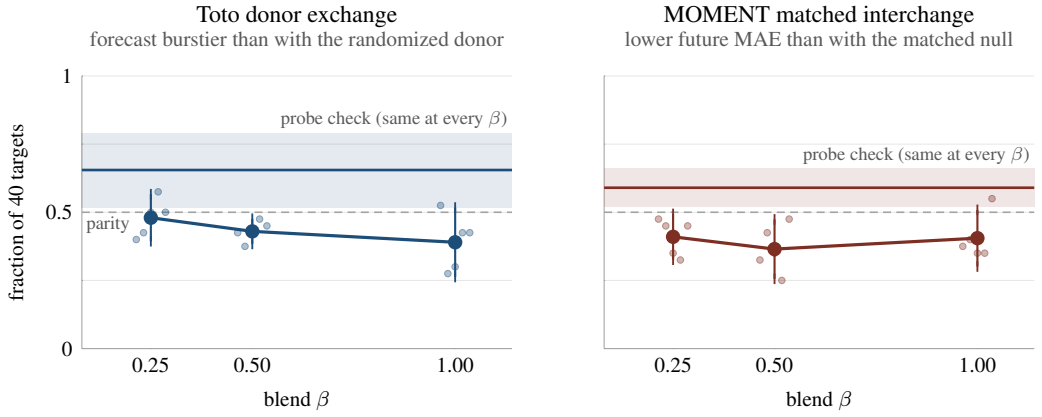

Figure 4: Probe check versus forecast endpoint across the residual blend β. The shaded band is the probe check (mean $\pm$ half-width), which is the same at every blend; points are resplits, and the line joins the forecast-endpoint means $\pm$ half-widths. The dashed line is parity.

Toto's future-burstiness probe reaches $R^2 = 0.106 \pm 0.100$ against 0.111 ± 0.066 for the six-statistic probe (above it in 3/5 resplits), so this label is about as recoverable from simple input statistics as from Toto's residuals. Each diagnostic also contrasts two exchanges with each other rather than with an intact forecast, so it measures relative forecast response, not a causal direction the model uses naturally.

5.5 Zero-shot coordination-probe transfer is negative

Applied without refitting, the BOOM-trained coordination probe has negative R^2 on both external suites, so it does worse than predicting the mean, by a much larger margin on LSTF than on FEV (Table 7). We report only coordination because several other external dynamic labels are nearly constant, which makes their R^2 unstable. This is a failure of zero-shot probe transfer: the fitted weights and BOOM standardization do not carry over to these series. It is not evidence that Toto's representations of external series lack coordination information, which only a probe refit on those series could test. The positive evidence in this paper is scoped to BOOM.

6 Discussion

A label-by-label profile

The results describe what is linearly recoverable from Toto's residuals, label by label, rather than giving a single verdict on what the model "knows". Cadence has the largest advantage over its strongest raw-window model, metric type a smaller but consistent one, domain is nearly tied, and cardinality is far better recovered from the raw window. These statements concern the evaluated representation, probe family, and baselines, not the total information the model contains: a nonlinear probe, another view, or a stronger raw-window model could shift any of them.

Why cardinality differs

In every resplit, Toto's validation-selected cardinality view is the layer-11, all-context, series-mean representation, which averages over variates and carries no explicit count, whereas the neural raw-window models see the count directly. Architecture, input format, and readout capacity also differ, so mean pooling is one candidate cause among several. The comparison shows that a label's accessibility can change sharply with the representation and input given to the evaluator.

What cadence may reflect

One hypothesis is that BOOM's Short/Medium metadata tracks temporal regularities visible across Toto's patch context. Since cadence is a metadata label rather than a spectral estimate, this mechanism is left to controlled tests.

Readout is not use

In both models, exchanges move the probe readout while forecasts do not respond consistently. Related TSFM work has used trend and periodicity concepts to steer forecasts [Wiliński et al., 2025], and synthetic-factor studies probe temporal representations under controlled conditions [Pandey et al., 2025]. Our exchange differs from concept-direction steering in both the concept (future burstiness) and the intervention (cross-series donor residuals), so its dissociation characterizes this exchange and endpoint and is compatible with prior steering results. The next Toto test would match donors on context and taxonomy as MOMENT's interchange does, preserve the target context while changing the intended factor, include sham exchanges, and report several forecast endpoints.

Limits and next steps

Five resplits of one corpus measure split variability, not uncertainty over independent observability data, and unknown pretraining overlap leaves the source of the signal unattributed. The cadence label covers Short and Medium series, and the common-support probes omit the raw-window models. A next study should repeat the paired design across independent telemetry corpora and TSFMs, with comparable raw-window models and label definitions, broader cadence classes, raw-window models refit within strata, and independently sourced series; controlled tests that vary context and patch scales would help explain the cadence result.

Recommendations for practice

A probe score on a forecasting model is interpretable only relative to the strongest baseline it beats. For each label, we recommend reporting the strongest input baseline, whether the readout also beats the same architecture without its trained configuration, and, separately, whether moving the readout moves the forecast; a claim should go no further than the comparisons it passes. Here, cadence and metric type pass the first two. The forecast test, run on future burstiness because its next-patch definition fixes an intended forecast direction, did not show a consistent forecast change.

Broader impacts

The audit helps researchers separate recoverable information from evidence of use: a probe score treated as a steering target could produce unintended forecast changes, and the donor-exchange readout is not a validated forecast-control knob. Reporting representation and forecast outcomes side by side keeps that boundary visible.

7 Conclusion

Across five series-disjoint BOOM resplits, short-vs-medium cadence and metric type are more linearly recoverable from Toto's frozen residuals than from every tested raw-window model, and their gaps over the six-statistic probe and random initialization persist in the tested strata. Domain is nearly tied with its strongest raw-window model, and cardinality is far better recovered from the raw window. MOMENT-base shows cadence, metric-type, and domain readouts above random initialization and the six-statistic probe on the same benchmark. Exchanging residuals moves the targeted readout without a consistent forecast change, and the BOOM-trained coordination probe has negative zero-shot R^2 on the tested external benchmarks. Reporting each readout against its strongest baseline, alongside forecast and transfer tests, keeps the line between what a model makes recoverable and what it uses explicit.

Code and audited results are publicly available at `https://github.com/DhyeyMavani2003/toto-interp/tree/neurips-2026-v1`.

Acknowledgments and Disclosure of Funding

Funding. This work received no external funding or third-party support.

Competing interests. The authors declare no competing interests.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL `https://arxiv.org/abs/1610.01644`.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, et al. Chronos: Learning the language of time series. In *Transactions on Machine Learning Research (TMLR)*, 2024. URL `https://arxiv.org/abs/2403.07815`.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. URL `https://doi.org/10.1162/coli_a_00422`.

Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*, 2024. URL `https://arxiv.org/abs/2407.07874`.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This time is different: An observability perspective on time series foundation models. *arXiv preprint arXiv:2505.14766*, 2025. URL `https://arxiv.org/abs/2505.14766`.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2126–2136, 2018. URL `https://aclanthology.org/P18-1198/`.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024. URL `https://arxiv.org/abs/2310.10688`.

Datadog. BOOM: Benchmark of observability metrics. Hugging Face dataset card, 2025. URL `https://huggingface.co/datasets/Datadog/BOOM`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://arxiv.org/abs/2010.11929`.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2024. URL https://proceedings.mlr.press/v236/geiger24a/geiger24a.pdf.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26: 1–63, 2025. URL https://www.jmlr.org/papers/volume26/23-0058/23-0058.pdf.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.03885.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2733–2743, 2019. URL https://aclanthology.org/D19-1275/.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. doi: 10.1016/j.apenergy.2021.116983.

Max Planck Institute for Biogeochemistry. Weather data, Jena. Weather station records, 2020. URL https://www.bgc-jena.mpg.de/wetter/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

MichalKecera. Rohlik Sales Forecasting Challenge. Kaggle, 2024. URL https://www.kaggle.com/competitions/rohlik-sales-forecasting-challenge-v2.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.14730.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Open Power System Data. Data package time series, version 2020-10-06 (primary data from the ENTSO-E transparency platform), 2020. URL https://doi.org/10.25832/time_series/2020-10-06.

Atharva Pandey, Abhilash Neog, and Gautam Jajoo. On the internal semantics of time-series foundation models. *arXiv preprint arXiv:2511.15324*, 2025. doi: 10.48550/arXiv.2511.15324. URL https://arxiv.org/abs/2511.15324.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. URL https://aclanthology.org/2020.tacl-1.54/.

Saman. Renewable energy and weather conditions. Kaggle dataset, 2023. URL https://www.kaggle.com/datasets/samanemami/renewable-energy-and-weather-conditions.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting, 2025. URL https://arxiv.org/abs/2509.26468.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601, 2019a. URL https://aclanthology.org/P19-1452/.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR)*, 2019b. URL https://openreview.net/forum?id=SJzSgnRcKX.

THUML. Time-series-library. GitHub repository, 2024. URL https://github.com/thuml/Time-Series-Library.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C58C86.

Alexander Matt Turner, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Nina Panickssery. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024. URL https://arxiv.org/abs/2312.06681.

Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring representations and interventions in time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 66861–66886, 2025. URL https://proceedings.mlr.press/v267/wilinski25a.html.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, 2021. URL https://arxiv.org/abs/2012.07436.

A Protocol Details

Data and splits

Each resplit (seeds 42–46) partitions the 2,807 BOOM series 70/15/15 by seeded shuffle with training capped at the first 500 shuffled training series, giving 500/421/422 train/validation/test series per resplit. Up to four windows per series are sampled at integer linspace starts over valid starts, excluding the latest start. Context is 1024 observations with a 64-observation patch; the target is the next patch. MOMENT-suite window counts per resplit span 1224–1296 train, 984–1077 validation, and 1044–1074 test windows; the Toto donor-exchange pool holds 1032–1068 test windows.

Views and probes

Candidate views are 13 layers (patch embedding plus blocks 0–11) by 3 token positions (all-context, final-context, first-decode) with series-mean pooling: 39 views for pretrained and random-initialization probes; the block-permutation grid additionally admits per-variate pooling (78 views). Categorical probes are logistic regression (balanced class weights, L2, $C = 1$, lbfgs, 4000 iterations); continuous probes are ridge ($\alpha = 1$). Features use train-split standardization (standard deviation floored at 10^{-6}; NaN to 0, infinities to $\pm 10^3$, clipped). Only the view is selected, by validation accuracy (macro-F_1 for cardinality); there is no regularization search.

Baselines

The six-statistic probe uses the mean, standard deviation, zero fraction, mean absolute value, and maximum absolute value over the last context patch, plus the standard deviation of per-variate means. The shuffled-label probe shuffles train-split targets with the resplit seed, keeps the same features and estimator, and is evaluated on unshuffled validation and test. Random initialization resets the full backbone with the resplit seed and is evaluated frozen. Block permutation reorders the 12 blocks by seeded randperm with identity rejection; runtime orders are 42: [6, 8, 1, 7, 0, 2, 10, 11, 4, 3, 5, 9]; 43: [4, 6, 7, 2, 9, 8, 1, 5, 0, 10, 11, 3]; 44: [4, 5, 9, 3, 7, 1, 8, 6, 11, 0, 10, 2]; 45: [11, 6, 3, 0, 4, 5, 7, 10, 2, 1, 8, 9]; 46: [9, 5, 4, 0, 6, 8, 2, 7, 3, 1, 11, 10]. Raw-window models (Table 4) train on padded contexts times a variate mask plus a coverage channel (neural models) or 15 features (GBDT: 6 raw, 6 spectral, 3 autocorrelation). Neural models use AdamW (lr 10^{-3}, weight decay 10^{-4}), 20 epochs, batch 16, weighted cross-entropy, and the best-validation-epoch checkpoint (macro-F_1 categorical, R^2 continuous). GBDT is HistGradientBoosting at a fixed 300 iterations with no validation selection.

Table 4: Raw-window baseline models. Neural models see the masked full context plus a coverage channel; GBDT sees 15 engineered features.

Model	Parameters	Input	Training	Checkpoint selection
CNN	39.5k	masked ctx + coverage	AdamW, 20 ep, batch 16	best val epoch
FNO	105.9k	masked ctx + coverage	AdamW, 20 ep, batch 16	best val epoch
Patch Transformer	246.7k	masked ctx + coverage	AdamW, 20 ep, batch 16	best val epoch
GBDT	–	15 eng. features	HistGradBoost, 300 iter	none (fixed budget)

Label definitions

Dynamic labels use all target variates of the series (which for BOOM windows is every variate; there are no held-out covariates); the last context patch is the final P observations ($P = 64$ Toto, $P = 8$ MOMENT); the robust scale is 1.4826 times the median absolute deviation, floored at 10^{-6}. Taxonomy: cardinality bins are 1, 2–8, 9–32, and above 32; metric type and domain keep singleton primary entries and map other/multilabel rows to auxiliary, which are dropped from fits; cadence is the raw BOOM frequency string (the Long bucket yields no 1024-context window).

Donor exchanges

Toto donor-exchange targets are test windows in the lowest future-burstiness quartile; high-burst donors are top-quartile windows from other series; randomized donors come from all eligible test windows of other series with no covariate or taxonomy matching. Each donor window is used at most once per arm; the same 40 triples are reused across blends. Forecasts use 16 samples over the

64-observation next patch with CPU/CUDA RNG state shared across arms. Forecast burstiness is the maximum over variates and steps of $|\text{forecast median} - \text{context median}|/\text{robust scale}$; WAPE is $\sum_{i,t} |y_{i,t} - \hat{y}_{i,t}| / \sum_{i,t} |y_{i,t}|$ over target variates i and forecast steps t (denominator floored at 10^{-6}). For BOOM windows the target variates are all variates of the series, with no held-out covariates. MOMENT's matched interchange instead patches its validation-selected layer-3 all-context view: real donors have future burstiness above the 0.75 quantile with the target's variate count; matched null donors are below it, drawn from the $K = 5$ nearest context-matched candidates with taxonomy penalties. Its forecast endpoint is future MAE from the official short-forecast head.

Zero-shot probe transfer

The BOOM-trained probe (weights plus BOOM-train standardization) is applied without refitting to external activations at the same view: one validation-selected probe per continuous label, context 1024, up to 100 series by 4 windows per dataset across 11 FEV configurations (`entsoe_15T`, `entsoe_30T`, `entsoe_1H`, `epf_be`, `epf_de`, `epf_fr`, `epf_np`, `epf_pjm`, `rohlik_sales_1D`, `solar_with_weather_15T`, `solar_with_weather_1H`) and 4 LSTF datasets (ETTh1, ETTh2, weather, electricity). The FEV configurations come from ENTSO-E Transparency Platform load data via Open Power System Data (`entsoe`) [Open Power System Data, 2020], the open-access day-ahead electricity-price benchmark (`epf`) [Lago et al., 2021], the Rohlik Sales Forecasting Challenge data (`rohlik_sales_1D`) [MichalKecera, 2024], and a Kaggle solar and weather dataset (`solar_with_weather`) [Saman, 2023]. ETTh1/ETTh2 come from ETDataset [Zhou et al., 2021], electricity from the UCI ElectricityLoadDiagrams20112014 data [Trindade, 2015], and weather from the Max Planck Institute for Biogeochemistry weather records in Jena [Max Planck Institute for Biogeochemistry, 2020]. Within each resplit, transfer averages dataset-level R^2 with equal weights.

Compute

The main five-resplit suite ran on a CPU cluster (16 CPUs; about 1:07–1:25 elapsed, 18.7–21.3 GB peak RSS). GPU stages on NVIDIA A10 nodes took 21–37 min (28.6–29.1 GB) for MOMENT activations, 3–4 min (6.7–9.0 GB) for the Toto donor exchange, and about 1–2 min (1.7–3.0 GB) for the MOMENT matched interchange.

B Within-BOOM Held-Out-Combination Stress Tests

We also evaluated rotated held-out-combination tests: in each resplit, a support-valid domain or taxonomy combination was withheld from both training and validation, then the matching test subset was evaluated. Because different values were withheld across resplits, the brackets in Table 5 are coverage ranges across held-out values, not confidence intervals for a shared population. These tests characterize within-BOOM coverage under the tested support.

Table 5: Rotated within-BOOM held-out-combination tests. Each bracket spans scores from distinct held-out values across the five resplits; it is not an uncertainty interval.

Target	Held-out support	Toto macro-F_1 range	Random init. range
Cadence	Domain × metric type	$0.865\ [0.627, 1.000]$	$0.541\ [0.250, 1.000]$
Cadence	Domain	$0.654\ [0.513, 0.766]$	$0.511\ [0.420, 0.575]$
Metric type	Domain × cadence	$0.258\ [0.130, 0.472]$	$0.179\ [0.135, 0.224]$
Metric type	Domain	$0.281\ [0.188, 0.505]$	$0.220\ [0.161, 0.266]$

C Additional Diagnostic Tables

Table 6 gives the Toto donor-exchange endpoints summarized in §5.4 and Figure 4; Table 7 gives the zero-shot probe transfer results of §5.5; Tables 8 and 9 report MOMENT's dynamic readouts and matched interchange.

Table 6: Toto donor exchange: fraction of the 40 targets per resplit whose forecast is burstier under the high-burst than under the randomized donor, and the median WAPE ratio (high-burst/randomized). Means $\pm$ half-widths across resplits; the same triples are used at every blend. Donors come from other series and are not matched on covariates or taxonomy. Probe check: 0.655 ± 0.136 at every blend.

Blend β	Forecast burstier	Resplits > 0.5	Median WAPE ratio
0.25	0.480 ± 0.086	1/5	0.969 ± 0.062
0.50	0.430 ± 0.046	0/5	0.952 ± 0.108
1.00	0.390 ± 0.127	1/5	0.903 ± 0.155

Table 7: Zero-shot coordination-probe transfer (BOOM-trained probe, no refitting). R^2 is macro-averaged over the datasets of each set within a resplit, relative to a constant-mean reference; means $\pm$ half-widths across resplits.

Benchmark set	Entries	Coordination R^2
FEV	11	-0.938 ± 0.654
LSTF	4	-17.668 ± 10.896

Table 8: MOMENT dynamic readouts on held-out BOOM series. Views are selected on validation R^2 and evaluated once on test series. The near-constant sparsity targets are omitted; their six-statistic R^2 values are unstable. Means $\pm$ half-widths (§4).

Dynamic label	MOMENT R^2	Random init.	Six-stat.	MOMENT above six-stat.
Coordination	0.092 ± 0.047	0.083 ± 0.041	0.071 ± 0.041	3/5
Current burstiness	0.024 ± 0.025	0.007 ± 0.018	0.057 ± 0.620	2/5
Future burstiness	0.015 ± 0.027	0.002 ± 0.021	-0.005 ± 0.082	2/5
Shift risk	-0.006 ± 0.067	-0.018 ± 0.062	0.021 ± 0.076	1/5

MOMENT's dynamic readouts are comparable to the six-statistic probe (Table 8). Coordination is above it in 3/5 resplits and future burstiness in 2/5; the six-statistic probe is highly variable. The cross-model readout claim therefore covers cadence, metric type, and domain. For the matched interchange (Table 9), the probe check is blend-invariant and reported in the caption; only the forecast endpoint varies by blend.

Table 9: MOMENT matched interchange. Each resplit evaluates 40 held-out target/real/matched-null triples with the same triples reused at each blend. Probe check (blend-invariant): 0.590 ± 0.071, above 0.5 in 5/5 resplits. The lower future-MAE fraction is the forecast endpoint. Means $\pm$ half-widths (§4).

Blend β	Lower future-MAE fraction	Resplits > 0.5
0.25	0.410 ± 0.084	0/5
0.50	0.365 ± 0.109	0/5
1.00	0.405 ± 0.104	1/5

D All Raw-Window and Backbone Baselines

Table 1 reports the strongest raw-window model for each label; Table 10 lists every neural and tree raw-window model alongside the two backbone baselines, random initialization and block permutation.

Table 10: Held-out test macro-F_1 of Toto and all raw-window and backbone baselines over five resplits. Neural raw-window checkpoints are selected on validation data. Means $\pm$ half-widths (§4).

Target	Toto	Random init.	Block perm.
Cadence	0.766 ± 0.024	0.547 ± 0.040	0.567 ± 0.020
Metric type	0.545 ± 0.037	0.398 ± 0.035	0.420 ± 0.016
Domain	0.484 ± 0.011	0.362 ± 0.017	0.394 ± 0.011
Cardinality	0.471 ± 0.029	0.250 ± 0.025	0.428 ± 0.127

Target	CNN	FNO	Patch Transformer	GBDT
Cadence	0.545 ± 0.019	0.633 ± 0.047	0.553 ± 0.027	0.580 ± 0.042
Metric type	0.431 ± 0.047	0.366 ± 0.061	0.398 ± 0.018	0.498 ± 0.007
Domain	0.432 ± 0.025	0.381 ± 0.069	0.437 ± 0.029	0.482 ± 0.040
Cardinality	0.959 ± 0.010	0.961 ± 0.016	0.948 ± 0.023	0.663 ± 0.038

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction state the claim as linear recoverability from Toto's frozen residuals, identify cadence and metric type as the labels more recoverable from those residuals than from the tested input and backbone baselines on BOOM, and scope the domain, cardinality, forecast, and zero-shot probe transfer claims by their baselines (§5).
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The paper scopes its claims to five resplits of one BOOM corpus with unknown pretraining overlap, states the narrower baselines of the common-support and MOMENT comparisons, and reports readout–forecast dissociation and negative zero-shot probe transfer (§3, §4, §5.2, §5.3, §5.4, §5.5, §6).
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors

should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [N/A]
   Justification: The paper is empirical and presents no formal theorems.
   Guidelines:
   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.
4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: The paper and App. A specify the model checkpoints, data sources, resplits and seeds, candidate views, probe settings, baselines, donor exchanges, and zero-shot probe transfer protocol needed to reproduce the main results.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: The code, audited result files, a CPU verifier for the reported tables and figures, and setup and GPU rerun instructions are publicly available at `https://github.com/DhyeyMavani2003/toto-interp/tree/neurips-2026-v1`. The repository contains no dataset files or model weights; its scripts download BOOM, FEV, Toto, and MOMENT from Hugging Face and LSTF from the Time-Series-Library links.
   Guidelines:
   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: The paper specifies the checkpoints, five seeded series-disjoint resplits, labels, activation views, probes, baselines, and donor exchanges (§3, §4, §5, App. A).
   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: The Student-$t(4)$ half-widths describe variation across five resplits of one BOOM corpus, not significance tests or independent-corpus intervals; donor triples are reused across blends (§4, §5, App. A).
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: The Compute paragraph of App. A reports the CPU and GPU allocations, elapsed times, and peak memory for the main five-resplit stages.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: The work uses existing model checkpoints and benchmark data, involves no human-subject research, and makes no deployment-safety or fairness guarantee.
   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [Yes]
    Justification: The discussion states that the audit helps separate recoverable information from evidence of use, that a probe score treated as a steering target could produce unintended forecast changes, and that the donor-exchange readout is not a validated forecast-control knob (§6).
    Guidelines:
    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?
    Answer: [N/A]
    Justification: The work introduces no new forecasting model or dataset; it evaluates published models, describes an audit protocol, and does not present its exchanges as deployment controls.
    Guidelines:
    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [Yes]
    Justification: All assets and their original sources are cited (§3, App. A). Licenses: Toto-Open-Base-1.0 and BOOM (Apache-2.0), MOMENT-1-base (MIT), ETDataset (CC BY-ND 4.0), and the UCI electricity and MPI-BGC Jena weather data (CC BY 4.0). The FEV configurations are used under FEV's research-use terms and credited to their original sources. No dataset or model weights are redistributed.
    Guidelines:
    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.
13. **New assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [Yes]
    Justification: We release analysis code and aggregate experimental results. The repository linked in item 5 documents the protocol, dependencies, execution steps, compute needs, limitations, and license terms. No new dataset or model weights are released.
    Guidelines:
    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.
14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    Answer: [N/A]
    Justification: The work includes no crowdsourcing or human-subject research.
    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work includes no human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the research method, data processing, or experiments described in the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.