THE PRICE OF A LABEL: STATIC IDENTITY IN HEADS ON FROZEN TIME-SERIES MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adding a store or platform label to a frozen time-series foundation model raises three separate questions: whether the label reaches the model (access), whether a fitted head gains from it (fitted value), and whether the gain carries to new data (transfer). We give an exact erasure condition: a static channel that enters only through centered values, with no bypass, loses its identity exactly. On 70 public retail series, three of five released forecasters give identical forecasts under every tested relabelling, and a fourth model's response is traced to unobserved entries and removed by neutralizing them. The standard permuted-label ablation can mislead: on one public retail subset at a frozen checkpoint, true store codes beat permuted codes over ten paired seeds yet lose to a label-free input code shared by all series, which keeps the lower mean loss in five further seeds, where true versus permuted is unresolved. On 3,087 development-seen fundraising campaigns, equal-width heads using true labels improve later seven-day path loss beyond group-mean and dispersion calibration by 0.0154 and 0.0048 nats on two frozen representations, with favorable adjusted seed intervals; for transfer, the same law yields no store-code contrast with a favorable adjusted interval on a public retail panel, so fitted value is population-specific. On 3,022 public crowdfunding campaigns, ordinary history recalibration is a strong reference, and adding centered category and country indicators lowers the mean loss by a further 0.03254 nats. A favorable true-versus-permuted contrast is therefore not by itself evidence that identity helps: label ablations need shared-code and ordinary-calibration references.

1 INTRODUCTION

Store, platform, department, category and country labels are often attached to frozen time-series foundation models, either through a covariate slot or through a trainable head. These are group metadata rather than unique entity identifiers, and the usual evidence that such a label helps is a permuted-label ablation: refit with shuffled labels and check that the true labels score better. That comparison bundles three separate questions. Can the label reach the model through its released interface (*access*)? Does a head fitted with it predict better than matched and ordinary controls (*fitted value*)? On which population does the gain hold (*transfer*)? We answer each with its own test, using exact interface interventions and matched readouts on frozen features.

Figure 1(c) shows why the separation matters. On one public retail subset at one frozen checkpoint, true store codes beat permuted codes but lose to a label-free input code shared by every series, measured by seven-day path negative log-likelihood. A true-versus-permuted comparison alone would credit the store label with a gain that a label-free shared input exceeds.

Our contributions are as follows.

- *An exact erasure condition for released interfaces* (Section 2). A covariate that enters only as its own centered sequence, with no other path, becomes the same input for every label value. A census of five released forecasters on 70 public retail series finds three that return bit-identical forecasts under every static-label change and a fourth whose response we trace to unobserved covariate entries that the condition excludes; neutralizing those entries removes the response. The fifth responds by about 2.5×10^{-5} of its mean forecast.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

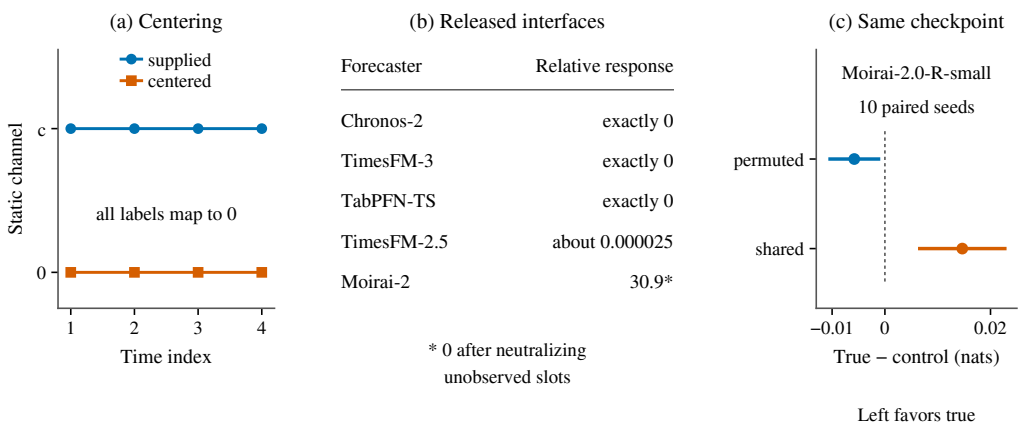


Figure 1: Label access and fitted value are separate questions. (a) Centering erases a static channel that has no other path; (b) relative forecast response to static relabelling (Table 1). (c) On one Moirai-2 checkpoint and retail subset, true codes beat permuted codes but lose to a shared code (seven-day path NLL, ten paired seeds, adjusted intervals).

- *A counterexample to the standard label ablation* (Section 5). The shared-code result above holds with adjusted paired-seed intervals on both contrasts over ten paired seeds, and both the original shared code and an equidistant shared code (Appendix G) keep the lower mean loss in five further seeds, although true versus permuted is unresolved in those seeds, so a label ablation needs a shared-input reference as well as a permuted one.

- *Fitted value beyond ordinary calibration* (Sections 3–4). Equal-width count heads read true, constant or permuted platform codes on frozen Chronos-2 and Moirai-1.1-R-large features (Ansari et al., 2025; Woo et al., 2024). On 3,087 development-seen fundraising campaigns, true codes improve later seven-day path loss over constant and permuted codes, and beyond group-mean and dispersion calibration by 0.0154 and 0.0048 nats on the two representations, with favorable adjusted seed intervals.

- *Ordinary calibration as a reference* (Section 5). On 3,022 public crowdfunding campaigns, ordinary history recalibration gives lower mean loss than an earlier metadata-only calibration, and adding centered category and country indicators lowers it further.

The practical message is direct: a favorable true-versus-permuted contrast is not by itself evidence that identity helps. Checking interface access, then comparing a fitted label against shared-code and ordinary-calibration references on the population of interest, shows whether a label reaches the model and where its fitted gain holds.

2 A STATIC LABEL CAN VANISH AT THE INPUT INTERFACE

Setting. A frozen trunk is a pretrained forecaster with unchanged weights. Its interface exposes target-history and possibly covariate slots. A trainable head reads its features and returns a forecast distribution. A static label, such as platform or department, stays constant over the history; it may enter through a covariate slot or reach the head directly.

Many released interfaces normalize each input channel by its own location and scale before the model sees it. Centering a constant channel removes its value unless another input path preserves it, which is panel (a) of Figure 1. A department code repeated over a history can therefore enter a forecasting interface without distinguishing departments in the predictor.

Proposition 1 (A sufficient condition for static invariance)**.** *Fix the target history, channel slots, observation masks and output coordinates. Suppose every static covariate enters the predictor only as its observed values minus a location statistic, divided by a positive scale, where the location statistic returns c whenever all observed values equal c, as the mean and the median do; unobserved entries are filled by a fixed rule after normalization, and a fixed map such as* $\operatorname{arcsinh}$ *may follow. If neither its identity, original value nor normalization statistics reaches the predictor by another path, changing that constant leaves the predictor unchanged.*

Centering makes every observed entry zero, so all labels produce identical inputs. This is an exact-arithmetic statement; stochastic predictions are compared in law or with a common random draw.

It is the static-covariate counterpart of channel identifiability (Lee et al., 2025): there, a model that treats channels alike gives identical outputs to channels with identical inputs, whereas here the interface itself turns every constant label into the same input, so no architecture behind it can recover the label.

An identity embedding, a retained channel mean or label-dependent padding violates the premise, and Appendix A gives the proof and counterexamples. A head-side embedding is one path that keeps the label, not the only one.

Categorical labels need a separate premise. Some interfaces encode a categorical covariate before normalizing it; Chronos-2, for example, fits a target encoder on each task's history when the target is univariate. An encoder that depends on the categorical sequence only through which positions share a category, and on the target values within those positions, cannot tell a category's name from any renaming of it, and a constant categorical sequence has a single such part, so its encoding cannot reveal which category it is. Equivariance alone would not suffice: a one-hot code read through distinct weights still separates every category (Appendix A).

What erasure can cost. Let B collect *all* nonlabel inputs available to the readout: frozen features, retained location and scale, calendar indicators and auxiliary history features. For a discrete target Y with finite conditional entropy, the unrestricted optimal log-score gain from additionally observing label R is

$$I(Y; R \mid B), \tag{1}$$

the information the label carries about the future beyond those inputs (Appendix A). This classical identity is not generally $I(Y; R \mid \phi(H))$; that expression requires $B = \phi(H)$. It is zero when B already reveals the label, while a restricted fitted head can still gain from an explicit code through its function class or estimation. We therefore measure the fitted gain directly.

What a matched shuffle preserves. For one-hot label matrices Z, Z', equality $ZZ^\top = Z'Z'^\top$ holds exactly when their observed groups agree up to a bijective renaming: each matrix entry records whether two rows share a label. A shuffle that breaks groups therefore changes the pooling kernel, even with the same label counts and raw parameter budget. The kernel matters because fitting cost depends on it. With M equally weighted contexts, each supplying r independent Gaussian training responses with means m and variance σ^2, a fixed linear smoother H applied to the training means has expected test squared error

$$R_H = \sigma^2 + \frac{\|(H - I)m\|^2}{M} + \frac{\sigma^2 \|H\|_F^2}{rM},$$

and for least squares $\|H\|_F^2 = \operatorname{rank}(H)$ (Appendix K). In a four-context example, equal-width linear heads with no label information still favor true over shuffled labels by $1/(8r)$ nats, because the shuffled design has rank three and the true design rank two. Conversely, a restricted head with the true code can beat the permuted code by $1/8$ nats although the label carries no conditional information, $I(Y; R \mid B) = 0$, because only the true code represents the mean. Fitted label contrasts therefore compare function classes and fitting costs as well as information, and classical risk correction accounts for the variance term. Head-side labels bypass normalization; constructed time-varying inputs are a separate intervention (Appendix G).

A census of five released forecasters. We passed static labels through the slots that five released forecasters expose, Chronos-2 (Ansari et al., 2025), TimesFM-3 (Jain & Sen, 2026), TimesFM-2.5

Table 1: Static-input responses on 70 distinct retail series, with thirty history days and seven forecast days. Each response is the largest absolute forecast change divided by that model's mean absolute forecast without covariates.

Forecaster	Response	Tested path
Chronos-2	0	Numeric and categorical; separate and cross-series
TimesFM-3	0	Native numeric covariate
TimesFM-2.5	2.5×10^{-5}	Covariate regression beside the forecaster
Moirai-2	30.9	Numeric covariate with unobserved entries
TabPFN-TS	0	Numeric and static-feature slots, official time-series checkpoint

(Das et al., 2024; Google Research, 2026), Moirai-2 (Liu et al., 2025) and TabPFN-TS (Hoo et al., 2026), holding the target histories fixed, on 70 public retail series that cover all ten stores and seven departments of the panel (Table 1 and panel (b) of Figure 1). Every model received the same covariates: constants, the department code held constant, time-varying controls and, for categorical or static slots, a placeholder label, the true departments, a reassignment of departments across series and a relabelling of their names. Response bands and time-varying controls were fixed before outputs existed (Appendix A). No future target is read, so the census measures sensitivity. None of the five interfaces has a static channel into the pretrained network: two state that static or all covariates are dropped, TimesFM-2.5 accepts static covariates only in a linear regression fitted beside the frozen forecaster, and the rest expose dynamic slots only.

Chronos-2, TimesFM-3 and TabPFN-TS return bit-identical forecasts under every tested static change, while time-varying controls move their forecasts, consistent with the proposition on these inputs. Moirai-2 responds through an identifiable unobserved-entry path: scaling turns each covariate fill into minus the label divided by $\sqrt{10^{-5}}$. A controlled intervention keeps the same 70 thirty-day histories and seven-day horizon, changing only the eleven unobserved covariate values, including left padding and future positions, to zero immediately before input projection. All four static contrasts then give bit-identical encoder representations and target forecasts. A sham intervention preserves every captured tensor, and the time-varying control still changes forecasts by up to 3.54, so the intervention isolates the combined path of these eleven entries (Appendix A). TimesFM-2.5 responds weakly on every tested path, by up to about 2.5×10^{-5} of its mean forecast; with per-series target normalization its static categorical response, 1.3×10^{-5}, is below the 1.9×10^{-5} produced by merely renaming the categories. Released paths, library versions and the complete census are in Appendix A.

Features as well as forecasts. An earlier probe of 48 retail windows from two distinct series finds exact equality in Chronos-2 embeddings as well as forecasts under the tested department relabellings, while dynamic controls change the forecasts (Appendix I).

3 THE PRICE OF A LABEL CARRIED IN THE HEAD

A sparse, bursty event panel. Our main panel is a proprietary daily record of 21,625 fundraising campaigns and 4,997,177 daily records, where the count is the number of new investors a campaign gains on a day. Every head reads daily increments of the recorded cumulative totals. In the eligible increment census, zero counts account for 86.58% of observations when campaigns are weighted equally (Appendix J). Events cluster in time: the chance of at least one event in the next seven days is 4.381 $[4.132, 4.654]$ times higher after a day with an event than after a quiet day, as the geometric mean of within-campaign risk ratios over the 3,506 campaigns for which the comparison is defined, with a 95% campaign-bootstrap interval.

Three labellings of one head. We fit one negative-binomial count head that reads frozen trunk features and assigns a probability to every seven-day vector of daily investor increments, each allowed to exceed one. Its inputs include an embedding table for the 37-slot training platform vocabulary, with 36 named categories and a shared `other` slot. Only 27 codes occur in the primary cohort (Table 10). Three arms differ only in the integer code handed to that table. The true arm uses the recorded platform. The constant arm assigns one code to every campaign and carries no between-campaign

platform information. The permuted arm randomly reassigns codes among campaigns, retaining the label multiset and table width while breaking the link between a campaign and its platform. Every arm has the same parameters and the same embedding width, and every arm is refitted rather than scored with permuted inputs, so the permuted and constant arms are refit-based importance controls (Hooker et al., 2021). Writing $S(\cdot)$ for the score of the head fitted and scored with a given labelling, the price of the label has two sides,

$$\Delta_{\text{const}} = S(\text{true}) - S(\text{constant}), \qquad \Delta_{\text{perm}} = S(\text{true}) - S(\text{permuted}), \tag{2}$$

and a negative value favours the true code. Δ_{const} is a refitted predictive-score contrast; Δ_{perm} preserves the scored population's label multiset and stored parameter budget while breaking the campaign-to-code link. On the later window the permutation is drawn within the scored campaigns, so that their label histogram under the permuted arm equals the true one code by code. An earlier permutation, drawn over all development campaigns, is reported beside it; on the scored campaigns it carries 37 distinct codes while their own true codes use 27, and the matched permutation removes this mismatch. The score is the negative logarithm of the probability the head assigns to the observed seven-day counts, in nats, averaged within each campaign and then equally across campaigns; lower is better. Each arm is fitted with ten paired seeds, using twenty epochs of AdamW at learning rate 3×10^{-4}, weight decay 10^{-4}, gradient clipping at one and batches of 2,048, and scored once at the fixed final checkpoint. Nine further permutations crossed with the same seeds measure the choice of permutation. Appendix B gives the complete protocol.

A later-time test of fitted label value. Protocols fixed before their respective fits refit every arm on the development data and score them, without further update, on a later window of the same source: 105,995 forecast origins from 3,379 campaigns, with origin days from 22 October to 1 December 2025. The primary population is the 3,087 of those campaigns in the development rosters, comprising 2,772 training campaigns and 315 validation-only campaigns; all 3,379 form a secondary population, and 607 later campaigns are absent from training, including 292 absent from both development rosters. Each cell is fitted once on Chronos-2 features (Ansari et al., 2025) and once on Moirai-1.1-R-large features (Woo et al., 2024). The refitted development arm means reproduce those of Appendix B to within 3.7×10^{-7} nats.

Table 2: Later-window scores on 3,087 campaigns over ten seeds, in nats per seven-day path. Negative differences and positive median gains favor true; seed intervals are paired and adjusted within each comparison family, and median intervals resample campaigns. Matched permutations preserve the scored label histogram; earlier ones use the development roster.

	Chronos-2	Moirai-1.1-R-large
Mean: true, constant, matched permuted	0.1088, 0.1237, 0.1221	0.0948, 0.1008, 0.1015
True minus constant, by seed	-0.0149 $[-0.0213, -0.0084]$	-0.0060 $[-0.0090, -0.0030]$
True minus matched permuted, by seed	-0.0133 $[-0.0252, -0.0014]$	-0.0067 $[-0.0093, -0.0042]$
True minus earlier permuted, by seed	-0.0147 $[-0.0283, -0.0012]$	-0.0060 $[-0.0086, -0.0034]$
True beats matched, of 100 fits	91	97
True beats earlier, of 100 fits	95	100
Median gain over constant, by campaign	0.00166 $[0.00164, 0.00167]$	0.00190 $[0.00188, 0.00191]$
Median gain over matched, by campaign	0.00119 $[0.00115, 0.00122]$	0.00146 $[0.00142, 0.00149]$
Campaigns where true beats constant	2,924	3,007

Favorable contrasts on both trunks. On Chronos-2 features the true code scores 0.0149 nats below the constant code and 0.0133 below the matched permuted code; on Moirai-1.1-R-large features it scores 0.0060 and 0.0067 below them (Table 2). All four of these adjusted seed intervals exclude zero. The true code beats the constant code in ten of ten seeds on both trunks, and beats the matched permuted code in nine of ten seeds on Chronos-2 features and ten of ten on Moirai features. Across ten permutations, 91 of 100 Chronos-2 fits and 97 of 100 Moirai fits favour the true code over the matched permuted one, with two-way seed-by-permutation intervals of $[-0.0170, -0.0060]$ and $[-0.0069, -0.0039]$; against the earlier permutation the counts are 95 and 100 of 100, with intervals of $[-0.0196, -0.0097]$ and $[-0.0077, -0.0049]$. Favorable signs occur across most campaigns: with the ten seeds averaged within each campaign, the true code scores lower than the constant code for 2,924 of the 3,087 campaigns on Chronos-2 features and for 3,007 on Moirai features, and the

campaign median gains are favorable with intervals that exclude zero. The true code also has the lowest mean in the secondary population of all 3,379 campaigns, with means of 0.3487, 0.3663 and 0.3986 for the true, constant and earlier permuted arms on Chronos-2 features. Appendix B retains all seeds, permutations and checks that the label path is active.

4 LABEL VALUE BEYOND ORDINARY CALIBRATION

A label can help a fitted head simply by shifting group means or dispersions, which ordinary calibration also does. We therefore compare the identity head with ordinary calibrations that adjust the means and dispersions of the constant head.

Beyond ordinary mean and dispersion calibration. A forty-fit comparison establishes gains beyond training-only platform mean offsets, with and without unit quadratic shrinkage: against group-mean offsets the head gains 0.0152 nats on Chronos-2 and 0.0053 on Moirai-large, with four-contrast-adjusted seed intervals and descriptive campaign intervals that exclude zero (Appendix C). A two-component Shapley accounting of the saved distributions assigns 0.00632 nats of the Chronos-2 margin over group calibration to the mean and 0.00890 to the dispersion, and 0.00176 and 0.00352 on Moirai-large (Appendix D), so we add dispersion to the control. We retain the constant heads and their mean offsets, then fit bounded global or encoded-platform log-dispersion shifts on training data only. At later scoring, a platform name absent by the training cutoff receives no group shift; all eligible later names were observed by then (Table 10). We fit 60 calibrators and report six prespecified populations. The full head beats both group and ridge mean-plus-dispersion controls on both feature extractors under the complete 36-comparison adjustment (Table 3); each subtraction uses the same feature extractor. Against the group control the margins are 0.0154 nats on Chronos-2 and 0.0048 on Moirai-large.

Table 3: Identity head minus sequential mean-and-dispersion calibration on 3,087 campaigns, in nats per seven-day path. Negative favors the head; intervals use ten paired seeds with Bonferroni adjustment over all 36 comparisons. Campaign medians average seeds first.

Features	Control	Mean	Adjusted seed interval	Campaign median
Chronos-2	Global	-0.01559	$[-0.02459, -0.00659]$	-0.00172
Chronos-2	Group	-0.01540	$[-0.02450, -0.00629]$	-0.00082
Chronos-2	Ridge	-0.01569	$[-0.02494, -0.00644]$	-0.00110
Moirai-large	Global	-0.00595	$[-0.01021, -0.00170]$	-0.00190
Moirai-large	Group	-0.00483	$[-0.00911, -0.00055]$	-0.00156
Moirai-large	Ridge	-0.00520	$[-0.00951, -0.00089]$	-0.00167

Graded labels. Coarsening the true code into 2, 4 or 8 groups, fixed in advance by development campaign counts, shows that true coarse labels can recover part of the gain, and the exact code scores lowest on both trunks (Appendix B, Table 6). Every coarse label is true, so this ladder varies label information without introducing wrong labels.

5 SHARED-INPUT AND ORDINARY-CALIBRATION REFERENCES

Under seed-unit intervals, the fundraising gains pass permuted, constant and calibration controls. Two further references on public data show what a label ablation should include: a label-free input shared by every series, and an ordinary recalibration that uses the same history.

True codes can lose to a shared code. On one frozen Moirai 2.0 checkpoint and a fixed subset of M5 retail (Makridakis et al., 2022), six equally sized heads use fixed histories and either an erased covariate, a head-side label or an input code injected *after* normalization. On the 1,750-series primary population, the true input code beats a permuted input code by 0.00580 nats per seven-day path, with four-comparison-adjusted paired-seed interval $[-0.01072, -0.00088]$ for true minus permuted. Yet true minus a label-free shared input code is $+0.01466$ $[+0.00627, +0.02304]$: all entities receive the

same codeword, which is also one of the ten class codewords (Appendix G); a reserved codeword at equal distance from all ten class codes also has lower mean loss than the true code in five further seeds (Appendix G.2). A true-versus-permuted ablation alone would credit the store codes with a gain that a shared, label-free input exceeds, so label ablations need a shared-input reference.

Ordinary history recalibration is a strong reference. On 3,022 MDCC crowdfunding development campaigns (Xu et al., 2023), whose seven-day target follows a first-week history, a retrospective mean-and-dispersion recalibration freezes a 301-parameter negative-binomial base whose 42-dimensional input contains an intercept, a 32-dimensional projection of frozen Moirai-2 features, seven log-count histories and native location and scale. Residual designs use an intercept alone, all 42 history features, or those features plus 18 category and country indicators centered on training campaigns. History features alone give a lower mean NLL than the earlier metadata-only calibration (Table 4), so recalibrating on the same history is itself a strong reference. Adding training-centered category and country indicators to that recalibration lowers the mean NLL by a further 0.03254 nats, with campaign percentile 95% interval $[-0.05622, -0.00593]$ (Appendix M).

Table 4: Calibration controls on 3,022 development campaigns. Lower NLL is better (nats per seven-day path). Parameters are incremental, except the frozen base.

Predictor	Parameters	NLL
Frozen history-only base	301	6.97896
Earlier metadata calibration	266	6.87159
Intercept-only recalibration	14	6.97859
History-feature recalibration	588	6.83551
History + metadata recalibration	840	6.80297

A three-part test for a static label. These results suggest a simple protocol. First, relabel the static input on fixed histories and compare outputs bit for bit, with a time-varying control that shows the path is live; bit-identical outputs mean the label did not influence the forecast through that path on the tested inputs and version (Section 2). Second, refit the head with true, constant, permuted and shared codes, and compare it with ordinary mean-and-dispersion recalibration on the same history (Sections 3–5). Third, report each contrast separately for the populations that matter, such as entities seen and unseen in training, since a fitted gain on one population need not carry to another.

6 Related work

Normalization and identity. RevIN restores instance statistics; Non-stationary Transformers addresses over-stationarization (Kim et al., 2022; Liu et al., 2022). Reversible normalization can also contain redundant or harmful components (Berthelier et al., 2026b). Channel Normalization uses channel-specific affine parameters; Adaptive CN conditions them on each input, and Prototypical CN supports unknown or varying channel counts and foundation-model use (Lee et al., 2025). STID trains identity embeddings to distinguish series (Shao et al., 2022). We instead test static-label erasure on specified released frozen paths and the fitted value of retained codes beyond ordinary controls.

Adding side information to pretrained forecasters. UniCA embeds heterogeneous covariates; ChronosX adapts them modularly (Han et al., 2026; Arango et al., 2025). Chronos-2, TimesFM and TabPFN-TS expose different covariate interfaces (Ansari et al., 2025; Das et al., 2024; Google Research, 2026; Jain & Sen, 2026; Hoo et al., 2026). Our TabPFN-TS census covers the v1.2.0 per-series forecasting path (Prior Labs, 2026). Chronos-2 group attention shares information among series; its group IDs specify membership rather than learned label semantics. Tiny Time Mixers embeds static categories in a fine-tuned decoder (Ekambaram et al., 2024; TSFM project contributors, 2026). MetaTST encodes metadata with a frozen language model and trains the forecasting Transformer (Dong et al., 2024). TFMAdapter adds historical and horizon covariates to a frozen forecaster, with an oracle future-target probe of covariate use and an uncertainty-based fallback (Dange & Sarawagi, 2025). ContextFormer admits time-invariant context while freezing the base except its final layer (Chattopadhyay et al., 2025). QKCV co-trains patching layers and a static-category encoder

in its main pretrained setting. Its item/store attributes and 28-day M5 point forecasts differ from our seven-day store-label path-loss comparison on frozen features (Wang & Ma, 2025). CoRA injects scored exogenous covariates into a frozen backbone (Qin et al., 2025). Internal conditioning and output-side adaptation are established. We test released-path retention and fitted gains against constant, multiset-matched permuted and shared codes rather than proposing a new embedding or placement ranking.

Measuring what reaches the predictor. A controlled covariate-use study supplies and withholds inputs on two interfaces we probe. In one constructed aggregate task, one model's error barely changes while the other's falls by orders of magnitude (Berthelier et al., 2026a). A 100-task benchmark includes 46 covariate tasks, 19 with static covariates, but ranks accuracy rather than interface response (Shchur et al., 2025). MoTime evaluates external modalities with varying histories and cold start (Zhou et al., 2025). We relabel a fixed history to separate input invariance from forecast accuracy. Entity embeddings, DeepAR's categorical embeddings with a negative-binomial likelihood, and the temporal fusion transformer's static encoders retain identity during training (Guo & Berkhahn, 2016; Salinas et al., 2020; Lim et al., 2021). We study concrete released paths of already frozen trunks.

Contribution relative to prior work. The contribution is the combined interface and fitted-value study: exact static-input invariance on specified released paths, a traced unobserved-entry exception, and controlled label contrasts under two frozen representations. Refitted variable importance already distinguishes information from fitted performance, including gains from noise features (Hooker et al., 2021; Mentch & Zhou, 2022), and leave-one-covariate-out and nonparametric importance measures refit models without a feature (Lei et al., 2018; Williamson et al., 2021). Probing research makes the same separation for learned representations: control tasks with random labels show that a probe can score well by memorizing (Hewitt & Liang, 2019), and usable information under computational constraints (Xu et al., 2020) and conditional probing beyond a baseline (Hewitt et al., 2021) show that a restricted readout's gain from a variable is not its conditional information. Global/local forecasting similarly depends on the permitted complexity (Montero-Manso & Hyndman, 2021). Our categorical-kernel observation and Gaussian risk examples specialize these principles to label ablations. The interface interventions, later-time gains and shared-code counterexample establish what specified label comparisons measure and where their fitted advantages hold.

7 Conclusion

A static label on a frozen time-series model raises three separate questions, and each has its own test. The erasure condition identifies, before any fitting, released paths through which a static label cannot act, and released-path interventions confirm erasure on three interfaces and trace the Moirai-2 exception to its unobserved entries. Matched heads show fitted gains beyond ordinary mean and dispersion calibration on 3,087 development-seen campaigns under two frozen representations. Shared-code and ordinary-recalibration references show when a favorable true-versus-permuted contrast would overstate the value of identity. Label ablations built this way tell practitioners whether a label reaches the model and where its fitted gain holds. Appendix H maps the main claims to their evidence and scope.

8 Limitations

Uncertainty unit and period. The fundraising intervals are seed-unit intervals conditional on the fixed 3,087 development-seen campaigns. After averaging seeds and resampling campaigns, the mean intervals include zero against the constant code on both trunks and against both the matched and the earlier permutation on Chronos-2, although the medians remain favorable; median campaign gains are small, 0.00119 to 0.00190 nats per path. On 1,917 earlier development-validation campaigns, Chronos-2's true code improves on the permuted code but loses to the constant code on the mean, and Moirai does not resolve a true-versus-permuted gain; we have not identified why the ordering changes. The later window had been opened once before, for a different comparison, and was scored under four preregistered protocols without cross-protocol adjustment; it is reused by the calibration controls, so it is not an untouched confirmation. Only 3,379 of its 9,150 campaigns satisfy the origin rules.

Where the gain holds. On Moirai features the mean gain is concentrated: removing one 72-campaign group leaves 16% and 2% of the equal-campaign gains over the matched-permuted and constant codes, although every single-group removal preserves both signs. All 24 calibration intervals on the full 3,379, validation-only 315, outside-development 292 and training-unseen 607 campaigns include zero. Among the 292, true versus constant is unresolved and misleading-label cost resolves only over Chronos-2 seeds. In the secondary population the constant arm beats the earlier permuted arm. Matched-permuted minus constant is unresolved on both trunks, with no registered equivalence margin (Appendix B), and the coarsening ladder is not monotone.

Controls. Equal stored parameters do not match effective function classes or optimization trajectories, and the matched permutation does not preserve training-only label counts or the pooling kernel. Neither contrast in Equation 2 isolates a causal mechanism, and splitting Δ_{const} into Δ_{perm} plus permuted minus constant does not separately identify information and estimation cost. On Chronos-2 features and the 3,087 development-roster campaigns, the training-fitted calibrated controls score worse than the uncalibrated constant head on the later window (Appendix C). The sequential calibrations are nonnested with the head, their finite search does not certify a global optimum, and the mean/dispersion decomposition is a Shapley accounting, not causal attribution; individually, only Moirai's dispersion component is resolved after its multiplicity correction.

Public transfer. On previously used M5 data with the same negative-binomial law, none of eight primary store-code contrasts over 30,490 series has an adjusted interval favoring the true code: two favor the group-rate controls, by 0.0093 nats on Chronos-2, and six are unresolved (Appendix F). An earlier 300-fit study with a different count law also found no store-code gain; its secondary department label improved on both controls only on Chronos-2 (Appendix I). In the same-checkpoint study, both comparisons with training-only store-rate calibration are unresolved; in a five-seed follow-up, true versus permuted is unresolved and the original and an equidistant shared code both have lower mean loss than true, descriptively (Appendix G). On MDCC, the nine-launch-date interval of the label increment, $[-0.06183, +0.00784]$, crosses zero; the recalibrations differ in capacity (588 versus 840 parameters), all nine fits miss the strict gradient diagnostic, and the intervals are descriptive on already-opened development data. A metadata-conditioned support-pair model loses to ordinary mean-and-dispersion calibration by 0.10484 nats per path, and only two of its five comparisons favor it (Appendix L).

Scope of the theory and interfaces. The erasure condition is sufficient, not necessary; the census covers the tested inputs and library versions, and the Moirai-2 intervention isolates the combined unobserved-entry path rather than each position; TimesFM-2.5's weak response is not traced to a mechanism. Pooled TabPFN-TS constructions, later checkpoint releases and metadata-based regrouping in Chronos-2 are untested. The identity of Equation 1 neither predicts nor bounds a fitted head's gain, and our theory is an exact sufficient condition with specializations of classical principles, not a new learning principle. The retail census and fundraising heads are not one end-to-end experiment. UniCA's trained static path is live but does not resolve a department-specific gain against reassignment in three seeds, and its backbone, loss and capacity differ from our heads (Appendix A). Fundraising uses one private source, one head, two trunks and ten seeds; its data cannot be redistributed, and twelve of 120 public checkpoints are replayed, not retrained.

REPRODUCIBILITY STATEMENT

The interface census needs no private data. It uses 70 series drawn from the public M5 competition record (Makridakis et al., 2022), five publicly released forecasters at the versions named in Appendix A, and the relabelling controls specified there. A third party holding only those public artifacts can rerun it end to end. Its verdicts are within-run comparisons between label arms on one machine, so they do not depend on reproducing our absolute forecasts: a forecaster either returns the same bytes under a changed static label or it does not. The public retail tests of Section 5 and Appendix F and the UniCA comparison use public-access M5 data, subject to its competition terms, and the crowdfunding study uses the public MDCC data (Xu et al., 2023). For the negative-binomial comparison, all 120 fits and 120 calibrations completed; twelve prespecified checkpoints reproduce their cached-input scores, and independent reconstruction covers all saved-score statistics. This is neither all-checkpoint replay nor independent retraining. The earlier retail study's release includes summaries of all 300 matched-width fits, an independent summary replay, the readout and training-loop implementations, and a data-access description. Reconstructing those summaries is not a rerun of the original fits or feature extraction; the source data and pretrained checkpoints must be obtained separately.

The same-checkpoint study in Appendix G adds 60 neural fits and 30 training-only calibrations on a fixed M5 subset. Six prespecified checkpoints reproduce their validation and later parameter surfaces; independent reconstruction checks all saved probability scores and summary statistics. This is not a replay of all 60 checkpoints or a repeat of their training. The public extension releases all 27 comparisons, seed values, store summaries and an aggregate replay without redistributing M5 source rows or third-party weights.

What needs the private source is the fundraising panel: the identity price of Section 3 and the calibration controls and coarsening ladder of Section 4. The appendices specify all prediction objects, fitted families, paired seeds, the identity protocol with its controls, and score decompositions. The additional mean-and-dispersion controls retain the saved constant heads and fit 60 training-only calibrators. The release supplies their numerical calibration kernels, generated-input checks and all 36 aggregate comparisons, not the private cache pipeline or neural retraining. That fundraising source cannot be redistributed without its provider's permission, so public data, implementation code and saved aggregate outputs do not by themselves reproduce those experiments. For the matched-width retail study, twelve CPU copy-versus-original code-path checks give identical per-series score hashes. They do not reproduce earlier GPU optimization trajectories. Independent saved-score reconstruction covers the primary and secondary labels and populations within the declared numerical tolerance. The public package reconstructs a specified subset of summaries, not every bootstrap or original fit. A fresh environment installation and external replication of the complete fitted study remain unverified.

The subsequent retrospective campaign control releases all nine saved recalibrators, their complete training and validation diagnostics, the three selected score vectors and a separate saved-model replay. It reuses the fixed base and original data roles, not a new source or a new fit of the old models.

AI USE STATEMENT

Large language model assistants were used as research and writing tools: for literature search, discussing ideas, writing code, checking derivations and editing the text. The authors directed the work, verified every proof, result, number and citation, and take full responsibility for the paper and its supplementary material.

ETHICS STATEMENT

The main panel is a proprietary daily record of fundraising campaigns provided by an industry partner and used here for research only; the public replication uses public-access retail data under the applicable competition terms. Campaigns are organizations rather than individuals, no individual-level or causal claim is made, and only task-required fields are retained under the source licenses and attribution rules. No deployment or commercial use is claimed, and the paper states throughout where its evidence stops.

REFERENCES

Anonymous authors. Cardinality-aware event forecasting, 2026. Unpublished manuscript.

Abdul Fatir Ansari, Caner Turkmen, Oleksandr Shchur, and Lorenzo Stella. Fast and accurate zero-shot forecasting with Chronos-Bolt and AutoGluon. AWS Machine Learning Blog, 2 December 2024, 2024. URL https://aws.amazon.com/blogs/machine-learning/fast-and-accurate-zero-shot-forecasting-with-chronos-bolt-and-autogluon/.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Henri Joseph Senetaire, Ali Caner Turkmen, Oleksandr Shchur, Danielle C. Maddix, Michael Bohlke-Schneider, Bernie Wang, and Syama Sundar Rangapuram. ChronosX: Adapting pretrained time series models with exogenous variables. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2242–2250, 2025. URL https://proceedings.mlr.press/v258/arango25a.html.

Gaspard Berthelier, Mariia Baranova, Andrei-Tiberiu Pantea, Etienne Le Naour, Adrien Petralia, Tahar Nabil, and Themis Palpanas. Investigating simple target-covariate relationships for Chronos-2 and TabPFN-TS. arXiv preprint arXiv:2605.12200, 2026a. URL https://arxiv.org/abs/2605.12200. ICLR 2026 Workshop on Time Series in the Age of Large Models.

Gaspard Berthelier, Tahar Nabil, Etienne Le Naour, Richard Niamke, Samir Perlaza, and Giovanni Neglia. On the role of reversible instance normalization. arXiv preprint arXiv:2603.11869, 2026b. URL https://arxiv.org/abs/2603.11869.

Sameep Chattopadhyay, Pulkit Paliwal, Sai Shankar Narasimhan, Shubhankar Agarwal, and Sandeep P. Chinchali. Context matters: Leveraging contextual features for time series forecasting, 2025. URL https://arxiv.org/abs/2410.12672. Preprint, version 5.

Afrin Dange and Sunita Sarawagi. TFMAdapter: Lightweight instance-level adaptation of foundation models for forecasting with covariates. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 498–507. ACM, 2025. doi: 10.1145/3746252.3761272.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Jiaxiang Dong, Haixu Wu, Yuxuan Wang, Li Zhang, Jianmin Wang, and Mingsheng Long. Metadata matters for time series: Informative forecasting with transformers, 2024. URL https://arxiv.org/abs/2410.03806.

Bradley Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004. doi: 10.1198/016214504000000692.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *Advances in Neural Information Processing Systems*, volume 37, pp. 74147–74181. Curran Associates, Inc., 2024. doi: 10.52202/079017-2359.

Google Research. TimesFM (time series foundation model) source code, release 3.0.1. GitHub repository google-research/timesfm, commit 45e0a3b, 2026. URL https://github.com/google-research/timesfm/tree/45e0a3bc7fc4acef17b7ba7910488be2159bae5f. forecast_with_covariates in src/timesfm/timesfm_2p5/timesfm_2p5_base.py, lines 198–212.

Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737, 2016. URL https://arxiv.org/abs/1604.06737.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Yinbo Sun, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, and De-Chuan Zhan. UniCA: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=I8q4MZb4OP.

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 and the 9th International Joint Conference on Natural Language Processing*, 2019.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, 2021. doi: 10.18653/v1/2021.emnlp-main.122. URL https://aclanthology.org/2021.emnlp-main.122/.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending TabPFN-v2 to time series forecasting, 2026. URL https://arxiv.org/abs/2501.02945v4. Version 4, January 26, 2026.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):82, 2021. doi: 10.1007/s11222-021-10057-z.

Ayush Jain and Rajat Sen. TimesFM-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, 31 August 2026, 2026. URL https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Channel normalization for time series channel identification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33655–33687, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37 (4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012. URL https://doi.org/10.1016/j.ijforecast.2021.03.012.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. arXiv:2511.11698, 2025. URL https://arxiv.org/abs/2511.11698.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0718.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022. doi: 10.1016/j.ijforecast.2021.11.013.

Lucas Mentch and Siyu Zhou. Getting better from worse: Augmented bagging and a cautionary tale of variable importance. *Journal of Machine Learning Research*, 23(224):1–32, 2022. URL https://www.jmlr.org/papers/v23/20-1264.html.

Pablo Montero-Manso and Rob J. Hyndman. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4):1632–1653, 2021. doi: 10.1016/j.ijforecast.2021.03.004.

Prior Labs. TabPFN time series: Releases v1.1.0 and v1.3.0. https://github.com/PriorLabs/tabpfn-time-series/releases, 2026. Versioned release notes and covariate documentation, accessed September 25, 2026.

Guo Qin, Zhi Chen, Yong Liu, Zhiyuan Shi, Haixuan Liu, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. CoRA: Covariate-aware adaptation of time series foundation models. arXiv preprint arXiv:2510.12681, 2025. URL https://arxiv.org/abs/2510.12681.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3): 1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 4454–4458. ACM, 2022. doi: 10.1145/3511808.3557702.

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. arXiv preprint arXiv:2509.26468, 2025. URL https://arxiv.org/abs/2509.26468.

TSFM project contributors. TSFM: Time series foundation models (granite-tsfm), release v0.3.9. GitHub repository ibm-granite/granite-tsfm, 2026. URL https://github.com/ibm-granite/granite-tsfm/tree/v0.3.9. Static categorical inputs to the TinyTimeMixer decoder first appear in release v0.2.10.

Hao Wang and Baojun Ma. QKCV attention: Enhancing time series forecasting with static categorical embeddings for both lightweight and pre-trained foundation models. arXiv preprint arXiv:2510.20222, 2025. URL https://arxiv.org/abs/2510.20222.

Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Xovee Xu, Jiayang Li, and Fan Zhou. MDCC: A multimodal dynamic dataset for donation-based crowdfunding campaigns. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5417–5421, 2023. doi: 10.1145/3583780.3615124.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10689.

Xin Zhou, Weiqing Wang, Francisco J. Baldán, Wray Buntine, and Christoph Bergmeir. MoTime: A dataset suite for multimodal time series forecasting. *arXiv preprint arXiv:2505.15072*, 2025. URL https://arxiv.org/abs/2505.15072.

RELATED MANUSCRIPTS AND REUSED MATERIAL

An earlier unpublished manuscript studies a cardinality-aware forecasting head (Anonymous authors, 2026). Its experiments and headline results are not evidence in this paper, and implementation reuse from it is disclosed and is not a contribution. The static-interface and identity studies reported here have their own evidence and assumptions. A separate concurrent submission reuses this paper's frozen representation cache and daily count score as fixed infrastructure in one appendix on readout geometry, and claims no contribution to the laws, heads or results reported here. No result from those studies is counted as an independent validation of the earlier matched-width public test in Appendix I.

A STATIC IDENTITY, NORMALIZATION, AND INFORMATIVE TEMPORAL CODES

This appendix separates three questions: whether an input interface preserves a label, whether a frozen encoder preserves the resulting distinction, and whether a fitted predictor benefits from it. The first does not imply the second or third. The arguments below concern the input representation, not a new probabilistic model or a guarantee of improved forecasting.

An exact condition for erasure. Let h denote the fixed target history and let r be a static label. For covariate channel j, the observed entries are $c_j(r)\mathbf{1}$ on a fixed nonempty mask. Suppose the complete predictor factors as

$$F(h, r) = D\big(h, N_1(c_1(r)\mathbf{1}), \ldots, N_d(c_d(r)\mathbf{1}), M\big), \qquad N_j(x) = \frac{x - \bar{x}_j\mathbf{1}}{s_j(x)}, \tag{3}$$

where $\bar{x}_j$ is a location statistic of the observed entries that returns c whenever all of them equal c, such as their mean or median, $s_j(x) > 0$, and M contains the fixed masks, channel slots, time coordinates, and output coordinates. Unobserved entries are replaced using the same fixed rule after normalization. No other input to D contains r, $c_j(r)$, or the covariate normalization statistics. Then $F(h, r)$ is independent of r. Indeed, every observed numerator is zero, so all inputs to D are identical. This proves Proposition 1. A fixed deterministic map such as $\operatorname{arcsinh}$ after normalization preserves the result. For stochastic inference, the statement is equality of conditional output laws, or equality under a common random draw.

The assumptions concern all observed parts of a channel. A channel that is constant in the past but changes to a label-dependent future value need not be erased. Changing channel count, channel-specific missingness, or output selection also changes the input outside the displayed constant values. The mathematical conclusion uses exact arithmetic; finite-precision mean computation need not produce a bit-identical zero for every real constant. The numerical measurements report exact equality for their tested inputs.

Categorical encoding requires a separate premise. Suppose a per-task encoder $E(c; h)$ depends on the categorical sequence c only through its equality partition and the target values h within the parts. For every bijection π of category names, $E(\pi \circ c; h) = E(c; h)$. A constant categorical sequence has one part, so its encoded sequence cannot reveal the category name. This is the relevant property of per-task target encoding; it is an invariance statement, not a consequence of equivariance alone.

For example, the one-hot encoder $E(r) = e_r$ is equivariant to category permutation. A readout $w^\top E(r)$ with distinct entries of w still distinguishes every category. Likewise, sorting category names before ordinal encoding need not preserve a time-varying numerical sequence under arbitrary renaming: the sequence (a, b, c) encoded as $(0, 1, 2)$ can become $(0, 2, 1)$. Centering and scaling do not identify these two sequences. A single-category sequence still becomes a constant under either encoding.

The implemented paths. In `chronos` 2.2.2, the Chronos-2 task formatter applies a new `TargetEncoder` to each categorical covariate when the target is univariate, using only that task's history. With a multivariate target it uses a new `OrdinalEncoder` instead. Thus the stronger arbitrary-name invariance above applies to the target-encoding branch, not universally to both branches. Both map a static category to a constant numeric channel. The model's context

preparation applies instance normalization independently to each variate, retains a fixed observation mask, and normalizes known future covariates using the same channel statistics. Covariate means are not concatenated to the target encoder tokens. Reversing normalization of the target output uses the target's statistics, which are fixed in the intervention. These statements follow the released task formatter, model, and normalizer; the inspected version, not an unversioned future release, defines the scope.

The native TimesFM-3 path in `timesfm` 3.0.1 uses channelwise running means and scales before its Transformer. Its preprocessing concatenates normalized values and masks, not the original covariate means. Target outputs are restored with the target statistics. The optional outer `use_znorm` argument defaults to false; the internal running normalization is a separate operation. Therefore the measured constant erasure must not be attributed to enabling that optional argument. See the versioned model and normalization utilities. These links pin the release's relevant pathway, not a proof of whole-wheel byte equality with the runtime installation. This statement does not cover the older TimesFM external-regression path, which explicitly accepts static numerical and categorical covariates in the documented external-regression wrapper.

The inspected `MoiraiForecast` interface in `uni2ts` 2.0.0 exposes dynamic real covariates, including past-only covariates, but no separate static-feature argument. This signature is not a proof of identity erasure: static information can be encoded into a dynamic channel, and a claim about its survival requires analysis or measurement of that path. See the released forecast interface.

The TabPFN-TS versioned documentation states that its static covariates are dropped. Its tabular base model's ability to consume categorical columns does not establish that the time-series wrapper forwards static metadata. The census below measures all three.

A measured census of five interfaces. A census fixed before any of its outputs existed measures these three cases together with Chronos-2 and TimesFM-3 on the same 70 series (Table 5). Covariates are identical for every model: constants 0 and 3, the department code held constant, two independent standard normal paths and, for categorical or static slots, a single placeholder label, the true departments, a reassignment of the true departments across series and a relabelling by a fixed seven-label bijection. Each entry divides the largest absolute change by the model's mean absolute forecast without covariates. The bands were fixed in advance. A path is invisible when every static change is exactly zero or at most 10^{-6} of that mean, used when one reaches 10^{-3} and weak in between, provided the time-varying control reaches 10^{-3}.

Four departures from the source inspection were fixed before the census. The `uni2ts` 2.0.0 predictor wrapper fails with the installed `gluonts` 0.16.2, so Moirai-2's own forward receives the tensors the wrapper would build. Moirai-2's report states that covariate support was dropped (Liu et al., 2025); the released code still accepts dynamic covariates, and the census measures that path as released. TabPFN-TS first ran with a local TabPFN v2.5 regressor after we wrongly assumed that its time-series checkpoint required a licence request; that checkpoint is in the model's public repository, and a recorded amendment reran the same contrasts with it, with the same result (Table 5). TimesFM-2.5's residual mode fits only on history after the first 32-step patch and has no training rows at thirty days, so only its default mode is measured: configuration A, the documented covariate settings, with and without per-series target normalization, and configuration B, which removes the nonnegativity guard. Moirai-1.1 returns samples, which would need a different equality test, and is excluded. The head-side experiments use Moirai-1.1-R-large only as a frozen feature extractor, and in them the label reaches the head's embedding directly, never the trunk's interface, so this exclusion does not bear on them.

TimesFM-2.5 falls in the weak band on all nine paths. Its normalized categorical path changes by 1.3×10^{-5} of the mean, below its relabelling control of 1.9×10^{-5}. The normalized numeric path changes by 2.5×10^{-5} and has no corresponding control. Raw categorical paths A and B change by 2.5×10^{-5} and 2.4×10^{-5}, compared with controls of 2.4×10^{-6} and 2.3×10^{-6}, respectively. Their cause remains unresolved; these measurements do not identify a rounding mechanism or establish useful label dependence. Its presence entries under configuration A, up to 7.6 times the mean, come from the nonnegativity guard: configuration B reduces them to 1.5×10^{-5}. Moirai-2 is in the used band. A diagnostic written after the census and run once suggested a padding route. With the thirty-day context, the scaled values at padded positions are -948.7 for the constant 3 and -1897.4 for department code 6, each minus the constant divided by $\sqrt{10^{-5}}$. With a sixteen-day

context, which needs no context padding, both static changes are exactly zero while the time-varying control still moves the forecast by 10.4 times the mean. This comparison also changes the history length, so it does not isolate padding as the cause. A second diagnostic, counting single-precision spacings, proved uninformative because forecasts near zero dominate it. Chronos-2 differs from the September 7 replication by at most 6.7×10^{-4} on the same node, and TimesFM-3 reproduces it exactly; within each run every static change is exactly zero.

A fixed-history intervention on unobserved covariate positions. We therefore tested the mechanism on the same 70 thirty-day histories with `Moirai-2.0-R-small` in `uni2ts` 2.0.0, retaining the seven-day horizon and patch size sixteen. For a constant covariate c, the scaler uses its 37 supplied values, including known future covariates but no future targets. Its location is c and its scale is $\sqrt{10^{-5}}$ in exact arithmetic. Centering thus makes valid values zero but transforms the two left-context fills and nine right-future fills into $-c/\sqrt{10^{-5}}$. The input projection receives these values alongside the unchanged observation mask; it does not first zero the unobserved values.

The intervention zeros only those eleven scaled covariate fills immediately before input projection. Target values, valid covariate values, masks, channel slots, time coordinates and model weights remain fixed. We compare the original call, a sham copy and this intervention for eight inputs: no channel, constants zero and three, true, reassigned and recoded departments, and two time-varying covariates. This gives 24 batched forward calls, without fitting. The sham is bit-identical to the original at every captured tensor. After neutralization, all four static contrasts, true versus constant, true versus reassigned, true versus recoded, and constant three versus zero, have bit-identical projection inputs, encoder representations and retained target forecasts. The dynamic contrast remains nonzero: its largest forecast change is 3.5438404, or 10.6729 times the average of the two runs' mean absolute forecasts. This symmetric denominator differs from the no-covariate denominator in Table 5.

An independent reconstruction from the fixed histories and metadata verifies the packed inputs, masks, scaling and reported comparisons; it is not a second model inference.

In exact arithmetic, zeroing the scaled fills makes every covariate input coordinate independent of c. With the target history and all other inputs fixed, the encoder and retained target forecast must therefore be unchanged. This does not apply to denormalized covariate outputs, whose own location can still depend on c, or guarantee bitwise equality for untested floating-point codes. The experiment establishes the mechanism for the tested release and inputs, not improved forecasting. It changes no future target, compares no forecast with a realized outcome, and does not show that each of the eleven fills separately affects the target forecast.

Passing a covariate channel at all changes the forecast even when its value is then erased: the Presence column reaches 420 times the mean absolute forecast on Chronos-2's numeric path. This effect does not depend on the label's value, so a relabelling test cannot detect it, and it is not label use. A practitioner comparing runs with and without a covariate argument should expect it and compare labellings within the same interface call instead.

Table 5: Complete interface census. Entries are the largest absolute change divided by the mean absolute forecast without covariates. Static is the largest static change on the path; Symmetry is the relabelling control of a categorical path; Dynamic is the change between two time-varying realizations of the covariate, a liveness control; Presence is the change from passing the covariate channel with a fixed placeholder value instead of no covariate. Band is invisible when every static change is exactly zero or at most 10^{-6}, used at 10^{-3} or more and weak between, with thresholds fixed before the census. A and B are TimesFM-2.5 configurations; normalized and raw refer to per-series target normalization in its regression. The last row is the historical sixteen-day Moirai-2 diagnostic, not the fixed-history intervention described above.

Interface	Setting	Static	Symmetry	Dynamic	Presence	Band
Chronos-2, numeric	separate series	0	–	76.3	420	invisible
Chronos-2, numeric	cross-learning	0	–	3.51	111	invisible
Chronos-2, categorical	separate series	0	0	90	97.5	invisible
Chronos-2, categorical	cross-learning	0	0	14.3	64.2	invisible
TimesFM-3, numeric	native	0	–	19	27.9	invisible
TimesFM-2.5, dynamic numeric	A, normalized	1.6×10^{-6}	–	0.354	7.58	weak
TimesFM-2.5, dynamic numeric	A, raw	1.4×10^{-5}	–	1.54	7.58	weak
TimesFM-2.5, dynamic numeric	B, raw	1.4×10^{-5}	–	1.46	1.5×10^{-5}	weak
TimesFM-2.5, static categorical	A, normalized	1.3×10^{-5}	1.9×10^{-5}	0.354	7.58	weak
TimesFM-2.5, static categorical	A, raw	2.5×10^{-5}	2.4×10^{-6}	1.54	7.58	weak
TimesFM-2.5, static categorical	B, raw	2.4×10^{-5}	2.3×10^{-6}	1.46	1.2×10^{-5}	weak
TimesFM-2.5, static numeric	A, normalized	2.5×10^{-5}	–	0.354	7.58	weak
TimesFM-2.5, static numeric	A, raw	2.2×10^{-5}	–	1.54	7.58	weak
TimesFM-2.5, static numeric	B, raw	2.1×10^{-5}	–	1.46	1.5×10^{-5}	weak
Moirai-2, numeric	30-step context	30.9	–	13.3	6.45	used
TabPFN-TS, numeric column	official TS-3	0	–	25.1	0	invisible
TabPFN-TS, static features	official TS-3	0	–	25.1	0	invisible
TabPFN-TS, numeric column	local v2.5	0	–	10.4	0	invisible
TabPFN-TS, static features	local v2.5	0	–	10.4	0	invisible
Moirai-2, numeric	16-step context	0	–	10.4	9.29	post hoc

A trained adapter on the same public panel. Native interfaces are not the only route: a trained adapter can learn to use a static label. Tiny Time Mixers belong here rather than in the census, because their static categorical path is a newly trained decoder input, switched off in the released checkpoints (Ekambaram et al., 2024; TSFM project contributors, 2026). We therefore ran the official UniCA code (Han et al., 2026), whose adapter embeds static categorical metadata, on a frozen Chronos-Bolt-base backbone (Ansari et al., 2024) and raw entries from the public panel: each concatenates thirty observed days and a seven-day target. Training uses 740,928 of the panel's 778,400 training windows, which is 23,154 of its 24,325 training series, and all 197,280 evaluation windows, with a further 1,000 windows reserved for early stopping. The training side is therefore a subset of the panel's training windows and the evaluation side is complete; the run record fixes all three counts. The static categories were store and department, with the department recorded, set to one constant category, or permuted among series within each store. Each arm used seeds 1729, 3253 and 5003 and UniCA's own unimodal settings: learning rate 5×10^{-5}, batch 128, 50 batches per epoch, at most 100 epochs and early stopping on 1,000 windows from held-out training series. The design was fixed before any evaluation output. Compatibility changes use a thirty-day context, UniCA's non-indexed training loader and a validation loader bounded to the batches covering its 1,000 windows. The installed random sampler generates all 31 admissible cutpoints of each 37-day entry in random order, with a seven-day future and no minimum observed past. Earlier cutpoints are left-padded to thirty days; only the final cutpoint gives the original thirty-day history and seven-day target. Early stopping ended every fit after 36 to 61 epochs of 50 batches of 128 sampled instances, or 230,400 to 390,400 optimizer presentations, with its lowest validation loss after 16 to 41 epochs. These counts are neither distinct raw-entry coverage nor passes through the 740,928-entry collection. Historical distinct-entry coverage was not logged. Evaluation uses the original thirty-day histories and seven-day targets.

Weighted quantile loss over the nine levels 0.1 to 0.9 is 0.558465 with true departments, 0.558534 with the constant category and 0.558458 with reassigned departments. True minus constant is -6.8×10^{-5} (97.5 percent series-cluster interval $[-1.39 \times 10^{-4}, -1.03 \times 10^{-5}]$), but the three seed differences are -4.9×10^{-4}, $+3.7 \times 10^{-5}$ and $+2.5 \times 10^{-4}$. True minus reassigned is $+7.4 \times 10^{-6}$, interval $[-3.4 \times 10^{-5}, 5.1 \times 10^{-5}]$. A diagnostic added after these results refitted the first true-label model, which reproduced its forecasts exactly, and changed only the department column at inference:

forecasts moved by at most 0.0035, 0.38 percent of the mean forecast, and by 1×10^{-6} on average. The static path is therefore live, but this bounded implementation does not resolve a department-specific gain against reassignment. Different training-instance construction, backbones, losses and capacities (UniCA trains about 20.3 million parameters) preclude an accuracy-matched comparison of input-side and head-side placement. This result does not establish that input-side adapters cannot exploit identity. Appendix I reports the earlier matched-width public head comparison. Its primary store-only and secondary department-only questions differ from this adapter's department comparison conditional on store; their effect sizes are not interchangeable.

Information loss is relative to the retained history. Let $A = E(H, R)$ be the complete available representation of the label after preprocessing. For a discrete target with finite conditional entropy, an unrestricted optimal log-score predictor has the exact excess risk

$$H(Y \mid H, A) - H(Y \mid H, R) = I(Y; R \mid H, A). \tag{4}$$

This is the conditional-entropy identity, since A is a deterministic function of (H, R). If A is constant in R at each history, the loss is $I(Y; R \mid H)$. It can be zero when history already determines the useful label information. Neither a finite fitted head gain nor a comparison with a different head class estimates this quantity without further assumptions. For the implemented readout, let B contain every nonlabel input, including the frozen representation, retained location and scale, calendar indicators and any auxiliary history features. Under finite conditional entropy, the unrestricted log-score improvement from adding R is exactly

$$H(Y \mid B) - H(Y \mid B, R) = I(Y; R \mid B).$$

Only in the special case $B = \phi(H)$ is this $I(Y; R \mid \phi(H))$. If $R = r(H)$ but B discards useful label information, $I(Y; R \mid H)$ can be zero while $I(Y; R \mid B)$ is positive. Conversely, if B reveals R, the latter is zero, yet a restricted fitted class can still benefit from an explicit label feature. Neither the approximation class nor finite-sample estimation error is controlled by this classical identity.

Codes must differ after normalization, not only before it. For a fully observed length-T numeric channel, write $P = I - \mathbf{1}\mathbf{1}^\top/T$ and, for nonconstant inputs, let $N(x) = \sqrt{T}Px/\|Px\|_2$. Then

$$N(x) = N(y) \quad \Longleftrightarrow \quad y = ax + b\mathbf{1} \quad \text{for some } a > 0, \tag{5}$$

because equality of normalized centered vectors is precisely positive collinearity of the centered vectors. Thus positive amplitude changes or offsets of one waveform cannot provide distinct codes through this normalizer. Negative scaling is not included in this equivalence. A scale floor or nonuniform mask can change the equivalence and must be analyzed for the actual implementation.

Choose distinct centered unit vectors $c_1, \ldots, c_G$ and define $\delta = \min_{r \neq s} \|\sqrt{T}c_r - \sqrt{T}c_s\|_2 > 0$. The identity is recoverable from the normalized code by nearest-neighbor decoding. More quantitatively, if the received channel is $x = ac_r + b\mathbf{1} + e$, with $a > 0$ and $\|Pe\|_2 \leq \epsilon < a$, then

$$\|N(x) - \sqrt{T}c_r\|_2 \leq \frac{2\sqrt{T}\epsilon}{a}. \tag{6}$$

To prove this, set $u = ac_r$ and $v = u + Pe$ and use $\|v/\|v\| - u/a\| \leq |1 - \|v\|/a| + \|v - u\|/a \leq 2\epsilon/a$. Consequently $\epsilon < a\delta/(4\sqrt{T})$ suffices for correct decoding. This is an elementary finite-code stability bound, not a forecasting-risk bound or a claim about a learned codebook.

A frozen encoder can erase an otherwise valid code. Let Φ_h be a frozen encoder and let $\Delta_h = \min_{r \neq s} \|\Phi_h(\sqrt{T}c_r) - \Phi_h(\sqrt{T}c_s)\|_2$. Distinct normalized codes imply no positive lower bound on Δ_h: the constant encoder is a counterexample. If $\Delta_h > 0$ and Φ_h is L_h-Lipschitz on the relevant perturbations, then $2L_h\sqrt{T}\epsilon/a < \Delta_h/2$ suffices for nearest-prototype recovery after encoding. Both the separation and the perturbation condition are additional assumptions, not properties inferred from one active covariate example. They suggest measuring minimum representation separation before fitting a code-conditioned forecasting head.

A global permutation of a one-to-one codebook preserves all label information. It tests code geometry for a frozen predictor, or optimization under retraining, not removal of identity. An information-removal control instead assigns codes independently of the labels at the entity level, with the marginal

code distribution, number of channels, masks and training budget held fixed. Finite samples can retain chance association, which requires repeated independent assignments and entity-level inference. Neither success nor failure of one codebook proves that head conditioning is the unique repair. Retaining normalization statistics, a static-feature bypass, and covariate adapters such as ChronosX are explicit alternatives (Arango et al., 2025). The empirical question is which repair yields reliable held-out gains under a stated information and compute budget.

B THE VALUE OF A CARRIED IDENTITY

Section 3 reports three contrasts between labellings of one count head. This appendix gives the protocol, the thirty cells and the controls. The protocol was fixed before the first fit was submitted.

Design. One negative-binomial count head on frozen Chronos-2 features, the development training and validation splits, and the same optimizer, batch size, epoch count, gradient clipping and scored checkpoint as the protocol of Section 3. Three arms differ only in the integer platform code handed to the head's existing 37-category embedding: the true code; one fixed permutation of the codes across campaigns, drawn once and applied to both splits, so the label multiset on their union is unchanged; and a single constant code. Every parameter, including the embedding table and its width, is identical across arms, so the arms are width-matched by construction. Ten paired seeds give thirty fits. The seed is the inferential unit and the metric is the equal-campaign mean negative log mass of the seven-day count path.

Faithfulness. At the five seeds of an earlier count-law comparison on the same data, the true-label arm returns 0.247147, 0.245612, 0.245035, 0.245404 and 0.262917 nats, each equal to the corresponding cell of that comparison, so an independently written trainer reproduces it.

Arm means over ten seeds. No code 0.242887; true code 0.249130; permuted code 0.255573.

Contrasts. True minus permuted, one permutation and ten seeds: mean -0.006444, standard deviation 0.006430, ten-seed interval $[-0.011043, -0.001844]$ on nine degrees of freedom, nine of ten seeds favourable, campaign median -0.001518 with interval $[-0.001594, -0.001446]$. Because one permutation leaves the draw itself unmeasured, nine further independent permutations were crossed with the same ten seeds, giving 100 fits: mean -0.006215 with a bootstrap interval over seeds and permutations $[-0.010217, -0.001928]$, 77 of 100 fits favourable, per-permutation means from -0.012977 to -0.002196 so that all ten draws agree in sign, and variance components 1.099×10^{-5} between permutations against 3.288×10^{-5} between seeds. True minus no code: mean $+0.006243$, interval $[+0.002551, +0.009935]$, campaign median -0.002146 with interval $[-0.002169, -0.002094]$. Permuted minus no code: mean $+0.012687$, interval $[+0.009181, +0.016192]$.

Controls. The largest difference among arms within a seed ranges from 0.004596 to 0.020416 nats, so the label pathway is active. The sorted label histograms of the true and permuted arms are identical on the development union for every seed, the constant arm uses one label, and the permutation leaves 386 of the $6{,}338$ distinct campaigns on their own label; those campaigns are the union of the two development splits, $5{,}985$ in training and $1{,}917$ in validation, $1{,}564$ of them in both. The contrasts above score the $1{,}917$ campaigns of the validation split. On the equal-campaign mean, where a lower score is better, the permuted arm scores highest and the no-code arm lowest, with the true arm between them, so a wrong label costs more than either the right label or none. That is the ordering a wrong label predicts.

What the permutation contrast does and does not show. A permuted code has the same multiset of labels over the development campaigns and the same embedding width as the true one, so their raw tensor dimensions and label multiplicities on that union are matched. This does not guarantee matching separately within training or validation, or after weighting by training origins. It does not match optimization paths or establish equal effective function classes on a scored subset. It can come from two sources we do not separate here. The true code may carry information the head uses, or the permuted code may act as a misleading feature that the head fits and is then penalised for. The

second reading is consistent with the arm means, since the no-code arm scores best of the three on the equal-campaign mean: reading the right label does not beat reading none, and what the contrast establishes is that reading a wrong one is worse than either. The later coarsening controls below provide additional evidence about these readings, but these development fits alone do not distinguish them.

The three arms on the frozen later window and on a second trunk. A second preregistration, fixed before any fit and amended once after a three-cell pilot, refits every arm with the protocol above and scores the fitted head, without further update, on the later window materialized on 2026-09-19: 105,995 origins from 3,379 of the window's 9,150 campaigns, with origin days from 2025-10-22 to 2025-12-01. The window was opened once before, for an earlier count-law comparison; this protocol is its second opening, and two later preregistered protocols, the information ladder and the multiset-matched control below, score it a third and a fourth time. Every result stands as written. The primary population is the 3,087 window campaigns that appear in the development rosters, so that every arm's label is defined identically. A metadata-only crosswalk separates 2,772 training-present and 315 validation-only campaigns within this population. Including the 292 absent from both rosters gives 607 training-unseen campaigns. The crosswalk alone adds no scores and changes neither the original population nor its scores; Appendix D reports the separate saved-distribution analysis of these groups. The 292 remaining campaigns receive, in the permuted arm, a label drawn from the window's own code histogram, and the population of all 3,379 is reported as secondary. Each cell is fitted on Chronos-2 features and again on Moirai-1.1-R-large features from the development caches the frozen configuration names; the head, its width and its optimizer are identical. Before any window score is read, the ten-seed development arm means must reproduce the values above: they do, to 4.7×10^{-8}, 3.5×10^{-7} and 3.7×10^{-7} nats.

On the window the ordering reverses. With the equal-campaign mean, ten seeds, and intervals adjusted for the six primary contrasts, on Chronos-2 features the true code scores 0.0147 nats below the permuted code, $[-0.0283, -0.0012]$, and 0.0149 below no code, $[-0.0213, -0.0084]$, while the permuted code and no code have no resolved mean difference, -0.0001 $[-0.0149, +0.0146]$; the arm means are 0.1088, 0.1236 and 0.1237. On Moirai-1.1-R-large features the true code scores 0.0060 below the permuted code, $[-0.0086, -0.0034]$, and 0.0060 below no code, $[-0.0090, -0.0030]$, with the permuted-code versus no-code interval again including zero, $[-0.0025, +0.0026]$; the arm means are 0.0948, 0.1008 and 0.1008. Across the nine further permutations, 95 of 100 Chronos-2 fits and 100 of 100 Moirai fits favour the true code, every permutation favours it on average, and the two-way seed-by-permutation intervals are $[-0.0196, -0.0097]$ and $[-0.0077, -0.0049]$. The campaign median favours the true code in every contrast on both trunks, by 0.0013 to 0.0019 nats, and the true arm wins in ten of ten seeds against no code on both trunks. In the secondary population of all 3,379 campaigns the true code again scores best, with means 0.3487, 0.3986 and 0.3663 on Chronos-2, but there the permuted code scores worse than no code; the difference comes from the 292 campaigns the head never saw, which the paragraph on unseen campaigns below scores separately. Controls: the largest within-seed difference across arms is at least 0.0103 nats on Chronos-2 and 0.0015 on Moirai, so the label pathway is live; the constant arm uses one label; and the permuted labels on the primary population are the development permutation restricted to it, which by construction has a different multiset from the subset's own true labels, 37 distinct against 27, a fact the amendment records.

What changes between the two surfaces is the value of the right code against none. On the development validation split, which follows the training period after a seven-day embargo, reading no code is best on the mean; on the later window the right code is best on the mean and at the median, on both trunks, and the wrong-code versus no-code mean interval includes zero. This is not an equivalence test. The development-split advantage of the true code over a permuted one holds on Chronos-2 features, where it scores 0.0062 below the permuted code over 100 fits, $[-0.0102, -0.0019]$, with 77 favourable, and does not appear on Moirai-1.1-R-large features, where it scores 0.0011 above it, $[-0.0007, +0.0027]$, with 35 favourable and 1 of the ten permutations favourable on average. Both surfaces lie after the training period, and we do not identify why the ordering changes between them.

The same contrasts with campaigns as the resampling unit. The single-permutation intervals above are Student intervals over ten paired seed differences, with nine degrees of freedom and the declared family-wise adjustment. The separate seed-by-permutation intervals use a two-way bootstrap. Campaigns are the population a reader cares about, so we report the complementary unit as

well, in an analysis added after the preregistered one: the ten seeds are averaged within each campaign first, and the 3,087 campaigns are then resampled 20,000 times under the same adjustment for six contrasts. No fit is repeated; these are the per-campaign scores each cell already stored. At that unit the campaign median favours the true code in every contrast on both trunks, by 0.00126 and 0.00166 nats on Chronos-2 features, $[0.00120, 0.00131]$ and $[0.00164, 0.00167]$, and by 0.00171 and 0.00190 on Moirai-1.1-R-large features, $[0.00168, 0.00178]$ and $[0.00188, 0.00191]$. The equal-campaign mean is less stable at this unit, as the heavy right tail of per-campaign scores would predict: of the four true-code mean contrasts only the Moirai true-against-permuted one excludes zero, at -0.0060 $[-0.0141, -0.0001]$, and the others straddle it. Two readings follow, and we give both. The gain the true code carries is robust at the campaign median on both trunks and in both contrasts; its size on the campaign mean is set by a small number of heavy campaigns, so the seed-unit Student mean interval conditions on this fixed campaign population. At the campaign median a permuted code is also slightly better than none, by 0.00023 and 0.00041 nats, which the seed-unit mean comparison does not resolve; the true code's gain over none is 7.2 and 4.6 times larger.

Encoded platform groups as the resampling unit. The 3,087 primary campaigns activate 27 codes in the 37-slot training vocabulary. The historical bootstrap groups campaigns by these true-arm codes, not by the 41 distinct source-name keys on this population (Table 10). We resample encoded groups, each with all its campaigns, and drop one group at a time, in an analysis added after the preregistered one that reuses the stored per-campaign scores and reports unadjusted 95 percent intervals. With 27 unequal encoded clusters this bootstrap is descriptive. On Chronos-2 features the equal-campaign mean gain excludes zero against the matched permuted code, -0.0133 $[-0.0342, -0.0016]$, and against no code, -0.0149 $[-0.0390, -0.0005]$; on Moirai-1.1-R-large features, where one encoded group carries 36 to 41 percent of the absolute differences, both intervals include zero, $[-0.0224, +0.0008]$ and $[-0.0229, +0.0028]$. The campaign median favours the true code with its encoded-group interval excluding zero in all four contrasts. Removing any one of the 27 encoded groups leaves all four mean gains negative: between -0.0144 and -0.0082 against the matched code and between -0.0162 and -0.0085 against no code on Chronos-2 features, and between -0.0077 and -0.0011 and between -0.0077 and -0.0001 on Moirai-1.1-R-large features. No single encoded group carries the sign of the gain, but its size is concentrated. On Moirai-1.1-R-large, removing the largest contributor, one encoded group with 72 campaigns, leaves 16 percent of the mean gain against the matched code and 2 percent against no code. On Chronos-2, removing each contrast's largest contributor leaves 62 and 57 percent, respectively; those encoded groups contain 72 and 5 campaigns and are not the same group. These are ratios of the remaining equal-campaign mean to the original mean, not net contribution shares, because removing a group changes the denominator.

Campaigns the head never saw. The 292 window campaigns absent from the development rosters are scored in every cell but enter only the secondary population, and in the permuted arm their labels are drawn afresh from the window's code histogram. We decompose that population after the preregistered analysis, with no new fit, and report unadjusted 95 percent intervals. These campaigns score an order of magnitude higher than the primary population, 2.885, 3.306 and 2.931 nats for the true, permuted and constant arms on Chronos-2 features, and a few dominate: the ten campaigns with the largest differences carry 66 to 78 percent of the absolute differences on Chronos-2 features and 41 to 61 percent on Moirai-1.1-R-large features. On Chronos-2 features a freshly drawn wrong label costs 0.375 nats against no code, $[0.151, 0.599]$ over seeds, but $[-0.033, 1.077]$ when campaigns are resampled, and the campaign median is essentially zero; on Moirai-1.1-R-large features it costs 0.030, $[-0.006, 0.067]$. The true code scores 0.421 below the permuted code on Chronos-2 features, $[-0.642, -0.200]$ over seeds and $[-1.109, -0.011]$ over campaigns, and 0.045 below it on Moirai-1.1-R-large features, $[-0.078, -0.012]$ over seeds, with a campaign interval that includes zero; it beats each of the ten permutations on average on both trunks. Against no code the true code is ahead by 0.046 and 0.014 nats, in six of ten seeds on each trunk and at the campaign median, but both intervals include zero. At these point estimates, the wrong-versus-constant contrast accounts for most of the true-versus-wrong difference. The true-versus-constant gain is unresolved, not zero; these subgroup comparisons do not identify a causal mechanism. The resolved seed-unit gains reported above concern the 3,087 campaigns seen in development.

Is the gap information, or the cost of a wrong label? A permuted code has the same development label multiset and embedding width as the true one. This matches raw tensor dimensions and union-

level counts, not every training exposure or effective class, and does not identify the source of the gap. It could be that the true code carries information the head uses, or that a permuted code acts as a misleading feature the head fits and is then charged for. A label whose information content is varied rather than permuted supplies an additional controlled comparison. We fix, in advance and without reference to any outcome, a coarsening that sends the 37 platform codes to g groups by ranking them on development campaign count and assigning rank i to group $i \bmod g$, and refit the whole protocol at $g = 2$, 4 and 8: sixty further cells, ten seeds and both trunks, with the constant arm and the true arm serving as the two ends of the ladder. Every coarse label is true, and none is wrong.

The exact label has the lowest score on both trunks, but the ordering is not monotonic: on Chronos-2 features the five rungs score 0.1237, 0.1206, 0.1124, 0.1169, 0.1088 nats; on Moirai-1.1-R-large features the five rungs score 0.1008, 0.1008, 0.0982, 0.0981, 0.0948 nats. The Spearman correlation between g and the score is -0.90 on both trunks; with five rungs the exact two-sided permutation p is $1/12$. The true code is the best rung on both, so no coarse label does better than the exact one. Individually, after the same adjustment, the four-group rung separates from the constant arm on Chronos-2 features, and the four-group and eight-group rungs do on Moirai-1.1-R-large features; the ordering between the four-group and eight-group rungs is not resolved. These outcomes support a predictive role for identity: some true coarse labels recover part of the gain, and the exact label recovers the most. They do not isolate that role from all optimization effects or establish that arbitrary wrong labels incur no cost.

A permuted control with the scored population's own labels. Restricted to the 3,087 scored campaigns, the permuted arm above carries 37 distinct codes while those campaigns' own true codes use 27, so it is not matched on the population that is scored. A further preregistered control removes the difference: the true codes of the scored campaigns are permuted among themselves and those of the remaining 3,251 development campaigns among themselves, so that on the scored population the arm's label histogram equals the true one code by code, and on the development set as a whole the multiset is unchanged. Reconstructing all ten assignments confirms these two equalities, but not equality of training-only, validation-only or training-origin-weighted label counts (Appendix D). This does not establish that the exposure difference causes the observed loss difference.

Everything else is as above, with ten seeds, both trunks and ten permutation indices, 200 fits in all; before the new arm is read, the same analysis reproduces the existing two-way interval of the earlier permuted arm exactly.

On Chronos-2 features the true code scores 0.0133 nats below the matched permuted code, $[-0.0252, -0.0014]$ after adjustment for the four new contrasts, with nine of ten seeds and 91 of 100 fits favouring it and a two-way seed-by-permutation interval of $[-0.0170, -0.0060]$; on Moirai-1.1-R-large features it scores 0.0067 below, $[-0.0093, -0.0042]$, with all ten seeds, 97 of 100 fits and $[-0.0069, -0.0039]$. The campaign median favours the true code against this arm on both trunks, by 0.00119 and 0.00146 nats, with intervals excluding zero; the campaign-level mean does so on Moirai-1.1-R-large features and not on Chronos-2 features, where the heavy tail described above widens it. The matched-permuted versus no-code seed-unit intervals include zero, -0.0016 $[-0.0125, +0.0094]$ and $+0.0007$ $[-0.0011, +0.0026]$; they do not establish equivalence. The true code's advantage therefore survives a wrong label that carries exactly the scored population's own multiset.

Table 6: Later-window scores along a ladder of true labels, in nats per seven-day count path, ten seeds per rung, primary population of 3,087 campaigns. A rung with g groups hands the head a true but coarser platform label; $g = 1$ is the constant code and $g = 37$ the exact code. The last two rows count the seeds, out of ten, in which a rung scores below the constant code.

Groups in the label, g	1	2	4	8	37
Chronos-2 features	0.1237	0.1206	0.1124	0.1169	0.1088
Moirai-1.1-R-large features	0.1008	0.1008	0.0982	0.0981	0.0948
Seeds beating $g = 1$, Chronos-2		6	8	5	10
Seeds beating $g = 1$, Moirai-1.1-R-large		5	10	9	10

Scope. One private source, one head family, an embedding of a static platform code, ten seeds, two trunks, a coarsening ladder fixed in advance, and a permuted control matched on the scored population's label multiset. The development origins had been seen by earlier work; the later window has been scored under four preregistered protocols, the first for a different comparison.

C IDENTITY READOUTS VERSUS GROUP-RATE CALIBRATION

This new development comparison asks whether ordinary group-rate calibration accounts for a fitted identity head's advantage. Its forty fresh fits reuse the previously opened training, validation and later-window caches; they are not a reopening of a sealed source or an untouched confirmation. The original four later-window protocols remain unchanged. This is an additional, prospectively specified comparison on that already used window. The complete family contains two representations, ten paired optimization seeds and constant and true-label heads. Every fit retains the twenty-epoch optimizer and complete nonlabel inputs described in Section 3. Actual weights and daily distribution parameters are saved; these are new fits, not reconstructed historical checkpoints.

Controls and their objective. Let η_{ot} and $k_{ot} > 0$ be the decoded log mean and dispersion of the fixed constant-label head for origin o and day t. Each calibrated law replaces η_{ot} by $\eta_{ot} + b_r$, leaving k_{ot} unchanged. One control shares a global b; two use the true platform code r, without and with unit quadratic shrinkage. The calibration objective is

$$\sum_c \frac{1}{|O_c|} \sum_{o \in O_c} \sum_{t=1}^{7} \ell_{\mathrm{NB}}(y_{ot}; \eta_{ot} + b_{r(c)}, k_{ot}) + \frac{\lambda}{2} \sum_r b_r^2, \qquad -5 \le b_r \le 5,$$

where λ is zero or one. The outer sum is not divided by the number of campaigns. All calibration uses training data only, with no validation-selected penalty or later-window optimization. The ordinary negative-binomial derivatives are

$$\ell'(b) = \frac{k(\mu - y)}{k + \mu}, \qquad \ell''(b) = \frac{k\mu(k + y)}{(k + \mu)^2} > 0, \qquad \mu = \exp(\eta + b).$$

These classical formulas permit bounded one-dimensional convex fitting. We retain boundary solutions for all-zero groups and require a projected gradient residual at most 10^{-8}. This floating-point diagnostic is not an interval-arithmetic certificate. At evaluation, a literal platform name absent from the source by the historical cutoff receives zero group offset, including when it shares the residual platform code. This source-name rule is not a claim that every supported name supplied an eligible training campaign. The global correction is unchanged by this fallback.

Complete results and uncertainty units. The primary population is the original 3,087 campaigns in the development rosters. Table 7 gives all four primary differences and their paired-seed Bonferroni intervals. Table 8 retains every arm and both secondary populations. There is no registered practical-equivalence margin; intervals containing zero remain unresolved. The campaign summaries average paired seeds first, then resample campaigns 20,000 times. They do not make platforms, repeated origins or shared calendar shocks independent.

Table 7: Full identity head minus group-mean calibration, in nats per seven-day path, on the original 3,087-campaign population. Dispersion is fixed in these controls. Negative favors the full head. Seed intervals use ten paired seeds and the original Bonferroni adjustment over four contrasts. Campaign intervals are marginal 95% descriptive intervals after averaging seeds, not correction for platform or calendar dependence.

Features	Calibration	Difference	Seed interval	Campaign interval
Chronos-2	Group	-0.0152	$[-0.0217, -0.0087]$	$[-0.0358, -0.0015]$
Chronos-2	Group + ridge	-0.0154	$[-0.0220, -0.0089]$	$[-0.0360, -0.0017]$
Moirai-large	Group	-0.0053	$[-0.0083, -0.0022]$	$[-0.0113, -0.0003]$
Moirai-large	Group + ridge	-0.0056	$[-0.0087, -0.0025]$	$[-0.0117, -0.0006]$

Table 8: Equal-campaign mean seven-day NLL, averaged over ten seeds. The full and development-unseen populations contain 3,379 and 292 campaigns. All arms are retained; lower is better.

Population	Features	Constant	Global	Group	Group + ridge	Identity
Development-roster	Chronos-2	0.12367	0.12382	0.12402	0.12423	0.10881
Development-roster	Moirai-large	0.10083	0.10074	0.10009	0.10040	0.09481
All later	Chronos-2	0.36627	0.36632	0.36547	0.36586	0.34871
All later	Moirai-large	0.30163	0.29990	0.29803	0.29840	0.29491
Development-unseen	Chronos-2	2.93098	2.93001	2.91800	2.92038	2.88493
Development-unseen	Moirai-large	2.42451	2.40543	2.39063	2.39166	2.41037

Table 9: Secondary full-head-minus-calibration differences and nominal 95% paired-seed intervals. None resolves the mean advantage on these populations. Favourable seeds have a negative difference, out of ten.

Population / features	Calibration	Difference	Seed interval	Seeds < 0
All later / Chronos-2	Group	-0.0168	$[-0.0370, +0.0035]$	7/10
All later / Chronos-2	Group + ridge	-0.0172	$[-0.0374, +0.0031]$	7/10
All later / Moirai-large	Group	-0.0031	$[-0.0086, +0.0023]$	7/10
All later / Moirai-large	Group + ridge	-0.0035	$[-0.0089, +0.0020]$	8/10
Development-unseen / Chronos-2	Group	-0.0331	$[-0.2313, +0.1652]$	6/10
Development-unseen / Chronos-2	Group + ridge	-0.0355	$[-0.2335, +0.1626]$	6/10
Development-unseen / Moirai-large	Group	$+0.0197$	$[-0.0246, +0.0641]$	4/10
Development-unseen / Moirai-large	Group + ridge	$+0.0187$	$[-0.0251, +0.0625]$	4/10

What this comparison establishes. The full head beats the two specified group-mean corrections on the fixed primary population, including under the separate descriptive campaign intervals. It need not do so on every campaign or population. The full head can alter dispersion and feature-dependent means, whereas the calibration keeps its learned base and dispersion fixed. Moreover, the full head bounds log means by 20 and post-decoding calibration permits values up to 25 in magnitude. Neither general class inclusion follows. These are fitted, nonnested controls, not an oracle decomposition or a theorem about the value of nonlinear interactions. Appendix E reports the additional sequential dispersion controls; they do not make these classes nested.

Independent reconstruction. A separately implemented check reads all forty saved daily-parameter surfaces and twenty calibration families. It reconstructs all 360 training, validation and later-window score arrays and their campaign, seed and bootstrap summaries. The largest path-score discrepancy is 1.94×10^{-11}; the largest independently recomputed projected gradient residual is below 10^{-8}. This checks probability arithmetic and inference from saved parameters, not neural checkpoint inference, fresh-environment training or independent-source replication.

D SAVED-DISTRIBUTION DECOMPOSITION AND TRAINING EXPOSURE

The identity head changes both the predicted mean and dispersion. We therefore separate their contributions to its loss difference from each saved group-mean control. This additional analysis was fixed after the complete calibration results were known. It uses the same forty neural fits, twenty calibration tasks, original targets, and ten optimization seeds; no model is refitted.

Table 10: Platform-key layers in the opened development data. The training vocabulary contains 36 named categories and one `other` slot. A source-name count is not a count of independent calibration parameters or encoded bootstrap clusters.

Layer	Key and analysis role	Count
Source names	Original source index; descriptive source omissions	41 original; 42 later
Training vocabulary	Model and group-calibration code	37 slots
Original-cohort groups	True-arm codes used by the historical bootstrap	27 active
Historical-name mask	Literal source name observed by the training cutoff	105,995 known later origins

Group mean and dispersion shifts are fitted by model code, not separately for the 41 or 42 source names. At later scoring, the historically known-name mask determines whether a group shift is applied. All 105,995 eligible later origins have a known name; that status does not imply an eligible training campaign, and this window cannot test an unseen literal source name. A target-free crosswalk from the pinned metadata establishes the counts without serializing source names or campaign identifiers.

Four legal distributions. Let $\ell(\mu, k)$ be the seven-day negative-binomial loss at one forecasting origin. Write (μ_0, k_0) for the calibrated constant head and (μ_1, k_1) for the full identity head. The four scores are $S_{ab} = \ell(\mu_a, k_b)$, for $a, b \in \{0, 1\}$. Their order-averaged contributions are

$$G_\mu = \tfrac{1}{2}\{(S_{00} - S_{10}) + (S_{01} - S_{11})\}, \tag{7}$$

$$G_k = \tfrac{1}{2}\{(S_{00} - S_{01}) + (S_{10} - S_{11})\}. \tag{8}$$

Thus $G_\mu + G_k = S_{00} - S_{11}$ at every origin and after averaging within campaigns and across campaigns or seeds. Each swapped parameter pair defines a normalized distribution, but is not a separately fitted model. This is the ordinary two-component Shapley identity, not causal attribution, model-class optimality, or evidence of a new interaction mechanism. Component medians need not sum to the median total gain.

Population and uncertainty. The additional family contains 36 later full-minus-control contrasts, 48 component estimates, and 72 earlier-to-later changes. Simultaneous intervals use all ten paired seeds, nine degrees of freedom, and a Bonferroni correction over the fixed 156 entries, including structurally empty intersections. These intervals describe optimization-seed variation conditional on the evaluated entities. The original four-comparison intervals are retained in the preceding calibration analysis and are not replaced by this larger family. Entity medians and platform omissions are descriptive; neither removes shared platform or calendar dependence. All 156 entries, paired-seed estimates and every platform omission accompany the tables as machine-readable data.

Among the original 3,087 campaigns, the mean and dispersion components against group calibration are respectively 0.00632 and 0.00890 nats for Chronos-2, and 0.00176 and 0.00352 for Moirai-large. Under the full additional multiplicity correction, Moirai's dispersion component remains positive, whereas its mean component and both Chronos components are unresolved individually. These fixed-distribution calculations motivated the ordinary dispersion controls subsequently fitted in Appendix E. The completed comparison does not identify a unique interaction mechanism.

The 607 training-unseen campaigns include 315 already present in validation. For that population, full minus group mean is -0.0139 with interval $[-0.2458, 0.2180]$ for Chronos-2 and 0.0115 with interval $[-0.0480, 0.0710]$ for Moirai-large.

Neither result establishes unseen-entity benefit or equivalence. The later effective origins run from October 22 to December 1, 2025, not the entire raw source window.

Earlier and later scores. We join identical entities within each saved model and seed. There are 2,772 training–later pairs and 1,767 validation–later pairs. The latter comprise 1,452 training-seen and 315 validation-only entities. Earlier training scores are in-sample; validation has already participated in development.

The changes therefore describe temporal sensitivity, not an effect of policy or financing stage. On all 1,767 validation pairs, the change in full-minus-group mean is 0.0684 for Chronos-2 and 0.0418 for Moirai-large, with simultaneous intervals $[0.0355, 0.1014]$ and $[0.0168, 0.0667]$. The corresponding ridge-control changes are -0.0078 and -0.0012, with intervals $[-0.0140, -0.0016]$ and $[-0.0059, 0.0035]$. A fixed model comparison can therefore depend on both the evaluation window and the regularization of its ordinary control.

What the permuted control matches. We reconstruct all ten original assignments from their fixed generator and check the saved audits and primary-population labels for all 200 existing cells. The historical runner did not serialize its complete development assignment map; this is a deterministic reconstruction, not recovery of an unavailable checkpoint. Every assignment preserves the campaign-label histogram on the 6,338-campaign development union and the 3,087-campaign primary population. None preserves that histogram on all 5,985 training campaigns, on the 4,421 training-exclusive campaigns, on the 1,917 validation campaigns, or on the 353 validation-only campaigns. Each assignment also changes the label distribution weighted by actual training origins. For validation-only campaigns, both training-origin counts are zero, so their equality says nothing about training balance. The original permuted experiment is retained with this exact scope; no assignment is changed and no new true-head score is paired with an old permuted model. The count tables use the full development populations, whose validation-only denominator 353 differs from the 315 later-window subset.

D.1 Populations and complete fitted contrasts

Table 11: The complete later population and its prespecified subsets. Source-name keys count distinct original source indices, not trained model codes. Pair columns count entities present in both the named earlier split and the later window. All are opened development data.

Later population	Entities	Origins	Source-name keys	Training pairs	Validation pairs
Original cohort	3087	101892	41	2772	1767
All later	3379	105995	42	2772	1767
Training-seen	2772	95355	38	2772	1452
Validation-only	315	6537	21	0	315
Outside development	292	4103	30	0	0
Training-unseen	607	10640	33	0	315

Table 12: Chronos-2: full minus control. Negative favors the identity head. Intervals use all ten paired seeds and the fixed 156-comparison adjustment. Entity medians are descriptive.

Population	Control	Mean and simultaneous interval	Entity median
Original cohort	Constant	-0.0149 $[-0.0257, -0.0040]$	-0.0017
Original cohort	Group mean	-0.0152 $[-0.0270, -0.0034]$	-0.0012
Original cohort	Ridge mean	-0.0154 $[-0.0273, -0.0035]$	-0.0015
All later	Constant	-0.0176 $[-0.0682, +0.0331]$	-0.0017
All later	Group mean	-0.0168 $[-0.0672, +0.0337]$	-0.0012
All later	Ridge mean	-0.0172 $[-0.0676, +0.0333]$	-0.0015

Continued on next page

Population	Control	Mean and simultaneous interval	Entity median
Training-seen	Constant	-0.0171 $[-0.0318, -0.0023]$	-0.0016
Training-seen	Group mean	-0.0174 $[-0.0329, -0.0018]$	-0.0011
Training-seen	Ridge mean	-0.0176 $[-0.0332, -0.0019]$	-0.0014
Validation-only	Constant	$+0.0044$ $[-0.0279, +0.0366]$	-0.0076
Validation-only	Group mean	$+0.0038$ $[-0.0278, +0.0355]$	-0.0074
Validation-only	Ridge mean	$+0.0036$ $[-0.0281, +0.0352]$	-0.0075
Outside development	Constant	-0.0461 $[-0.5534, +0.4613]$	-0.0026
Outside development	Group mean	-0.0331 $[-0.5268, +0.4606]$	-0.0019
Outside development	Ridge mean	-0.0355 $[-0.5287, +0.4578]$	-0.0026
Training-unseen	Constant	-0.0199 $[-0.2592, +0.2194]$	-0.0073
Training-unseen	Group mean	-0.0139 $[-0.2458, +0.2180]$	-0.0071
Training-unseen	Ridge mean	-0.0152 $[-0.2469, +0.2165]$	-0.0071

Table 13: Moirai-large: full minus control. Negative favors the identity head. Intervals use all ten paired seeds and the fixed 156-comparison adjustment. Entity medians are descriptive.

Population	Control	Mean and simultaneous interval	Entity median
Original cohort	Constant	-0.0060 $[-0.0110, -0.0011]$	-0.0019
Original cohort	Group mean	-0.0053 $[-0.0108, +0.0002]$	-0.0016
Original cohort	Ridge mean	-0.0056 $[-0.0112, +0.0000]$	-0.0018
All later	Constant	-0.0067 $[-0.0161, +0.0026]$	-0.0019
All later	Group mean	-0.0031 $[-0.0167, +0.0105]$	-0.0016
All later	Ridge mean	-0.0035 $[-0.0170, +0.0101]$	-0.0018
Training-seen	Constant	-0.0070 $[-0.0120, -0.0021]$	-0.0019
Training-seen	Group mean	-0.0063 $[-0.0113, -0.0013]$	-0.0016
Training-seen	Ridge mean	-0.0066 $[-0.0117, -0.0015]$	-0.0017
Validation-only	Constant	$+0.0031$ $[-0.0086, +0.0147]$	-0.0056
Validation-only	Group mean	$+0.0039$ $[-0.0133, +0.0210]$	-0.0051
Validation-only	Ridge mean	$+0.0035$ $[-0.0136, +0.0206]$	-0.0052

Continued on next page

Population	Control	Mean and simultaneous interval	Entity median
Outside development	Constant	-0.0141 $[-0.0841, +0.0558]$	-0.0074
Outside development	Group mean	$+0.0197$ $[-0.0907, +0.1302]$	-0.0058
Outside development	Ridge mean	$+0.0187$ $[-0.0904, +0.1278]$	-0.0062
Training-unseen	Constant	-0.0052 $[-0.0407, +0.0303]$	-0.0056
Training-unseen	Group mean	$+0.0115$ $[-0.0480, +0.0710]$	-0.0051
Training-unseen	Ridge mean	$+0.0108$ $[-0.0480, +0.0696]$	-0.0052

D.2 Mean and dispersion contributions

Table 14: Chronos-2: order-averaged loss reductions with simultaneous 156-comparison seed intervals. Positive values favor the full-head parameter component; neither component is a causal effect.

Population	Control	Mean component	Dispersion component
Original cohort	Group mean	$+0.0063$ $[-0.0007, +0.0134]$	$+0.0089$ $[-0.0050, +0.0228]$
Original cohort	Ridge mean	$+0.0064$ $[-0.0006, +0.0135]$	$+0.0090$ $[-0.0050, +0.0229]$
All later	Group mean	$+0.0153$ $[-0.0114, +0.0420]$	$+0.0015$ $[-0.0299, +0.0328]$
All later	Ridge mean	$+0.0155$ $[-0.0112, +0.0421]$	$+0.0017$ $[-0.0297, +0.0330]$
Training-seen	Group mean	$+0.0074$ $[-0.0010, +0.0158]$	$+0.0100$ $[-0.0053, +0.0253]$
Training-seen	Ridge mean	$+0.0075$ $[-0.0009, +0.0159]$	$+0.0101$ $[-0.0053, +0.0254]$
Validation-only	Group mean	-0.0031 $[-0.0347, +0.0285]$	-0.0007 $[-0.0119, +0.0104]$
Validation-only	Ridge mean	-0.0030 $[-0.0346, +0.0286]$	-0.0006 $[-0.0117, +0.0106]$
Outside development	Group mean	$+0.1101$ $[-0.1702, +0.3903]$	-0.0770 $[-0.3428, +0.1889]$
Outside development	Ridge mean	$+0.1113$ $[-0.1680, +0.3906]$	-0.0758 $[-0.3421, +0.1905]$
Training-unseen	Group mean	$+0.0513$ $[-0.0786, +0.1813]$	-0.0374 $[-0.1636, +0.0888]$
Training-unseen	Ridge mean	$+0.0520$ $[-0.0775, +0.1815]$	-0.0368 $[-0.1632, +0.0896]$

Table 15: Moirai-large: order-averaged loss reductions with simultaneous 156-comparison seed intervals. Positive values favor the full-head parameter component; neither component is a causal effect.

Population	Control	Mean component	Dispersion component
Original cohort	Group mean	$+0.0018$ $[-0.0027, +0.0062]$	$+0.0035$ $[+0.0003, +0.0068]$

Continued on next page

Population	Control	Mean component	Dispersion component
		$+0.0020$	$+0.0036$
Original cohort	Ridge mean	$[-0.0025, +0.0065]$	$[+0.0003, +0.0069]$
		$+0.0033$	-0.0001
All later	Group mean	$[-0.0058, +0.0123]$	$[-0.0067, +0.0064]$
		$+0.0035$	-0.0000
All later	Ridge mean	$[-0.0055, +0.0125]$	$[-0.0066, +0.0066]$
		$+0.0016$	$+0.0047$
Training-seen	Group mean	$[-0.0028, +0.0060]$	$[+0.0017, +0.0078]$
		$+0.0018$	$+0.0048$
Training-seen	Ridge mean	$[-0.0027, +0.0063]$	$[+0.0017, +0.0079]$
		$+0.0032$	-0.0071
Validation-only	Group mean	$[-0.0108, +0.0172]$	$[-0.0158, +0.0017]$
		$+0.0034$	-0.0069
Validation-only	Ridge mean	$[-0.0106, +0.0173]$	$[-0.0156, +0.0019]$
		$+0.0192$	-0.0390
Outside development	Group mean	$[-0.0588, +0.0973]$	$[-0.0879, +0.0100]$
		$+0.0196$	-0.0383
Outside development	Ridge mean	$[-0.0570, +0.0962]$	$[-0.0877, +0.0110]$
		$+0.0109$	-0.0224
Training-unseen	Group mean	$[-0.0312, +0.0530]$	$[-0.0476, +0.0028]$
		$+0.0112$	-0.0220
Training-unseen	Ridge mean	$[-0.0302, +0.0526]$	$[-0.0474, +0.0034]$

Table 16: Chronos-2: all four legal negative-binomial distributions, scored with equal-campaign weights. No hybrid distribution was fitted.

Population	Control	S_{00}	S_{10}	S_{01}	S_{11}
Original cohort	Group mean	0.12402	0.12057	0.11799	0.10881
Original cohort	Ridge mean	0.12423	0.12057	0.11802	0.10881
All later	Group mean	0.36547	0.35275	0.36656	0.34871
All later	Ridge mean	0.36586	0.35275	0.36659	0.34871
Training-seen	Group mean	0.09953	0.09529	0.09268	0.08215
Training-seen	Ridge mean	0.09973	0.09529	0.09272	0.08215
Validation-only	Group mean	0.33954	0.34302	0.34067	0.34335
Validation-only	Ridge mean	0.33978	0.34302	0.34064	0.34335
Outside development	Group mean	2.91800	2.80736	2.99440	2.88493
Outside development	Ridge mean	2.92038	2.80736	2.99448	2.88493
Training-unseen	Group mean	1.57992	1.52850	1.61726	1.56599
Training-unseen	Ridge mean	1.58119	1.52850	1.61728	1.56599

Table 17: Moirai-large: all four legal negative-binomial distributions, scored with equal-campaign weights. No hybrid distribution was fitted.

Population	Control	S_{00}	S_{10}	S_{01}	S_{11}
Original cohort	Group mean	0.10009	0.10029	0.09852	0.09481
Original cohort	Ridge mean	0.10040	0.10029	0.09864	0.09481
All later	Group mean	0.29803	0.29785	0.30126	0.29491
All later	Ridge mean	0.29840	0.29785	0.30134	0.29491
Training-seen	Group mean	0.07832	0.07884	0.07571	0.07201
Training-seen	Ridge mean	0.07862	0.07884	0.07583	0.07201
Validation-only	Group mean	0.29164	0.28904	0.29930	0.29550
Validation-only	Ridge mean	0.29199	0.28904	0.29931	0.29550

Continued on next page

Population	Control	S_{00}	S_{10}	S_{01}	S_{11}
Outside development	Group mean	2.39063	2.38639	2.44456	2.41037
Outside development	Ridge mean	2.39166	2.38639	2.44435	2.41037
Training-unseen	Group mean	1.30137	1.29798	1.33129	1.31287
Training-unseen	Ridge mean	1.30205	1.29798	1.33119	1.31287

D.3 Matched entities across earlier and later windows

Table 18: Chronos-2: change in full-minus-control loss from training to the later window, on identical entities and saved models. Negative is a more favorable later contrast. Training comparisons include in-sample scores. Empty intersections remain not applicable.

Later population	Paired entities	Control	Later minus earlier
Original cohort	2772	Constant	$+0.0078$ $[-0.0077, +0.0234]$
Original cohort	2772	Group mean	$+0.0052$ $[-0.0117, +0.0221]$
Original cohort	2772	Ridge mean	$+0.0055$ $[-0.0114, +0.0223]$
All later	2772	Constant	$+0.0078$ $[-0.0077, +0.0234]$
All later	2772	Group mean	$+0.0052$ $[-0.0117, +0.0221]$
All later	2772	Ridge mean	$+0.0055$ $[-0.0114, +0.0223]$
Training-seen	2772	Constant	$+0.0078$ $[-0.0077, +0.0234]$
Training-seen	2772	Group mean	$+0.0052$ $[-0.0117, +0.0221]$
Training-seen	2772	Ridge mean	$+0.0055$ $[-0.0114, +0.0223]$
Validation-only	0	Constant	Not applicable
Validation-only	0	Group mean	Not applicable
Validation-only	0	Ridge mean	Not applicable
Outside development	0	Constant	Not applicable
Outside development	0	Group mean	Not applicable
Outside development	0	Ridge mean	Not applicable
Training-unseen	0	Constant	Not applicable
Training-unseen	0	Group mean	Not applicable
Training-unseen	0	Ridge mean	Not applicable

Table 19: Moirai-large: change in full-minus-control loss from training to the later window, on identical entities and saved models. Negative is a more favorable later contrast. Training comparisons include in-sample scores. Empty intersections remain not applicable.

Later population	Paired entities	Control	Later minus earlier
Original cohort	2772	Constant	$+0.0143$ $[+0.0087, +0.0199]$
Original cohort	2772	Group mean	$+0.0094$ $[+0.0035, +0.0153]$

Continued on next page

Later population	Paired entities	Control	Later minus earlier
Original cohort	2772	Ridge mean	$+0.0097$ $[+0.0039, +0.0155]$
All later	2772	Constant	$+0.0143$ $[+0.0087, +0.0199]$
All later	2772	Group mean	$+0.0094$ $[+0.0035, +0.0153]$
All later	2772	Ridge mean	$+0.0097$ $[+0.0039, +0.0155]$
Training-seen	2772	Constant	$+0.0143$ $[+0.0087, +0.0199]$
Training-seen	2772	Group mean	$+0.0094$ $[+0.0035, +0.0153]$
Training-seen	2772	Ridge mean	$+0.0097$ $[+0.0039, +0.0155]$
Validation-only	0	Constant	Not applicable
Validation-only	0	Group mean	Not applicable
Validation-only	0	Ridge mean	Not applicable
Outside development	0	Constant	Not applicable
Outside development	0	Group mean	Not applicable
Outside development	0	Ridge mean	Not applicable
Training-unseen	0	Constant	Not applicable
Training-unseen	0	Group mean	Not applicable
Training-unseen	0	Ridge mean	Not applicable

Table 20: Chronos-2: change in full-minus-control loss from validation to the later window, on identical entities and saved models. Negative is a more favorable later contrast. Training comparisons include in-sample scores. Empty intersections remain not applicable.

Later population	Paired entities	Control	Later minus earlier
Original cohort	1767	Constant	-0.0074 $[-0.0133, -0.0015]$
Original cohort	1767	Group mean	$+0.0684$ $[+0.0355, +0.1014]$
Original cohort	1767	Ridge mean	-0.0078 $[-0.0140, -0.0016]$
All later	1767	Constant	-0.0074 $[-0.0133, -0.0015]$
All later	1767	Group mean	$+0.0684$ $[+0.0355, +0.1014]$
All later	1767	Ridge mean	-0.0078 $[-0.0140, -0.0016]$
Training-seen	1452	Constant	-0.0079 $[-0.0141, -0.0016]$
Training-seen	1452	Group mean	$+0.0858$ $[+0.0482, +0.1234]$
Training-seen	1452	Ridge mean	-0.0071 $[-0.0126, -0.0017]$
Validation-only	315	Constant	-0.0052 $[-0.0367, +0.0263]$
Validation-only	315	Group mean	-0.0116 $[-0.0429, +0.0198]$
Validation-only	315	Ridge mean	-0.0108 $[-0.0422, +0.0205]$

Continued on next page

Later population	Paired entities	Control	Later minus earlier
Outside development	0	Constant	Not applicable
Outside development	0	Group mean	Not applicable
Outside development	0	Ridge mean	Not applicable
Training-unseen	315	Constant	-0.0052 $[-0.0367, +0.0263]$
Training-unseen	315	Group mean	-0.0116 $[-0.0429, +0.0198]$
Training-unseen	315	Ridge mean	-0.0108 $[-0.0422, +0.0205]$

Table 21: Moirai-large: change in full-minus-control loss from validation to the later window, on identical entities and saved models. Negative is a more favorable later contrast. Training comparisons include in-sample scores. Empty intersections remain not applicable.

Later population	Paired entities	Control	Later minus earlier
Original cohort	1767	Constant	$+0.0000$ $[-0.0040, +0.0040]$
Original cohort	1767	Group mean	$+0.0418$ $[+0.0168, +0.0667]$
Original cohort	1767	Ridge mean	-0.0012 $[-0.0059, +0.0035]$
All later	1767	Constant	$+0.0000$ $[-0.0040, +0.0040]$
All later	1767	Group mean	$+0.0418$ $[+0.0168, +0.0667]$
All later	1767	Ridge mean	-0.0012 $[-0.0059, +0.0035]$
Training-seen	1452	Constant	-0.0029 $[-0.0059, +0.0002]$
Training-seen	1452	Group mean	$+0.0496$ $[+0.0204, +0.0788]$
Training-seen	1452	Ridge mean	-0.0029 $[-0.0058, +0.0001]$
Validation-only	315	Constant	$+0.0133$ $[-0.0018, +0.0284]$
Validation-only	315	Group mean	$+0.0055$ $[-0.0205, +0.0315]$
Validation-only	315	Ridge mean	$+0.0063$ $[-0.0197, +0.0323]$
Outside development	0	Constant	Not applicable
Outside development	0	Group mean	Not applicable
Outside development	0	Ridge mean	Not applicable
Training-unseen	315	Constant	$+0.0133$ $[-0.0018, +0.0284]$
Training-unseen	315	Group mean	$+0.0055$ $[-0.0205, +0.0315]$
Training-unseen	315	Ridge mean	$+0.0063$ $[-0.0197, +0.0323]$

D.4 PERMUTED-LABEL EXPOSURE COUNTS

Table 22: Development union: maximum absolute difference between assigned and true counts over all 37 labels. Complete per-label counts are released alongside this table. Zero training origins make origin equality vacuous.

Permutation	Entities	Training origins	Largest label count change	Largest origin count change
1	6338	461324	0	28828
2	6338	461324	0	27105
3	6338	461324	0	29153
4	6338	461324	0	27209
5	6338	461324	0	27645
6	6338	461324	0	30356
7	6338	461324	0	27490
8	6338	461324	0	28887
9	6338	461324	0	27587
10	6338	461324	0	27550

Table 23: Original later cohort: maximum absolute difference between assigned and true counts over all 37 labels. Complete per-label counts are released alongside this table. Zero training origins make origin equality vacuous.

Permutation	Entities	Training origins	Largest label count change	Largest origin count change
1	3087	284653	0	20196
2	3087	284653	0	19887
3	3087	284653	0	20548
4	3087	284653	0	19725
5	3087	284653	0	21032
6	3087	284653	0	19819
7	3087	284653	0	19718
8	3087	284653	0	21403
9	3087	284653	0	22109
10	3087	284653	0	21051

Table 24: All training: maximum absolute difference between assigned and true counts over all 37 labels. Complete per-label counts are released alongside this table. Zero training origins make origin equality vacuous.

Permutation	Entities	Training origins	Largest label count change	Largest origin count change
1	5985	461324	203	28828
2	5985	461324	206	27105
3	5985	461324	208	29153
4	5985	461324	197	27209
5	5985	461324	203	27645
6	5985	461324	205	30356
7	5985	461324	210	27490
8	5985	461324	213	28887
9	5985	461324	218	27587
10	5985	461324	206	27550

Table 25: Training exclusive: maximum absolute difference between assigned and true counts over all 37 labels. Complete per-label counts are released alongside this table. Zero training origins make origin equality vacuous.

Permutation	Entities	Training origins	Largest label count change	Largest origin count change
1	4421	311064	133	26171
2	4421	311064	134	24394
3	4421	311064	128	25877
4	4421	311064	130	25176
5	4421	311064	124	24866
6	4421	311064	120	28789
7	4421	311064	144	24660
8	4421	311064	128	26831
9	4421	311064	132	24832
10	4421	311064	132	25132

Table 26: All validation: maximum absolute difference between assigned and true counts over all 37 labels. Complete per-label counts are released alongside this table. Zero training origins make origin equality vacuous.

Permutation	Entities	Training origins	Largest label count change	Largest origin count change
1	1917	150260	133	22476
2	1917	150260	134	20141
3	1917	150260	128	17012
4	1917	150260	130	19782
5	1917	150260	124	19473
6	1917	150260	120	19841
7	1917	150260	144	18629
8	1917	150260	128	21052
9	1917	150260	132	20605
10	1917	150260	132	17914

Table 27: Validation only: maximum absolute difference between assigned and true counts over all 37 labels. Complete per-label counts are released alongside this table. Zero training origins make origin equality vacuous.

Permutation	Entities	Training origins	Largest label count change	Largest origin count change
1	353	0	203	0
2	353	0	206	0
3	353	0	208	0
4	353	0	197	0
5	353	0	203	0
6	353	0	205	0
7	353	0	210	0
8	353	0	213	0
9	353	0	218	0
10	353	0	206	0

E ORDINARY MEAN AND DISPERSION CALIBRATION

The saved-parameter decomposition motivates a familiar stronger control: can ordinary distribution calibration account for a gain that is not explained by adjusting means alone? We retain the previously fitted constant-label neural heads and their global, group-specific and ridge group-specific mean offsets. With those means fixed, we fit an additional global or group-specific log-dispersion shift using training data only. The ridge control uses the same unit quadratic penalty for the new shifts. All shifts lie in $[-5, 5]$; no decoded value is clipped after the shift. These are ordinary sequential mean-plus-dispersion calibrations, not a new architecture or joint maximum-likelihood estimator.

The prospective family contains 20 saved-head cells, two feature extractors and ten seeds, with three calibrators in each cell. The first training-only cost fit is reused as its declared family member instead of being fitted again. Every one-dimensional objective is evaluated at 101 fixed grid points, with the declared neighboring intervals refined by bounded local minimization. The best finite feasible grid or refined candidate is retained, together with all local stopping failures. This finite search does not establish convexity or global optimality. No validation or later score chooses a grid, penalty, group, endpoint or model.

The full comparison family has 36 contrasts: two feature extractors, six fixed later populations and three controls. We retain the original 3087-campaign population and all 3379 later, 2772 training-seen, 315 validation-only, 292 outside-development and 607 training-unseen campaigns. Negative values indicate lower seven-day negative log-likelihood for the identity-aware readout. Intervals use ten paired optimization seeds with Bonferroni adjustment over all 36 contrasts and condition on these fixed campaigns. Campaign medians and complete platform omissions are descriptive, not an independent-sample or joint-dependence guarantee. This additional family does not replace the previously reported narrower mean-control comparisons or their adjustment.

All four specified group and ridge comparisons favor the identity-aware readout after the full adjustment. This supports fitted benefit beyond these two sequential calibration procedures on the original cohort, not a unique interaction mechanism or an oracle function-class gap.

Independent saved-parameter reconstruction covers all 60 later calibrated probability surfaces, all 36 comparisons and all selected training objectives. The complete 101-point training grid is independently evaluated only for the prospectively selected first cell, including every group of all three controls. The probability calculation uses a separate log-gamma expression, shared with the producer's numerical sanity check rather than its primary beta-function score. The reconstruction shares immutable input loaders and is neither a repeated neural fit nor a fresh-environment reproduction.

On all 3,379 later campaigns, the 315 validation-only campaigns, the 292 outside-development campaigns and the 607 training-unseen campaigns, all 24 simultaneous intervals include zero. These populations do not establish superiority or equivalence. The 12 favorable comparisons are confined to the original 3,087 campaigns and the 2,772 training-seen subset.

E.1 COMPLETE POPULATIONS AND CONTROLS

Table 28: Chronos-2: all six prespecified populations and all three ordinary controls. Differences and uncertainty units match Table 3. The 607 training-unseen campaigns combine 315 validation-only and 292 outside-development campaigns; none is an untouched confirmation. All signs are retained.

Population	Control	Mean and adjusted seed interval	Campaign median
Original cohort (3087)	Global	-0.01559 $[-0.02459, -0.00659]$	-0.00172
Original cohort (3087)	Group	-0.01540 $[-0.02450, -0.00629]$	-0.00082
Original cohort (3087)	Ridge	-0.01569 $[-0.02494, -0.00644]$	-0.00110

Continued on next page

Population	Control	Mean and adjusted seed interval	Campaign median
All later (3379)	Global	-0.01992 $[-0.06037, +0.02052]$ -0.02364	-0.00173
All later (3379)	Group	$[-0.06470, +0.01741]$ -0.02361	-0.00084
All later (3379)	Ridge	$[-0.06467, +0.01745]$ -0.01769	-0.00114
Training-seen (2772)	Global	$[-0.02982, -0.00556]$ -0.01781	-0.00167
Training-seen (2772)	Group	$[-0.02989, -0.00572]$ -0.01806	-0.00056
Training-seen (2772)	Ridge	$[-0.03029, -0.00584]$ $+0.00289$	-0.00098
Validation-only (315)	Global	$[-0.02184, +0.02761]$ $+0.00580$	-0.00790
Validation-only (315)	Group	$[-0.02008, +0.03169]$ $+0.00524$	-0.00567
Validation-only (315)	Ridge	$[-0.02052, +0.03100]$ -0.06572	-0.00613
Outside development (292)	Global	$[-0.46787, +0.33643]$ -0.11082	-0.00283
Outside development (292)	Group	$[-0.52475, +0.30312]$ -0.10735	-0.00191
Outside development (292)	Ridge	$[-0.51954, +0.30484]$ -0.03012	-0.00227
Training-unseen (607)	Global	$[-0.21829, +0.15806]$ -0.05030	-0.00756
Training-unseen (607)	Group	$[-0.24512, +0.14452]$ -0.04892	-0.00539
Training-unseen (607)	Ridge	$[-0.24295, +0.14510]$	-0.00592

Table 29: Moirai-large: all six prespecified populations and all three ordinary controls. Differences and uncertainty units match Table 3. The 607 training-unseen campaigns combine 315 validation-only and 292 outside-development campaigns; none is an untouched confirmation. All signs are retained.

Population	Control	Mean and adjusted seed interval	Campaign median
Original cohort (3087)	Global	-0.00595 $[-0.01021, -0.00170]$ -0.00483	-0.00190
Original cohort (3087)	Group	$[-0.00911, -0.00055]$ -0.00520	-0.00156
Original cohort (3087)	Ridge	$[-0.00951, -0.00089]$ -0.00488	-0.00167
All later (3379)	Global	$[-0.01512, +0.00536]$ -0.00472	-0.00191
All later (3379)	Group	$[-0.01478, +0.00533]$ -0.00487	-0.00159
All later (3379)	Ridge	$[-0.01481, +0.00507]$ -0.00708	-0.00170
Training-seen (2772)	Global	$[-0.01112, -0.00305]$ -0.00546	-0.00188
Training-seen (2772)	Group	$[-0.00927, -0.00165]$	-0.00120

Continued on next page

Population	Control	Mean and adjusted seed interval	Campaign median
Training-seen (2772)	Ridge	-0.00591 $[-0.00978, -0.00203]$ $+0.00399$	-0.00141
Validation-only (315)	Global	$[-0.00809, +0.01607]$ $+0.00072$	-0.00571
Validation-only (315)	Group	$[-0.01420, +0.01564]$ $+0.00104$	-0.00318
Validation-only (315)	Ridge	$[-0.01363, +0.01571]$ $+0.00646$	-0.00351
Outside development (292)	Global	$[-0.07553, +0.08844]$ -0.00360	-0.00724
Outside development (292)	Group	$[-0.08339, +0.07618]$ -0.00140	-0.00337
Outside development (292)	Ridge	$[-0.07895, +0.07614]$ $+0.00518$	-0.00437
Training-unseen (607)	Global	$[-0.03847, +0.04883]$ -0.00136	-0.00573
Training-unseen (607)	Group	$[-0.04525, +0.04253]$ -0.00014	-0.00319
Training-unseen (607)	Ridge	$[-0.04289, +0.04261]$	-0.00353

F PUBLIC IDENTITY AND CALIBRATION COMPARISONS UNDER ONE COUNT LAW

This comparison holds the daily probability family fixed across the public retail and private fundraising studies: every head predicts seven ordinary negative-binomial daily masses, and the score is their summed negative log mass. It does not assert that the two sources have the same data-generating process. The earlier retail experiment in Appendix I uses a different count law and remains a separate complete comparison.

Population and inputs. The input caches and M5 series rosters are those described in Appendix I. There are 24,325 training series, 32 origins per series and 778,400 training rows. The development population contains 6,165 different series at the same 32 origins. The later window contains all 30,490 series at 38 origins each; its 6,165 training-held-out series are a secondary population. Histories contain thirty daily counts and targets seven. M5 is previously opened development data, not an independent new source. The nonlabel readout inputs are 768 frozen representation coordinates, two location/scale coordinates and seven weekday indicators. The eight-dimensional store or department embedding yields input width 785.

Complete fitted family. The two fixed representations are Chronos-2 and Moirai-1.1-R-large. For each representation and label vocabulary, the true, constant and permuted arms share the same width-256 constructor and seed-specific initialization. There are 470,656 trainable parameters for store and 470,632 for department; the constant arm retains the full embedding table. Equal raw parameter counts do not imply equal effective classes. One multiset-preserving permutation is fixed separately within training and held-out series, and remains fixed across seeds, origins and epochs. All 120 neural fits use twenty epochs, 7,620 updates, AdamW learning rate 3×10^{-4}, weight decay 10^{-4}, gradient clipping at one and batches of 2,048. Only the final checkpoint is scored; validation is not used for stopping or selection.

Each constant head supplies three additional training-only mean corrections: a global offset, label-specific offsets and label-specific offsets with unit quadratic shrinkage. Offsets lie in $[-5, 5]$, are added after decoding and leave dispersion unchanged. This gives 120 calibrations. The objective sums each series' origin-mean seven-day negative log mass, with the stated penalty. The offset optimum is checked against the box-constrained first-order condition at tolerance 10^{-8}. The neural fits do not carry an analogous global optimality claim. Offset log means can reach magnitude 25, whereas the neural decoder bounds them by 20; these are generally nonnested classes.

Uncertainty and complete comparisons. Scores average origins within each series and then give each series equal weight. The primary family compares true store identity to permuted, constant, group-rate and shrunk-group-rate controls on both representations, on all 30,490 later series. Ten paired seeds supply Student intervals with nine degrees of freedom and Bonferroni adjustment across these eight contrasts. Each of the other label/population combinations is a separate, secondary eight-contrast family. None can replace the primary family. The primary result is zero favorable, two unfavorable and six unresolved contrasts, as shown in Table 30.

Table 30: Public store-identity comparisons under a common negative-binomial law. Differences are true identity minus reference, in nats per seven-day path; negative favors true identity. Intervals use ten paired seeds, with Bonferroni adjustment over all eight comparisons on the fixed 30,490-series population.

Backbone	Reference	Difference	Simultaneous 95% CI
Chronos-2	Permuted	0.0015	$[-0.0030,\ 0.0060]$
Chronos-2	Constant	0.0065	$[-0.0006,\ 0.0136]$
Chronos-2	Group rate	0.0093	$[0.0037,\ 0.0149]$
Chronos-2	Shrunk group	0.0093	$[0.0037,\ 0.0149]$
Moirai 1.1	Permuted	-0.0186	$[-0.0643,\ 0.0271]$
Moirai 1.1	Constant	0.0014	$[-0.0402,\ 0.0430]$
Moirai 1.1	Group rate	0.0146	$[-0.0147,\ 0.0440]$
Moirai 1.1	Shrunk group	0.0146	$[-0.0147,\ 0.0440]$

The series summaries first average paired differences over seeds, then use 20,000 shared bootstrap draws of series to report marginal 95% mean and median intervals. These are descriptive, without eight-contrast adjustment or correction for dependence through items, stores and dates. The median is of paired differences, not a difference of two model medians. Neither a zero-containing seed interval nor a descriptive series interval establishes equivalence. Absolute scores for all six arms are reported separately.

Reconstruction scope. All 120 saved model publications, 480 saved score vectors and 120 checkpoint identities are checked before complete-family interpretation. A prespecified set of twelve checkpoints, one seed across all representation, label and neural-arm combinations, is replayed on the cached inputs, including all associated calibrated scores. The replay uses the same architecture and input loader, with an independent negative-binomial mass calculation. The other 108 neural checkpoints are not replayed. Separate all-family reconstruction checks 5,046 statistical quantities with maximum absolute difference 1.32×10^{-13}.

This covers saved-score statistics and the declared inference subset, not raw M5 rematerialization, independent training or a fresh-environment replication.

Table 31: Descriptive series uncertainty. All 30,490 series in the later window, store identity. Ten paired seed differences are averaged within each series before taking the mean or median across series. The median is of paired differences, not a difference of model medians. Brackets are marginal 95% intervals from 20,000 shared series bootstrap draws, without eight-contrast adjustment or correction for shared items, stores, or overlapping dates. These intervals do not replace the primary seed analysis.

Reference	Series mean [95% CI]	Series median [95% CI]
Chronos-2		
Permuted identity	0.0015 [0.0007, 0.0024]	−0.0015 [−0.0020, −0.0009]
Constant identity	0.0065 [0.0057, 0.0073]	0.0012 [0.0006, 0.0017]
Group rate	0.0093 [0.0085, 0.0101]	0.0025 [0.0020, 0.0028]
Shrunk group rate	0.0093 [0.0085, 0.0101]	0.0025 [0.0020, 0.0028]
Moirai 1.1 large		
Permuted identity	−0.0186 [−0.0195, −0.0178]	−0.0146 [−0.0152, −0.0141]
Constant identity	0.0014 [0.0003, 0.0025]	0.0019 [0.0014, 0.0024]
Group rate	0.0146 [0.0137, 0.0156]	0.0126 [0.0121, 0.0132]
Shrunk group rate	0.0146 [0.0137, 0.0156]	0.0126 [0.0121, 0.0132]

Table 32: Secondary comparison. All 30,490 series in the later window, department identity. Negative differences favor true identity. Scores are nats per seven-day path, averaged over origins within each series and equally over series. The intervals use ten paired optimization seeds, Bonferroni-adjusted over the eight contrasts in this label and population at nominal familywise 95 percent; they condition on the fixed data. No interval containing zero establishes equivalence.

Backbone	Reference	Mean difference	Bonferroni 95% CI	Seeds < 0
Chronos-2	Permuted identity	−0.0113	[−0.0210, −0.0016]	10/10
Chronos-2	Constant identity	−0.0090	[−0.0200, 0.0021]	8/10
Chronos-2	Group rate	−0.0042	[−0.0130, 0.0046]	8/10
Chronos-2	Shrunk group rate	−0.0042	[−0.0130, 0.0046]	8/10
Moirai 1.1 large	Permuted identity	−0.0157	[−0.0454, 0.0139]	7/10
Moirai 1.1 large	Constant identity	−0.0105	[−0.0460, 0.0251]	7/10
Moirai 1.1 large	Group rate	0.0028	[−0.0256, 0.0312]	6/10
Moirai 1.1 large	Shrunk group rate	0.0028	[−0.0256, 0.0312]	6/10

Table 33: Descriptive series uncertainty. All 30,490 series in the later window, department identity. Ten paired seed differences are averaged within each series before taking the mean or median across series. The median is of paired differences, not a difference of model medians. Brackets are marginal 95% intervals from 20,000 shared series bootstrap draws, without eight-contrast adjustment or correction for shared items, stores, or overlapping dates. These intervals do not replace the primary seed analysis.

Reference	Series mean [95% CI]	Series median [95% CI]
Chronos-2		
Permuted identity	−0.0113 [−0.0126, −0.0100]	−0.0122 [−0.0131, −0.0113]
Constant identity	−0.0090 [−0.0103, −0.0077]	−0.0113 [−0.0121, −0.0105]
Group rate	−0.0042 [−0.0055, −0.0030]	−0.0062 [−0.0069, −0.0054]
Shrunk group rate	−0.0042 [−0.0055, −0.0030]	−0.0062 [−0.0069, −0.0054]
Moirai 1.1 large		
Permuted identity	−0.0157 [−0.0172, −0.0144]	−0.0112 [−0.0122, −0.0102]
Constant identity	−0.0105 [−0.0119, −0.0090]	−0.0074 [−0.0084, −0.0065]
Group rate	0.0028 [0.0013, 0.0042]	0.0074 [0.0065, 0.0082]
Shrunk group rate	0.0028 [0.0013, 0.0042]	0.0074 [0.0064, 0.0082]

Table 34: Secondary comparison. The 6,165 validation series in the development window, store identity. Negative differences favor true identity. Scores are nats per seven-day path, averaged over origins within each series and equally over series. The intervals use ten paired optimization seeds, Bonferroni-adjusted over the eight contrasts in this label and population at nominal familywise 95 percent; they condition on the fixed data. No interval containing zero establishes equivalence.

Backbone	Reference	Mean difference	Bonferroni 95% CI	Seeds < 0
Chronos-2	Permuted identity	-0.0042	$[-0.0088, 0.0005]$	8/10
Chronos-2	Constant identity	-0.0006	$[-0.0073, 0.0061]$	5/10
Chronos-2	Group rate	0.0015	$[-0.0029, 0.0059]$	1/10
Chronos-2	Shrunk group rate	0.0015	$[-0.0029, 0.0059]$	1/10
Moirai 1.1 large	Permuted identity	-0.0150	$[-0.0442, 0.0141]$	7/10
Moirai 1.1 large	Constant identity	-0.0083	$[-0.0443, 0.0277]$	5/10
Moirai 1.1 large	Group rate	0.0033	$[-0.0201, 0.0266]$	3/10
Moirai 1.1 large	Shrunk group rate	0.0033	$[-0.0201, 0.0266]$	3/10

Table 35: Descriptive series uncertainty. The 6,165 validation series in the development window, store identity. Ten paired seed differences are averaged within each series before taking the mean or median across series. The median is of paired differences, not a difference of model medians. Brackets are marginal 95% intervals from 20,000 shared series bootstrap draws, without eight-contrast adjustment or correction for shared items, stores, or overlapping dates. These intervals do not replace the primary seed analysis.

Reference	Series mean [95% CI]	Series median [95% CI]
Chronos-2		
Permuted identity	-0.0042 $[-0.0056, -0.0028]$	-0.0034 $[-0.0041, -0.0027]$
Constant identity	-0.0006 $[-0.0018, 0.0006]$	-0.0012 $[-0.0019, -0.0007]$
Group rate	0.0015 $[0.0003, 0.0026]$	-0.0004 $[-0.0010, 0.0002]$
Shrunk group rate	0.0015 $[0.0003, 0.0026]$	-0.0004 $[-0.0010, 0.0002]$
Moirai 1.1 large		
Permuted identity	-0.0150 $[-0.0161, -0.0139]$	-0.0093 $[-0.0100, -0.0088]$
Constant identity	-0.0083 $[-0.0098, -0.0067]$	-0.0010 $[-0.0017, -0.0004]$
Group rate	0.0033 $[0.0019, 0.0047]$	0.0067 $[0.0062, 0.0074]$
Shrunk group rate	0.0033 $[0.0019, 0.0047]$	0.0067 $[0.0062, 0.0074]$

Table 36: Secondary comparison. The 6,165 validation series in the development window, department identity. Negative differences favor true identity. Scores are nats per seven-day path, averaged over origins within each series and equally over series. The intervals use ten paired optimization seeds, Bonferroni-adjusted over the eight contrasts in this label and population at nominal familywise 95 percent; they condition on the fixed data. No interval containing zero establishes equivalence.

Backbone	Reference	Mean difference	Bonferroni 95% CI	Seeds < 0
Chronos-2	Permuted identity	-0.0203	$[-0.0301, -0.0105]$	10/10
Chronos-2	Constant identity	-0.0172	$[-0.0267, -0.0078]$	10/10
Chronos-2	Group rate	-0.0130	$[-0.0203, -0.0058]$	10/10
Chronos-2	Shrunk group rate	-0.0130	$[-0.0203, -0.0058]$	10/10
Moirai 1.1 large	Permuted identity	-0.0196	$[-0.0376, -0.0017]$	9/10
Moirai 1.1 large	Constant identity	-0.0206	$[-0.0485, 0.0073]$	8/10
Moirai 1.1 large	Group rate	-0.0081	$[-0.0335, 0.0173]$	6/10
Moirai 1.1 large	Shrunk group rate	-0.0081	$[-0.0335, 0.0173]$	6/10

Table 37: Descriptive series uncertainty. The 6,165 validation series in the development window, department identity. Ten paired seed differences are averaged within each series before taking the mean or median across series. The median is of paired differences, not a difference of model medians. Brackets are marginal 95% intervals from 20,000 shared series bootstrap draws, without eight-contrast adjustment or correction for shared items, stores, or overlapping dates. These intervals do not replace the primary seed analysis.

Reference	Series mean [95% CI]	Series median [95% CI]
Chronos-2		
Permuted identity	-0.0203 $[-0.0223, -0.0183]$	-0.0144 $[-0.0157, -0.0129]$
Constant identity	-0.0172 $[-0.0191, -0.0154]$	-0.0121 $[-0.0134, -0.0109]$
Group rate	-0.0130 $[-0.0148, -0.0112]$	-0.0088 $[-0.0099, -0.0076]$
Shrunk group rate	-0.0130 $[-0.0148, -0.0112]$	-0.0088 $[-0.0099, -0.0076]$
Moirai 1.1 large		
Permuted identity	-0.0196 $[-0.0214, -0.0180]$	-0.0129 $[-0.0146, -0.0117]$
Constant identity	-0.0206 $[-0.0225, -0.0188]$	-0.0126 $[-0.0137, -0.0116]$
Group rate	-0.0081 $[-0.0100, -0.0063]$	-0.0018 $[-0.0028, -0.0006]$
Shrunk group rate	-0.0081 $[-0.0100, -0.0063]$	-0.0018 $[-0.0028, -0.0006]$

Table 38: Secondary comparison. The same 6,165 training-held-out series in the later window, store identity. Negative differences favor true identity. Scores are nats per seven-day path, averaged over origins within each series and equally over series. The intervals use ten paired optimization seeds, Bonferroni-adjusted over the eight contrasts in this label and population at nominal familywise 95 percent; they condition on the fixed data. No interval containing zero establishes equivalence.

Backbone	Reference	Mean difference	Bonferroni 95% CI	Seeds < 0
Chronos-2	Permuted identity	0.0007	$[-0.0043, 0.0057]$	4/10
Chronos-2	Constant identity	0.0064	$[-0.0007, 0.0136]$	1/10
Chronos-2	Group rate	0.0091	$[0.0032, 0.0149]$	1/10
Chronos-2	Shrunk group rate	0.0091	$[0.0032, 0.0149]$	1/10
Moirai 1.1 large	Permuted identity	-0.0187	$[-0.0641, 0.0267]$	7/10
Moirai 1.1 large	Constant identity	0.0016	$[-0.0398, 0.0431]$	5/10
Moirai 1.1 large	Group rate	0.0147	$[-0.0141, 0.0436]$	3/10
Moirai 1.1 large	Shrunk group rate	0.0147	$[-0.0141, 0.0436]$	3/10

Table 39: Descriptive series uncertainty. The same 6,165 training-held-out series in the later window, store identity. Ten paired seed differences are averaged within each series before taking the mean or median across series. The median is of paired differences, not a difference of model medians. Brackets are marginal 95% intervals from 20,000 shared series bootstrap draws, without eight-contrast adjustment or correction for shared items, stores, or overlapping dates. These intervals do not replace the primary seed analysis.

Reference	Series mean [95% CI]	Series median [95% CI]
Chronos-2		
Permuted identity	0.0007 $[-0.0010, 0.0025]$	-0.0010 $[-0.0022, 0.0004]$
Constant identity	0.0064 $[0.0046, 0.0083]$	0.0022 $[0.0011, 0.0033]$
Group rate	0.0091 $[0.0072, 0.0109]$	0.0031 $[0.0022, 0.0042]$
Shrunk group rate	0.0091 $[0.0072, 0.0109]$	0.0031 $[0.0022, 0.0042]$
Moirai 1.1 large		
Permuted identity	-0.0187 $[-0.0207, -0.0167]$	-0.0149 $[-0.0160, -0.0137]$
Constant identity	0.0016 $[-0.0008, 0.0041]$	0.0021 $[0.0011, 0.0032]$
Group rate	0.0147 $[0.0126, 0.0169]$	0.0133 $[0.0121, 0.0145]$
Shrunk group rate	0.0147 $[0.0126, 0.0169]$	0.0133 $[0.0121, 0.0145]$

Table 40: Secondary comparison. The same 6,165 training-held-out series in the later window, department identity. Negative differences favor true identity. Scores are nats per seven-day path, averaged over origins within each series and equally over series. The intervals use ten paired optimization seeds, Bonferroni-adjusted over the eight contrasts in this label and population at nominal familywise 95 percent; they condition on the fixed data. No interval containing zero establishes equivalence.

Backbone	Reference	Mean difference	Bonferroni 95% CI	Seeds < 0
Chronos-2	Permuted identity	-0.0119	$[-0.0214, -0.0024]$	10/10
Chronos-2	Constant identity	-0.0090	$[-0.0201, 0.0021]$	10/10
Chronos-2	Group rate	-0.0044	$[-0.0134, 0.0046]$	6/10
Chronos-2	Shrunk group rate	-0.0044	$[-0.0134, 0.0046]$	6/10
Moirai 1.1 large	Permuted identity	-0.0144	$[-0.0431, 0.0143]$	6/10
Moirai 1.1 large	Constant identity	-0.0096	$[-0.0448, 0.0256]$	7/10
Moirai 1.1 large	Group rate	0.0037	$[-0.0245, 0.0319]$	6/10
Moirai 1.1 large	Shrunk group rate	0.0037	$[-0.0245, 0.0319]$	6/10

Table 41: Descriptive series uncertainty. The same 6,165 training-held-out series in the later window, department identity. Ten paired seed differences are averaged within each series before taking the mean or median across series. The median is of paired differences, not a difference of model medians. Brackets are marginal 95% intervals from 20,000 shared series bootstrap draws, without eight-contrast adjustment or correction for shared items, stores, or overlapping dates. These intervals do not replace the primary seed analysis.

Reference	Series mean [95% CI]	Series median [95% CI]
Chronos-2		
Permuted identity	-0.0119 $[-0.0151, -0.0088]$	-0.0115 $[-0.0140, -0.0097]$
Constant identity	-0.0090 $[-0.0120, -0.0060]$	-0.0108 $[-0.0126, -0.0084]$
Group rate	-0.0044 $[-0.0073, -0.0015]$	-0.0058 $[-0.0074, -0.0041]$
Shrunk group rate	-0.0044 $[-0.0073, -0.0015]$	-0.0058 $[-0.0074, -0.0041]$
Moirai 1.1 large		
Permuted identity	-0.0144 $[-0.0174, -0.0115]$	-0.0110 $[-0.0129, -0.0084]$
Constant identity	-0.0096 $[-0.0127, -0.0065]$	-0.0076 $[-0.0097, -0.0057]$
Group rate	0.0037 $[0.0006, 0.0067]$	0.0072 $[0.0057, 0.0092]$
Shrunk group rate	0.0037 $[0.0006, 0.0067]$	0.0072 $[0.0057, 0.0092]$

Table 42: Absolute negative log scores for every arm. All 30,490 series in the later window, store identity. All six arms are retained. Each score sums seven daily losses, averages origins within a series, and then averages ten fitted seeds. The two columns summarize that seed-averaged series score by its mean and median. Lower is better. The arm medians are not paired contrast medians. Global and group rate corrections are fitted only on training data after the constant-identity head; they do not count as additional neural fits.

Backbone	Readout	Series mean	Series median
Chronos-2	Constant identity	8.5400	7.7556
Chronos-2	True identity	8.5464	7.7603
Chronos-2	Permuted identity	8.5449	7.7572
Chronos-2	Global rate	8.5372	7.7534
Chronos-2	Group rate	8.5371	7.7546
Chronos-2	Shrunk group rate	8.5371	7.7546
Moirai 1.1 large	Constant identity	8.5551	7.7703
Moirai 1.1 large	True identity	8.5565	7.7688
Moirai 1.1 large	Permuted identity	8.5751	7.7822
Moirai 1.1 large	Global rate	8.5418	7.7675
Moirai 1.1 large	Group rate	8.5418	7.7679
Moirai 1.1 large	Shrunk group rate	8.5418	7.7679

Table 43: Absolute negative log scores for every arm. All 30,490 series in the later window, department identity. All six arms are retained. Each score sums seven daily losses, averages origins within a series, and then averages ten fitted seeds. The two columns summarize that seed-averaged series score by its mean and median. Lower is better. The arm medians are not paired contrast medians. Global and group rate corrections are fitted only on training data after the constant-identity head; they do not count as additional neural fits.

Backbone	Readout	Series mean	Series median
Chronos-2	Constant identity	8.5424	7.7545
Chronos-2	True identity	8.5334	7.7568
Chronos-2	Permuted identity	8.5447	7.7582
Chronos-2	Global rate	8.5385	7.7545
Chronos-2	Group rate	8.5377	7.7597
Chronos-2	Shrunk group rate	8.5377	7.7597
Moirai 1.1 large	Constant identity	8.5562	7.7691
Moirai 1.1 large	True identity	8.5457	7.7557
Moirai 1.1 large	Permuted identity	8.5614	7.7627
Moirai 1.1 large	Global rate	8.5442	7.7583
Moirai 1.1 large	Group rate	8.5429	7.7606
Moirai 1.1 large	Shrunk group rate	8.5429	7.7606

Table 44: Absolute negative log scores for every arm. The 6,165 validation series in the development window, store identity. All six arms are retained. Each score sums seven daily losses, averages origins within a series, and then averages ten fitted seeds. The two columns summarize that seed-averaged series score by its mean and median. Lower is better. The arm medians are not paired contrast medians. Global and group rate corrections are fitted only on training data after the constant-identity head; they do not count as additional neural fits.

Backbone	Readout	Series mean	Series median
Chronos-2	Constant identity	6.2521	5.1258
Chronos-2	True identity	6.2515	5.1238
Chronos-2	Permuted identity	6.2557	5.1314
Chronos-2	Global rate	6.2502	5.1248
Chronos-2	Group rate	6.2501	5.1238
Chronos-2	Shrunk group rate	6.2501	5.1238
Moirai 1.1 large	Constant identity	6.2812	5.1389
Moirai 1.1 large	True identity	6.2729	5.1364
Moirai 1.1 large	Permuted identity	6.2879	5.1484
Moirai 1.1 large	Global rate	6.2698	5.1381
Moirai 1.1 large	Group rate	6.2697	5.1376
Moirai 1.1 large	Shrunk group rate	6.2697	5.1376

Table 45: Absolute negative log scores for every arm. The 6,165 validation series in the development window, department identity. All six arms are retained. Each score sums seven daily losses, averages origins within a series, and then averages ten fitted seeds. The two columns summarize that seed-averaged series score by its mean and median. Lower is better. The arm medians are not paired contrast medians. Global and group rate corrections are fitted only on training data after the constant-identity head; they do not count as additional neural fits.

Backbone	Readout	Series mean	Series median
Chronos-2	Constant identity	6.2546	5.1272
Chronos-2	True identity	6.2373	5.1247
Chronos-2	Permuted identity	6.2577	5.1325
Chronos-2	Global rate	6.2516	5.1266
Chronos-2	Group rate	6.2504	5.1280
Chronos-2	Shrunk group rate	6.2504	5.1280
Moirai 1.1 large	Constant identity	6.2802	5.1410
Moirai 1.1 large	True identity	6.2596	5.1321
Moirai 1.1 large	Permuted identity	6.2792	5.1366
Moirai 1.1 large	Global rate	6.2696	5.1353
Moirai 1.1 large	Group rate	6.2677	5.1401
Moirai 1.1 large	Shrunk group rate	6.2677	5.1400

Table 46: Absolute negative log scores for every arm. The same 6,165 training-held-out series in the later window, store identity. All six arms are retained. Each score sums seven daily losses, averages origins within a series, and then averages ten fitted seeds. The two columns summarize that seed-averaged series score by its mean and median. Lower is better. The arm medians are not paired contrast medians. Global and group rate corrections are fitted only on training data after the constant-identity head; they do not count as additional neural fits.

Backbone	Readout	Series mean	Series median
Chronos-2	Constant identity	8.4912	7.6722
Chronos-2	True identity	8.4976	7.6627
Chronos-2	Permuted identity	8.4968	7.6726
Chronos-2	Global rate	8.4885	7.6713
Chronos-2	Group rate	8.4885	7.6651
Chronos-2	Shrunk group rate	8.4885	7.6651
Moirai 1.1 large	Constant identity	8.5046	7.6878
Moirai 1.1 large	True identity	8.5062	7.6942
Moirai 1.1 large	Permuted identity	8.5249	7.7000
Moirai 1.1 large	Global rate	8.4913	7.6752
Moirai 1.1 large	Group rate	8.4915	7.6730
Moirai 1.1 large	Shrunk group rate	8.4915	7.6730

Table 47: Absolute negative log scores for every arm. The same 6,165 training-held-out series in the later window, department identity. All six arms are retained. Each score sums seven daily losses, averages origins within a series, and then averages ten fitted seeds. The two columns summarize that seed-averaged series score by its mean and median. Lower is better. The arm medians are not paired contrast medians. Global and group rate corrections are fitted only on training data after the constant-identity head; they do not count as additional neural fits.

Backbone	Readout	Series mean	Series median
Chronos-2	Constant identity	8.4935	7.6724
Chronos-2	True identity	8.4845	7.6722
Chronos-2	Permuted identity	8.4964	7.6711
Chronos-2	Global rate	8.4897	7.6714
Chronos-2	Group rate	8.4889	7.6699
Chronos-2	Shrunk group rate	8.4889	7.6699
Moirai 1.1 large	Constant identity	8.5057	7.6791
Moirai 1.1 large	True identity	8.4961	7.6777
Moirai 1.1 large	Permuted identity	8.5105	7.6857
Moirai 1.1 large	Global rate	8.4936	7.6721
Moirai 1.1 large	Group rate	8.4924	7.6793
Moirai 1.1 large	Shrunk group rate	8.4924	7.6793

G INFORMATION ACCESS AND FITTED PREDICTION ON ONE CHECKPOINT

Question and construction. Does making a label accessible to a frozen encoder improve a fitted forecast, rather than merely change its representation? We hold a Moirai 2.0 R small checkpoint fixed and compare six pathways on a metadata-ranked, balanced subset of the already-opened M5 data. The training cohort contains 1400 series and the held-out cohort 350 series, with respectively 20 and 5 series in each of 70 store–department strata. The later population contains all 1750 selected series; its 350 training-unseen series are reported separately. This is a constructed development experiment, not independent-source confirmation or the full M5 benchmark.

For the erased-covariate pathway, normalized covariate values are replaced by zero before the encoder input projection. True, constant and fixed-permuted identities then give identical cached representations on all 122,500 evaluated origins. For the input-code pathways, a fixed equal-norm code occupies the 37 observed covariate entries. The 11 unobserved entries are fixed to zero, and the observation mask is held fixed. The constant, true and permuted codes use the same construction; this is an explicit intervention after normalization, not a change to the native software interface. The head-label pathways use the erased representation and receive either the true or fixed-permuted store label directly in the readout. Permutations preserve store frequencies separately in the selected training and held-out cohorts.

Matched fitting and controls. All six pathways use the same 371,584 stored trainable parameters, 256-wide readout, seven daily negative-binomial factors, initialization, row order and 440 optimizer updates across 20 epochs. There are ten seeds per pathway and no endpoint selection by validation or later scores. Matching stored parameter counts does not match active parameters, information or effective function classes. Three ordinary controls fit global, store-specific and ridge-regularized store-specific mean offsets to the erased-covariate head using training data only, with its dispersion fixed. All 60 neural fits and 30 calibrations are retained.

Complete primary result. On the 1750-series later population, true minus permuted input code is -0.00580 nats, with paired-seed interval $[-0.01072, -0.00088]$. However, it is worse than the constant code by $+0.01466$ nats, with interval $[+0.00627, +0.02304]$. Its differences from store-rate and ridge store-rate calibration are unresolved: $+0.00137$ with $[-0.00897, +0.01170]$ and $+0.00137$ with $[-0.00896, +0.01170]$, respectively. Thus one of the four primary comparisons is favorable, one is reversed and two are unresolved. Representational response alone does not establish a predictive gain for this fitted procedure.

On the full later population, the true head label has no resolved advantage over its permuted counterpart and is worse than the erased head and both store-rate controls. Among the 350 training-unseen later series, true head labels outperform permuted labels, but their differences from the erased head and both calibrations remain unresolved. That subgroup finding does not replace the prespecified primary population or establish source transfer. Unresolved intervals are not equivalence claims; these comparisons do not rule out other input encodings, training procedures or identity-aware models.

Uncertainty and reconstruction. Each score averages seven-day negative log mass over origins within a series and then gives series equal weight. The primary input-code contrasts and secondary head-label contrasts have separate four-comparison Bonferroni adjustments within each population. Their Student-t intervals use ten paired optimization seeds and condition on the fixed series population. The separate series bootstrap averages seeds first and is descriptive because shared store and calendar dependence is not corrected. All populations, signs, intervals and store omissions are released. Saved-parameter reconstruction checks 7,441,000 path scores with maximum discrepancy 2.84×10^{-13} and 1449 statistical quantities with maximum discrepancy 4.44×10^{-15}. All six preselected seed-1729 checkpoints reproduce their complete validation and later parameter surfaces exactly; the erased-head training export is also replayed. These checks share the architecture and input loader and do not rerun training or constitute a fresh-environment reproduction.

G.1 COMPLETE METHOD MEANS AND UNCERTAINTY

Table 48: Same-checkpoint evaluation of all six neural pathways and three rate calibrations. Entries are mean negative log-likelihood in nats per seven-day path, averaging origins within series and then weighting series equally. Lower is better.

Method	All later 1750	Earlier held out 350	Later unseen 350
Erased covariate	8.01432	5.55981	7.84475
Constant input code	7.99774	5.55417	7.82891
True input code	8.01240	5.56748	7.83590
Permuted input code	8.01819	5.57234	7.85264
True head label	8.02907	5.57239	7.85117
Permuted head label	8.03359	5.57688	7.86454
Global-rate calibration	8.01138	5.55735	7.84144
Store-rate calibration	8.01103	5.55764	7.84140
Ridge store-rate calibration	8.01103	5.55764	7.84140

Table 49: All later series (1750). Negative differences favor the first method. The first four and next four contrasts each use their own prespecified Bonferroni-four family over ten optimization seeds. The final cross-pathway contrast is descriptive with a nominal 95% interval. Seed inference conditions on these fixed series; families are not pooled across populations.

Comparison	Mean	Paired-seed interval
True input code − Constant input code	$+0.01466$	$[+0.00627, +0.02304]$
True input code − Permuted input code	-0.00580	$[-0.01072, -0.00088]$
True input code − Store-rate calibration	$+0.00137$	$[-0.00897, +0.01170]$
True input code − Ridge store-rate calibration	$+0.00137$	$[-0.00896, +0.01170]$
True head label − Permuted head label	-0.00452	$[-0.01234, +0.00331]$
True head label − Erased covariate	$+0.01476$	$[+0.00469, +0.02482]$
True head label − Store-rate calibration	$+0.01804$	$[+0.00879, +0.02729]$
True head label − Ridge store-rate calibration	$+0.01804$	$[+0.00880, +0.02729]$
True input code − True head label (descriptive)	-0.01667	$[-0.02466, -0.00869]$

Table 50: Earlier held-out series (350). Negative differences favor the first method. The first four and next four contrasts each use their own prespecified Bonferroni-four family over ten optimization seeds. The final cross-pathway contrast is descriptive with a nominal 95% interval. Seed inference conditions on these fixed series; families are not pooled across populations.

Comparison	Mean	Paired-seed interval
True input code − Constant input code	$+0.01331$	$[+0.00901, +0.01760]$
True input code − Permuted input code	-0.00486	$[-0.00916, -0.00056]$
True input code − Store-rate calibration	$+0.00983$	$[+0.00381, +0.01586]$
True input code − Ridge store-rate calibration	$+0.00983$	$[+0.00381, +0.01586]$

Continued on next page

Comparison	Mean	Paired-seed interval
True head label − Permuted head label	-0.00449	$[-0.01212, +0.00314]$
True head label − Erased covariate	$+0.01258$	$[+0.00486, +0.02030]$
True head label − Store-rate calibration	$+0.01475$	$[+0.00702, +0.02247]$
True head label − Ridge store-rate calibration	$+0.01475$	$[+0.00702, +0.02247]$
True input code − True head label (descriptive)	-0.00491	$[-0.01088, +0.00105]$

Table 51: Later training-unseen series (350). Negative differences favor the first method. The first four and next four contrasts each use their own prespecified Bonferroni-four family over ten optimization seeds. The final cross-pathway contrast is descriptive with a nominal 95% interval. Seed inference conditions on these fixed series; families are not pooled across populations.

Comparison	Mean	Paired-seed interval
True input code − Constant input code	$+0.00699$	$[-0.00146, +0.01544]$
True input code − Permuted input code	-0.01674	$[-0.02202, -0.01147]$
True input code − Store-rate calibration	-0.00550	$[-0.01606, +0.00505]$
True input code − Ridge store-rate calibration	-0.00550	$[-0.01605, +0.00506]$
True head label − Permuted head label	-0.01337	$[-0.02288, -0.00386]$
True head label − Erased covariate	$+0.00642$	$[-0.00339, +0.01623]$
True head label − Store-rate calibration	$+0.00977$	$[-0.00103, +0.02058]$
True head label − Ridge store-rate calibration	$+0.00978$	$[-0.00103, +0.02058]$
True input code − True head label (descriptive)	-0.01527	$[-0.02693, -0.00362]$

Table 52: All later series (1750). These are marginal descriptive 95% bootstrap intervals from 20,000 resamples of series after averaging the ten seeds. They are separate from the paired-seed inference and do not account for shared store or calendar dependence.

Comparison	Series mean and interval	Series median and interval
True input code − Constant input code	$+0.01466$ $[+0.01057, +0.01876]$	$+0.00328$ $[+0.00070, +0.00614]$
True input code − Permuted input code	-0.00580 $[-0.01035, -0.00129]$	-0.00293 $[-0.00684, +0.00013]$
True input code − Store-rate calibration	$+0.00137$ $[-0.00259, +0.00540]$	$+0.00047$ $[-0.00232, +0.00406]$
True input code − Ridge store-rate calibration	$+0.00137$ $[-0.00259, +0.00540]$	$+0.00047$ $[-0.00233, +0.00407]$
True head label − Permuted head label	-0.00452 $[-0.00911, +0.00007]$	-0.00196 $[-0.00490, +0.00152]$
True head label − Erased covariate	$+0.01476$ $[+0.01074, +0.01899]$	$+0.00677$ $[+0.00384, +0.00947]$

Continued on next page

Comparison	Series mean and interval	Series median and interval
True head label − Store-rate calibration	$+0.01804$ $[+0.01411, +0.02217]$	$+0.01129$ $[+0.00828, +0.01427]$
True head label − Ridge store-rate calibration	$+0.01804$ $[+0.01411, +0.02218]$	$+0.01130$ $[+0.00829, +0.01427]$
True input code − True head label (descriptive)	-0.01667 $[-0.02068, -0.01268]$	-0.01347 $[-0.01677, -0.01038]$

Table 53: Earlier held-out series (350). These are marginal descriptive 95% bootstrap intervals from 20,000 resamples of series after averaging the ten seeds. They are separate from the paired-seed inference and do not account for shared store or calendar dependence.

Comparison	Series mean and interval	Series median and interval
True input code − Constant input code	$+0.01331$ $[+0.00509, +0.02334]$	$+0.00312$ $[-0.00236, +0.00882]$
True input code − Permuted input code	-0.00486 $[-0.01192, +0.00233]$	-0.00293 $[-0.00872, +0.00287]$
True input code − Store-rate calibration	$+0.00983$ $[+0.00238, +0.01760]$	$+0.00255$ $[-0.00358, +0.00756]$
True input code − Ridge store-rate calibration	$+0.00983$ $[+0.00238, +0.01760]$	$+0.00255$ $[-0.00357, +0.00757]$
True head label − Permuted head label	-0.00449 $[-0.01104, +0.00188]$	$+0.00036$ $[-0.00783, +0.00689]$
True head label − Erased covariate	$+0.01258$ $[+0.00702, +0.01822]$	$+0.00930$ $[+0.00243, +0.01477]$
True head label − Store-rate calibration	$+0.01475$ $[+0.00917, +0.02041]$	$+0.01028$ $[+0.00497, +0.01676]$
True head label − Ridge store-rate calibration	$+0.01475$ $[+0.00918, +0.02041]$	$+0.01031$ $[+0.00499, +0.01677]$
True input code − True head label (descriptive)	-0.00491 $[-0.01266, +0.00288]$	-0.00357 $[-0.00882, +0.00154]$

Table 54: Later training-unseen series (350). These are marginal descriptive 95% bootstrap intervals from 20,000 resamples of series after averaging the ten seeds. They are separate from the paired-seed inference and do not account for shared store or calendar dependence.

Comparison	Series mean and interval	Series median and interval
True input code − Constant input code	$+0.00699$ $[-0.00331, +0.01665]$	$+0.00111$ $[-0.00465, +0.00705]$

Continued on next page

Comparison	Series mean and interval	Series median and interval
True input code − Permuted input code	-0.01674 $[-0.02858, -0.00579]$	-0.01039 $[-0.01704, -0.00052]$
True input code − Store-rate calibration	-0.00550 $[-0.01359, +0.00235]$	-0.00446 $[-0.01347, +0.00263]$
True input code − Ridge store-rate calibration	-0.00550 $[-0.01359, +0.00235]$	-0.00447 $[-0.01348, +0.00266]$
True head label − Permuted head label	-0.01337 $[-0.02361, -0.00374]$	-0.00660 $[-0.01460, +0.00020]$
True head label − Erased covariate	$+0.00642$ $[-0.00142, +0.01419]$	$+0.00402$ $[+0.00062, +0.00871]$
True head label − Store-rate calibration	$+0.00977$ $[+0.00220, +0.01714]$	$+0.00870$ $[+0.00413, +0.01299]$
True head label − Ridge store-rate calibration	$+0.00978$ $[+0.00220, +0.01714]$	$+0.00868$ $[+0.00411, +0.01299]$
True input code − True head label (descriptive)	-0.01527 $[-0.02318, -0.00734]$	-0.01394 $[-0.02139, -0.00453]$

G.2 An equidistant shared-code control

The preceding constant input code was the codeword assigned to class zero. It was therefore at distance zero from that class and at a positive distance from the other nine classes. To check whether this asymmetry alone explains the adverse result, we added a codeword that is not assigned to any class. For $z = 0, \ldots, 10$, let $c_z = \sqrt{37/2}(e_{2z} - e_{2z+1})$ in the 37 observed covariate positions; c_{10} is reserved for the shared-code control. All eleven codewords have zero coordinate mean and squared norm 37, and c_{10} is at squared distance 74 from every class code. The other eleven padded input positions and their masks remain unchanged.

The identity transform and three fixed products of two Householder reflections act on the same 37 positions for every code arm. Each transform fixes the all-ones direction, preserves all codeword distances, and differs from the identity by rank at most two. We refit true, fixed-permuted, old class-zero shared, and reserved shared input-code heads under each transform, with five new paired optimization seeds. Five additional heads receive no input code; each is followed by two training-only negative-binomial recalibrations, one with fourteen daily mean/dispersion offsets and one with thirty-two global and group offsets. Thus all 80 code-head fits, five no-code fits and ten ordinary recalibrations are included. The same frozen encoder, later population, seven-day path loss, head width and 440-update schedule are used throughout. The ordinary recalibrations are not equal-capacity nonlinear heads.

Table 55: New five-seed development comparison on the 1,750-series later population. Entries are mean seven-day negative-log-likelihood differences, true input code minus the named control, in nats per path; positive values favor the control. These six sets of contrasts are descriptive and are not jointly multiplicity-corrected.

Transform	Permuted	Old shared	Reserved	No code	Global cal.	Group cal.
Identity	-0.00158	$+0.01593$	$+0.01259$	-0.00741	$+0.00242$	$+0.00296$
First	-0.00097	$+0.01835$	$+0.01207$	-0.00792	$+0.00191$	$+0.00245$
Second	-0.00137	$+0.02074$	$+0.01648$	-0.00473	$+0.00510$	$+0.00564$
Third	-0.00326	$+0.02054$	$+0.01732$	-0.00427	$+0.00555$	$+0.00609$

For the identity transform, the new five-seed true-minus-permuted interval is $[-0.00824, +0.00508]$ around a mean of -0.00158; it does not reproduce the resolved contrast from the separate historical ten-seed experiment. The true-minus-reserved-shared mean is $+0.01259$, with an unadjusted five-seed interval $[+0.00582, +0.01936]$. We do not pool the two seed families after seeing their results. The three prespecified differences between each rotated and unrotated true-minus-reserved-shared contrast are -0.00052 $[-0.00304, +0.00200]$, $+0.00388$ $[-0.00356, +0.01133]$, and $+0.00473$ $[-0.00145, +0.01092]$, respectively; these intervals use Bonferroni correction across the three tests. All cross zero, and all absolute means fall below the predeclared 0.005-nat threshold. The tested low-rank orientation-dependence gate therefore fails. Both shared codes have lower mean loss than the true code under all four transforms in these five-seed fits. These ordinary contrasts are descriptive, not four multiplicity-adjusted wins; the old code's asymmetry does not by itself account for the observed mean-loss ordering. This is an added experiment on an already-opened development cohort and one encoder, not a test of arbitrary rotations, independent-source transfer or a training-free identity selector.

H CLAIMS AND THEIR EVIDENTIAL BOUNDARIES

The status of a broad claim can differ from the status of one of its comparisons. Here "supported" is always restricted to the stated population, law, procedure and uncertainty unit. "Contradicted" rejects the indicated universal claim, not every possible identity-aware forecaster. "Unresolved" means that a measured interval includes zero, not equivalence. "Not identified" marks an attribution the design cannot establish; "Not evaluated" marks a claim without the required test.

Table 56: Completed experiment families and the precise claims they bear on. Counts are experiment cells or comparisons, not independent datasets.

Claim	Evidence, denominator and boundary	Status
Specified input paths can erase a static label	A sufficient structural condition proves exact erasure. On 70 public series, five released interfaces were probed; three were bit-invariant and a fourth response was traced to unobserved entries in a separate 24-call batched intervention covering the same 70 series. These finite checks do not characterize all inputs or interface versions.	Supported
The fitted full head can beat ordinary calibration on the original fundraising cohort	On 3,087 development-seen campaigns, all four specified group and ridge mean-plus-dispersion contrasts favor the full head after adjustment across the 36-comparison family. The additional controls comprise 20 saved base cells and 60 training-only calibrators. Their sequential finite search is not an oracle joint optimum.	Supported
Saved mean and dispersion swaps isolate a unique identity mechanism	The complete saved-law family contains 156 registered entries: 36 later fitted contrasts, 48 component estimates and 72 earlier-to-later changes; 24 are structurally inapplicable. The component sum is an exact Shapley accounting identity, not causal attribution, and it does not exclude a misspecified common base.	Not identified
The private gain transfers uniformly to the public primary label	The same-law public family completed 120 neural fits, 120 training-only calibrations and 48 comparisons. Among eight adjusted primary store-label contrasts, none supports the identity head, two favor group-rate calibration and six include zero. Department is a separately fitted secondary label.	Contradicted

Continued on next page

Claim	Evidence, denominator and boundary	Status
A true input code is best among matched same-checkpoint routes	One public subset retains 60 neural fits, 30 training-only calibrations and 27 complete comparisons. Of its four primary input comparisons, one favors true over permuted, one favors the shared label-free code over true, and two are unresolved. The input-code arms inject codes after normalization on one fixed checkpoint; they do not repair the native input interface.	Contradicted
The benefit holds for training-unseen entities	The 607 training-unseen fundraising campaigns include 315 already in validation; all corresponding adjusted ordinary-calibration intervals include zero.	Unresolved
The benefit transfers to an independent source	The public source was previously used in development, and no new source has passed the full eligibility and prior-use checks.	Not evaluated
A rule can decide whether to pay for an identity-conditioned fit before new outcomes arrive	The current comparisons evaluate already fitted candidates. They do not validate a pre-fit, cost-aware three-way choice among ignoring a label, ordinary calibration and conditional fitting.	Not evaluated

I EARLIER PUBLIC IDENTITY COMPARISON WITH A DIFFERENT COUNT LAW

This experiment asks whether a prediction head benefits from true store identity rather than a wrong or constant identity at exactly the same width. The preregistered primary store test does not reproduce the fundraising advantage on either frozen representation. Department identity, specified as a secondary test, helps on Chronos-2 but remains uncertain across fitted seeds on Moirai. An interface can discard a label without that label necessarily improving a fitted head.

Data and prediction target. The M5 retail panel (Makridakis et al., 2022) contains 30,490 daily item–store series over 1,941 days, with seven departments and ten stores. The training roster has 24,325 series at 32 origins, or 778,400 training rows. Development evaluation uses the other 6,165 series at those same 32 origins, or 197,280 rows. Each input contains thirty history days and each target seven daily counts. Training origins are spaced sixty days apart, with their last targets ending at day 1897 in the one-based data columns.

The later period uses every series at 38 daily origins, or 1,158,620 rows. Its seven-day targets span 9 April through 22 May 2016. The primary population is all 30,490 series: their identity assignments are defined in the existing training or development rosters. The secondary population is the 6,165 series never used to fit a head. At each origin, the score is the negative log mass of the seven-day integer-count path. Scores are averaged over the 38 origins within a series, then equally over series. No amount, revision or deadline term enters this retail score.

These targets follow the training period, but M5 is not an untouched dataset at project level. Earlier studies used this panel for development, and other project lines scored some of these later dates with other models. This is a prospectively specified comparison on a previously used public panel, not independent-source confirmation.

Matched prediction heads. The frozen representations are Chronos-2 and Moirai-1.1-R-large, computed from the same thirty history days and accompanied by location and scale coordinates. For a given label, the three arms share the same constructor, hidden width of 256, and seed-specific initial parameters; only the integer code supplied to the eight-dimensional embedding changes. Store heads have 427,989 trainable parameters, and department heads have 427,965. Equality holds within each label comparison, not between label vocabularies. A constant code retains the same embedding table and prediction head; it does not remove parameters.

This retail head uses a hurdle, history-dependent daily count law with a fitted tail, with its amount, revision and survival losses masked. It is not the negative-binomial head used in the primary

fundraising identity comparison. The three identity arms are controlled within each dataset and label; this study does not hold the head family fixed across fundraising and retail.

The true arm receives its series' store or department. The wrong-code arm uses one fixed permutation of labels among series, separately within the training and development-held-out rosters. Every series keeps its assigned code at every origin. This preserves the label multiset on all series, both rosters and all scored rows. The constant arm receives one code for every series. Store is the primary label, chosen as the retail analogue of the fundraising platform before fitting; department is secondary. Neither test estimates the incremental value of department conditional on an already store-aware head.

All fits use AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 2,048, gradient-norm clipping at one, and twenty fixed epochs. The tail threshold is the 95th percentile of positive training counts. Only the final weights are scored; neither early stopping nor validation selection is used. The ten seeds are 1729, 3253, 5003, 7919, 104729, 20260920, 20260921, 20260922, 20260923 and 20260924. On each representation, the primary store arms use one wrong-code permutation and ten seeds. Nine additional store permutations form a secondary sensitivity analysis. The department test uses the three arms and the same ten seeds, with only its first permutation. All 300 planned fits completed: 240 store fits and 60 department fits.

Declared decision and uncertainty. The contrasts are true minus wrong code, true minus constant code, and wrong minus constant code. Negative values favor the first arm. For the primary decision, each seed contributes one paired difference of equal-series means. Student intervals use nine degrees of freedom and a Bonferroni adjustment over the three contrasts on two representations. The fundraising result is called replicated only if true identity beats both controls under that rule on both representations. The department test, held-out population, medians and additional permutations cannot change this decision.

A complementary analysis averages the ten seeds' scores within each series, then resamples series 20,000 times. The tables give these mean intervals beside the seed intervals, with the same six-comparison adjustment. The series intervals condition on the fitted seeds; they do not capture both seed and series variation jointly. Different series can share an item, store or date, so this resampling is not a multiway dependence analysis. Per-series medians, all per-seed differences, and nominal intervals are retained in the released score summaries.

Primary store result. Table 57 gives all six declared primary contrasts. On Chronos-2, true minus wrong code is $+0.00101$ nats with adjusted seed interval $[-0.00116, +0.00318]$. True minus constant is $+0.00445$ $[+0.00154, +0.00736]$: the true code is worse in all ten seeds. On Moirai, the corresponding differences are $+0.00165$ $[-0.01561, +0.01891]$ and -0.00910 $[-0.02474, +0.00654]$. The registered replication criterion therefore fails on both representations, despite live label pathways and matched widths.

The two uncertainty units also give different answers. The series intervals place true minus wrong code above zero on both representations, whereas both seed intervals include zero. On Moirai, true minus constant favors the true code under series resampling but remains uncertain across fitted seeds. Reporting only the series intervals would conceal this distinction.

Table 57: Primary store-identity test on all 30,490 later-period series. Entries are paired negative-log-mass differences in nats per seven-day path; negative favors the first arm. Both interval columns are adjusted over six contrasts. Seed intervals use ten paired fits; series intervals condition on their average. Shuffled denotes the fixed multiset-matched wrong code.

Representation	Contrast	Mean	Seed interval	Series interval
Chronos-2	True − shuffled	+0.00101	[-0.00116, +0.00318]	[+0.00022, +0.00176]
Chronos-2	True − constant	+0.00445	[+0.00154, +0.00736]	[+0.00373, +0.00518]
Chronos-2	Shuffled − constant	+0.00344	[-0.00015, +0.00702]	[+0.00308, +0.00382]
Moirai	True − shuffled	+0.00165	[-0.01561, +0.01891]	[+0.00107, +0.00221]
Moirai	True − constant	-0.00910	[-0.02474, +0.00654]	[-0.00981, -0.00842]
Moirai	Shuffled − constant	-0.01075	[-0.02944, +0.00795]	[-0.01125, -0.01026]

Secondary department result. Table 58 retains the same three controls and uncertainty units for department. True department improves on both controls with Chronos-2, with all ten paired seeds favorable in each contrast. Moirai's mean differences favor true department, but neither adjusted seed interval excludes zero. Favorable series-unit mean intervals on both representations do not overturn the seed uncertainty or rescue the primary store result.

Table 58: Secondary department-identity test on all 30,490 later-period series. Definitions match Table 57; the department contrasts form their own six-comparison family. This is not department added to a store-conditioned head.

Representation	Contrast	Mean	Seed interval	Series interval
Chronos-2	True − shuffled	-0.01107	[-0.01576, -0.00639]	[-0.01269, -0.00950]
Chronos-2	True − constant	-0.00770	[-0.01166, -0.00375]	[-0.00929, -0.00614]
Chronos-2	Shuffled − constant	+0.00337	[-0.00209, +0.00883]	[+0.00303, +0.00371]
Moirai	True − shuffled	-0.00269	[-0.04023, +0.03486]	[-0.00430, -0.00113]
Moirai	True − constant	-0.00948	[-0.05168, +0.03272]	[-0.01116, -0.00787]
Moirai	Shuffled − constant	-0.00679	[-0.03790, +0.02431]	[-0.00776, -0.00581]

Series excluded from fitting. Tables 59 and 60 repeat the declared analyses on the 6,165 series excluded from all head fitting. Store identity still does not satisfy the primary criterion; department still favors true labels across seeds only on Chronos-2. These series share the same public source, stores and calendar with the fitting population, so their separation is not independence from the data-generating setting.

Table 59: Store identity on the secondary population of 6,165 series excluded from head fitting. Intervals, signs and inferential units follow Table 57. This population cannot replace the preregistered primary population.

Representation	Contrast	Mean	Seed interval	Series interval
Chronos-2	True − shuffled	+0.00155	[-0.00066, +0.00377]	[-0.00030, +0.00340]
Chronos-2	True − constant	+0.00497	[+0.00176, +0.00817]	[+0.00333, +0.00664]
Chronos-2	Shuffled − constant	+0.00342	[-0.00036, +0.00719]	[+0.00252, +0.00455]
Moirai	True − shuffled	+0.00194	[-0.01491, +0.01878]	[+0.00077, +0.00307]
Moirai	True − constant	-0.00812	[-0.02395, +0.00771]	[-0.00959, -0.00665]
Moirai	Shuffled − constant	-0.01006	[-0.02862, +0.00850]	[-0.01116, -0.00892]

Table 60: Department identity on the secondary population of 6,165 series excluded from head fitting. Definitions follow Table 58. Department and the held-out population are both secondary analyses.

Representation	Contrast	Mean	Seed interval	Series interval
Chronos-2	True − shuffled	-0.01025	[-0.01490, -0.00560]	[-0.01383, -0.00672]
Chronos-2	True − constant	-0.00666	[-0.01075, -0.00256]	[-0.01019, -0.00323]
Chronos-2	Shuffled − constant	+0.00359	[-0.00171, +0.00888]	[+0.00281, +0.00441]
Moirai	True − shuffled	-0.00118	[-0.03725, +0.03489]	[-0.00481, +0.00223]
Moirai	True − constant	-0.00794	[-0.04932, +0.03344]	[-0.01164, -0.00436]
Moirai	Shuffled − constant	-0.00676	[-0.03788, +0.02437]	[-0.00893, -0.00461]

Sensitivity to the wrong-code assignment. Across ten seeds and ten store permutations, the mean true-minus-wrong difference is $+0.00126$ $[-0.00026, +0.00277]$ for Chronos-2 and -0.00087 $[-0.00931, +0.01084]$ for Moirai. These are the declared 10,000-draw two-way seed-by-permutation 95 percent intervals, not the six-comparison primary intervals. Respectively 33 and 59 of the 100 paired differences favor the true code. Every permutation mean is positive on Chronos-2; eight of ten are negative on Moirai. These secondary summaries remain consistent with uncertainty in the primary true-versus-wrong contrast.

Earlier development period. Table 61 reports the fixed heads on the earlier 6,165-series development roster. These rows were already opened by earlier work and serve as context, not an independent test. The later-period decision is not chosen from these scores.

Table 61: Development negative-log-mass means on 6,165 series, averaged equally over the ten fixed fits. Store is the primary label and department secondary. Lower is better. These means are descriptive, not the later-period replication criterion.

Representation	Label	True	Shuffled	Constant
Chronos-2	Store	6.22694	6.22909	6.22649
Moirai	Store	6.25157	6.25293	6.25969
Chronos-2	Department	6.21369	6.22848	6.22625
Moirai	Department	6.24534	6.24704	6.26158

Implementation and reproducibility checks. All 300 fits used a single CPU thread. Before analysis, a code-path control compared the copied training pipeline with the unchanged original pipeline on the same CPU node for three reference seeds, both labels and both representations. The twelve pairs produced identical saved development-score bytes and matching parameter counts. This checks the copied implementation, not bitwise reproduction of older V100 fits. CPU versus V100 differences were reported separately; floating-point changes across many optimizer steps can change fitted predictions.

The label pathway is declared inert only if the largest arm difference is below 10^{-9} nats at every seed. Neither label is inert on either representation. Input identities, parameter equality, initialization, permuted-label histograms and constant-label shape also pass their declared checks. An independent implementation reconstructed both labels and both populations from the saved per-origin scores, with maximum numerical discrepancy below 2×10^{-13} nats; this verifies score arithmetic, not a new fit or dataset.

The anonymous artifact provides M5-derived per-series score summaries and a relocated replay of 362 reported seed, median, development and permutation quantities. It does not reproduce all 300 training runs, per-origin scoring or the series-bootstrap intervals from those lightweight files. The fuller internal reconstruction covers the last two arithmetic steps. Raw retail data must be obtained under the M5 competition's access terms; public availability does not automatically permit unrestricted redistribution of the original records or pretrained model weights. A fresh-environment, independent end-to-end training reproduction remains unverified.

Static identity in native covariate interfaces. The representations above use history alone. Chronos-2 accepts covariates, Moirai accepts dynamic covariates, and TimesFM-3, released 2026-08-31, supports multivariate past and known-future covariate paths. Interface sensitivity is therefore a separate question from the forecast accuracy of conditioning placements.

An initial probe holds histories fixed and changes only department content on 48 cached M5 windows with 30 history days and a 7-day horizon. Chronos-2 receives department through three paths: fixed-count one-hot variates for `embed`, a constant numeric past covariate for `predict`, or a constant categorical past covariate. Replacing department by another department or a permuted assignment gives an exact zero maximum change in every tested embedding or forecast coordinate across the 48 rows. TimesFM-3 also gives exact equality between a constant department-code covariate and a constant zero, despite forecasts of order one.

Controls establish sensitivity to other input changes: adding a covariate changes TimesFM-3 forecasts by 4.2; a time-varying covariate changes Chronos-2 forecasts by 10.8 to 16.1 and TimesFM-3 by 3.9. Renaming a time-varying categorical pattern with seven labels changes nothing, whereas changing the pattern does. Centering removes constant numeric content when no identity-bearing statistic bypasses it (Proposition 1). The categorical result concerns the tested single-target path, not every categorical encoder or multivariate path. All findings are specific to the library versions and input treatments. Head-side labels, covariate adapters, retained normalization statistics and dynamic codes are possible alternatives, not ranked by this probe. The separate TimesFM external-regression interface is outside the native TimesFM-3 comparison.

These 48 windows contain two distinct series from one store and one department, not 48 independent series. Table 1 reports a stratified replication on 70 distinct series across ten stores and seven departments that retains the static-input equality. Distinct series can share products and dates; neither probe provides statistical inference about forecast accuracy.

Interpretation. The interface probe establishes sensitivity or erasure under specified input paths. The fitted identity comparisons establish predictive value under a specified architecture, optimization budget and population. Those are different statements: the primary public store test fails, while a secondary department test helps on one representation. Neither a live input pathway nor static-label erasure alone determines which labels a finite fitted model will use profitably.

J CHARACTERIZATION OF THE FUNDRAISING RECORDS

The development records contain heterogeneous, mostly inactive count paths. All measurements in this section use opened development data. The broad census contains 21,625 campaigns and 4,997,177 daily records. Campaign is the statistical unit; gaps, duplicate dates, noninteger cumulative counts, and downward revisions break a segment rather than becoming zero increments. The fitted model experiment uses a stricter chronological origin roster, reported separately.

Inactivity is pervasive. Equal-campaign zero fractions are 86.58% for counts and 31.36% for amounts; mean campaign-median zero runs are 15.15 and 11.06 days. Seven-, 14-, and 30-day count Fano factors are 33.08, 45.57, and 67.08. The next-seven-day event risk is 4.381 $[4.132, 4.654]$ times higher after an event-day origin, as the geometric mean of within-campaign risk ratios over 3,506 eligible campaigns with a within-campaign continuity correction and a 95% campaign-bootstrap interval, yet positive-event gaps have $\mathrm{CV}^2 = 0.479$. Long inactive periods, calendar and lifecycle position, platform effects, and campaign heterogeneity remain alternatives to self-excitation. These observations support testing a history-dependent daily law, not a causal or continuous-time Hawkes mechanism.

The observation process is also structured. By development end, 33.51% of campaigns have no observed deadline, 35.89% have a missing or post-development deadline, and 16.84% show more than one deadline value. These measurements require as-of inputs, observation masks, and explicit treatment of revised or incomplete records. They do not identify why a deadline is missing or justify independent-censoring or inverse-probability claims. The source reveals daily interval increments but no within-day event times.

K WHAT FITTED LABEL CONTRASTS CAN IDENTIFY

An observed loss difference compares specified predictors, not information alone. This section gives elementary finite-design witnesses, drawing on established cautions about refitted variable importance (Hooker et al., 2021; Mentch & Zhou, 2022) and classical risk estimation (Efron, 2004). The examples neither estimate the mechanism of the fundraising or retail results nor provide a new forecasting method.

The categorical partition. Let each row of Z and Z' contain exactly one unit entry; unused columns are allowed. Then $ZZ^\top = Z'Z'^\top$ if and only if a bijection between their used categories maps every row's original label to its new label. Indeed, the (i, j) Gram entry is the indicator that rows i, j share a category. Equality makes the row-defined map well-defined, injective and surjective; the converse follows immediately. The averaging operator $Z(Z^\top Z)^+ Z^\top$ has the same partition as its nonzero support. Matching a label histogram does not match this operator.

This is a statement about the observed rows and raw one-hot codes, not arbitrary learned representations. A train-only kernel does not fix future group assignments. Apply the same renaming across training and testing; algorithmic invariance additionally requires an equivariant model class, regularizer, initialization and update rule. A frozen nonlinear encoder need not preserve predictions under an orthogonal rotation of its input codes.

Exact finite-training risk. There are M equally weighted input contexts. Each supplies r independent training responses with mean m_i and Gaussian variance σ^2; all errors are independent. A fresh test response is drawn independently at a uniformly sampled context. A fixed linear smoother H, chosen without outcomes, predicts from the vector $\bar{Y}$ of training means. Its expected test squared error is

$$R_H = \sigma^2 + \frac{\|(H - I)m\|^2}{M} + \frac{\sigma^2 \|H\|_F^2}{rM}. \tag{9}$$

To prove this, substitute $\bar{Y} = m + \bar{\epsilon}$ into the squared prediction error. The cross term has mean zero and $\mathrm{Cov}(\bar{\epsilon}) = \sigma^2 I/r$. For a Gaussian predictive law with the known variance, NLL is $\frac{1}{2}\log(2\pi\sigma^2) + R_H/(2\sigma^2)$. For least squares, H is the orthogonal projection and $\|H\|_F^2 = \mathrm{rank}(H)$; equal raw widths need not match this variance. Ridge smoothers instead require the squared Frobenius norm.

Four exact witnesses. Use $M = 4$, $\sigma^2 = 1$, and a head with three columns $[1, B, C]$. The code C is true, permuted or constant; all arms use least squares with the Moore–Penrose inverse. Supplied codes have the same multiset within each true/permuted pair. All differences below are expected Gaussian NLL, averaged over fresh outcomes and training draws.

1. **No information, yet true beats permuted.** Let $B = (-1, -1, 1, 1)$, true $C = (0, 0, 1, 1)$, permuted $C' = (0, 1, 0, 1)$ and $m = 0$. The true and constant designs have rank two; the permuted design has rank three. True minus permuted is $-1/(8r)$, although Y is independent of both labels and history.

2. **Decodable identity can help a restricted model.** Let $B = (-3, -1, 1, 3)$, $C = (1, 0, 0, 1)$, $C' = (1, 1, 0, 0)$ and $m = C - 1/2$. Both coded designs have rank three, but only the true design represents m. True minus permuted is $-1/8$ although $I(Y; C \mid B) = 0$. The ordinary history-only head $[1, B, B^2]$ has the same prediction span as the true head since $m = (B^2 - 5)/8$.

3. **Information need not pay its fitting cost.** Let $B = (-1, 1, -1, 1)$, $C = (-1, -1, 1, 1)$, $C' = B$ and $m = C/4$. True minus constant is $1/(8r) - 1/32$: positive at $r = 1$, zero at $r = 4$ and negative at $r = 16$. Conditional identity information is positive in all three cases; only the training amount changes.

4. **An additive head can miss the available information.** Keep the previous inputs but set $m = BC$ elementwise. True minus constant is $1/(8r) > 0$, since both additive heads miss the interaction. The four-column control $[1, B, C, BC]$ represents it, at a larger model-class and estimation budget. This control is not parameter matched.

A classical correction and its limits. Writing $\tau^2 = \sigma^2/r$, the fixed-smoother risk estimate

$$\widehat{R}_H = \sigma^2 + \frac{\|(H-I)\bar{Y}\|^2 + 2\tau^2 \operatorname{tr}(H) - M\tau^2}{M} \tag{10}$$

is unbiased for equation 9; expand the square and use $\|H-I\|_F^2 = \|H\|_F^2 - 2\operatorname{tr}(H) + M$. This is classical covariance-penalty estimation, not a new diagnostic method. For two independent mean vectors u, v, the product $((H-I)u)^\top((H-I)v)/M$ also estimates the approximation term without bias, but requires twice the observations. Neither correction turns an approximation advantage into conditional information, as the second witness shows. Unknown noise, outcome-selected smoothers, temporal dependence and changing test populations require additional arguments. In particular, the number of evaluation campaigns is not the independent training sample size.

Complete controlled check. All four laws were fixed before generating outcomes, each with $r \in \{1, 4, 16\}$ and 512 independent repetitions. The three code arms, ordinary quadratic-history control and larger interaction control were fitted on every training set: 6,144 training sets and 30,720 fitted predictions. A second independent response stream served only the cross-replica calculation. Fresh-test NLL was integrated analytically. Independent reconstruction regenerated every draw and all 48 paired contrast summaries; maximum numerical discrepancy was 3.11×10^{-15}. This is controlled implementation evidence, not a real-source test. All predictions, errors and descriptive intervals are released; no condition or fit failed.

Table 62: Fixed-law comparisons at one training response per context. NLL differences in nats per scalar outcome; lower favors the first arm. Monte Carlo means use 512 independent training sets. Parentheses contain descriptive 95% intervals, not simultaneous significance claims.

Condition and contrast	Exact expectation	Monte Carlo mean (interval)
No information: true–permuted	-0.12500	-0.11953 $(-0.13385, -0.10521)$
Decodable: true–permuted	-0.12500	-0.11503 $(-0.13792, -0.09213)$
Weak information: true–constant	$+0.09375$	$+0.08845$ $(+0.07412, +0.10277)$
Interaction: true–constant	$+0.12500$	$+0.12671$ $(+0.11161, +0.14181)$

The ordinary history-only quadratic head matches the true head in the decodable example, so no new information is needed for that gain. The classical risk correction changes the apparent no-information contrast from $+0.11953$ to -0.13047, consistent with its -0.125 expected test contrast, but that corrected advantage still does not establish information. These checks support the restricted warning; they do not establish an improved method beyond classical risk estimation.

For any three fully specified fitted procedures, true minus constant equals true minus permuted plus permuted minus constant. That arithmetic identity does not identify either summand as pure signal or pure cost. The examples show why raw width and a matched label multiset are insufficient.

Vanishing code amplitude. A fixed continuous forward map converges when the code amplitude tends to zero. A refitted unconstrained linear coefficient can instead grow as the inverse amplitude: for every nonzero a, $\{waC : w\} = \{vC : v\}$. With ridge penalty λw^2, this reparameterization gives $\lambda v^2/a^2$. Forward continuity, achievable risk and regularized fitting therefore make different claims.

L A FIRST-FORTNIGHT RECORDED-EVENT DEVELOPMENT STUDY

Does conditioning a joint activity law on campaign metadata improve on ordinary calibration and shared dependence? We answer this narrower question on the Multimodal Dynamic Dataset for Donation-based Crowdfunding Campaigns (MDCC; Xu et al., 2023), a previously inspected public source, using a new outcome-independent roster and a fixed first-fortnight target. This is a development experiment with classical conditional exponential-family heads, not a new architecture or an untouched-source confirmation.

Population, target and collection assumptions. The raw release has 14,961 rows and 14,859 distinct campaign IDs. There are 99 repeated IDs, comprising 96 pairs and three triples, not 102 repeated campaigns. We exclude all 99 before interpreting event arrays; no repeated donation lists are merged. The remaining 14,760 singletons are assigned by launch date, without filtering on donation count, amount or future activity. Training launches fall in October 16–24, 2022, validation in November 1–4, and evaluation in November 12–20. These give 4,037, 1,923 and 3,022 campaigns; 5,778 other singletons are time-purged. There is one forecast per campaign. The gaps put preceding labels before subsequent forecast origins in the recorded, timezone-unspecified coordinate. Cross-country UTC order is not independently verified.

The authors describe launch-time category and country and at least one month of tracking per campaign. We use that documentation, not a per-row crawl log, to support the first fourteen days. Following the historical data interpretation, numeric event times are treated as elapsed seconds since launch and binned into fixed 24-hour intervals. The first seven counts are history; the following seven are the target. Counts refer to archived records, not unique donors. A zero means no recorded event in that bin under the collection assumption, not proof of no donation or no collection outage. An empty valid list can produce zeros; malformed or missing lists cannot. All selected rows pass list, length and finite-value checks. The original time-unit conversion and row-level collection coverage remain unverified.

One frozen representation and seven heads. All arms use the same Moirai-2.0-R-small checkpoint with exactly seven observed history entries. Padding is unobserved; forecast placeholders never contain target values. We standardize its 384-dimensional representation on training data, apply a fixed 32-dimensional Gaussian-QR projection, append seven $\log(1 + \text{count})$ histories and native location and scale, standardize again, and add an intercept. This gives $B \in \mathbb{R}^{42}$. Metadata I consists of training-fitted category and country indicators, eighteen columns in total; an unseen level contributes zeros. Raw history is available to every head. Campaign ID is never a fitted feature.

The base product-negative-binomial head predicts seven means from B and seven constant dispersion parameters. Ordinary metadata calibration fits additive log-mean and log-dispersion shifts to that fixed base, including intercept shifts. That original family had no matched no-metadata recalibration arm, so its baseline contrast does not isolate metadata. The later controls in Appendix M address this gap without refitting the original family. The remaining heads share the base's predicted active-day count law $\pi_0(K \mid B)$ and positive-count law g_0, but change the conditional support distribution:

$$q_\eta(C \mid B, I) = \pi_0(K \mid B) q_\eta(S \mid K, B, I) g_0(C_S \mid S, B),$$
$$q_\eta(s \mid k, B, I) \propto q_0(s \mid k, B) \exp\{f_\eta(s, B, I)\}.$$

Normalization enumerates the 128 supports. Shared pairs use 21 coefficients on $s_i s_j$. History pairs make them linear in B; metadata pairs make them linear in $[B, I]$. A stronger ordinary control jointly fits history pairs and separate additive-day fields for each K, linear in $[B, I]$. Full support fits all 128 residual logits with ridge shrinkage toward the base. These parameterizations do not form a single nested model ladder. They share K and positive marks, not the conditional daily inclusion rates; on a fixed $K \geq 2$ slice, $\sum_{j \neq i} s_i s_j = (K - 1) s_i$. Thus pair coefficients can represent position effects, and a pair-model gain would not by itself isolate higher-order dependence.

Fitting, selection and precision. Each arm uses the same ridge grid $\{0.001, 0.01, 0.1\}$, one initialization, L-BFGS-B and a 300-iteration cap. Training minimizes mean full-path NLL plus the squared-coefficient penalty. Validation selects one base, which is fixed for all other arms, then one model per arm; there is no train-plus-validation refit. All 21 candidates terminate and all seven selections are saved before evaluation scoring. All are finite and report SciPy success, but only four

meet the additional gradient-infinity-norm diagnostic of 10^{-6}; one of the seven selected models does. The other seventeen stop by relative objective reduction, not the iteration cap. We retain every model and residual and make a finite-budget comparison, not a global-optimum claim.

The clipped negative-binomial objectives are not asserted convex. All fits together take 137.17 wall seconds and 506.12 CPU seconds on four CPU threads; the separate L40S history cache takes 5.55 seconds. No training saving is claimed. The archive distinguishes cache, fitting and evaluation costs.

Table 63: First-fortnight models on 3,022 evaluation campaigns. Lower NLL is better, in nats per seven-day path. Parameters exclude the common frozen base except in its own row; λ is selected on validation only. Calibration versus baseline is descriptive, not an isolated metadata effect.

Head	Parameters	λ	NLL
History-only negative binomial	301	.001	6.97896
Metadata mean/dispersion calibration	266	.001	6.87159
Shared pairs	21	.001	6.96972
History-conditioned pairs	882	.1	6.97992
Metadata-conditioned pairs	1,260	.1	6.97644
History pairs + metadata position fields	4,242	.1	6.98443
Shrunk full support	7,680	.1	6.97701

The prespecified comparison is not uniformly favorable. Table 64 gives all five primary comparisons. Two favor metadata pairs under paired campaign bootstrap intervals and Holm-adjusted one-sided tests. Ordinary mean-and-dispersion calibration is better by 0.10484 nats per path. Comparisons with shared pairs and shrunk full support remain unresolved. The required all-five condition therefore fails. The apparent advantage over the combined position control also becomes unresolved when resampling launch-date blocks. There are only nine such blocks, so that sensitivity is imprecise. Neither set of intervals is a finite-sample coverage guarantee or a cross-source uncertainty statement.

Table 64: Metadata-pair minus comparator NLL. Negative favors metadata pairs. Both intervals use 5,000 paired bootstrap draws at nominal 99% confidence; campaign intervals adjust five comparisons by Bonferroni. Block intervals resample nine launch dates and remain descriptive.

Comparator	Mean	Campaign interval	Date-block interval
Mean/dispersion calibration	$+.10484$	$[+.06484, +.14588]$	$[+.05038, +.17830]$
Shared pairs	$+.00672$	$[-.00121, +.01500]$	$[-.00216, +.01603]$
History pairs	$-.00348$	$[-.00568, -.00139]$	$[-.00615, -.00142]$
History pairs + position	$-.00799$	$[-.01435, -.00148]$	$[-.01441, +.00005]$
Shrunk full support	$-.00058$	$[-.00812, +.00685]$	$[-.01082, +.01032]$

Interpretation and reproduction. Metadata pairs improve conditional-support NLL from 1.74091 to 1.73839 relative to the history-only base, descriptively. Yet ordinary calibration changes all three components of the path law: its count, conditional-support and positive-mark NLLs are 1.78463, 1.74114 and 3.34583. It is not a matched-component test of dependence or an isolated metadata effect. The practical result is that a richer support law did not replace ordinary distributional correction in this specified budget and population. It does not show that dependence is absent, that metadata is useless, or that these heads cannot improve with different training. The original family was not refitted or repaired after its results; Appendix M reports a separate, explicitly retrospective set of ordinary controls.

An independent NumPy/SciPy implementation reconstructs all 21 saved models, training objectives, validation rows, seven evaluation score vectors and both bootstrap schemes. The maximum discrepancy in gradient infinity norms is 2.50×10^{-16}. This is floating-point saved-model verification, not a fresh training rerun. Full model, row and input hashes, all optimization statuses, source assumptions and data-access instructions are retained.

M RETROSPECTIVE CONTROLS FOR HISTORY-DEPENDENT RECALIBRATION

The preceding experiment compared metadata calibration with an unchanged history-only base. That contrast also changes fitting and distributional flexibility. We therefore add a separate retrospective comparison: how much does ordinary recalibration achieve using the same history features, and what further fitted difference follows from supplying group metadata? The original twenty-one models, their selected parameters and their failed five-comparison criterion are unchanged. This is not a new probability family, an outcome-blind confirmation or a repair of that experiment's primary test.

Fixed base, information and population. We retain all 4,037 training, 1,923 validation and 3,022 evaluation campaigns and the exact first-week-to-next-week target of Appendix L. The base is its already selected 301-parameter negative-binomial model with ridge coefficient 0.001. Its 42-dimensional input contains an intercept, the saved 32-dimensional projection of frozen Moirai-2 features, seven log-count histories and native location and scale, with the original training standardizations. We recompute neither the projection nor the base. Before any new fitting, zero residuals recover all saved validation scores within 1.43×10^{-13}, below the fixed absolute tolerance 2×10^{-8}.

Three residual designs use an intercept alone, all 42 history features, or those features plus 18 category/country indicators. The latter indicators are centered using only training campaigns. An unknown level is encoded as zero before centering and therefore becomes minus the training mean, not zero after centering. The residual parameter counts are 14, 588 and 840. Centering removes, on the training design, the constant direction supplied by the uncentered indicator blocks and its particular effect on the intercept penalty. It does not equalize capacity, feature correlations, redundant coordinates or optimization. Category and country are group metadata, not unique campaign identifiers.

Probability law and training rule. Let ℓ^0_{ij} and s^0_{ij} be the base's log mean and log size, already clipped to $[-15, 15]$ and $[-6, 10]$. For a residual design row x_i and coefficient matrices U_μ, U_r, the fitted law is

$$q_U(y_i \mid x_i) = \prod_{j=1}^{7} \mathrm{NB}\Big(y_{ij}; e^{\mathrm{clip}(\ell^0_{ij} + x_i^\top U_{\mu,j}, -15, 15)}, e^{\mathrm{clip}(s^0_{ij} + x_i^\top U_{r,j}, -6, 10)}\Big).$$

The second negative-binomial parameter is size, not variance. Each factor is normalized on all nonnegative integers. The residual is added *after* clipping the base; clipping the summed raw logits would be a different model. These controls can change count, support and positive-magnitude distributions. They are not fixed-marginal tests of residual temporal dependence.

For each design we minimize training mean full-path negative log-likelihood plus $\lambda \|U\|_F^2/2$ over the same three values $\lambda \in \{0.001, 0.01, 0.1\}$. Every coefficient, including each intercept, is penalized. All nine fits start at zero residual and use float64 L-BFGS-B, at most 300 iterations and 40 line-search steps, relative-objective tolerance 10^{-10} and gradient tolerance 10^{-6}. There are no restarts. Only after all nine finite terminal outcomes does validation select one candidate per design. Ties within 10^{-12} prefer the larger penalty. The three selections are saved before computing their new evaluation scores; there is no train-plus-validation refit. The selected penalties are 0.001 for the intercept and 0.01 for each feature-dependent design.

The source and its outcomes were already opened before this question was chosen. The existing outcome-array member is deserialized in full to obtain training and validation rows. This is not an unread-target claim; only the new evaluation predictions and comparisons await the complete selection. The earlier metadata calibrator contributes its already verified, paired score vector, not a new fit or a new from-parameter replay in this follow-up. Its uncentered indicator coding also differs from the new design.

Every finite fit, including incomplete stationarity. All nine models are finite. Eight stop by relative objective reduction; the richest design at $\lambda = 0.001$ reaches the 300-iteration limit. None passes the additional gradient-infinity-norm diagnostic of 10^{-6}. All three selected models report relative-objective success, but their gradients still exceed that threshold. These are retained finite-budget solutions, not stationary or globally optimal estimates. The objective is clipped and is not asserted convex. Tables below retain all candidates and their selection, losses, objectives, gradients, iteration counts and costs.

Table 65: All nine calibration fits. The objective is training path NLL plus the ridge penalty; a star marks the validation-selected fit. All finite outcomes are retained.

Features	λ	Train NLL	Train objective	Validation NLL
Intercept only	0.001^*	7.742744	7.742745	7.448231
Intercept only	0.01	7.742745	7.742747	7.448346
Intercept only	0.1	7.742752	7.742759	7.448846
History	0.001	7.493461	7.504834	7.309907
History	0.01^*	7.522640	7.548743	7.274892
History	0.1	7.586356	7.615760	7.298956
History + metadata	0.001	7.404441	7.421974	7.235563
History + metadata	0.01^*	7.447624	7.486415	7.196681
History + metadata	0.1	7.544146	7.588956	7.247575

Table 66: Stopping diagnostics and fitting costs for every new candidate. Times are seconds; CPU time sums four threads. Eight fits stop by relative objective decrease (Rel.), one at the 300-iteration cap (Cap). None meets the 10^{-6} gradient-norm diagnostic. Total fitting costs are 42.45 wall seconds and 169.01 CPU seconds; these do not include the inherited cache or evaluation.

Features	λ	Steps	$\|\nabla\|_\infty$	Stop	Wall	CPU
Intercept only	0.001^*	8	2.82×10^{-6}	Rel.	0.49	1.94
Intercept only	0.01	7	1.74×10^{-5}	Rel.	0.37	1.49
Intercept only	0.1	6	1.21×10^{-6}	Rel.	0.43	1.73
History	0.001	282	1.50×10^{-5}	Rel.	11.94	47.56
History	0.01^*	119	1.48×10^{-5}	Rel.	4.42	17.61
History	0.1	40	2.44×10^{-5}	Rel.	1.97	7.82
History + metadata	0.001	300	2.86×10^{-5}	Cap	14.84	59.08
History + metadata	0.01^*	123	7.28×10^{-6}	Rel.	4.62	18.42
History + metadata	0.1	40	4.48×10^{-6}	Rel.	3.36	13.36

Complete comparisons and uncertainty. Table 4 reports all three selected procedures, the frozen base and the earlier metadata calibrator on exactly the same evaluation campaigns. The latter's original advantage over the base is 0.107362309 nats per path, not a pure metadata gain. Recalibration using history features alone achieves an even lower mean score. Its contrast with the intercept also changes fitting, the penalty center and history-dependent mean and size adjustments; it does not isolate heteroskedasticity alone.

The declared new contrast adds centered metadata to the history-feature recalibrator. Both its mean and campaign interval favor that addition, but its launch-date sensitivity spans zero. The tables below report every prespecified point difference, both primary intervals and all date-specific counts and differences. Only the primary contrast receives these intervals; they are not intervals for the old metadata comparisons or an adjusted family of new tests.

Each interval uses 5,000 paired bootstrap draws, with campaign seed 2026092601 and date-block seed 2026092602. Campaign resampling gives a percentile 95% interval. Date resampling draws from the nine launch dates, retaining a pooled campaign-weighted mean rather than an equal-date estimand. Dates are the first ten characters of the recorded launch timestamp; no new UTC convention is inferred. Temporally adjacent dates need not be independent. Neither sensitivity accounts for choosing this research question after seeing the old outcomes, refitting uncertainty or all cross-date dependence. A zero-spanning interval proves neither equivalence nor absence of metadata information.

Table 67: All nine planned retrospective contrasts, in nats per path. Negative favors the first fitted procedure. Only history-plus-metadata versus history has the two descriptive interval summaries below; these are not causal or equal-capacity contrasts.

Comparison	Mean difference
History + metadata − History	-0.03254
History − Intercept only	-0.14308
History + metadata − Intercept only	-0.17562
Intercept only − Frozen base	-0.00037
History − Frozen base	-0.14344
History + metadata − Frozen base	-0.17599
Earlier metadata − History	$+0.03608$
Earlier metadata − History + metadata	$+0.06862$
Earlier metadata − Frozen base	-0.10736

Table 68: History-plus-metadata minus history recalibration. The two percentile 95% intervals use 5,000 paired draws each. Nine launch dates provide a descriptive dependence sensitivity, not a finite-sample coverage or transfer guarantee.

Mean	Campaign interval	Launch-date interval
-0.03254	$[-0.05622, -0.00593]$	$[-0.06183, +0.00784]$

Table 69: Launch-date counts and mean history-plus-metadata minus history NLL. Dates use the recorded naive calendar, not a verified UTC convention. All nine dates are included.

Launch date	Campaigns	Mean difference
2022-11-12	344	-0.05979
2022-11-13	387	-0.09035
2022-11-14	506	-0.03801
2022-11-15	445	-0.04054
2022-11-16	414	-0.07899
2022-11-17	424	-0.00848
2022-11-18	303	$+0.09935$
2022-11-19	180	$+0.01002$
2022-11-20	19	-0.06075

Interpretation and reconstruction. The useful conclusion is that an ordinary metadata-free recalibrator changes what the old metadata-versus-base contrast can establish. The additional metadata comparison has a favorable mean but unresolved date sensitivity, different capacities and finite optimization. No practical decision loss, cross-source stability, causal information effect or new methodology is established. The original source and collection assumptions still apply.

A separate NumPy/SciPy implementation reconstructs feature designs from the saved transforms, all nine training and validation loss vectors, objectives and analytic gradients; it independently reproduces validation selection, the three selected evaluation vectors and both bootstrap arrays. All 232 declared checks pass. It uses a beta-function probability expression rather than the producer's gamma-function expression and imports neither fitting module. This is saved-model arithmetic, not independent training. Old metadata scores are explicitly paired-row reuse. The anonymous supplement preserves all nine parameter arrays and diagnostics, the public protocol conversion, input identities and an executable reconstruction. Runtime totals are sums of recorded fit times, not a separately measured speedup.