THE WINDOW LEVER: OBSERVATION-WINDOW LENGTH OUTWEIGHS OBSERVED SCALE AND ARCHITECTURE EFFECTS IN CUSTOMER BEHAVIOR FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting a customer's next year: add customers, grow the model, switch architectures, or look further back? We test all four, one knob at a time. Looking back wins. Stretching the observation window from 8 to 16 weeks adds $+0.07$ to $+0.107$ log-space R^2 in all eight models at 1M, and the simplest gains most. Test-set composition is the obvious suspect, so we fix test customers and targets across windows (training cohorts still grow with the window). The gain gets bigger, not smaller, and the matched transformer gain beats all 1,000 permutation-null draws. The direction replicates on a public dataset. For sequence models, the other levers move an order of magnitude less. $10\times$ the customers shifts R^2 by at most 0.006, as does a $54\times$ larger transformer (three-seed mean), and on one seed, five sequence architectures land within 0.0052 at a matched objective at 16 weeks. All three are one-knob contrasts in small-model recipes; joint scaling is untested. We internally registered four predictions, with falsifiers, before the deciding evaluations. Three held. The one fully blind prediction failed: an activity-mixture reading called for $+0.014$ from 24 to 32 weeks; we measured -0.0004. The pooled weekly-aggregate probe plateaus near 22 to 24 weeks, but activity-specific refits keep climbing: no information ceiling is established. One limit remains: every test anchor falls in one eight-week band, so window length and calendar coverage move together. The direction is robust; the magnitude is not identified. Observation-window length belongs beside model selection as a first-class design axis: what a sequence model is shown matters more here than what it is.

1 INTRODUCTION

Forecasting a customer's future activity from their event history is a workhorse industrial task (Fader et al., 2005; Wang et al., 2019; Li et al., 2022), and the default recipe for improving it is the general sequence-modeling one: more training data, a bigger model, a newer architecture (Kaplan et al., 2020; Hoffmann et al., 2022). Whether any of those is the binding constraint is a question with prior art — history length was tested for churn (Ballings and Van den Poel, 2012), and observation-window sufficiency is under active study (Appendix M) — but it is rarely answered under controls: interventions ship sequentially, on moving populations, and improvements get attributed to whatever changed last. Our contribution is the control, not the question: this paper holds the population fixed, runs the interventions one knob at a time, and preregisters where the window bends — four falsifiable commitments (Table 1), of which the one fully blind prediction failed on its own internally registered falsifier. We fix the test cohorts, the target (364-day forward activity, $\log(1+\max(0, y))$-transformed; log-space R^2 throughout), and the evaluation, then sweep the axes the field usually credits: model family (eight), population scale (1M/5M/10M tiers), capacity, input content, training objective, and the length of the observation window itself.

The result is lopsided. Doubling the observation window from 8 to 16 weeks — which, on this end-anchored export, also shifts which calendar months the input covers (§5) — moves log-space R^2 by $+0.07$–0.107 at 1M in all eight families we train, from a single-feature linear model through boosted trees (Ke et al., 2017) to a ContiFormer variant (Chen et al., 2023b) and a decay-gated recurrent

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.

model (Mei and Eisner, 2017), both time-inert here (§2.5) — while every other lever is bounded, not null (Figure 1). Ten times more customers moves sequence-model levels by ≤ 0.006 (seed-42 basis; the 6-feature linear floor is the one exception, dipping 0.022–0.028 at 10M with its window gain intact, §2.5), and the window gain itself is scale-invariant. Doubling width or depth moves ≤ 0.005 — both an order of magnitude under the window. These bounds are one-knob-at-a-time measurements within the fixed d128, 2-layer recipe: customers scale at fixed capacity, capacity at fixed data (1M tier), never together, so no compute-optimal cell is measured (§5). At a matched training objective, five sequence architectures — including two continuous-time families whose inter-event interval was identically zero, so their time mechanisms were inert (Appendix H) — land within ~ 0.005 of each other (basis flags: §2.5). The lever is what the model is shown more than what the model is.

Longer windows admit different customers, which confounds window with composition. On *matched cohorts* (same customer-weeks at every window, identical targets, constant denominator) the window gain gets *larger*: the raw comparison understates it in every objective and family measured, at both population scales, and this replicates at three seeds (deltas in §2.2). We cite the matched form.

If the window is the lever of the title, the next questions are where it stops and for whom; preregistered commitments answer the first, activity-specific refits the second: the 24-week measurement favors a saturating law, per-decile refits show its bend moving with customer activity (consistent with a mixture), a preregistered 32-week extension bounds it at $\sim$5–6 months on the pooled weekly-aggregate probe, and a companion preregistered transformer test corroborates the deep model decelerating onto the same form (single-seed, confound-qualified; §2.3, §2.4, Appendix N.1).

The machinery behind these numbers is built to be checked: matched-cohort cross-window comparisons, bit-exact build anchors (R^2 to 15–17 digits), and preregistered saturation analysis. Our contributions:

- **A controlled quantification of the observation-window lever.** A grid over eight model families at 1M shows $+0.07$–0.107 log-space R^2 from window extension (direction replicating on a second dataset, linear and deep families, windows-as-carved — the external windows also enlarge the eligible training set, §2.6) while scale, capacity, architecture (at matched objective), and content — each varied one knob at a time within the fixed recipe — move the sequence families ≤ 0.007 on 3-seed means (content is a single-seed 8-week cell; the $54\times$-parameter capacity probe reads -0.006, 3-seed, Appendix A; linear-floor exception disclosed in §2.5, §2.1).

- **A matched-cohort framework for window comparisons.** Population composition inflates full-set error more than the R^2 denominator as the window grows; matching shows raw deltas *understate* the gain, at both 1M and 10M scale (§2.2).

- **An internally registered saturation analysis and a bounded allocation result.** The pooled weekly information curve bends at 22–24 weeks; the bend moves with activity (consistent with a mixture; allocate within-window context inversely to activity), and a preregistered 32-week extension bounds the law: the pooled fit saturates at $\sim$5–6 months on this instrument, the form replicating on a second dataset (§2.3, §2.4, §2.6).

- **A scored preregistration record.** Four quantitative commitments with falsifiers, committed before their deciding evaluations with varying blindness (disclosed per commitment): one hit; one hit as committed (single-seed, confound-qualified, §2.4); one level-space hit across a disclosed cohort shift (dual-scored, its cohort-invariant scoring a nearer miss); and one miss, the only fully blind prediction, reported at the same volume as the hits (Table 1).

Scope: the primary evidence is one proprietary population. Within it, the 1M/5M/10M tiers are customer-disjoint cohorts (non-overlapping sampling-bucket ranges, §3.1), so cross-tier agreement — the window gain, its matched form (1M and 10M), the sequence-family scale bound — is replication across three non-overlapping customer cohorts of that population rather than resampling of one. The external check is direction-only, in both the linear and deep families (the deep leg is a preregistered three-seed end-anchored test, §2.6; levels near zero, never comparable across datasets). The M8/M12/M16 grid rungs are 3-seed at all three scales; the M24 scale rungs are 3-seed with wide per-seed spread (ranges 0.0094/0.0072 at 5M/10M); the 1M transformer headline claims and the matched 1M Huber ladder are 3-seed; baseline-family rows (GRU/LSTM Huber, ContiFormer, Neural Hawkes) are seed-42 and flagged where cited.

Table 1: The four preregistered commitments and their outcomes. Each prediction and its falsifier were committed before the deciding evaluation ran, same-day for (i) and (ii); the last column states, in brief, what was already known at commit time — the commitments are not equally blind, and saying so is part of the protocol. Three of the predictions are curve fits to already-banked points (the first is a threshold); what each registration adds is the discriminating criterion, fixed in advance (Appendix N.2). †Commitment (ii) is a level-space hit across a disclosed cohort shift; the cohort-invariant gain scoring is a nearer miss — both scorings reported (observed gain $+0.015$ vs. saturating -0.008, short of the log-linear prediction; Appendix L). ‡Commitment (iv) is scored as committed; the deciding measurement is single-seed and carries a disclosed train-carve confound — an unmeasured transformer-side offset of the -0.033 class observed on the ridge sub-ladder would move the observation toward the log-linear prediction (§2.4). §This interval, like every primary-population matched-delta interval, is row-sampling uncertainty at the three shipped test anchors, not an interval over calendar anchors (Appendix B).

Commitment	Prediction (falsifier)	Observed	Verdict	Known at commit time
(i) Composition survival	matched Δ(M16−M8) $\geq +0.05$	$+0.096$ $[0.094, 0.098]$	Hit ($\sim 2\times$)§	Bar set before the deciding transformer re-evaluation; full-set deltas ($\approx +0.075$) already banked.
(ii) Saturation form, 24 wk	saturating $R^2(24)$=0.4345 vs. log-linear 0.4952	0.4296 (Δ 0.005 vs. 0.066)	Hit†	Forms fit before any 24-week evaluation.
(iii) Mixture tail, 32 wk (committed 07-15)	$\Delta(24\to32)=+0.0142$ $[0.0126, 0.0231]$; falsifier < 0.007	-0.0004 $[-0.0014, +0.0004]$	**Miss**	Fully blind: no 32-week job had ever run at commit (two days before (iv)).
(iv) Deep-model bend, 32 wk (committed 07-17)	saturating $+0.012$ vs. log-linear $+0.041$; midpoint 0.026	$+0.011$ $[0.010, 0.013]$	Hit‡	Committed before any matched-cohort number existed for it; the confounded full-set value was banked and disclosed.

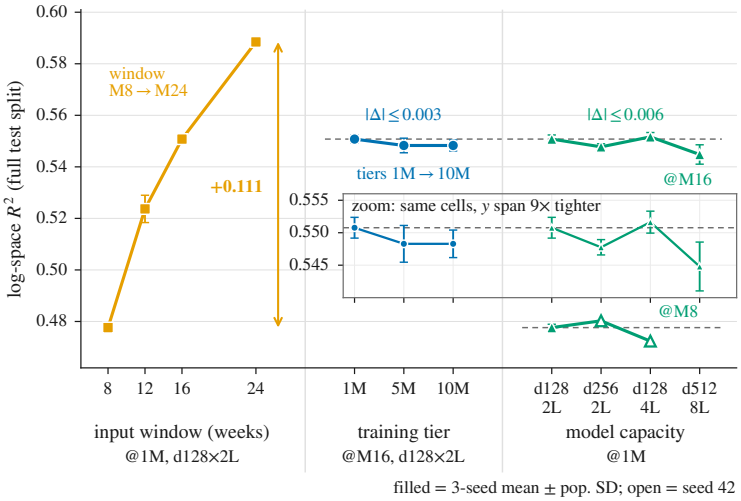

Figure 1: **The observation window is an axis compute-optimal studies hold fixed in training (Brüel Gabrielsson, 2026).** Same recipe; only the named knob is set per ladder — window length also shifts calendar coverage (§5). Window M8→M24 buys $+0.111$; at M16, $10\times$ customers and $4\times$ width/depth move at most 0.003/0.006 (3-seed-mean basis). Filled = 3-seed mean ± SD, open = seed 42; matched window climb is larger still ($+0.131$, 3-seed, §2.2).

2 Results

Setup. All results are on the canonical customer cohorts: disjoint samples at nominal 1M/5M/10M tiers, three anchor weeks, and an input window of M weeks of event history preceding the anchor

Table 2: Input-window sweep at 1M scale, log-space R^2 on the matched cohort (fixed population; §2.2); full-split and pinball rows: Table 3, Appendix A. Both matched rows' M8–M16 cells are 3-seed mean ± population SD; the matched M24 entries (4-way cohort: present in all four windows' test sets) and the matched Δ column († on both matched rows) are the seed-42 paired evaluations (CIs in Appendix A), so neither matched row's Δ need equal the difference of its printed 3-seed cells; the raw full-split M24 cells are composition-inflated and not comparable to M16 (see text) — the matched rows are the like-for-like read. *Train-side eligibility:* longer-window cells also train on up to 10% more, colder customers (Appendix J); matching fixes test composition, not this. The scale bound does not isolate its *size*: $10\times$ the customers moves sequence-family levels ≤ 0.006, an order below the $+0.070$–$+0.096$ window deltas, but on disjoint training cohorts, each arm on its own test carve, so it bounds cohort and carve together, not size alone, nor the *composition* of the colder added stratum. Every contrast here is "window as carved" (§5).

Objective	**Split**	**M8**	**M12**	**M16**	**M24**	Δ(M16−M8)
Huber	matched (3-seed)	$0.4777 \pm .0012$	$0.5368 \pm .0058$	$0.5716 \pm .0012$	$0.6100^\dagger$	$+0.0959^\dagger$
MSE	matched (3-seed)	$0.5075 \pm .0010$	$0.5606 \pm .0007$	$0.5945 \pm .0004$	$0.6265^\dagger$	$+0.0863^\dagger$

(M8/M12/M16/M24 = 8/12/16/24 weeks). The target is total customer activity over the 364 days following the anchor, $\log(1+\max(0, y))$-transformed; the load-bearing metric is log-space R^2 on the held-out test split ($n_{\text{test}} \approx 480$K per 1M-scale cell, ≈ 2.4M at 5M, ≈ 4.8M at 10M), with no target winsorization. The transformer base recipe is d128, 2 layers, 20 epochs, rotary positional encoding; the swept levers are the input window, the training objective $\in \{\text{Huber}, \text{MSE}, \text{pinball}\}$, capacity, and scale. Cells marked 3-seed report mean ± population SD over seeds $\{42, 123, 456\}$; single-seed cells are flagged. Build seams are controlled directly: cited cells carry bit-exactness build anchors (R^2 to 15–17 digits) and the pipeline is deterministic end-to-end at fixed seed (Appendix H). Vocabulary (carve, leg, rung, seam, banked): Appendix N.

2.1 The observation window is the lever

At a fixed 8-week window the standard levers move little: four of the five sequence families at the matched (Huber) objective land in a narrow band (0.4734–0.4776; per-cell values: Appendix A), with the decay-gated recurrent leg starting ~ 0.011 below it and converging as the window grows (§2.5); capacity, content, and $10\times$ customers move the sequence families by at most $|0.007|$ on 3-seed means (content is a single-seed 8-week cell; the $54\times$-parameter capacity probe reads -0.006, 3-seed, Appendix A; the 6-feature linear floor alone dips 0.022–0.028 at 10M, §2.5). The capacity probe reuses the d128 recipe unchanged, so it bounds capacity *at fixed recipe* rather than measuring compute-optimal scaling. A retune of that cell has since run and returns no verdict — selection was on validation only and the confirmation arm is not yet at its full seed count. The training objective moves the number most (~ 0.04, second-order to what follows) — an apparent task ceiling. It is not: it is a conditioning-set ceiling. Extending the input window — matched test cohorts and targets, with window-dependent training eligibility and calendar coverage — moves log-space R^2 more than every architectural, scale, capacity, and content intervention we measured combined (Table 2, Figure 3).

Three-seed at 1M (Table 3), the climb from M8 to M16 is $+0.0701$–$+0.0800$ log-space R^2 by objective — monotone at every objective and every seed, with the M8–M16 climb an order of magnitude above seed dispersion. Extending to M24 adds a further $+0.035$–$+0.038$ raw, though at a visibly decelerating per-week rate (§2.3), and the raw M24 column is denominator-inflated by cohort composition (full-set zero-fraction 11.9% vs. 4.8% matched; target log-SD 2.80 vs. 2.16).

2.2 Window gains survive cohort matching

Longer windows admit different customers: eligibility widens with the window, longer windows sweep in lapsed, colder customers, and the full-set target variance — the R^2 denominator — grows with the window from composition alone (composition statistics: Appendix B). We therefore re-evaluate on matched cohorts: the customer-weeks present in every window's test set, with numerically identical targets and a constant denominator by construction (§3.3; cohort construction, Appendix B), so any climb is error reduction on the same customers. The result, for every family and objective tested: the matched climb is larger than the raw climb — Huber Δ(M16−M8) $+0.0959$ matched (seed-42) vs.

$+0.0731$ full-set (3-seed), all three per-objective matched deltas carrying a paired, customer-clustered bootstrap CI bounded far from zero (per-objective and per-family deltas: Appendix B). Composition was hiding part of the gain rather than manufacturing it, and the matched climb clears our preregistered survival threshold ($+0.05$) by $\sim 2\times$. The result replicates at 10M (Figure 5), where composition deflates every rung: Δ(M16$-$M8) is $+0.0934 \pm 0.0005$ matched (3-seed, per-seed clustered CIs exclude zero) vs. $+0.0753$ full (seed-42), in line with the 1M matched ladder's $+0.0939 \pm 0.0022$. The matched window gain is therefore scale-invariant at three seeds (10M cohort and per-rung detail: Appendix B, Appendix A). The full-set M8–M16 deltas are conservative at both scales, and the matched form is the safe one to cite across windows.

The mechanism is where the errors shrink. Stratifying the matched M16-vs-M8 squared-error reduction of the linear leg by input-window activity decile shows the bottom three deciles carry 69.8% of the total (the lowest decile alone 40.3%). An eight-week window starves low-activity customers of signal, and extending the window helps the customers with the least of it most. Alignment (bit-exact) and full-set reproduction gates precede every matched transformer number (Appendix N); the matched 1M and 10M ladders are both 3-seed (per-seed clustered CIs exclude zero; 1M rungs within 0.004 of cited seed-42 values); matched Δ(M24$-$M8) $= +0.1313 \pm 0.0022$ on the 4-way cohort (3-seed; every seed's clustered CI excludes zero).

2.3 Where the pooled curve bends

Granted the lever, where does it stop? We committed falsifiable predictions *before* any 24-week evaluation existed, registered internally rather than in a public registry: fit to the weekly information curve — a nested ridge on $k = 1..16$ weeks of weekly aggregate features, matched-cohort denominators — a saturating-exponential and a log-linear form predict matched $R^2(24) = 0.4345$ and 0.4952 respectively, a 0.06 separation that one M24 measurement resolves.

The M24 cohort favors saturation (Figure 2a). Observed matched weekly-ridge $R^2(24) = 0.4296$: within 0.005 of the saturating prediction, 0.066 from log-linear. The curve bends at roughly 22–24 weeks: the pooled weekly ridge gains little beyond this point — with one diagnostic: a single exponential *underfits* the pooled curve, exactly as a mixture of per-customer saturation rates predicts (Appendix N.1; carve-splice control, marginal-week values, and fit detail in Appendix C).

2.4 The prediction we failed

We tested whether the pooled curve keeps climbing past this bend with a preregistered extension to a 32-week carve (Appendix N.2): before running it we committed to the mixture reading's prediction of a $+0.0142$ pooled gain from week 24 to 32, with a falsifier of < 0.007. The observed gain was -0.0004 (95% CI $[-0.0014, +0.0004]$, customer-clustered bootstrap; the interval excludes the entire preregistered range $[0.0126, 0.0231]$ and its upper bound sits below the 0.007 discriminator; Figure 2b) — the mixture extrapolation is rejected on its own falsifier, and the saturated-pooled null survives: weeks 25–32 add nothing to the pooled weekly-ridge instrument, which plateaus at $k \approx$ 21–22 on the M32-carve ladder. The pooled ridge curve stays flat through 32 weeks, though activity-specific refits keep improving. The plateau is no carve-pipeline artifact: an independently produced re-cut of the same export (the producer's own merge pipeline, which also deduplicates exact-duplicate events; measured per-k footprint $< 10^{-3}$ R^2) reproduces the k=1..32 ladder verdict on its own train/test against three criteria committed before it was scored — saturating fit beats log-linear (RMSE 0.011 vs. 0.020), flat tail (-0.0005, below the $+0.007$ discriminator), late marginal under the committed 20% of k=16's — a pipeline replication of the same export, not an independent dataset.

The transformer tells the same story in rate terms: its seed-42 matched marginal gain decelerates $+0.0156 \rightarrow +0.0083 \rightarrow +0.0046 \rightarrow +0.0014$ R^2/week across M8$\rightarrow$M32, while its matched M24 level (0.6100, seed-42) sits far above the weekly-ridge asymptote (a cross-instrument, cross-objective comparison, but the gap is $4\times$ the largest objective effect we measure, so the reading is basis-robust): the deep model decelerates over the same window, on training cohorts that grow with it, but extracts far more from it. Commitment (iv) of Table 1 makes this preregistered: the observed matched within-cohort gain $\Delta(24\rightarrow32) = +0.011$ (customer-clustered 95% CI $[0.010, 0.013]$, n=351,174 — 94.2% of the 4-way cohort — single seed) lands on the saturating prediction ($+0.012$) and excludes the log-linear one ($+0.041$). The saturating fit's asymptote is 0.629 against a fitted $R^2(32)$ of 0.622, leaving

fitted headroom beyond 32 weeks of at most 0.007 — on par with the bounded axes. One confound qualifies this single-seed contrast: the 24- and 32-week checkpoints were trained on different carves (the 32-week model on a re-carved M32 train basis with its own feature scaler), and the ridge sub-ladder measured a -0.033-class cross-carve level offset; a transformer-side offset of that class would put the true gain near the log-linear prediction, though the truncation control below does not identify the transformer's own basis offset. The pooled-probe verdict (iii) is within-carve and immune to this seam; the input-cap exclusion leaves the gain unchanged (§5); and a within-carve truncation control (same checkpoint, train basis, scaler, and rows; inputs cut at evaluation to the most recent 24 weeks) loses 0.023 R^2 (clustered 95% CI $[0.021, 0.024]$): the gain is not a train-basis artifact alone. Evaluation-time truncation adds a distribution shift of unknown bias direction, so the control corroborates without sharpening the estimate; the deep-model reading stays confound-qualified.

2.5 Window gains across the tested implementations

Every family climbs — eight of them (Table 6, Appendix A). At 1M, Δ(M16$-$M8) runs from $+0.070$ for a continuous-time attention model (ContiFormer) to $+0.107$ for a single-feature linear model ($\alpha x + \beta$ on input-window activity) — broadly, the less a family can represent, the more the window is worth to it, though not monotonically (no single feature carries the boosted-tree climb: drop-one ablation, Appendix A). At M16 the transformer, GRU, LSTM, and simplified re-implementations of ContiFormer and Neural Hawkes all land within ~ 0.005 of each other on the same objective; the decay-gated recurrent leg starts 0.015 below the pack at M8 and converges by M16 — window information washes out even the largest small-window architecture gap (M24 continuation: Appendix A). The continuous-time rows are seed-42 point estimates on an anchored build, trained at the baseline legs' own budget, and both ran with an inter-event interval of identically zero, so their time mechanisms were inert (Appendix H) — they are not part of the loss/budget equalization below; the window gain tracks the conditioning set rather than an architectural capability, with window-dependent training eligibility and calendar coverage unresolved (Table 2, §5).

Scale stays inside a narrow band: each 10M transformer cell (Table 8, Appendix A) sits within 0.006 of its 1M twin. A bounded band is not the same as no effect (Appendix A). The window gain is scale-*invariant* in both forms ($+0.0753$ vs. $+0.0754$ full, seed-42; $+0.0934 \pm 0.0005$ vs. $+0.0939 \pm 0.0022$ matched, 3-seed); $10\times$ more customers moves the gain by ≤ 0.002. This scale bound is a one-knob-at-a-time result within the fixed d128, 2-layer recipe (§5). The M8/M12/M16 rungs are 3-seed at all three scales, and the recurrent grids — at a fixed cached data exposure per tier (Appendix A) — tell the same story. So does ridge: at 10M it climbs steepest ($+0.1050$), though its *level* dips 0.022–0.028 below the 1M twin at every window — a window-stable linear-floor offset that is the one family exception to the scale bound. The M16-10M ridge cell is an off-pipeline recomputation gated to 8 decimals on the adjacent M12-10M cell (full protocol with Table 8, Appendix A). The sequence-model history gain is $+0.069$–$+0.085$ at every scale we test, endpoints seed-42: window length, not customer count, is the axis that matters on this task.

Architecture differences are small under incompletely matched recipes. At fixed window and architecture the training objective reorders the leaderboard: MSE $>$ Huber $>$ pinball at every window, a $\sim$0.03–0.05 spread — an ordering the metric is not neutral about, log-space R^2 being affine in log-space MSE, so only the squared-error arm is scored on its own training criterion (Appendix H). The spread is second-order against the window effect but decisive for cross-family comparisons, where it initially masquerades as an architecture effect. Under the Huber head the transformer at M16 (0.5508 ± 0.0016) ties the GRU (0.5522, seed 42 — within one seed SD of the transformer's 3-seed mean); under the MSE head it clears the same GRU row by $+0.035$. Neither read survives loss equalization: retrained at the matched objective the three sequence families climb together, at the fully equalized recipe every family$\times$window cell lands within ± 0.005 of the transformer's 3-seed mean, and run the other way — the transformer down-budgeted to the baselines' recipe — the one out-of-band gap is 0.0060 at M16, *in the baselines' favor*: at neither operating point does any window show a transformer advantage outside the band (cells, gaps, and the full retraining protocol: Table 7, Appendix A). The earlier $+0.035$ margin was a loss confound, and we retire it; the equalized cells ran on builds anchored bit-exactly on both training paths (Appendix H), so the build seam is zero. What survives is the converse: the objective and the conditioning set carry the result, and architecture is second-order.

2.6 External validity: a second population, direction only

Everything above is one proprietary population. As an external check we replicate the window intervention on the public Alibaba Taobao UserBehavior dataset (Zhu et al., 2018) — a 9-day span, so windows are days not weeks, and the target is $\sim$69% zero: an intentionally hostile, much sparser regime. The 3/5/7-day window carves are end-anchored to a common date (identical target period, only the length of observed history varies; full specification in Appendix E). Ridge on the $\log(1+y)$ target, matched cohort of 6,985 users present in all three window carves: log-space $R^2 = 0.0213 \rightarrow 0.0305 \rightarrow 0.0374$, monotone, with all three paired window-delta bootstrap CIs excluding zero (Δ(D7−D3) $= +0.0162$ $[+0.0090, +0.0229]$). A day-resolution ladder replicates the §2.3 *shape*, with the bend at or beyond the observable range — sparser signal bends later; no universal horizon constant (Appendix E). The transformer leg replicates the direction under a preregistered three-seed test on the same carves: Δ(D7−D3) $= +0.047$ $[+0.035, +0.059]$ (customer-clustered; all seeds positive, each seed's CI excluding zero; longer end-anchored carves also grow the eligible train set, so the intervention is windows-as-carved rather than pure history; full design, power disclosure, and the superseded same-start null in Appendix E). Absolute levels stay near zero and we make no magnitude claim: the external-validity evidence is *direction-only*, in both the linear and deep families.

3 Data and methods

3.1 Cohorts, carves, and splits

The primary dataset is customer event streams from a single consumer platform, exported as fixed cohort carves: input window $M \in \{8, 12, 16, 24\}$ weeks $\times$ population tier (1M/5M/10M); a row carries M weeks of history before its anchor and the target after it. No grid carve was ever re-cut (a later 32-week carve adjudicates the preregistered persistence test (Appendix N.2), validated same-family at 2.3×10^{-9} on shared keys). Tiers are non-overlapping sampling-bucket ranges — no customer appears in more than one — and tier names are bucket labels, not counts (split sizes: Appendix J).

Splits are by customer *and* time simultaneously (disjoint calendar ranges, $\approx$8-week buffers, identical in every audited carve): target weeks run 2024-01-29 to 2024-09-09 in train, 2024-11-04 to 2024-12-30 in validation, and 2025-02-24 to 2025-04-21 in test — three test anchor weeks, four weeks apart. Windows are anchored at the target week and grow backwards, so this fixed test band is what ties window length to calendar coverage (§5). Disjointness is verified by direct join, not assumed: across six carves and twelve pairwise checks, train$\cap$test and train$\cap$val share **zero** customers and **zero** rows (Appendix J) — but the 364-day target horizons overlap in calendar time, so the splits are out-of-time in anchor weeks, not in labels.

Era discipline. Every cell trains and evaluates on this one export; no number is compared, even qualitatively, across exports. The public dataset for the external check is described where used (§2.6).

3.2 Target and evaluation

The target is total customer activity over the 364 days following the anchor week, a real-valued total rather than a count, modeled as $\log(1+\max(0, y))$ with no winsorization. All R^2 values are computed in this transformed space ("log-space R^2") on the full test split unless a matched cohort is stated: the raw target is heavy-tailed, so back-transformed R^2 is dominated by a few extreme customers and reorders models across seeds. Every reported number states its space; no comparison mixes spaces. Two common alternatives are addressed: on the matched cohort the denominator is constant, so ΔR^2 is proportional to ΔMSE — the matched log-space MSE column (Table 4) is the squared-error reading of the same result — and MAPE is degenerate on this target, which is exactly zero for 4.8%–30.4% of rows depending on the population slice evaluated.

Objective-matched comparison rule. The training objective moves log-space R^2 by 0.03–0.05 at fixed architecture — more than any architecture intervention we measured — so a cross-family delta is admitted only when both cells match on objective family, transform, cohort/era, evaluation protocol, and build (full rule: Appendix H); cells failing any leg are within-family curves, never cross-family wins — the rule that retires the transformer-over-RNN margin in §2.5. *Normalization provenance.* Feature-normalization statistics are train-split-derived per carve: the M8–M24 carves

use the full-upstream train scaler, while the 32-week carve's scaler is derived from its own merged train — train-only in both cases, with no leakage path, but a distinct normalization basis for the M32 cells, which we disclose.

3.3 MATCHED-COHORT EVALUATION

Any intervention that changes data eligibility — here, lengthening the input window — changes *who* is in the test set along with *what* the model sees (§2.2); a raw cross-window delta confounds information with composition. A negative control is a no-effect condition the observed estimate must beat (Petersen, 2025): with targets bit-identical across windows, customer-level swaps of the M8/M16 prediction vectors form an exact null under which the estimate centers at zero — the observed $+0.0959$ exceeds all $1{,}000$ permuted draws (Monte Carlo permutation $p{=}1/1001$), and the clustered-CI false-positive rate on permuted datasets is consistent with the nominal 5% (Appendix K). The matched form is the only form in which we cite a cross-window delta; full-set deltas are reported as the raw read, empirically conservative on the M8–M16 rungs.

4 DISCUSSION

What the falsification bought. The mixture model was a defensible reading of the within-window evidence; its extrapolation was quantitative, committed in advance, and wrong by its own falsifier. What survives is stronger for it. Of the four preregistered commitments (Table 1), the only fully blind one could have failed and did; the other three are weaker instruments that held, one of them a dual-scored level-space hit across a disclosed cohort shift. The discriminating weight sits on the blind registration, not spread evenly over four, yet the set is still harder to attribute to analyst degrees of freedom than any single test. The record is internal and self-attested, and its table discloses how blind each commitment was. The saturation reading extends past the linear probe, though not independently of it: commitment (iv), committed before any matched-cohort number existed for the 32-week transformer, landed on the saturating form rather than the log-linear one (§2.4). Both curves decelerate, but a shared saturation law is not established.

5 LIMITATIONS

Window length and calendar coverage are not separated (§1). The largest open control concerns the lever itself: all test target weeks sit in one $\approx$8-week band, so extending M adds specific calendar months — 16 weeks first reaches the prior November (earliest anchor only; the latest reaches it at 24), 24 and 32 weeks the preceding autumn and summer. One seasonal mechanism is a candidate explanation for both the M8$\rightarrow$M16 step size and the 22–24-week bend (where the last anchor finishes covering that season), and nothing here discriminates it from a pure history effect: every sweep shares those three anchors, and the anchor-varying control we ran is void under its own criterion (every cell's gain is several times larger at the earliest anchor than the latest, but per-anchor R^2 spread at $\mathrm{M8}$ is, on MSE, 5.6–$6.2\times$ the permitted bound, and baseline difficulty is ordered with calendar position). The magnitude is therefore not identified; the sign is positive in every arm tested, though not analytically identified. The needed control is anchor variation, not end-anchoring: end-anchored re-carving fixed the target period on the external dataset (§2.6), but the primary ladder is *already* end-anchored, which is why coverage is not separated here; a calendar-balanced anchor set cannot be recovered from this export (each row carries only its own anchor's target) and needs a new extraction plus a ladder re-run. A calendar-*balanced* average is the stronger estimand and remains unmeasured.

No cost normalization. 8$\rightarrow$16 weeks changes what history the input carries, not its length, at a fixed 750-event cap; 1M$\rightarrow$10M is ten times the data; the capacity probe is $\approx 54\times$ the parameters ($491{,}248 \rightarrow 26{,}479{,}216$, Table 12). We report per-cell wall-clock classes, not total GPU-hours, and ran no cell at matched cost across axes, so "outweighs" is a statement about effect size at the recipes tested, not about budget-optimal allocation.

Two bounded axes lack matched schedules. The transformer legs ran at an effectively constant learning rate while the recurrent and continuous-time legs annealed in full, and the optimizer-step count grows with the training set while the cosine period does not — so the architecture band spans families optimized differently, and the scale bound compares two optimization regimes rather than

one knob. The learning-rate asymmetry pushes toward manufacturing the architecture band rather than concealing one, and on the baseline-only three-seed basis, where that schedule asymmetry does not apply, the architecture spread is wider than the seed-42 figure the abstract reports. We state both as measured and do not claim either would survive matched schedules. Mechanism, counter-argument, and what these asymmetries do *not* reach: Appendix I.

The primary evidence is one platform, one target, one export era, by design — era discipline is what makes the within-paper deltas clean, and it is also the boundary of the claims. Within that platform, the 1M/5M/10M tiers are customer-disjoint cohorts (§3.1), so findings that agree across tiers replicate across three non-overlapping customer cohorts — internal replication, not an independent data source. The external replication is direction-only over a far shorter horizon, and its intervention is windows-as-carved — longer end-anchored windows also enlarge the eligible training set, so it does not separate more history from more training rows; its deep-model leg replicates the direction (three seeds, preregistered) but was underpowered at the preregistered effect — the verdict stands on a $\sim 3\times$-larger observed effect; one of nine runs collapsed (excluded from monotonicity legs only), and absolute levels on the public dataset are near zero (power disclosure and full design: Appendix E). The saturation and allocation results are measured on a weekly-aggregate linear probe, corroborated for saturation by a preregistered matched-transformer test at 32 weeks (single seed, pooled). Several grid cells at 5M/10M scale are seed-42 point estimates (flagged wherever cited), and two cells are off-pipeline recomputations disclosed where they appear. Matching fixes test composition but conditions on survival into the intersection — customers present at every window — which is the population the window decision actually concerns, but not the full platform population. The bounded-axis claims are one-knob-at-a-time results inside the fixed d128, 2-layer recipe: capacity is probed only at the 1M tier (3-seed at 16 weeks, single-seed at 8) and without re-tuned hyperparameters, and the joint capacity$\times$data cell of compute-optimal scaling is not tested, so these bounds are claims about this operating regime, not about jointly scaled models. The ablated content field, product category, is missing for most events in this export, so its 8-week cell establishes unavailability, not uninformativeness. Finally, a fixed 750-event input cap attenuates event-weighted readings of the longest-window gains (3.0% of the matched cohort was already cap-hit at M24; 24.2% of the incremental $24\rightarrow32$-week event mass is truncated, concentrated in the heavy tail); the saturation verdict does not rest on capped inputs — the uncapped ridge instrument replicates it, a direct exclusion test leaves the matched gain unchanged, and a preregistered cap-raise to $1{,}500$ events (single seed) shifts it inside its prediction band (full cap analysis and per-window cap-hit fractions: Appendix H).

6 Conclusion

Held to fixed cohorts, targets, and matched evaluation, customer-behavior forecasting has one dominant input-side lever: the observation window, confounded with calendar coverage (§5). It holds across eight model families and a $10\times$ span of population scale, grows once composition is controlled, and is bounded by a saturation law — five to six months on the pooled weekly-aggregate probe — that we predicted, measured, and then failed to extrapolate on our own internally registered falsifier. The machinery that made that negative informative rather than ambiguous — matched cohorts, bit-exact build anchors, preregistered predictions with discriminating criteria — is the transferable part.

Use of AI assistants

We disclose against the task list in the venue's AI policy.

Used. An AI coding assistant wrote and edited research code: model variants, the data and evaluation harnesses, the scripts that derive the window carves from a superset extract, and the analysis scripts behind the tables and figures. It drafted and edited prose. It surfaced candidate related work, which we read and cited ourselves. It produced intermediate analyses and summaries of run outputs, and it assisted in interpreting results: it proposed readings of run outputs and of the fitted saturation law, which we checked against the banked artifacts and either adopted, narrowed, or discarded. Where a reading could not be sourced to those artifacts, we narrowed or dropped it. Because the carve and harness code reformats data, we count this as assistance with cleaning and reformatting datasets. And one of the paper's organizing ideas came out of that assisted analysis rather than from us unaided — that the pooled saturation law is a mixture, so context is worth most where activity is lowest — so we

count this as help developing a conceptual framework. We re-derived it from the banked artifacts before adopting it, and the claim and its stated scope are ours.

Not used. We did not use it to generate or augment training or evaluation data, to propose or refine the hypotheses, to design the experiments or set the preregistered thresholds and falsification rules, for translation, or for qualitative or thematic analysis. The predicted levels those rules score against came from the assisted per-decile analysis disclosed above. No number reported here was produced by a language model: every statistic comes from a run recorded in our own artifacts and was verified against them before it was cited.

Not applicable. The paper states no theorems and contains no proofs, so formulating mathematical claims, supplying ingredients for proofs, and writing proofs do not arise. The two fitted functional forms, the saturating and log-linear laws, were specified by us in advance of the fit.

Where the assistant produced analysis, the readings drawn from it are ours and were checked against the banked artifacts before use. The experimental design, the preregistered criteria, and all claims and their stated scopes are ours, as is responsibility for this paper.

ETHICS STATEMENT

This is aggregate behavioral forecasting on platform event data: every reported result is a population-level statistic, and no data, models, or individual-level predictions are released. The findings measure whether observation history carries information, and for whom; the allocation result cuts both ways: context is worth most for low-activity customers, which can guide better service for lightly observed users or, equally, more effective targeting of them. What the paper measures is a ceiling and a budget, and both point toward using *less* history rather than more. The pooled weekly-history fit saturates: on the pooled within-carve probe — the one instrument here whose 24-to-32-week comparison carries no train-basis seam — extending weekly history from 24 to 32 weeks returns -0.0004 (95% CI $[-0.0014, +0.0004]$), which is nothing. The measured saturation point ($\sim$5–6 months) is therefore a retention *ceiling* on these instruments rather than a license to keep more — a ceiling on the pooled fit, not on the signal: per-decile refits of the same weekly family still gain $+0.012$–$+0.021$ from week 24 to 32 in every activity decile, every clustered CI excluding zero (Appendix F). And because the bend moves with activity, routing context inversely to activity retains $\geq 99\%$ of the fixed-24-week value at 49.2% of the feature-week budget on the pooled-of-decile-refits instrument, not against the pooled probe above; that is an in-sample reading, since the allocation is searched on the rows it is scored on (Appendix F). The routing total is itself a full-24-week activity scalar, uncounted in that budget, and routing alone retains 90.0% — so about half the detailed weekly features, for substantially the same accuracy.

The residual tension is real and we state it rather than resolve it: conditioning on months of behavioral history sits in tension with data minimization, and the customers this result helps most are the least-observed ones, who are also the least placed to anticipate it. The paper measures that concentration directly — the bottom three activity deciles carry 69.8% of the error reduction — and it is why the allocation reading should be taken as *where context is scarce*, not as a reason to collect more of it.

One mitigation we would have claimed does not survive our own measurements, so we withdraw it rather than qualify it: routing on coarse activity aggregates is *not* a substitute for event-level history. The aggregate probe reaches $R^2(24) = 0.4296$ where the event-level model reaches 0.6100 on the same matched cohort (§2.4) — a gap far larger than any bounded effect in this paper. On the pooled-of-decile-refits instrument the budget result uses (uniform-24 $= 0.601$, Appendix F) the gap is 0.009, still above every bounded axis, and the budget result is measured on aggregates alone and does not show that event-level data is dispensable.

Data handling. The primary corpora are internal platform event data. No raw events, model weights, or individual-level predictions leave that boundary or are released here, and every quantity reported in this paper is an aggregate over a population rather than a statement about any customer. The public dataset used for the external check (§2.6) is a publicly released corpus, obtained as released and used for research only; we redistribute no part of it.

The contribution is measurement discipline — what the lever is, where its pooled fit saturates, how to attribute it cleanly — not a deployment capability. Deploying longer observation windows is a

platform-governance decision, with the same obligations as any other use of customer behavioral data.

REPRODUCIBILITY STATEMENT

Determinism and build anchoring are gated as described in the §2 setup (twin-dispatch proofs and the anchor ledger: Appendix H). Seed policy: 3-seed ($\{42, 123, 456\}$) mean $\pm$ population SD wherever marked; remaining seed-42 point estimates are flagged per cell and never anchor a claim seed dispersion could plausibly overturn, with one disclosed exception — the seed-42 architecture band, which does not hold on the baseline-only three-seed basis (Appendix I).

Two baseline legs are simplified re-implementations rather than the published models: the ContiFormer leg is a causal attention stack with an additive per-head time-decay bias and no ODE solver, and the Neural Hawkes leg is a decay-gated LSTM cell with no intensity likelihood, and on this export both legs ran with an inter-event interval of identically zero, so their time mechanisms contributed nothing — the architecture band of §2.5 compares this transformer against these implementations, not against the published ones, and does not test continuous-time mechanisms at all (all specified in Appendix H). The swept "objective" likewise names the loss on the aggregate prediction alone: a fixed-weight auxiliary per-event term rides alongside it in every cell and the sweep never moves it (Table 11).

A reproducibility protocol for multi-build experiment campaigns. Bit-exact anchor cells certify container builds seam-free before their results are cited (Table 10 lists seven bit-exactness proofs across five successive builds, agreeing to 15–17 digits), dispatches are deterministic end-to-end at fixed seed, and the saturation predictions were committed before their deciding evaluations ran, on internal self-attested records rather than third-party timestamps (§2, Appendix H, Appendix L).

REFERENCES

Ibram Abdelmalak, Kiran Madhusudhanan, Jungmin Choi, Christian Klötergens, Vijaya Krishna Yalavarthi, Maximilian Stubbemann, and Lars Schmidt-Thieme. Channel dependence, limited lookback windows, and the simplicity of datasets: How biased is time series forecasting? In *PAKDD*, pages 585–597, 2026. doi: 10.1007/978-981-92-1462-4_46.

Rishi Ahuja, Kumar Prateek, Simranjit Singh, and Vijay Kumar. Retrieval mechanisms surpass long-context scaling in time series forecasting, 2026. arXiv:2605.08217.

Michel Ballings and Dirk Van den Poel. Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications*, 39(18):13517–13522, 2012. doi: 10.1016/j.eswa.2012.07.006.

Luca Bertinetto, João F. Henriques, Samuel Albanie, Michela Paganini, and Gül Varol. Preface. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, volume 148 of *Proceedings of Machine Learning Research*, page i. PMLR, 2021.

Rickard Brüel Gabrielsson. Scaling laws for behavioral foundation models over user event sequences, 2026. arXiv:2606.05257.

Luca Butera, Giovanni De Felice, Andrea Cini, and Cesare Alippi. Why do time series models need long context windows?, 2026. arXiv:2606.01999.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023a.

Yiwen Chen, Joshua Ainslie, Krzysztof Choromanski, Xiang Gao, Su-Lin Wu, Yiping Yuan, and Qian Sun. ClockRoPE: Random Fourier rotations for temporal routine modeling, 2026. arXiv:2607.26369.

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. ContiFormer: Continuous-time transformer for irregular time series modeling. In *NeurIPS*, pages 47143–47175, 2023b. doi: 10.52202/075280-2042.

Hien Dang, Pratik Patil, and Alessandro Rinaldo. Prediction-only distillation in linear and logistic regression, 2026. arXiv:2607.15450.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *ICML*, pages 11091–11104, 2024.

Peter S. Fader, Bruce G. S. Hardie, and Ka Lok Lee. "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, 24(2):275–284, 2005. doi: 10.1287/mksc.1040.0098.

Xiao Han, Yao Xiao, Chenyu Wu, and Tongchen Zhang. How early is early enough? Design-dependent observation-window sufficiency in subscription churn prediction, 2026. arXiv:2607.00473.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *NeurIPS*, pages 30016–30030, 2022. doi: 10.52202/068431-2176.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *NeurIPS*, 2017.

Kiwan Kwon, Kangmin Kim, Hojin Lee, Yeseong Jung, Hyeongwoo Kong, Vamsi K. Potluru, Saerom Park, and Yongjae Lee. Seq2Synth: Benchmarking temporal fidelity in synthetic sequential tabular data, 2026. arXiv:2607.15606. Extended version of a paper accepted at CIKM 2026.

Jiacheng Li, Yujie Wang, and Julian McAuley. Time interval aware self-attention for sequential recommendation. In *WSDM*, pages 322–330, 2020. doi: 10.1145/3336191.3371786.

Jingteng Li, Alexander Capstick, Louise Rigny, Iona Biggart, Neil J Sebire, and Payam Barnaghi. LLM4EHR: Aligning clinical time series with medical event sequences via large language models, 2026. arXiv:2607.15447.

Kunpeng Li, Guangcui Shao, Naijun Yang, Xiao Fang, and Yang Song. Billion-user customer lifetime value prediction: An industrial-scale solution from Kuaishou. In *CIKM*, pages 3243–3251, 2022. doi: 10.1145/3511808.3557152.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Hongyuan Mei and Jason Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *NeurIPS*, 2017.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *ICLR*, 2024.

Anne Helby Petersen. Are you doing better than random guessing? A call for using negative controls when evaluating causal discovery algorithms. In *UAI*, pages 3465–3479, 2025.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, 2008. doi: 10.7551/mitpress/9780262170055.001.0001.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. In *IJCAI*, pages 4585–4593, 2021. doi: 10.24963/ijcai.2021/623.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. Scaling law for time series forecasting. In *NeurIPS*, pages 83314–83344, 2024. doi: 10.52202/079017-2650.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Ramin Soleimani, Andrea Visentin, and Dirk Pesch. Behaviour-conditioned neural processes for adaptive residential short-term load forecasting, 2026. arXiv:2607.16168.

Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010. doi: 10.1214/09-STS313.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In *NAACL*, pages 613–623, 2021. doi: 10.18653/v1/2021.naacl-main.51.

Xiaojing Wang, Tianqi Liu, and Jingang Miao. A deep probabilistic model for customer lifetime value prediction. *arXiv preprint arXiv:1912.07753*, 2019.

Xiaokai Wei, Jiajun Wu, Daiyao Yi, Reza Shirkavand, and Michelle Gong. Rotate both ways: Time-and-order RoPE for generative recommendation, 2025. arXiv:2510.20455.

Tingsong Xiao, Zelin Xu, Wenchong He, Zhengkun Xiao, Yupu Zhang, Zibo Liu, Shigang Chen, My T. Thai, Jiang Bian, Parisa Rashidi, and Zhe Jiang. XTSFormer: Cross-temporal-scale transformer for irregular-time event prediction in clinical applications. In *AAAI*, pages 28502–28510, 2025. doi: 10.1609/aaai.v39i27.35073.

Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. In *KDD*, pages 1079–1088, 2018. doi: 10.1145/3219819.3219826.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In *ICML*, pages 11692–11702, 2020.

A Extended results and ablations

Window sweep, pinball objective. The pinball rows of the input-window sweep (Table 3) show the same pattern as the Huber and MSE rows of Tables 2 and 3: monotone climb, matched delta larger than full.

Table 3: Input-window sweep at 1M scale: full-test-split rows (all three objectives) and the pinball matched row (companion to Table 2; same conventions).

Objective	Split	M8	M12	M16	M24	Δ(M16−M8)
Huber	full (3-seed)	0.4776±.0013	0.5237±.0053	0.5508±.0016	0.5885±.0015	+0.0731
MSE	full (3-seed)	0.5075±.0010	0.5557±.0010	0.5875±.0007	0.6242±.0005	+0.0800
pinball	full (3-seed)	0.4698±.0007	0.5178±.0012	0.5400±.0025	0.5746±.0006	+0.0701
pinball	matched (s42)	0.4707	0.5341	0.5656	0.5973	+0.0948

Matched-cohort linear leg. Table 4 reports the six-feature ridge instrument per window on the fixed cohort: the matched climb exceeds the raw climb, the same composition-deflation direction as the three sequence families with a matched leg (§2.2) — transformer, GRU and LSTM; the two continuous-time families have none.

Table 4: Matched-cohort evaluation, linear leg (6-feature ridge, 1M): full-test-set vs. matched R^2 with the constant-denominator MSE column (log-target space). The matched climb exceeds the raw climb.

Window	Full R^2	Matched R^2	Matched MSE (log)
M8	0.4373	0.4369	2.627
M12	0.4962	0.4993	2.336
M16	0.5365	**0.5412**	2.140

Objective-matched baseline comparison, in full. Table 5 is the cross-family comparison backing §2.5: every row optimizes squared error on the log target and is scored on the same matched cohort, so the remaining ordering is not an objective or composition artifact. The gradient-boosted baseline is a fixed-hyperparameter ensemble (400 trees, 63 leaves). The time-representation widening in full: on the full test split the stacked cell (MSE objective with MS-RoPE, a multi-scale rotary position encoding over four physical-time bands, initialized at an hour, a day, a week and a year and then learned, with learned mixing; Appendix G) clears GBM by $+0.0250$ where MSE alone cleared it by $+0.0065$ — a $\sim 4\times$ widening from a representation change, not a capacity change; on the matched-cohort instrument the widening is $+0.0185$ over $+0.0078$ (3-seed transformer basis; the banked seed-42 pairing read $+0.0082$), $\sim 2.3\times$ — conservative. The GBM comparator is fixed-hyperparameter and untuned (Table 11); a tuned GBM would shrink both margins and, shrinking the smaller one proportionally more, would raise the multiplier — so read $\sim 2.3\times$ as a ratio against *this* baseline, not as a tuned-tabular ceiling. The MSE transformer's full-set margins over GBM shrink monotonically with window on 3-seed means. The recurrent matched columns on the Huber basis: GRU 0.5729 / LSTM 0.5715 vs. 0.5731 for the Huber transformer — identical cohort and denominator, seed 42, full-split values reproducing the banked cells exactly.

Table 5: Objective-matched comparison at M16, 1M, matched cohort (n=407,338), log-space R^2. All rows optimize squared error on the log target. Sequence models see raw event streams; GBM and ridge see six aggregate features. GBM is an offline fit on a harness that reproduces the banked ridge cells to within 1.1×10^{-5} (3-seed SD $\leq 1.4\times10^{-4}$). ‡matched-cohort value from the banked per-row predictions, 3-seed mean ± population SD (full-test-split 0.6060 ± 0.0006, 3-seed); the Transformer (MSE) row is likewise 3-seed (seed-extension evaluations gated bit-exact against the banked cell); the LightGBM row is the banked offline fit (seed 42; its 3-seed mean matches at display precision); the ridge row is deterministic. The stacked cell is the only transformer cell where matched R^2 sits marginally below full-split at every seed — the full-split proxy was not conservative here, which is why we score the matched leg directly.

Model	Input	Matched R^2 ↑
Transformer (MSE + MS-RoPE)	raw events	$\mathbf{0.6052\pm0.0007}^{\ddagger}$
Transformer (MSE)	raw events	0.5945 ± 0.0004
LightGBM	6 aggregate features	0.5867
Ridge	6 aggregate features	0.5412

Per-family window deltas. Table 6 tabulates the eight-family window lever of §2.5. The $+0.069$–$+0.085$ range quoted there is a hull over families and tiers rather than a per-tier span: its endpoints are two different models at two different scales — the 5M transformer-Huber cell ($0.5443 - 0.4758$) and the 1M decay-gated recurrent cell ($0.5474 - 0.4620$), both at seed 42.

Loss- and recipe-equalized retraining, in full. Table 7 backs the architecture paragraph of §2.5; the retraining protocol in full: retraining the recurrent baselines under the same MSE objective at their own budget, at 3 seeds and *every* window, all three sequence architectures climb $\sim+0.08$ together (seed SDs ≤ 0.0011 throughout). Equalizing the training *recipe* as well — both families retrained at the transformer's budget (8,192-effective batch, 20 epochs, near-full training stream; 3 seeds) — lands every family×window cell within ±0.005 of the transformer's 3-seed mean (the M16 sign flip of the transformer−GRU gap within seed noise; spread ~ 0.001 at M16; per-cell values and gaps in Table 7); down-budgeting the transformer to the baselines' recipe instead (batch 128, 10 epochs; three seeds at M16, single seed elsewhere) converges at M8/M12, with both recurrent nets — retrained at this same down-budgeted operating point, so these cells are *not* the Table 7 rows —

Table 6: The window lever in every baseline family: Δ(M16$-$M8) log-space R^2 at 1M, full test split unless marked matched. Transformer deltas are in Table 3 ($+0.0701$–$+0.0800$, 3-seed, by objective). Bases as flagged in the text: recurrent and continuous-time rows are seed-42 Huber-recipe cells; GBM is a 3-seed offline fit; the linear rows are deterministic. The two continuous-time rows (the ContiFormer and Neural Hawkes legs in the text) are simplified re-implementations rather than the published models, and both ran with an inter-event interval of identically zero (Appendix H). [a] $+0.089$ on the matched cohort.

Family	**Input**	Δ(M16$-$M8)
GRU (2-layer)	raw events	$+0.076$
LSTM (2-layer)	raw events	$+0.077$
Causal attention variant (time bias inactive)	raw events	$+0.070$
LSTM variant (decay inactive; no point-process likelihood)	raw events	$+0.085$
Gradient-boosted trees (Ke et al., 2017)	6 aggregate features	$+0.084^a$
Ridge	6 aggregate features	$+0.099$
Single-feature linear ($\alpha x + \beta$)	input-window activity total	$+0.107$

landing 0.0060 *above* it at M16 (GRU 0.59019, LSTM 0.59013 vs. transformer 0.58417 ± 0.0016; 3-seed means both sides) — outside the band, in the baselines' favor. The two operating points (different batch/epochs/topology/build) are reported separately, not pooled; at neither budget does any window show a transformer advantage outside the band. The equalized cells ran on two further builds anchored bit-exactly on both the transformer and the recurrent paths (Appendix H).

Table 7: Architecture converges to a narrow band at matched objective — a bound on the family spread, not a resolved zero: loss- and recipe-equalized retraining at 1M, MSE objective, log-space R^2 on the full test split (3-seed means; columns are M8/M12/M16). *Own budget* is each recurrent family's original training budget (batch 128, single process, 10 epochs, fixed cached draw); *transformer budget* is the transformer's ($8{,}192$-effective batch, 20 epochs, near-full stream). Seed SDs are ≤ 0.0011 (own budget) and ≤ 0.0002 (recurrent, transformer budget). The M16 sign flip in the transformer$-$GRU gap is within seed noise.

Cell (MSE objective)	**M8**	**M12**	**M16**
Transformer	0.5075	0.5557	0.5875
GRU, own budget	0.5038	0.5533	0.5863
LSTM, own budget	0.5024	0.5520	0.5850
GRU, transformer budget	0.5055	0.5547	0.5879
LSTM, transformer budget	0.5033	0.5537	0.5869
Transformer $-$ GRU (transformer budget)	$+0.0020$	$+0.0010$	-0.0004
Transformer $-$ LSTM (transformer budget)	$+0.0042$	$+0.0020$	$+0.0006$

10M-tier window ladders. Table 8 tabulates the 10M rows cited in §2.5.

Table 8: 10M-tier window ladders, log-space R^2 on the full test split (seed 42). Each transformer cell sits within 0.006 of its 1M twin; the ridge *level* dips below its 1M twin at every window — the single family-level exception to the scale bound. The two tiers are *disjoint* customer carves and each model is scored on its own carve's test split, so a carve-level offset rides every cross-scale row here: the val$\rightarrow$test drop is larger on the 10M carve by $+0.0024/+0.0044/+0.0029$ at M8/M12/M16, measured on the 1M side at seed 42 and on the 10M side as a 3-seed mean, so the two sides of that offset do not share a seed basis. Subtracting each window's own offset leaves transformer residuals of $+0.0010/-0.0014/+0.0014$, mixed in sign and all under 0.0015. The ridge M16-10M cell is an off-pipeline recomputation with an 8-decimal reproduction gate (see the *Off-pipeline recomputation protocol* paragraph).

Model (10M tier)	**M8**	**M12**	**M16**	**M24**
Transformer (Huber)	0.4757	0.5199	0.5510	0.5860
Ridge (6-feature)	0.4095	0.4732	0.5146	—

Off-pipeline recomputation protocol. The M16-10M ridge cell was recomputed off-pipeline after its training job twice exceeded the wall clock; the recomputation reproduces the adjacent M12-10M

cell's pipeline value end-to-end from raw data to 3.9×10^{-9} (and to 1.7×10^{-11} when refitting on the pipeline job's own saved features). The M12-10M value is read from the completed training step of a job whose redundant eval step failed.

Recurrent and linear grids across scale. GRU at M8 scores $0.4760/0.4744/0.4751$ at 1M/5M/10M, LSTM $0.4734/0.4717/0.4730$. Across the full window grid the GRU and LSTM rows have a maximum within-window spread of 0.004 (18 cells). These cells train from a fixed-size cached draw at every tier (Appendix H) — population composition at fixed exposure, not response to more data; the recipe-equalized 1M retrain at matched MSE (Table 7) sits at most $+0.002$ above the capped own-budget cells. Ridge is within 0.002 from 1M to 5M (5M: $0.4353/0.4951/0.5348$) but not to 10M. At 3 seeds the transformer-Huber M24 rung matches across scales (1M 0.5885 ± 0.0015, 5M 0.5832 ± 0.0039, 10M 0.5885 ± 0.0033, 3-seed mean $\pm$ population SD; per-seed ranges at 5M/10M are wide, $0.0094/0.0072$, so flatness here carries that width). The linear-floor level dip at 10M relative to 1M is window-*stable* across the three windows scored at both tiers ($-0.028/-0.023/-0.022$ at M8/M12/M16), i.e. a scale offset in the linear floor rather than an M8-specific composition artifact. It is the one exception to the scale bound: at 0.022–0.028 it exceeds the ≤ 0.006 sequence-family bound, so the scale bound is a sequence-model result — what stays scale-invariant for the linear family is its window *gain*, not its level. This exception survives the carve correction of Table 8. The dips are $-0.02777/-0.02303/-0.02196$; subtracting each window's own carve offset — the offset is measured on the transformer ladder, so this transfers it across families — leaves $-0.02533/-0.01863/-0.01901$, i.e. the offset absorbs only $8.8/19.1/13.4\%$ of the dip. At 0.0186–0.0253 the corrected gap still clears the ≤ 0.006 bound at every window.

A bounded band is not a resolved zero. Everything above bounds how far a $10\times$ change in training cohort moves the metric; for the sequence families none of it fixes a sign. Each model is scored on its own carve's test split, so every cross-scale row carries the carve-level offset of Table 8, and subtracting that offset leaves transformer residuals mixed in sign and under 0.0015. Separating a cohort effect from the offset needs both arms scored on one common evaluation surface; we do not report such a contrast here, so the sign stays open in this paper. These cross-cohort differences do not bound the isolated effect of training-set size.

GBM feature ablation. No single feature carries the boosted-tree window gain: dropping any one of the six keeps the matched gain at $+0.090$–$+0.098$, while the window-extended activity total *alone* delivers $+0.110$ — the lever rides on more weeks of signal, not feature richness.

Capacity and content detail. The M8 capacity/content cells behind §2.1, all seed 42: d256 0.4803 vs. Huber 0.4771 ($+0.003$); 4 layers 0.4724 (-0.005); +category features 0.4783 ($+0.001$). The same category ablation at 12 and 16 weeks reads 0.5128 and 0.5481 against Huber 0.5256 and 0.5525 (-0.013 and -0.004, seed 42, same carves), so the 0.007 content bound holds at 8 and 16 weeks and fails at 12. The $4\times$-wide, $4\times$-deep M16 variant (d512, 8 layers, effective batch preserved) scores 0.5448 ± 0.0038 (3-seed), 0.006 below the d128 3-seed mean, with the widest seed spread of the capacity cells — same anchored build, hyperparameters not re-tuned for the larger model (§5).

Matched-delta clustered CIs (1M). Paired, customer-clustered bootstrap CIs for matched Δ(M16$-$M8) at 1M (B=1000), all three objectives: Huber $[0.0940, 0.0978]$, MSE $[0.0846, 0.0879]$, pinball $[0.0929, 0.0968]$.

10M matched per-rung detail. Matched levels $0.4757/0.5318/0.5698$ at M8/M12/M16 (seed-42); the 3-seed extension gives 0.47455 ± 0.00083 / 0.53439 ± 0.00193 / 0.56795 ± 0.00134 (3-seed mean $\pm$ population SD). Every rung is composition-deflated: Δ(M16$-$M8) $+0.0941$ $[0.0935, 0.0948]$ matched vs. $+0.0753$ full, Δ(M12$-$M8) $+0.0561$ $[0.0555, 0.0566]$ vs. $+0.0442$, Δ(M16$-$M12) $+0.0381$ $[0.0375, 0.0386]$ vs. $+0.0311$ — as at 1M.

M24 family convergence. Neural Hawkes starts 0.015 below the pack at M8 (0.4620 vs. $\sim$$0.477$; ContiFormer 0.4773) and converges by M16 (0.5474 vs. 0.548–0.552); the convergence continues at 24 weeks — the five recurrent, continuous-time and linear baseline families enumerated here land within $\sim$$0.008$ of the transformer's M24 cell (GRU 0.5874, LSTM 0.5857, ContiFormer 0.5810, Neural Hawkes and ridge both 0.5796 vs. transformer-Huber 0.5869; seed 42, cross-build — within-family curves, not cross-family wins, per §3.2).

B MATCHED-COHORT EVALUATION, IN FULL

Composition statistics. Longer windows sweep in lapsed customers whose only activity sits 58–113 days back: the M16-only 15% of the test set is 30.4% zero-target (vs. 4.8% shared; median 6 events vs. 60), and target log-SD grows $2.16 \to 2.39 \to 2.56$ — $+41\%$ target variance at M16 from composition alone, inflating the R^2 denominator. The swap-in is not one-directional — each window's split also carries rows the other two lack, in near-equal number in both directions (Appendix J) — so the matched cohort is the intersection of three splits rather than the smallest of them.

Cohort construction. The 1M three-window cohort is the 407,338 customer-weeks present in all three M8/M12/M16 test sets (84.7% of each; a 372,778-row 4-way variant adds M24), keyed on (customer, target week), with numerically identical targets across windows (max discrepancy 8×10^{-8}); on the fixed population the denominator is constant by construction (target log-SD 2.1597 at every window).

Per-objective and per-family deltas. Matched vs. full-set Δ(M16−M8): Huber $+0.0959$ vs. $+0.0731$, MSE $+0.0863$ vs. $+0.0800$, pinball $+0.0948$ vs. $+0.0701$ (matched deltas seed-42; full deltas 3-seed, Table 3), ridge $+0.1044$ vs. $+0.0992$ (no bootstrap CI), with matched log-space test MSE dropping monotonically (Table 4). Each of the three per-objective matched deltas carries a paired, customer-clustered bootstrap CI (same customers resampled at both windows, B=1000; Huber $[0.0940, 0.0978]$, all objectives bounded far from zero; 1M CI list in the "Matched-delta clustered CIs" paragraph of Appendix A).

10M cohort. The 10M matched cohort holds n=4,069,475 rows (84.8% of each window); every rung is composition-deflated, with M12 rungs matching (per-rung detail: Appendix A). On the primary population, matched-delta CIs are paired, customer-clustered bootstraps: they are row-sampling uncertainty for fixed trained models at the three shipped test anchors, not intervals over training seeds or calendar anchors. Seed variation is not negligible beside them: at both scales the 3-seed mean matched Huber Δ(M16−M8) sits just below the seed-42 interval ($+0.0939\pm0.0022$ vs. $[0.0940, 0.0978]$ at 1M; $+0.0934\pm0.0005$ vs. $[0.0935, 0.0948]$ at 10M), so each interval is conditional on its seed.

C SATURATION CURVE DETAIL

To rule out carve-splice artifacts we rebuilt the entire curve from the M24 carve alone (a single train set and matched 4-way cohort, with no splice; curve values in Figure 2a), monotone through the old boundary. The marginal week is worth $\sim+0.0001$ R^2 by $k = 22$–24, down from $+0.0039$ at $k = 16$. On the clean curve the saturating form again beats log-linear (fit RMSE 0.0103 vs. 0.0162); the single exponential's underfit (asymptote 0.4227 below the observed 0.4296) is the mixture diagnostic cited in §2.3.

D PER-DECILE REFIT MECHANICS

The within-decile saturation curves (Appendix N.1; Figure 2a inset) are a ridge fit per (decile, k) on train rows binned by input-window activity, evaluated within the matching test decile, so each curve is calibrated to its own subpopulation. Fitted saturation rates are monotone in activity: deciles 6–7 fit $k_{1/2} = 27.1/24.7$ weeks (bends beyond the observable window), deciles 8–10 fit $k_{1/2} = 10.0/5.2/3.9$ weeks, and deciles 1–5 have rate $c \leq 0.004$: their saturating fits degenerate to near-linear (rise $\approx$0.09–0.12 over k=1..16), i.e. no detectable bend within the window. The per-decile refit is the *primary* instrument; a secondary global-model-scored variant is bias-dominated (global calibration artifacts produce spurious early bends in the bottom deciles) and is not used for any claim.

E EXTERNAL-VALIDITY DETAIL (PUBLIC DATASET)

Carve specification. The 3/5/7-day windows are end-anchored to 2017-12-02 00:00 China Standard Time: input windows D3 = [11-29, 12-02), D5 = [11-27, 12-02), D7 = [11-25, 12-02), every carve predicting the identical two-day target — the user's total purchase count over [12-02, 12-04). Users enter a carve with a minimum of three events in its input window (so longer windows admit more users; the matched cohort intersects all three). The raw span opens 2017-11-25 00:00; a sparse noise tail of records past 12-04 is excluded. The linear leg is ridge on four aggregate features — the log-scaled input-window activity total, event count, purchase count, and purchase rate — standardized, with the regularization strength cross-validated over 25 log-spaced values in $[10^{-3}, 10^{3}]$; the target is $\log(1+y)$, the primary transform (its clamp at zero cannot bind on a count, where the primary target is a real-valued total). The transformer legs reuse the primary per-process recipe (2 layers, d_{model}=128, batch 256, 20 epochs, AdamW with cosine schedule; Table 11) with an MSE — and, for the robustness leg, Huber — objective on the log target, trained single-node, three seeds per window.

Day-scale ladder. A day-resolution ladder on the public dataset replicates the saturating *shape* of §2.3: the saturating form beats log-linear ($4.5\times$ lower fit RMSE on the full test split; $2.9\times$ on the matched fit, where the fitted bend moves later) with the bend at or beyond the edge of the observable range ($k_{1/2}\approx$6 days on a 7-day ladder, full-split fit) — the form transfers, but there is no universal horizon constant: sparser signal bends later, consistent with Appendix N.1.

Superseded same-start carves. An earlier version of the window carves shared a start date instead of an end date, letting the 2-day target window shift with the carve; its confounded estimate (Δ(D7−D3) = $+0.0154$) matches the fixed-target one.

Deep-model replication, in full (preregistered, three seeds per window). The transformer leg replicates the direction under a preregistered three-seed design on the end-anchored carves (three seeds per window, matched cohort n=6,985): Δ(D7−D3) = $+0.047$ $[+0.035, +0.059]$ (customer-clustered paired bootstrap). Every seed is positive ($+0.062/+0.019/+0.059$), each with its own clustered CI excluding zero. After excluding one run whose training collapsed to the constant-zero predictor from its first epoch, the window response is monotone D3<D5<D7; the D7−D3 contrast touches no D5 leg and is identical with or without the exclusion. The earlier same-start transformer null is not contradicted — it was a single seed on carves whose target period shifted with the window. This replication carries two caveats. The observed seed spread ($\text{sd}=0.024$) came in $3\times$ the design's power gate, so the test was underpowered at the preregistered, ridge-derived effect ($+0.0154$; post-hoc power 0.29); it lands only because the observed effect is $\sim 3\times$ larger (one-sided t_2 p=0.04 — the inference rests on the clustered CIs, not the seed count). And absolute levels stay near zero (D3 seed-mean -0.007): a window-direction replication on a hostile $\sim$2-day-horizon, 69%-zero target, not evidence of useful absolute accuracy. As carved, longer end-anchored windows also bundle a larger eligible training set (610K→686K rows): the contrast is "longer windows as carved," not a pure-history effect. The nine cells of this design — three windows × three seeds — lie outside the build-anchor ledger of Appendix H: they ran on a single container build, and they are gated instead by an off-pipeline re-score. The evaluation legs retained no per-row predictions, so predictions were regenerated from each cell's own checkpoint; checkpoint identity and elementwise row alignment were then verified on all nine cells, and the regenerated predictions reproduce each cell's banked full-set log-space R^2 bit-exactly on seven of the nine and to within 6×10^{-8} on the other two — a CPU-versus-GPU floating-point seam, inside the re-score's own 10^{-6} specification. The matched-cohort levels and paired contrasts reported in this paragraph are computed from those regenerated predictions.

Objective robustness of the external direction. The direction also survives a training-objective swap: retraining the same cells under a Huber objective — a single-knob change against the same image, data, carves, and seeds — replicates the healthy-cells contrast, Δ(D7−D3) = $+0.045$ $[+0.030, +0.060]$ (customer-clustered paired bootstrap), excluding zero. That paired contrast rests on a single healthy seed-pair; on the fuller healthy basis including replacement seeds, the unpaired window-mean contrast attenuates to $+0.022$ $[+0.011, +0.033]$, direction and significance intact. The catch: 4 of 9 Huber cells collapsed to the constant-zero predictor (vs. 1 of 9 under MSE), worst at the shortest window (3 of 5 D3 seeds, including replacements), so the Huber leg rests

on fewer healthy seeds than the MSE leg. The collapse-rate contrast is itself a finding about loss choice on a $\sim$69%-zero target. Direction-only, as everywhere on this dataset: Huber levels are never compared to MSE levels, and no level is compared across datasets.

Raw-target objective MS-RoPE probe, in full. Retrained under an alternative raw-target objective (MSE on the raw target with $p_{0.5}/p_{99.5}$ winsorization) and scored in a clipped-raw R^2, MS-RoPE underperforms RoPE by 0.054 (M8) and 0.075 (M16) — two-seed means, sign-opposite to its $+0.021$ log-space gain. Per-event *test* loss is essentially unchanged (1296.6 MS-RoPE vs. 1322.1 RoPE at M8; 1271.9 vs. 1268.1 at M16; seed 42); the deficit is concentrated in aggregate calibration, measured as the *dispersion ratio* (prediction SD over target SD, raw space): MS-RoPE's predictions are underdispersed relative to RoPE's at both windows (0.69 vs. 0.77 at M8; 0.70 vs. 0.86 at M16; seed 42) — consistent with an objective–calibration interaction rather than a degraded temporal representation. The window lever transfers under both encodings at both seeds ($+0.070/+0.071$ RoPE vs. $+0.052/+0.047$ MS-RoPE, M8$\rightarrow$M16, in that statistic), attenuated under MS-RoPE in each. At a second seed the reversal keeps its sign and ordering, at 26–66$\times$ the RoPE anchor cells' seed spread; the basis is two seeds, not a full seed characterization. Training on the primary-dataset pipeline is bit-deterministic (the image-swap anchors reproduce their parents exactly). The reversal is the paper's own target-space discipline made concrete: an intervention's sign can depend on the statistic it is scored in.

Transformer null, in full. Scored on a matched cohort of the earlier same-start carves (n=6,822 vs. the 6,985 of the end-anchored cohort — levels are not comparable across cohorts), the transformer reads $0.0472/0.0516/0.0430$ at D3/D5/D7 with every paired window delta's bootstrap CI spanning zero ($+0.0044\,[-0.014, +0.023]$, $-0.0043\,[-0.020, +0.011]$, $-0.0086\,[-0.027, +0.009]$): on this 6,822-row matched cohort at a single seed the deep model is underpowered, and shows no significant window trend in either direction. Its full-test-set numbers appear to *decrease* with window, but that ordering does not survive matching — a cohort-composition artifact of the non-matched carves, not an anti-lever.

F ALLOCATION: FULL PER-DECILE MACHINERY

The activity-stratified curves suggest heterogeneity but do not establish a mixture law. Activity-specific refits keep improving past the pooled plateau, and transformer comparisons change training cohorts (Appendix J), so the pooled plateau does not establish an information ceiling. Refitting the weekly ladder within activity deciles of the matched cohort — a ridge fit per (decile, k), evaluated within its own test decile (mechanics in Appendix D) — shows the saturation rate monotone in activity: for the top decile the half-saturation point is $k_{1/2} = 3.9$ weeks, rising through 5.2 and 10.0 weeks over the next two deciles, while the bottom half of the cohort shows no detectable bend within 16 weeks — still climbing at the window edge (a shape claim, not a level claim: within-decile R^2 levels are much lower than pooled, since activity absorbs most between-customer variance). The pooled 22–24-week bend is driven by the high-activity minority: the highest-activity decile reaches half of its fitted asymptotic gain at 3.9 weeks, light users still gain from longer context within the measured range, and context is worth allocating inversely to activity — the customers a fixed short window starves are the least-observed ones (§2.2). The 32-week extension bounds the prescription rather than extending it: gains from week 24 to 32 are small in every decile ($+0.012$–$+0.021$; every decile's clustered CI excludes zero, including the heaviest). The lowest three deciles average $+0.018$ against the mixture's predicted $+0.045$–$+0.059$. Pooled flatness is therefore shared-model fit allocation, not exhaustion: a single pooled model turns mid-deciles 4–8 significantly negative (inverted-U, vanishing under per-decile refit), unlike inverse context-scaling in time-series forecasting (Ahuja et al., 2026). Process-identification theory offers a possible explanation: longer histories can reduce uncertainty about the generating process even beyond its dependence horizon (Butera et al., 2026). How quickly the residual uncertainty shrinks with further window growth we do not fit. The prescription has a budget reading. Routing by the full-window activity total plus broadly inverse-to-activity feature weeks retains $\geq$99% of the pooled-of-decile-refits instrument's R^2 at 49.2% of the budget; the routing total is uncounted in the budget, and routing alone retains 90.0%. These are instrument-internal values (uniform-24 $= 0.601$, never against the pooled ridge 0.430). The allocation is found by a greedy search on the same matched test rows that score it, so the $\geq$99% is an optimistic reading, and it clears 99% by only $\sim 1\times10^{-4}$ R^2. The payoff is budget, not accuracy (the per-decile oracle

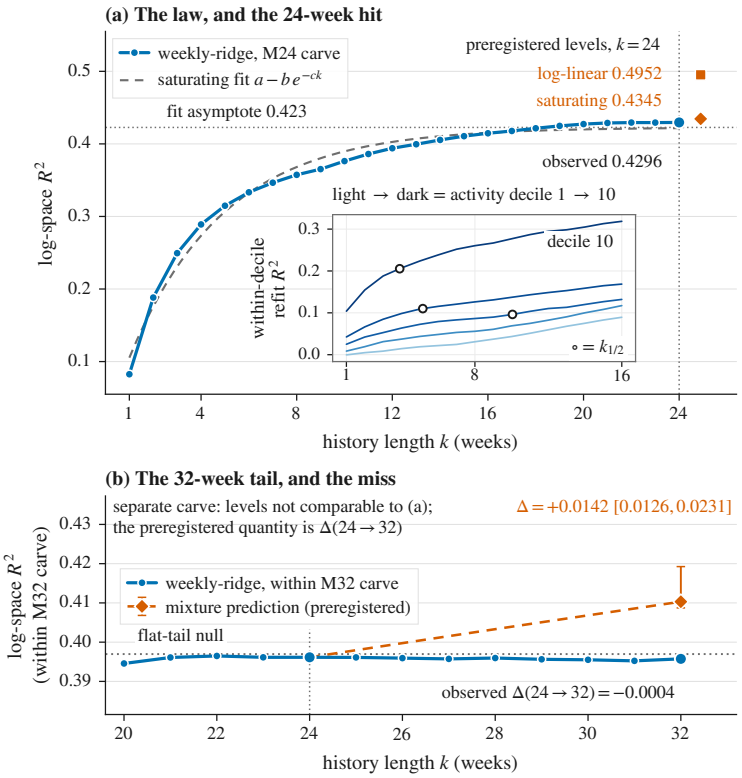

Figure 2: **The preregistration arc: the saturating law, the 24-week hit, and the 32-week miss.** **(a)** Weekly-ridge log-space R^2 vs. history length k=1..24, from the M24 carve alone on the 4-way matched cohort (n=372,778; no train-carve seam). Both preregistered levels sit *above* the outcome: the observed 0.4296 falls 0.005 under the saturating prediction and 0.066 under log-linear (Appendix L). Those levels come from forms fit on the 3-way cohort, not from this curve's own fit — which is why the saturating marker sits above the plotted 0.423 asymptote. Inset: per-decile refit $R^2(k)$ shaded by prior-activity decile; circles mark the half-rise $k_{1/2}$ (3.9/5.2/10.0 weeks for the top three deciles; the bottom half still climbs at the window edge — a shape claim, Appendix N.1). **(b)** The falsification region of the 32-week extension (within-M32-carve instrument, n=351,174; the level offset from (a) is the M32 train basis): predicted $\Delta(24{\to}32){=}{+}0.0142$ $[0.0126, 0.0231]$ (the preregistered band, the only interval drawn; its endpoints are the calibrated and raw 90% bootstrap bounds, taken as the wider envelope, so it is not a single 90% interval), observed -0.0004 (95% CI $[-0.0014, +0.0004]$) from the paired customer-clustered bootstrap of the same instrument (reported, not drawn). That observed interval is row-sampling uncertainty at the three shipped test anchors, not an interval over calendar anchors (Appendix B). The curve walks into the flat-tail null and the interval excludes the entire preregistered range (threshold 0.007; mechanism check fails in the same direction, Appendix F).

coincides with fixed-24), and it is a within-window payoff: by ∼six months the marginal week is worth ≤ 0.003 R^2 in every decile on the 32-week per-decile refit ladder.

G MS-RoPE MECHANISM READOUTS

A dedicated ablation wave at M16 confirms the model-side flatness on the anchored build (all schema content -0.004, anchor feature off -0.003, fully learnable positions -0.0073, a dual mean+median head -0.000; seed 42). The unconstrained control ties one learnable frequency scale to each rotary frequency; MS-RoPE uses learned softmax mixing over four shared, calendar-initialized bands.

Relation to prior time encodings. Prior work encodes time in the attention construction directly: Chen et al. (2026) and Wei et al. (2025) extend the rotary position embeddings of Su et al. (2024) — random Fourier rotations over temporal routines, and joint time-and-order rotary for generative

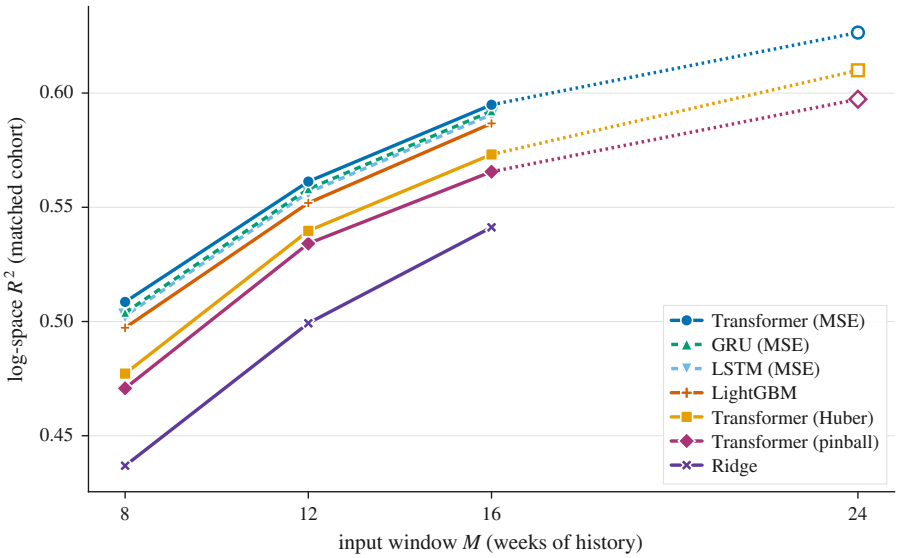

Figure 3: **The input window is the lever, across families.** Matched-cohort log-space R^2 (M8∩M12∩M16, n=407,338; constant target denominator) vs. input window at 1M scale. Every matched family climbs $+0.086$ to $+0.104$ from M8 to M16 (the loss-equalized RNN full-set curves gain $+0.083$), while at any fixed window the spread across the three sequence architectures with a matched leg under a matched objective is an order of magnitude smaller; ridge is the only plotted family meaningfully below the pack (the single-feature linear model has no matched leg and is not plotted). Open markers extend the transformer curves to M24 on the 4-way matched cohort (n=372,778). GRU/LSTM (dashed) are the loss-equalized MSE recurrent families on the *same* matched cohort (3-seed: GRU $0.5039/0.5580/0.5921$, LSTM $0.5025/0.5565/0.5906$; fixed-exposure basis; full-split curves are the conservative companion). Transformer matched points are seed-42 evaluations of the banked cells (the MSE curve's M8–M16 levels are 3-seed-confirmed within 0.0025, Table 2); GBM seed SD is $\leq 1.4\times10^{-4}$. Curves differ in training objective; cross-family *level* orderings are meaningful only within an objective family — GBM optimizes L2 and sits above the Huber/pinball transformer curves but below the MSE transformer, an objective effect, not a family effect (§2.5, Table 5).

Table 9: Time-representation sweep at M16, 1M, Huber objective: 3 seeds per arm, same build (bit-exactness anchored), positional encoding is the only knob varied. The structured physical-time bands beat the integer-position baseline and sit above the unconstrained learnable control on 3-seed means (a margin the control's own seed spread absorbs, so that ordering is not seed-resolvable), with $5.8\times$ lower seed SD than the control (a descriptive point estimate: at n=3 per arm the variance ratio is $F(2,2)\approx34$, two-sided $p\approx0.06$, so the stability ordering is suggestive, not statistically resolved; all $\pm$ population SD, house convention; the Δ column is the mean of paired per-seed deltas $\pm$ population SD).

Time encoding	R^2 **(3-seed)**	Δ vs. RoPE
RoPE (integer positions)	0.5508 ± 0.0016	—
Learnable per-frequency (unconstrained)	0.5628 ± 0.0070	$+0.0120\pm0.0079$
MS-RoPE (4 physical-time bands)	$\mathbf{0.5716\pm0.0012}$	$+0.0208\pm0.0014$

recommendation — while Xiao et al. (2025) combine cycle-aware time encoding with hierarchical multiscale attention; interval-aware self-attention predates the rotary line (Li et al., 2020). MS-RoPE differs from these in its band structure, not in rotating over physical time or in targeting calendar periods: four time-scale bands, initialized at hour, day, week and year and then learned, each rescale elapsed time on every rotary frequency, and each frequency pair superposes the four resulting rotations with softmax-normalized amplitudes. Chen et al. (2026) instead rotate each frequency pair by a single harmonic of a routine period (daily or weekly in their experiments), sampled once from the Fourier coefficients of a periodic attention-modulation kernel and held fixed (a learnable variant in their appendix performs similarly); Wei et al. (2025) give each rotary plane a single time frequency

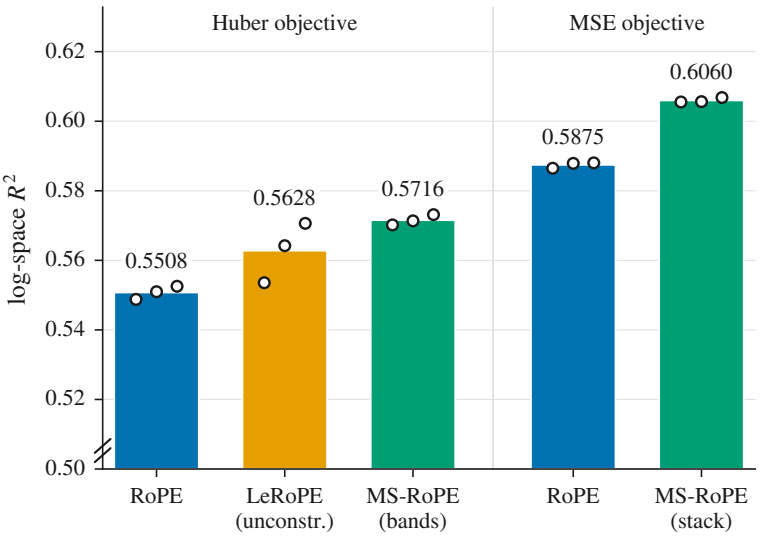

Figure 4: **Structured physical-time position encoding, and its stack with the MSE objective.** M16-1M cells; bars are 3-seed means, dots individual seeds $\{42, 123, 456\}$, including the stack cell (0.6060 ± 0.0006). Under the Huber objective, MS-RoPE's semantic frequency bands beat RoPE ($+0.021$) and sit above the unconstrained learnable control (the "LeRoPE (unconstr.)" bar) on means ($+0.009$, inside the control's seed-to-seed variation), and carry a lower seed SD than the control (σ 0.0012 vs. 0.0070 at 3 seeds, both population SD; a descriptive point estimate, not resolvable at n=3 — Table 9). Stacked on the MSE objective, the two gains are nearly orthogonal (0.6060 observed vs. 0.6083 fully additive, 96% of the additive gain, 3-seed), giving the best M16 cell of this era (the M24 stack sibling reaches 0.6419 ± 0.0024, 3-seed). The y-axis starts at 0.50 (break marked); bar labels carry the values.

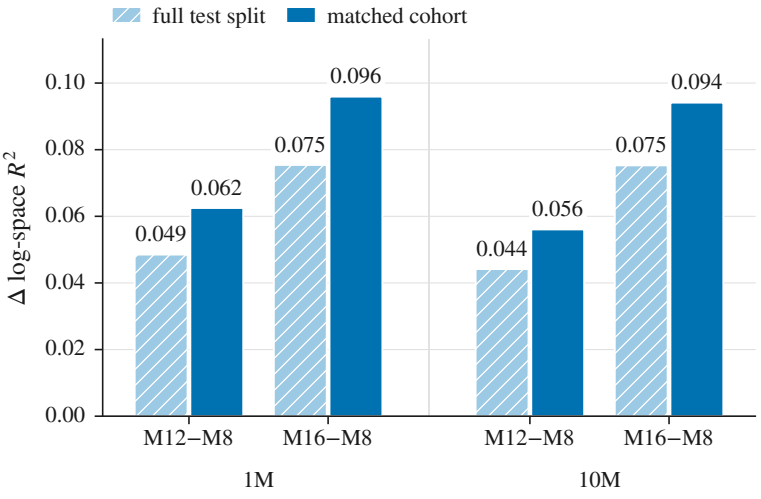

Figure 5: **Cohort composition deflates, not inflates, the window effect.** Full-test-split vs. matched-cohort window deltas (transformer, Huber, seed 42) at 1M and 10M. At both scales the matched delta exceeds the full-set delta on every rung (e.g. Δ(M16−M8) $+0.094$ matched vs. $+0.075$ full at 10M), because longer-window cohorts admit colder customers who raise the full-set squared error (numerator) proportionally more than its target variance (denominator) — so raw cross-window deltas *understate* the window gain, and the matched form is the composition-safe estimate.

from a fixed or learnable bank (e.g., a geometric ladder, optionally with planes dedicated to daily or weekly periods), with time and index fused in the angle or split across planes or heads; and Xiao et al. (2025) learn free frequencies initialized on a DFT grid inside an additive, non-rotary time encoding, taking their temporal scales from hierarchical clustering of event times. MS-RoPE's band structure is

co-observed with the stability result, not shown to cause it. Learnable frequencies alone buy $+0.012$; the band *structure* a further $+0.009$ on 3-seed means — a margin within the control's own seed spread, so the ordering over the control is not seed-resolvable at n=3. Free per-frequency scalars wander from seed to seed; structured physical-time encoding is at least as accurate on 3-seed means as learning the frequencies freely, and lower in seed SD — a point estimate, not resolvable at n=3. The stacked cell reaches its 0.6060 ± 0.0006 against 0.5875 for MSE alone and 0.5716 for MS-RoPE alone (Figure 4; 96% of the additive gain); it is the best M16 cell in this data era, and its M24 sibling under the same stack reaches 0.6419 ± 0.0024 (3-seed). The stack's matched-cohort value sits below its full-split value at every seed — direct evidence the matched protocol is not a one-way ratchet.

On the three MS-RoPE M16-1M checkpoints (one shared tied module, mixing = softmax verified against the forward pass), the learned mixing concentrates on the day and week bands (0.37 and 0.38 of the mass; hour 0.06, year 0.19), and each band's learned timescale stays within a factor of two of its physical prior across all three seeds (contracting modestly and sign-consistently; cross-seed mixing SD ≤ 0.021 per band). Structure and stability are co-observed, so we read the tied band structure as a plausible regularizer rather than a proven-causal one (the control's stability is measured only at the R^2 level). The low hour-band weight is not a timestamp-granularity artifact (events carry real intraday variation), though hour-comb decoherence over multi-day gaps is untested. At the R^2 level the encoding gain persists across the whole window ladder ($+0.015$–0.022 over plain RoPE across the ladder, smallest at M8; seed 42) and mildly amplifies the window gain itself ($+0.116$ M8$\rightarrow$M24 under MS-RoPE vs. $+0.110$ under RoPE, seed 42 both legs): complementary levers, measured jointly. Across the window ladder the mixing shifts from the day band to the week band on the Huber ladder ($0.50 \rightarrow 0.26$ vs. $0.30 \rightarrow 0.51$, strictly monotone, seed 42, single-seed ladder) and the slowest band's timescale tracks the observable Δt range (134d at M8 $\rightarrow$ 241d at M24) — the encoding adapts its resolution to the window, while the plain-RoPE leg bounds this adaptation's contribution to the window gain at $\sim+0.006$ of $+0.116$ (seed 42).

H Reproducibility

Determinism. The pipeline is deterministic end-to-end at fixed seed: independent dispatches of the same cell return bit-identical results. A repeated linear cell reproduces its R^2 to 17 significant digits, and six independently dispatched twin groups across the main grid are bit-identical; twins are deduplicated, never double-counted in a mean.

Build anchoring. Training environments change over a long study; we treat every environment change as untrusted until proven otherwise. Each build in the ledger below ships with a bit-exactness proof against a reference run: a fixed reference cell is retrained on the new build and must reproduce the banked R^2 bit-for-bit (15–17 significant digits) before any cell trained on that build is citeable. Five successive builds passed this gate during the study, so the tested reference metrics agree exactly across builds; this does not establish equivalence for untested configurations. One later build failed the gate — its linear reference cell missed bit-exactness by 3.4×10^{-6} in R^2 — and every cell trained on that build was quarantined and is not cited here; the gate is what makes "untrusted until proven otherwise" falsifiable. Among the anchored cells the one nonzero seam we measured — floating-point reduction order across distributed-evaluation world sizes — is 1.5×10^{-8}, below citation precision; anchor comparisons are world-size-matched regardless. Table 10 is the ledger: the reference values were banked on the study's first build, and each subsequent build in the ledger must reproduce them before any of its cells are cited. Not every cited cell was produced on a ledger build: the transformer cells of the external check (§2.6) ran on a single container build and are gated instead by the off-pipeline re-score described in Appendix E, and the linear and day-scale probes there are offline fits rather than pipeline training runs.

Objective-matched comparison rule, in full. A cross-family delta is admitted without qualification only when both cells match on all five legs: (i) *training-objective family* — at fixed architecture the objective moves log-space R^2 by 0.03–0.05 (M16-1M, 3-seed: MSE 0.5875 vs. Huber 0.5508 vs. pinball 0.5400), more than any architecture intervention we measured. Part of that spread is definitional, and a reader should see which part before interpreting it: on a fixed cohort log-space R^2 is an affine transform of log-space MSE, so the reported metric *is* the squared-error objective's own criterion up to that map. Only that arm is scored on the statistic it was trained to optimize; the Huber and pinball arms are scored on one they were not, and the ordering the spread shows is the

Table 10: Anchor ledger. Each row is one bit-exactness proof: the named build retrained the fixed reference cell and reproduced the banked R^2 to every reported digit (15–17 significant digits, world-size-matched). Reference values were banked on the study's first build; the transformer reference cell is Huber M16-1M (Huber M8-1M for the second build), the recurrent reference GRU M16-1M.

Build	Anchor path	Reference R^2 (banked)	Reproduced
2	transformer (Huber M8-1M)	0.47712987661361694	bit-exact
3	transformer (Huber M16-1M)	0.5525479912757874	bit-exact
3	recurrent (GRU M16-1M)	0.552227258682251	bit-exact
4	transformer (Huber M16-1M)	0.5525479912757874	bit-exact
4	recurrent (GRU M16-1M)	0.552227258682251	bit-exact
5	transformer (Huber M16-1M)	0.5525479912757874	bit-exact
6	transformer (Huber M16-1M)	0.5525479912757874	bit-exact

ordering this metric is built to show. We therefore use the spread only as a comparability fence and make no claim that squared error produces the better model; establishing that would require each arm scored in the statistic its own objective elicits, which is outside this paper. The remaining legs: (ii) *target transform* (log-space throughout; no comparison mixes spaces); (iii) *cohort and era* (one export, matched cohort wherever the delta crosses windows); (iv) *evaluation protocol*, including the distributed reduction order, whose world-size seam is the 1.5×10^{-8} measured above; and (v) *training build*, via the anchor ledger of Table 10. Cells failing any leg are reported as within-family curves rather than as cross-family wins, and where we read across a failed leg anyway we label the failure at the site; this rule is what retires the apparent transformer-over-RNN margin in §2.5.

Seeds. On the primary dataset the seed set is fixed at $\{42, 123, 456\}$. On the external dataset the MSE leg cited in §2.6 is three seeds per window, with one cell collapsed to the constant-zero predictor (excluded from the monotonicity leg only) and its window contrast paired; on the Huber robustness leg four cells collapsed and were replaced at seeds 789 and 1042, so that leg's healthy bases differ per window and its fuller-basis window contrast is reported unpaired, attenuating to about half the paired point. Per cell class: the M8, M12, and M16 window rungs are 3-seed at all three population tiers; the M24 rung is 3-seed at all three tiers (at 5M/10M with wide per-seed ranges, 0.0094/0.0072); the 32-week extension cells are cited at seed 42 only, and the capacity probe is 3-seed at 16 weeks and single-seed at 8. On the matched cohort, the headline Δ(M16−M8) contrast is 3-seed at both 1M and 10M (as is the stack cell); the MSE matched transformer columns are 3-seed (seed-extension evaluations gated bit-exact against the banked cells); the Huber matched M8–M16 cells in Table 2 are 3-seed; the pinball matched columns, the matched M24 entries, and the matched Δ(M16−M8) pairings remain seed-42 evaluations of the banked cells. Cells marked 3-seed are 3-seed mean $\pm$ population SD over $\{42, 123, 456\}$, printed with $\pm$ where the per-cell SD is given and with the spread stated in the caption otherwise; all remaining cells are single-seed point estimates, flagged as such where cited. At n=3 the printed population SD is $\sqrt{2/3}\approx0.82\times$ the sample SD and is not a standard error: a two-sided 95% t_2 interval on a 3-seed mean has half-width $\approx3.0\times$ the printed SD. Seed 42 is shared by every model family, no cell is re-run for a better draw, and single-seed cells never anchor a claim that seed dispersion could plausibly overturn — with one disclosed exception, the seed-42 architecture band, which does not hold on the baseline-only three-seed basis (Appendix I).

Training recipe. Table 11 collects the fixed recipe behind the reported transformer grid, the baseline legs' differences, and the tabular baselines' inputs. No value was tuned per cell: within a comparison only the stated knob under study (window, objective, positional encoding, tier, seed) is set differently. Where setting that knob also moves something we did not set — the realized learning-rate schedule across population tiers — Appendix I says so. The recipe was fixed before the grid was run and held constant across it, so no architecture or optimizer value could be chosen in light of a result reported here.

Parameter counts. Table 12 reports per-family parameter counts, read directly from the banked `best_model.pt` state dicts of the cited M16-1M cells (counts are architecture properties, invariant across window, seed, and objective; totals include small non-trainable buffers such as rotary frequency caches, $<0.1\%$). The recurrent baselines are 2.4–3.1$\times$ *smaller* than the transformer and Neural Hawkes smaller still — so the cross-family convergence band is not a parameter-parity artifact in the transformer's favor.

Table 11: The fixed training recipe. One recipe serves every transformer cell in the reported window, objective, encoding and tier grid; the capacity probe of Table 12 varies width and depth at the same effective batch. The recurrent and continuous-time baseline legs differ as shown (their as-dispatched budget difference is disclosed in the compute paragraph below); the tabular baselines take six aggregate features over the input window, the boosted-tree legs using unnormalized counts and sums where the linear legs use rates, means and SDs.

Transformer cells	
Architecture	d_{model}=128, 2 layers, 4 heads, dropout 0.1, attention pooling
Optimizer	AdamW, learning rate 5×10^{-4}, weight decay 0.01
Schedule	100-step linear warmup followed by a cosine schedule whose period overshot the realized 1M step count by $\sim10\times$ (Appendix I), preceded by a 200-step calibration phase in which only the input-activity baseline parameters receive gradient; gradient-norm clip 1.0; no mixed precision
Batch / epochs	256 per process (32 ranks, $\approx$8,192 effective), 20 epochs
Huber δ	1.0 (in the log-target space)
Pinball τ	0.5 (the median; the same τ serves the pinball-objective cells and the auxiliary per-event head)
Training objective	a fixed-weight sum of the swept loss on the aggregate prediction and an unswept auxiliary per-event loss; the "objective" lever moves the first term only
Event featurization	per event: a type code and five categorical attribute codes (embedded at dimensions 4–16) plus two numeric per-event value fields, concatenated with two declared time-feature channels that are identically zero on this export (continuous-time paragraph below); most recent 750 events (cap paragraph below)
Recurrent / continuous-time baseline legs	
Optimizer / schedule	AdamW, learning rate 3×10^{-4}, weight decay 0.01, cosine annealing
Batch / epochs	128, single process, 10 epochs
Tabular baselines (six aggregate features)	
Features	input-window activity total, event count, per-event activity sum, purchase rate, mean and SD of per-event value — all computed over the input window
Ridge	features standardized; regularization strength cross-validated over 25 log-spaced values in $[10^{-3}, 10^{3}]$, closed form
GBM	gradient-boosted trees, 400 trees, learning rate 0.05, 63 leaves (fixed, untuned)

Table 12: Per-family parameter counts, read from the banked checkpoints (M16-1M cells). Tabular models listed by their fitted degrees of freedom; LightGBM is fixed-hyperparameter and untuned.

Family	**Parameters**
Transformer (d128, 2 layers)	491,248
+ MS-RoPE variant	491,284
GRU (2-layer)	158,851
LSTM (2-layer)	209,027
ContiFormer	405,523
Neural Hawkes	93,443
Capacity probe (d512, 8 layers)	26,479,216
Ridge (6 features + intercept)	7
Single-feature linear	2
LightGBM (400 trees $\times$ 63 leaves; fixed, untuned)	$\leq$25,200 leaf values

The two continuous-time legs are simplified re-implementations. ContiFormer here is a causal attention stack with an additive per-head exponential time-decay bias on the attention logits: there is no ODE solver, and so no solver tolerances. Neural Hawkes here is a decay-gated LSTM cell: there is no intensity function, no likelihood term and no Monte Carlo integration, so it is a decay-gated recurrent regressor rather than a point process. Both train on the same aggregate regression loss as the other baselines. Read the architecture band of §2.5 accordingly: it compares this transformer against these implementations, not against the published models.

Both continuous-time legs ran with no inter-event interval. There is a stronger simplification than either of those, and it is not a design choice: on this export both legs ran with $\Delta t \equiv 0$. The only time channel either one reads is a two-column per-event tensor projecting a day-of-week and a week-index field, and neither field exists in this extract's event struct, so the tensor is allocated and then filled with zeros on every event. We also verified this empirically on the 16-week 1M carve: over $15{,}041$ events across 200 sequences both columns are identically zero. The other carves were not probed; the absence of any producer for either field in the code path is carve-independent. The export does carry a per-event timestamp, but neither leg consults it. The consequences are exact rather than approximate. ContiFormer's bias, $\exp(-|\alpha_h|\,|\Delta t|/3600) + \beta_h$ for learned per-head α_h, β_h and $|\Delta t|$ in seconds, becomes $\exp(0) + \beta_h = 1 + \beta_h$: a per-head constant added to every logit in a row, which cancels in the softmax, and a bias with no α_h dependence left in it. Both receive identically zero gradient, and what remains is a plain causal attention stack. Neural Hawkes' decay factor becomes $\exp(0) = 1$, so the hidden state is not decayed between events and the decay gate's parameters are likewise gradient-dead, leaving a standard LSTM cell. So the architecture band of §2.5 does not test continuous-time mechanisms on this data. What it reports is that two architectures whose time-handling was inert still land within ~ 0.005 of the others at 16 weeks on the seed-42 basis, where all five arms are matched — a weaker claim than the family names suggest, and it does not hold on the baseline-only three-seed basis, where GRU and ContiFormer bind a 0.0059 spread (Appendix I). No measured value moves: the channel is equally dead in all five families, the transformer included, whose own event-time information comes from rotary position encoding over event order rather than from these channels. Populating that channel is the obvious extension, and it is not done here.

Compute resources. With the disclosed off-pipeline recomputations as exceptions (§2.5), cells ran as managed cloud training jobs on nodes of 8 V100 GPUs each. Primary-population *transformer* training legs use four-node distributed data parallelism (32 ranks); the recurrent and continuous-time baseline legs (GRU, LSTM, ContiFormer, Neural Hawkes) ran single-node, single-process, as did the external-dataset cells, evaluation legs, and the 10M linear refits — which also fixes the reduction order behind the world-size seam above. The as-dispatched training budgets differ across the compared legs on three axes: batch, epochs, and data exposure. Transformer cells train at a per-rank batch of 256 over 32 ranks ($\approx 8{,}192$ effective) for 20 epochs, streaming the full train set over training (a rotating near-full draw each epoch). The baseline legs train at batch 128 in a single process for 10 epochs, reading only a fixed RAM-cached draw of at most $384{,}000$ training rows ($3{,}000$ cached batches, reused every epoch, at every population tier — $\sim 1/13$ of the 1M-tier train set and a smaller fraction at 5M/10M). The data-exposure axis biases *against* the baselines — they land within ~ 0.005 of the transformer at 16 weeks on a thirteenth of the data, Neural Hawkes ~ 0.015 below it at 8 weeks — so the cross-family convergence band survives it. At the 1M tier we have since removed the axis directly: retraining both recurrent families at the transformer's full budget and near-full data exposure moves each family$\times$window cell by at most $+0.002$ (§2.5). Calibrated cross-family deltas at 5M/10M and the baseline scale comparisons (Appendix A) still inherit the fixed exposure cap alongside the batch and epoch differences. Wall-clock scales with population tier: at the primary recipe (d128, 2 layers), 20-epoch 1M-tier training cells complete in 2.2–2.6 hours at 16–24-week windows (5-epoch variants in under an hour), with evaluation legs of 11–15 minutes; 10M-tier transformer cells run to roughly a day under a 30-hour wall-clock limit and the single-node 10M linear refits take ~ 20 hours. The full campaign comprises several hundred training and evaluation jobs in this configuration space; we do not report a total-GPU-hours figure, but every cited pipeline cell's execution identity, build digest, and configuration hash are retained with the experiment records backing the anchoring protocol above.

Input truncation cap. Event sequences are truncated to 750 events at every input window — the most recent 750 for every model trained through the data pipeline. The cap binds for a window-increasing fraction of test customers — 0.22% (M8), 0.84% (M16), 2.37% (M24), 3.93% (M32) — and on the matched cohort 3.0% of customers were already cap-hit at M24, receiving no new input from the longer window. On a three-of-forty-bucket sample of the narrower M24$\cap$M32 cohort, 24.2% of the incremental $24\rightarrow32$-week event mass is truncated, concentrated in the heavy tail; that sample is less active than the scored cohort, so it likely understates the truncated share. A fixed input cap therefore attenuates event-weighted readings of the longest-window gains (§5); the ridge instrument is uncapped and replicates the saturation. Two direct tests bound the cap's effect: removing the capped customers leaves the matched gain unchanged ($+0.0115$, customer-clustered 95% CI $[0.0101, 0.0130]$, vs. $+0.0112$ full), and a preregistered cap-raise intervention bounds it from the other side — retraining both cells with the cap doubled to $1{,}500$ events shifts the matched $\Delta(24\rightarrow32)$ by $+0.0020$ (customer-clustered 95% CI for the shift $[0.0013, 0.0026]$), inside its ± 0.004 prediction

band, leaving residual cap-hit of $0.48\%/0.99\%$ (single seed; the pair trains at a smaller effective batch, so only the within-pair contrast is read).

Code and data availability. The 1M/5M/10M event-sequence corpora are proprietary customer behavioral data and are not releasable; the Taobao corpus is public. No data, models, or individual-level predictions are released — every reported result is a population-level aggregate, and the tables and figures print those aggregates. Some claims summarize a group of cells rather than printing each cell's value. The four preregistration artifacts are content-addressed, with SHA-256 digests printed in Appendix L.

I SCHEDULE MATCHING ACROSS THE BOUNDED AXES

The architecture band compares families that did not share a learning-rate schedule. The band is a spread across families, one of which was optimized differently. The transformer legs ran at an effectively constant learning rate: a cosine decay was configured, but the period it was given overshot the realized optimizer-step count by about ten times, so the schedule ended close to its peak rate. The recurrent and continuous-time legs set their cosine period to the epoch count and step once per epoch, so they annealed in full. Annealing generally helps a final metric, so the asymmetry handicaps the transformer against the families the band brackets it with. That pulls the families together, which is the direction that would manufacture this band rather than conceal one. Against it, the transformer trains twice as many epochs, so the net sign is not settled by argument. We report the band as measured under the schedules that ran, and do not claim it would survive a matched one. The two continuous-time families' three-seed extensions landed after this bound was committed, and they do not leave it intact. Restricted to the four baseline families, which are mutually matched, the three-seed spread is 0.0059 — wider than the ≤ 0.0052 band the abstract reports, with GRU and ContiFormer binding it. Two things follow, and the first is the uncomfortable one. The schedule exculpation just given does not cover that pair: both annealed per epoch and both were scored at the last epoch, so the asymmetry that would pull the transformer toward the baselines is absent exactly where the bound fails. Second, the transformer's own three-seed cell cannot simply be added to that spread. It ran on a different training container build from the four baseline arms, and that is a property of how the job was launched rather than of the recipe, so a transformer-against-baselines three-seed contrast is not a matched comparison to begin with — which is why the figure above is the four-family one and why we do not print a five-family three-seed spread at all. We report the band on the seed-42 basis, where all five arms are matched, and record here that it does not hold on the baseline-only three-seed basis. It is a measured interval either way, not one of the four registered commitments of Appendix N.2.

Checkpoint selection does not use the reported metric, and at its widest that channel exceeds these bounds. The epoch each number comes from was chosen by minimizing validation error on the *original* scale, while every figure we report is in log space. The log-space validation series was recorded at every epoch and was not the criterion, and re-selecting on it moves cells by $+0.002$ to $+0.009$ — the top of that range wider than any bound in this paper. One cell was re-evaluated end to end against the full test split to check this is not an artifact of the validation data, and the two checkpoints differ there by $+0.0022$, tracking the validation-side estimate. The effect does not cancel out of the contrasts, because it is not equal across cells. Two defenses apply and one does not. On the scale contrast the offset is nearly common-mode — on validation the two tiers of the 8-week cell move by $+0.0043$ and $+0.0041$ — so the ≤ 0.006 sequence-family bound is close to selection-invariant; and a retrain that corrected the schedule and switched selection to log space together leaves the capacity magnitude bound unchanged, with the sign not reproducing on three-seed means. The architecture band has no such defense: the recurrent and continuous-time legs select no epoch at all, persisting their final one, and the tabular legs have no epoch to select, so that band compares a validation-selected transformer against unselected comparators. The one counterfactual reachable without a baseline-side series — stripping selection from the transformer rather than granting it to the comparators, since the comparator cancels from the difference — puts the as-run rule's worth at no more than $+0.0004$ on validation over the six M16-1M cells (two objectives, three seeds), negative on three of them. We therefore read the architecture band as inclusive of selection sensitivity rather than corrected for it.

The population-scale comparison is not schedule-matched either. The same period is used at every population scale, but the number of optimizer steps is not: it grows with the training set. At 1M the

schedule covers a tenth of its period; at 10M it covers all of it and then overruns, and because the cosine is periodic rather than clamped the rate turns back up over the final epochs. So the scale bound compares two optimization regimes rather than one knob. The schedule mismatch has an unmeasured effect on the scale comparison; its direction requires a matched re-run. We nonetheless state it for what it is, an unmatched contrast. Three things this does *not* reach. The window lever is measured within a scale, and window length does not change the number of training rows — the 8- and 16-week cells at 1M run the same number of steps and end within four significant figures of the same rate — so every window delta is schedule-matched. So are the capacity cells inside the transformer family, which share both period and effective batch, and the tabular baselines, which have no learning-rate schedule at all — ridge is closed form, and the boosted trees use a fixed, unannealed shrinkage rate.

J COHORT DISJOINTNESS AUDIT

Split sizes. Test splits hold 480,570/480,748/480,673/480,652 rows at 1M (M8/M12/M16/M24), $\approx$2.4M rows at 5M, and 4,799,661/4,799,870/4,800,091/4,800,107 at 10M (same window order) — the per-cell n_{test} figures of §2. Split sizes in *customers* are given as ratios rather than absolute counts, matching the convention that tier names are bucket labels and not counts (§3.1). Within the 1M tier the test split holds $\approx 1/6$ as many unique customers as the training split, at every M. Both splits widen as the window grows, so a longer window sweeps in lapsed, colder customers (§2.2), and from M8 to M24 the training split's unique-customer count grows by $\approx 10\%$ and the test split's by $\approx 4\%$. The training-side figure is the one the headline table's train-side eligibility caveat cites; the test-side drift is why every cross-window contrast is read on the matched cohort — the intersection of the window carves — rather than on each carve's own test split. The 5M and 10M tiers carve on the same recipe with the same split boundaries and are audited directly below.

Not nested. The window splits are not nested: keyed on (customer, target week), the 1M M8 split holds 73,232 rows the M16 split lacks and M16 holds 73,335 that M8 lacks, with the same signature at 5M and 10M. Two strata of near-equal size in opposite directions, against a row total flat across windows while the eligible pool grows, is the signature of a draw to a fixed size per window rather than of an eligibility rule monotone in M, which could only move rows in one direction. We did not read the producer's row-selection step, so we report the join and not the rule; the three-window intersection is therefore a fraction of *each* split rather than all of the M8 one (Appendix B), and every matched-cohort claim conditions on the enumerated intersection of the windows under comparison (§5).

Direct-join audit. We verified split disjointness by direct join rather than trusting the pipeline: over all four 1M carves plus 5M and 10M spot checks, train$\cap$test and train$\cap$validation contain zero shared customers and zero shared (customer, target week) rows — twelve pairwise checks, all exactly zero — with zero duplicate rows inside any split. An independent audit of the M8/M16/M24 splits reproduced the result (overlap 0, duplicates 0, full join, no sampling). Tier disjointness holds by construction: the tiers are non-overlapping sampling-bucket ranges, so no customer can appear in two of them.

K MATCHED-ESTIMATOR NEGATIVE CONTROL, IN FULL

Over 1,000 customer-level swaps of the M8/M16 prediction vectors the estimated Δ(M16$-$M8) centers at zero (mean 6×10^{-6}, SD 0.0011) while the observed $+0.0959$ exceeds every permuted draw (Monte Carlo permutation $p{=}1/1001$), and the clustered-CI procedure ($B{=}500$ per dataset) excluded zero on 7.5% of 200 permuted datasets — consistent with the nominal 5% (exact $p{=}0.078$), conditional on the fixed seed-42 pair: the control validates the estimator's calibration, not seed variability.

L PREREGISTRATION MECHANICS

What each commitment could have shown. The four are not equal instruments. Commitment (i) is a threshold rather than a form comparison, so it can only fail low, and the full-set deltas already banked at commit time were above the bar (Table 1); on the M8–M16 rungs the full-set read is the conservative one, so the matched read the threshold was scored on is the larger of the two. Commitment (ii) separates two forms that were nearly tied on the data they were fit to, so it tests

whether the better in-sample form extrapolates rather than which form fits, and its cohort-invariant gain scoring is a nearer miss. Commitment (iii) was fully blind — no 32-week job had run at commit — and it is the one the data falsified. Commitment (iv) is scored as committed on a single-seed measurement carrying the train-carve seam of §2.4, where an offset of the class measured on the ridge sub-ladder would put the true gain near the log-linear prediction. The discriminating weight of the record therefore sits on the blind registration; we do not spread it evenly over four.

Commitment (ii): dual-basis scoring. The saturating and log-linear forms were fit on the 3-way matched cohort and scored on the 4-way cohort, whose k=16 level sits 0.028 lower ($0.4426 \rightarrow 0.4146$). In level space the observation lands within 0.005 of the saturating prediction; on the cohort-invariant *gain* k=16→24 the observation ($+0.015$) sits between the saturating (-0.008) and log-linear ($+0.053$) predictions, nearer the saturating form, and the within-carve k=22→24 marginal ($+0.0001$/week) is unambiguous on either scoring. We report both scorings rather than the more flattering one.

Commitment (iv): registration disclosure and outcome. At registration time the 32-week transformer's composition-confounded full-set R^2 was already banked; this is disclosed in the registration, and no matched-cohort number existed for it. The observation, $+0.011$ (single seed), lands on the saturating form. Unmatched, the full-set gain reads $+0.024$: composition inflates it two-fold, the mirror image of the M8→M16 case where composition *masked* the gain — the matched framework corrects in whichever direction composition biases.

Registration provenance. The four registration artifacts are content-addressed; their SHA-256 digests (16-hex prefixes; full digests retained with the experiment records) are `504e309ee7896795` (i, 2026-07-11), `13cb1b424dcbc911/1992d7c8fbc0d9e4` (ii, 2026-07-13), `367b1328d37ee12b` (iii, 2026-07-15), and `b8bf45d1fa5bac72` (iv, 2026-07-17). The registrations are *internal and self-attested*: records retained with the experiment logs, not deposits with an external registry, and not third-party timestamped. The printed digests make each commitment verifiable against those records, and the disclosure column of Table 1 states what was already known at each commit time, with commitment (i)'s matched-ridge value recorded in Appendix N.2; readers should weigh the record accordingly. The count of four is the count of *content-addressed* commitments, each with a printed digest, and not a census of every prediction recorded over the study. Predictions made without such an artifact are not scored in Table 1, and a reader counting commitments should read the table as that narrower thing.

M Related work

Matched evaluation and matching estimators. Matching in causal inference builds comparable groups so an effect estimate is not a composition artifact (Rosenbaum and Rubin, 1983; Stuart, 2010); shift corrections reweight training (Shimodaira, 2000; Quiñonero-Candela et al., 2008). We borrow the diagnosis, not the estimator: exact test-set intersection, identical targets, no counterfactual claim (§3.3, Appendix B). Concurrent and prior work documents evaluation protocols that flip comparisons: a window-sufficiency curve on subscription churn data inverts when cohort and target move with the window but rises when both are held fixed (Han et al., 2026), untuned look-back inverts architecture rankings (Abdelmalak et al., 2026), and the same measure-on-the-time-axis instinct reranks generators for synthetic sequential data (Kwon et al., 2026).

Context length. Context extension asks how a trained model accepts longer inputs (Chen et al., 2023a; Peng et al., 2024; Ding et al., 2024) and how models use nominal length (Liu et al., 2024); our question is prior to both — whether the extra input is worth conditioning on at all. The horizon question is not new (Ballings and Van den Poel, 2012), but the lifetime-value literature (Fader et al., 2005; Wang et al., 2019; Li et al., 2022) typically fixes the feature window and varies what sits downstream of it; our delta is the control, not the question (§2.1).

Scaling laws. The scaling-law literature fits smooth forms to performance as parameters, data, and compute grow (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022) — exactly derivable for linear models — with context at most one capacity knob among many; the behavioral foundation-model scaling study closest to ours sweeps parameters, data, and sampling, fixes training context length, and examines evaluation-time sensitivity to this axis (Brüel Gabrielsson, 2026). Related

regression theory derives risk decompositions and asymptotic risk formulas for prediction-only self-distillation (Dang et al., 2026). We retrain across calendar observation windows at fixed capacity, recipe and target (the window as carved: a longer window also admits colder customers, §2.2), and the better-fitting of the two preregistered forms saturates: the pooled weekly information curve bends at 22–24 weeks and is flat through 32; Shi et al. (2024) derive regime-dependent relationships among optimal look-back, dataset size and model capacity, and our window gain is scale-flat within a $\sim$0.005 R^2 resolution over the tested decade — with the unbounded log-linear alternative disfavored by criteria committed before the deciding evaluations ran (§2.3). Our fixed-window contrasts do not directly estimate changes in the optimal look-back.

Preregistration in machine learning. Preregistration has been proposed for NLP (van Miltenburg et al., 2021) and piloted in a dedicated workshop (Bertinetto et al., 2021), largely as registered hypotheses and analysis plans; our four commitments are quantitative — a survival threshold and three numeric predictions, scored against criteria committed in advance (one hit dual-scored across a disclosed cohort shift) — and the one that missed is reported against its own falsifier rather than refitted around (Table 1, §2.4).

Event-sequence models. The natural models for irregular event streams are temporal point processes and continuous-time sequence models (Mei and Eisner, 2017; Zuo et al., 2020; Chen et al., 2023b; Shchur et al., 2021); such streams are also pretraining substrate for clinical foundation models (Li et al., 2026). The point-process and continuous-time families enter our grid as baselines on the same cohorts — the two continuous-time ones as simplified re-implementations whose inter-event interval was identically zero on this export, leaving their time mechanisms inert (Appendix H) — and family choice is second-order to the conditioning set: five sequence families land within $\sim$0.005 at 16 weeks, and the decay-gated recurrent leg starts 0.015 below the pack and converges as the window grows (§2.5).

N Conventions, allocation, preregistration, and the matched-cohort protocol

Vocabulary and conventions. Recurring terms, defined once: a *carve* is a fixed cohort extract cut from the platform's raw event-stream export at a given window and tier (§3.1); a *cell* is one training or evaluation run at one grid point (family, window, tier, objective) at a given seed; a *rung* is one point on a sweep ladder; a *leg* is one arm of a comparison or of a pipeline (a training leg, an evaluation leg); *banked* means recorded in the study's provenance ledger when first produced, and cited from that record; an *instrument* is the measurement model a curve is read from (e.g. the weekly-aggregate ridge probe of §2.3); a *seam* is any environment or data-basis difference — container build, evaluation world size, train carve — between compared cells. Values listed as $x/y/z$ (or $x/y/z/w$) are at M8/M12/M16 (M24) in that order unless labeled otherwise.

N.1 Where the history goes: the bend moves with activity

Within the window the pooled saturation law behaves as a mixture, and the bend moves with activity: per-decile refits give the top decile a half-saturation point of $k_{1/2}$=3.9 weeks, rising through 5.2 and 10.0, while the bottom half of the cohort shows no detectable bend within 16 weeks (a shape claim, not a level claim) — so within the window, context is worth allocating inversely to activity. The 32-week extension bounds the prescription (gains are small but positive in every decile, $+0.012$–$+0.021$, clustered CIs excluding zero), and an activity-routed window policy (an external, hard-partition analogue of behavior-conditioned forecasting; Soleimani et al., 2026) retains $\geq$99% of this instrument's own fixed-24 value (uniform-24 $= 0.601$, not the pooled ridge 0.430) at 49.2% of the feature-week budget, with routing alone already retaining 90.0% on a routing scalar the budget does not count — a budget payoff, not an accuracy one. Full per-decile machinery, the pooled-model interference check, the process-identification reading, and the budget reading's mandatory disclosures: Appendix F.

N.2 PREREGISTRATION

Four commitments were registered before their deciding evaluations ran (van Miltenburg et al., 2021); each fixed a discriminating criterion in advance, and all four outcomes — predictions, falsifiers, observed values, verdicts, and what was known at each commit time — are scored against those pre-committed thresholds in Table 1, including commitment (iii), whose prediction the data *falsified*. In brief: (i) a composition-survival threshold for the matched window gain ($\geq +0.05$ log-space R^2, set before the deciding transformer re-evaluation; the registration file records the bar together with a matched ridge $+0.104$ that already cleared it, and self-describes as updated after that ridge re-evaluation landed); (ii) saturating vs. log-linear predictions for matched $R^2(24)$, fit before any 24-week evaluation (fit on the 3-way cohort, scored on the 4-way — a disclosed shift, both scorings reported, Appendix L); (iii) the mixture reading's prediction of a $+0.0142$ pooled gain from week 24 to 32 (predicted band $[0.0126, 0.0231]$, its endpoints the calibrated and raw 90% bootstrap bounds), with a falsifier of < 0.007 that would reject the mixture in favor of a flat saturated tail; and (iv) saturating ($+0.012$) vs. log-linear ($+0.041$) forms for the 32-week transformer's within-cohort gain, committed before any matched-cohort number existed for it. Registration mechanics, including commitment (iv)'s pre-banked full-set disclosure, are in Appendix L.

What to do with the lever. Extending the observation window outweighs everything else measured — at every scale, in every family. Where both sides are measured, at 1M, the window gain is $\sim$10–20$\times$ the capacity and customer-scale bounds and $\sim$3–5$\times$ the best time-representation intervention (the MS-RoPE gain of $+0.0208 \pm 0.0014$ on three-seed means at M16, against the window's $+0.07$–0.107; Table 9) — and the gain is information, not fit: it grows under matched evaluation, and the pooled probe saturates where predicted. Lengthen the window toward the bend, match the objective to the evaluation statistic; expect more customers, parameters, and depth to buy very little — a bounded band, not a measured zero (Appendix A).

Allocation, within its window. The top activity decile half-saturates within a month, the next two at 5.2 and 10.0 weeks; light users keep gaining through at least 16 weeks; past 24 weeks residual gains are small everywhere — so inverse-to-activity allocation is a first-few-months rule, not an unbounded prescription.

Matched-cohort protocol. We key every test row on (customer, target week), intersect the test sets of all windows under comparison, and evaluate every window's model on exactly that fixed set — n=407,338 rows (84.7% of each 1M test split — the splits are not nested, Appendix J) for the three-window cohort, a 372,778-row four-window variant, and n=4,069,475 (84.8%) at 10M. Targets agree across carves to floating-point noise on the three-window cohorts (max discrepancy 8×10^{-8} at 1M and 1.1×10^{-7} at 10M; on the four-window variant 109 of 372,778 rows carry an M24-versus-M8 carve drift up to 0.31 in log space, worth 6×10^{-7} on the delta, and re-scoring the Huber delta on both target vintages leaves it unchanged at four decimals) and the denominator is constant by construction (target log-SD 2.1597 at every window on the 1M three-window cohort, 2.1592 on the four-window variant), so any climb is error reduction on the same customers with the same denominator.

Two gates precede every matched number computed from saved per-row predictions: *alignment* — per-row prediction arrays are re-keyed by reconstructing the evaluation order, and recomputed targets must equal the saved arrays elementwise (bit-exact: max difference 0.0) — and *reproduction* — full-set R^2 recomputed from those arrays must return the banked training-run value (within 4×10^{-8} at 1M; within 5×10^{-9} at 10M) *before* the cohort mask is applied. A matched evaluation failing either gate is discarded, not repaired. The ridge and GBM matched cells are offline refits with no retained per-row arrays; they are gated instead by full-set reproduction of the banked cells — ridge within 1.1×10^{-5}, GBM as described in Appendix A.