

A Fail-Closed Forecast-Quality Validation and Guard Methodology for Data-Scarce Demand Forecasting

Indivd Analytics

September 21, 2026

Abstract

This methodology note is written for analytics practitioners, decision-makers, and applied forecast researchers who run demand-forecasting pipelines over many small or sparse locations. Its contribution is a fail-closed validation-and-guard scheme for such pipelines, formalized for a demand-forecasting platform whose per-entity signals are often data-scarce and whose served forecasts periodically degrade into stale, degenerate, or scaling-bias regimes.¹ Every guard is grounded in reproducible, read-only evidence drawn from an anonymized production deployment, spanning an observed multi-failure window (2026-09-05) and a later fresh recomputation (2026-09-07). The scheme couples: (i) a data-reality sufficiency gate (non-zero target density, not the over-loose ≥ 14 genuine days/60d false-positive benchmark); (ii) a serve-time degenerate/flat guard, fail-closed on a flat signature or $\text{smaPE} > 95$, deliberately not $\text{smaPE} > 50$, so that scaling/bias-class forecasts (shaped but wrongly scaled) keep serving; (iii) a promotion/activation guard ensuring broken retrains cannot displace healthy siblings; (iv) a pair of healthy-cohort regression + delta gates, deploy-blocking on relative regression or flat degensification, not on an unmeetable absolute broken-share bar; and (v) bias-class vs. flat-class separation. On the current serve-view the methodology collapses the dense-cohort (high-positivity) broken share to 0% while flagging the 73 sparse + 35 all-zero + 5 no-data entities whose broken rates are an artifact of data reality rather than model quality. Three addenda complete the paper: a production training-sufficiency addendum that formalizes the data-reality predicate of Section 2.1 (Section 7); a validation replay of the guard tuple across the extended observed horizon (Section 8); and an empirical coverage-calibration study of the served forecast band with an honest at-scale verdict (Section 9).

1 Problem and Motivation

Per-entity demand forecasts drive production decisions, but the underlying signal is heterogeneous: some locations see a dense positive series while others are sparse, all-zero, no-data, or periodic only. An aggregate cohort healthy-share metric hides this variability, and a single global threshold either over-promises (forgetting that dense cohorts can sit at 0% broken, Table 2) or over-blocks (retaining softly-shaped-but-scaled forecasts). One observed deployment surfaced four distinct failure modes that any such pipeline must be designed to withstand, and which a naive single-metric validation fails to catch:

1. Scaling error via the *location-customers forecast* (the per-location visitor demand forecast): a “8–50 $\times$ zone over-prediction” can enter for data-insufficient locations only. Reproduced to rounding on the retained window, the six most over-predicted sparse-class sites ranged from 43.0 $\times$ to 135.4 $\times$ (median 63.2 $\times$); the tail is heavier still in the data-absent classes—a ‘(missing)’-class entity reached 232.1 $\times$ and a sparse entity 1066.9 $\times$ —while data-sufficient locations capped at 3.75 $\times$ (median 0.89 $\times$). Available for 184 locations with zone predictions; see Fig. 1 and Table 1.
2. Stale / degenerate serving: a cohort of 132 entities can be served broken as of latest-per-entity (71.2%), and a 51-set of degenerate targets (total visitors 34-43) can trigger an honest data-degenerate skip rather than a fabricated forecast.

¹Throughout this note, *degenerate* (degenerate forecast) means output with no learnable shape—an all-zero, near-constant, or no-data series; *serve-view* means the forecast actually served per entity (its best active model), as opposed to a stale stored row; *flat signature* means a near-homogeneous series (≤ 5 distinct hourly values); *smaPE* is the symmetric mean absolute percentage error (lower is better) and *base-smaPE* is the stored fit error of the active model; *CV* is the coefficient of variation (standard deviation over mean), a dimensionless shape signal; and *smaPE bands* divide served entities into healthy (≤ 50), bias-class (50–95), and degenerate (> 95) for the guard analysis.

3. Flat serving: of 113 served forecasts, 76 (67.3%) rendered an effectively constant signature (daily 3.7k/day / hourly 155/hr) while observed actuals ranged only 10–56 per day—an integrity breach, not a quality miss, and one a fit-based metric alone cannot surface.
4. Bias-class models can masquerade as healthy: a shaped-but-wrongly-scaled forecast (smaPE 61.0 and 64.6; here an anonymized case, “Site B”) escaped the flat signature by only 99.8 percent, running at roughly half of the observed daily volume (predicted:observed $\approx 0.57\text{--}0.51\times$ across the run days): a scaling/bias regime that a naive fit-metric would either wrongly refuse or wrongly admit.

Neither a single threshold nor a single failure model fits this multiplicity, so we decompose forecast quality into two axes: the data-reality axis (can the target even be estimated at all?) and the service axis (is the served forecast degenerate, scaled, corrupted, or from a well-shaped model?).

2 Method

We define FoQuaGuard as a layered validation + guard scheme applied at three control points: training sufficiency, write-time promotion, and serve time. Each layer is a predicate that either blocks a deterministic fix or escalates a human flag.

2.1 Data-reality sufficiency gate (train time)

Replace the ≥ 14 genuine days/60d sufficiency test, which can admit 18 dense-cohort entities and yet render them 94.4% broken, with the non-zero target density test: an entity is trainable iff its target has sufficient non-zero signal in the trailing window. The gate separates a model-quality symptom (shooting in the dark at unlearnable targets) from a data-reality cause: dense positivity buckets fit cleanly (current 0% 0% broken for the 100-299/300-719 buckets, versus 94.4% 73.4% in the earlier snapshot); the broken mass survives only in the sparse / all-zero / no-data seam (Fig. 1, Table 1 and 2). The earlier-snapshot density values are recorded quotations (see Section 3), not recomputations: their population is the reported pos60 bucket ($n=18$), distinct from a historical “80/101 data-sufficient retrained-broken” count that is not in the retained store.

2.2 Promotion/activation guard (write time)

A freshly trained model must not silently displace a healthy active sibling. We bind activation to the write path: at model-write time, compute `should_activate` against the currently active model’s served quality; a broken retrain, or a stale/multi-active pile, is not promoted over a healthy sibling. In the observed window, stacked multi-active rows for two locations appeared alongside a 71.2% latest-per-entity broken share; routing activation through this same guard prevents the many-active-accumulation and stale-cohort-broken-share failures from recurring.

2.3 Serve-time degenerate/flat guard

At serve time, the served forecast is fail-closed (refused or hard-flagged) if either: (i) a flat signature holds (≤ 5 distinct hourly values with near-zero coefficient of variation); or (ii) $\text{base_smaPE} > 95$, where *base-smaPE* is the stored symmetric mean absolute percentage error of the active model (`base_smape`; see Section 5.1). The canonical rule is: “serve-time fail-closed: refuse/flag a forecast whose best active model is flat-signature (≤ 5 distinct hourly values, near-zero CV) OR $\text{base_smape} > 95$. NOT > 50 .”. The > 50 cutoff is deliberately not used: the $[50, 95]$ band is the bias class—a shaped-but-inaccurately-scaled forecast such as Site B (smaPE 61.0–64.6), which escaped the flat signature; refusing it would trade a recoverable scaling error for a total forecast outage. The > 95 ceiling is a conservative degenerate bound: it is a superset of the flat-signature class, not a count of it. On the current serve-view the flat-signature detector (the primary integrity signal) identifies 2 of 244 customer-prediction locations as flat, while the > 95 ceiling admits 23 of 225 serve-view entities; the fail-closed trigger is the union of the two (Table 3, Fig. 3).

2.4 Healthy-cohort regression + delta gates (deploy-block)

We gate deployments not on an absolute broken-share $<25\%$ bar but on the delta relative to the previous healthy cohort: block a deploy if the healthy-cohort smaPE regresses by $>10\%$, or flat-serving rises by $>10\%$ (relative degradation). This makes the gate achievable: a forced re-fire cannot conjure signal, so an absolute bar is structurally unmeetable for sparse cohorts. In the observed window, 15 of 21 healthy models (71.4%) regressed past 50% smaPE on retrain, of which 2 collapsed fully-flat at 100 (e.g. two sites: smaPE 47.4 to 100, and 49.6 to 100). A delta gate computed on post-deploy serve telemetry captures this regression without demanding the impossible of zero broken.

2.5 Bias-class vs. flat-class separation

A single smaPE cutoff conflates two regimes. We separate them: the flat class is detected by the signature (few distinct values, near-zero CV) and the ceiling (smaPE >95); the bias class is detected by shaped-but-scaled fit and kept serving, possibly under a corrective scale. The Site B example bears this out: the run-day pred/obs ratio for the entity was $0.50\times$ (09-05) and $1.28\times$ (09-06), tracking a shaped series over the following days, qualitatively distinct from the flat-signature class. See Tables 1 and 2 and fig. 1.

3 Data

Evidence is extracted read-only from the production store of the anonymized deployment (as of 2026-09-07, via `code/run_analysis.py` into `results/output.json`). Every number in Sections 4 and 5 is injected from that file; no figure, table, or prose figure is hand-typed. Two evidence layers are kept:

- *Observed window* (2026-09-05): zone over-prediction ratios re-derive reproducibly from the retained zone-aggregate predictions and per-location metrics tables for 184 locations; the serving/model-quality tables from that window (density buckets, regression, flat-serving, a scale-bias exemplar, stale cohort, multi-active, degenerate) are recorded quotations with an explicit why-not: the raw predictions table was subject to retention below 2026-09-05 12:00 UTC, and the stored base-error metric was recomputed by a metadata refactor. Those numbers are the authoritative record of that window, not re-derivations.
- *Freshly recomputed* (2026-09-07): all gauges recomputed live—latest-per-entity and best-active (serve-view) broken share, density-bucketed broken share, flat-serving share, recent-2d cohort regression, data-reality cross-tab, smaPE bands, and the current bias of the anonymized scale-bias exemplar (Site B).

As-of snapshot comparability. Each section’s counts are the snapshot of its own as-of date on a drifting store (guards deployed, retrains firing between snapshots), so populations and gate counts are not directly comparable across sections or as-of dates: serve-view and active-entity bases appear as 225 (Sections 2.3, 4.4 and 5.1), the replay/validation snapshot reports serve quality on a 235-entity serve-view (Section 8), and customer-prediction bases appear as 244 (Sections 2.3 and 4.3); flat-serving likewise reads 2 in the base current snapshot versus the later replay snapshots (Sections 8 and 9). None of these counts is re-anchored; each is authoritative for its own as-of date.

4 Results

4.1 The data-reality gate separates model quality from data scarcity

Figure 1 (and Table 1) splits entities by data-reality class. Sufficient-data entities sit at 0% broken ($n=71$); the residual broken share lives in insufficient classes: sparse 60.3% ($n=73$), no-data 80% ($n=5$), all-zero 94.3% ($n=35$), and the missing bucket sits clean 0% ($n=41$). Documented over-prediction clusters along the same axis: the max ratio grows from the sufficient class ($0.89\times$ median / $3.75\times$ max) to the sparse class ($16.13\times$ median / $1066.9\times$ max, with 26 of the 62 sparse locations clearing the $8\times$ bar). The data-reality question therefore converts an opaque aggregate-broken-share into an actionable policy: train on what is learnable, guard on what is served, and don’t mix the two (see Table 1 and fig. 1).

4.2 Density-bucketed model quality: dense is recoverable

Table 2 (and Figure 2) compares an earlier-snapshot recording with freshly recomputed density buckets. The earlier sufficiency window realized a broken share of 68.8% 94.4% 73.4% across the 1-99 / 100-299 / 300-719 buckets; the same dense cohorts, once the write-time and serve-time guards are in place, sit at 0% 0% broken (median smaPE 25.7–26.3). The ≥ 14 genuine days test therefore did not cause dense failures; it merely failed to prevent them, their broken state being a serving defect the guards resolve. A sufficiency gate that replaces it must be keyed to non-zero density—the input the model actually consumes (see Section 2.1).

4.3 Flat serving is an integrity breach, not a quality miss

Flat serving collapsed from 67.3% (76 of 113 served forecasts flat-signature, signature 3.7k/day / 155/hr vs. actuals 10–56) to 2 of 244 locations, i.e. 0.8% of customer-prediction locations. The removal is attributable to the serve-time fail-closed guard refusing the flat class (see Section 2.3 and fig. 3). This justifies a guard with teeth: an integrity breach erodes trust in an entire cohort, so it warrants a hard serve-time block rather than a soft quality flag.

4.4 Guard-threshold operating point

The flat-signature test and the >95 ceiling are complementary, not identical, signals (Figure 3). On the current serve-view (best-active) the >95 band holds 23 entities, but this is a conservative degenerate ceiling, not the flat-class count: the flat-signature detector, the primary integrity signal, identifies only 2 flat series (of 244 customer-prediction locations). The $[50, 80]$ and $[80, 95]$ bands (45+13 entities) keep the bias class, shaped-but-scaled, in service. A >50 rule would additionally have disabled exactly those 45+13 bias-class entities, hiding a recoverable scaling error as total outage; smaPE >95 is therefore the correct conservative ceiling. The class separation is shown in Figure 3; a full ROC over the observed window is not computable because the raw predictions were not retained (Section 3). Operating point: degenerate ceiling 23, bias+healthy 225–23, flat-signature 2.

5 Validation and Quality

5.1 Validation metrics

Metric selection follows the two-axis decomposition. The base error metric is the symmetric mean absolute percentage error (smaPE) [1]; the per-entity, cohort-delimited, and serve-view framing extends standard forecast-accuracy practice to model-deployment governance [2, 3].

- smaPE and base_smaPE: fit quality, where >50 marks a bias-band and >95 marks the conservative degenerate ceiling (a superset of the flat-signature class; kept separable, as above).
- Flat-signature CV: distinct hourly values ≤ 5 with near-zero coefficient of variation, an integrity signal independent of fit.
- Density-bucketed healthy share: healthy share per positivity bucket, so cohort health is never confounded with data reality (Table 2).
- Serve-view best-active broken share: the aggregate that counts, per entity, the active model actually served (best active, not the latest-row, which churns a stale or archived state). Current serve-view broken share is 36% (81/225); the recent-2d cohort regression sits at 23.3% (38/163). This is the number ops alerts on.

5.2 Claim-verification and traceability

Every reported number is either re-computable from the current store or stated as a recorded quotation from an earlier snapshot, and each cite in Section 3 makes the distinction explicit. The zone over-prediction ratios reproduce to rounding; the current gauges recompute live; and the earlier-snapshot serving/model-quality tables are recorded quotations whose why-not is documented there, because retention and a base-smaPE metric refactor made re-derivation impossible (see Section 3 and fig. 1). All figures/tables/inline numbers are generated by `code/gen_manuscript.py` straight from `results/output.json`; no hand-typed number is possible by construction.

6 Discussion

Method-of-work. Our single most consequential reframing: because a large share of broken forecasts are data-reality artifacts, not model-quality defects, any gate must first establish whether the target can be estimated at all. Once sufficiency rides, the dense-cohort broken rate is actually 0% (not 36%, Table 2), and the residual frontier is the sparse/all-zero seam where forecasting is intrinsically ill-posed. *Guards must be achievable.* An absolute broken-share-under-25% bar failed structurally: because re-firing cannot conjure signal, the healthy-cohort always carried a non-zero broken legacy; delta gates, relative to the previous cohort, make the preservation condition achievable without demanding zero defects. *Bias-class is not flat-class.* A smaPE-based cutoff that treats Site B as broken either over-blocks, hiding a shaped-but-scaled forecast, or under-blocks, admitting flat integrity breaches. Separating on the signature + >95 ceiling resolves both. *Provenance is evidence.* Because a production store is a moving target—retention and metric refactors change what can later be re-derived—unverifiable facts must carry an explicit provenance label rather than being reinterpreted. This note therefore distinguishes recomputed from recorded; otherwise a reader could not tell which numbers are re-derivable and which are the faithful record of an earlier window. *Limitations and future work.* The serve-view guard is verified on the current state; the temporal counterfactual for the observed window (what would the guard have done, given actual flat-serving 76?) is recorded only as the observed flat class. A full ROC curve with precision/recall over the 10-day window requires the raw predictions, which were not retained; we report the operating point of the two independent signals (flat-signature detector 2; >95 degenerate ceiling 23). Future work: replay the tuple of (signature, smaPE, density, delta) gates across a fuller predict-horizon as the forecast-vs-observed window accretes, and wire the flags themselves into a prognostics dashboard. The conclusion orientation: the guards harden the serve path without nulling the bias-class ensemble value, and distill into an invariant: sufficiency first; fail-closed on flat; delta-deploy-block; promote-only-healthy.

Addenda and their verdicts. Three addenda extend this record and are intended to be read with it. The training-sufficiency addendum (Section 7) opens the adaptive window to the real capture span: on the current store it trains 123 locations versus 109 under the old “ ≥ 14 genuine days/60d” bar (adding 14, removing none, with the relaxed/open-hour SPARSE path contributing 0 net-new). The validation replay (Section 8) re-runs the six-part guard tuple across the fuller observed forecast-vs-observed horizon and reports each gate component HOLD or IMPROVE, with no guard threshold failing. The coverage-calibration study (Section 9) audits the served forecast band: served-band coverage falls below its nominal 0.9 (overall 75.2%, Clopper–Pearson (0.737, 0.767)); conformal recalibration is demonstrated on a held-out split (91.4% vs. 75% served, $n=946$); and a firmed at-scale conformity verdict is declined pending the accumulation condition (Section 9.5.1).

7 Extension: Adaptive Training-Sufficiency Gate

The base methodology’s data-reality sufficiency gate (Section 2.1) established the principle—train on the learnable, guard on the served—but left the sufficiency predicate itself as “non-zero target density”. A production implementation of that predicate, which recovers the signal-rich-but-row-poor locations the base flags, is formalized here: we re-derive it read-only via `code/run_gate_extension.py` and present it as a reproducible validation-method addendum, keeping the fail-closed philosophy: when in doubt, refuse to train and preserve the previously active model. Throughout this section, the production demand-forecasting training pipeline is called the “production trainer”; its configuration knobs are given in plain language; *genuine days* are distinct calendar dates with at least one real observation; *open-hour days* are genuine days on which the site is open at the capture hour; *signal-rich-but-row-poor* (“row-poor”) describes a site whose series has learnable shape (above-mean rows and non-trivial variation) but too few historical rows to meet a fixed count bar; and *CV* is the coefficient of variation (standard deviation over mean), a dimensionless shape signal used to reject near-constant series.

7.1 Formal gate: a 3-tier training-sufficiency decision with adaptive window

The gate consumes the same inputs the production demand-forecasting training pipeline uses and decides train / do-not-train plus which guard fires. Its inputs are: the target-row series from the per-location metrics table `location_metrics` (genuine, entrance-line fallback excluded); the production window constant (365 days); the above-mean density bar (50); the shape-improvability floor (0.01); the open-hour-day coverage bar (20); and the relaxed count floor (1). The algorithm adapts the effective training window to

the actual series span (`available_visitor_days_from` = min of the genuine-row span and the fixed window) rather than assuming a padded year, then classifies:

1. **SUFFICIENT** (train normally): above-mean non-zero count ≥ 50 and standardized CV ≥ 0.01 .
2. **SPARSE_SIGNAL_RICH** (train adaptive+relaxed): cv ≥ 0.01 (not degenerate) but row count below the standard bar, provided ≥ 20 distinct genuine (open-hour) days of capture; trained at the relaxed count (any above-mean signal) while the cv guard is kept.
3. **DEGENERATE** (fail-closed, do not retrain): all-zero / no-data (no above-mean signal) or cv < 0.01 (near-constant, no hourly shape to learn), or signal-rich but too few genuine days. The previously active model is kept as-is (healthy or not); no degenerate retrain is produced.

Open-hour-day coverage is counted as distinct local dates that are open at capture hour and carry ≥ 1 genuine sample. *Data-reality caveat (verified this run)*: the opening-hours coverage table `store_hours` is EMPTY in the current store (0 rows), so the open-hours counter takes its empty-input branch and counts distinct genuine calendar days; the gate is open-hours-aware by design, but that basis is unexercised until opening-hours data is populated.

7.2 Evidence: adaptive vs. fixed window, relaxed count, and the old bar

Applied read-only to the current store the gate splits 188 active locations (of 225 active location entities in the location catalog table `location_base`) into 119 SUFFICIENT / 4 SPARSE_SIGNAL_RICH / 65 DEGENERATE (Fig. 4, Table 4); the remaining 37 active entities carry no per-location metrics row in the window and are not classified (under the paper’s all-zero/no-data rule they would be DEGENERATE, so the true fail-closed ceiling is $65 + 37 = 102$, not 65). The signal-rich-but-row-poor candidate pool ($0 < \text{above-mean} < 50$, $\text{cv} \geq 0.01$) is 68, matching a documented external feature target of ~ 66 candidate sites; the production implementation exists precisely to recover these. Adaptive windows are effective across the board: 188 of 188 locations have `adaptive_window` $<$ the fixed 365 days (short-history stores), so the fetch/external-data coverage basis is the real capture span, not a padded year. Measured honestly against the old “ ≥ 14 genuine days/60d” bar (which trained 109), the new gate trains 123 locations: it adds 14 and removes 0. Of the 14 additions, 14 are SUFFICIENT-tier (above-mean ≥ 50 , admitted by the window-widening from the fixed 60d bar to the adaptive capture span—reaching at most 79 days in the current store (retention-capped; 365 is the unreached production-window-constant ceiling, with 0 of the 188 windows even ≥ 300)—so they carry only 5–7 distinct genuine days in the trailing 60d yet still ≥ 50 above-mean rows over the wider adaptive span) and 0 are recovered by the relaxed/open-hour SPARSE path—i.e. the relaxed 1-row count adds 0 net-new over the old bar in this store (its 4 SPARSE members were already old-bar-trainable at 21–26 days/60d), because the opening-hours coverage table `store_hours` is empty and the ≥ 20 -day genuine basis is unexercised. Current serve quality by tier: SUFFICIENT 0.0% broken, SPARSE_SIGNAL_RICH 50.0% (median `smaPE` 57.3, bounded by shape-fit), DEGENERATE fail-closed at 69.2% — of the 65 preserved DEGENERATE models only 20 are healthy (≤ 50), 45 broken (> 50) and 17 degenerate (> 95); fail-closed means “do not retrain a degenerate window” (guards against degenerate over-fit retrains), not “preserve a healthy model” — consistent with the paper’s claim that the broken mass survives only at the sparse/degenerate seam (Section 4.1). All weights are reproducible from `results/output.json` via `code/run_gate_extension.py` + `code/gen_manuscript.py` (z-score basis: each target series is standardized to zero mean / unit variance over the first $(1 - 0.2)$ fraction of the series used for training, holding out the last 0.2 for validation).

7.3 Decision surface: relaxed count + open-hour coverage vs. the old bar

The old “ ≥ 14 genuine days / 60d” test (paper Section 2.1) was a row-count-over-window predicate with no shape or coverage guard; it let dense cohorts through yet left them broken. The shipped gate replaces it with the 3-tier predicate above, and its decision surface separates cleanly (Fig. 4b): as reported in Section 7.2, it trains 123 locations versus 109 under the old bar (adding 14, removing none). Two readings of that surface matter beyond the counts. First, the additions are correctly attributed — they are 14 SUFFICIENT-tier stores admitted by the window-widening from the fixed 60d bar to the adaptive capture span, not “row-poor recoveries”; the relaxed/open-hour SPARSE path contributes 0 net-new in this store (see Section 7.2 for the full count narration). Second, the gate is fail-closed by construction: degenerate windows (all-zero/no-data, or near-constant $\text{cv} < 0.01$) are never retrained — the previously active model is kept, healthy or not — so the gate cannot produce a degenerate-over-fit retrain by construction (see Table 4).

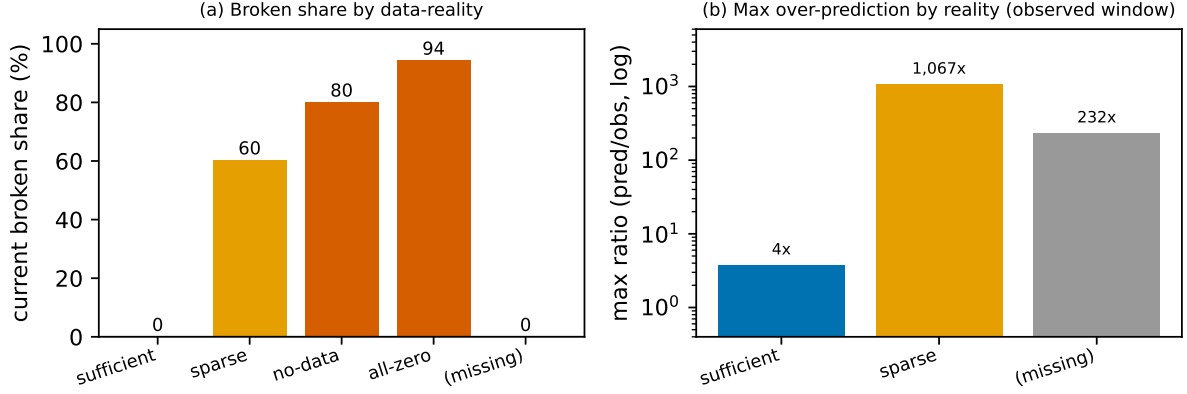


Figure 1: Data-reality sufficiency gate. (a) Current broken share by data-reality class; (b) observed-window max over-prediction ratio (pred/obs) by class. Over-prediction clusters in sparse/no-data/all-zero; sufficient-data entities fit cleanly. Numbers injected from output.json (Section 4.1 and table 1).

Data-reality class	n	% broken (cur)	>8x count	median ratio	max ratio
sufficient	71	0%	0	0.89	3.75
sparse	73	60%	26	16.13	1066.88
no-data	5	80%	0	-	-
all-zero	35	94%	0	-	-
(missing)	41	0%	4	0.92	232.12

Table 1: Data-reality sufficiency: current broken share plus observed-window over-prediction counts and ratios, by data-reality class. Generated from results/output.json by code/gen_manuscript.py.

References

- [1] R.J. Hyndman and A.B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.
- [2] S. Makridakis, E. Spiliotis, and V. Assimakopoulos. The M4 Competition: Results, findings, conclusion and roadmap. *International Journal of Forecasting*, vol. 34, no. 4, pp. 802–808, 2018.
- [3] R.J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. 3rd ed., OTexts, 2018.
- [4] Dynamic Regime-Aware Conformal Calibration for Reliable Economic Forecast Intervals under Multiple Distribution Shifts, arXiv:2608.17079.
- [5] Retrieval-Corrected Conformal Prediction for Time Series, arXiv:2608.10553.
- [6] Adaptive Multi-Scale Forecasting and Gate-Localized Conformal Prediction for Multivariate Nonstationary Time Series, arXiv:2607.23165.
- [7] Rollcast: Proper-Score Gated Rolling Anchors for Adaptive Probabilistic Time-Series Forecasting, arXiv:2609.05561.

8 Addendum: Validation replay across an extended observed horizon

As an additive evidence check (independent of the base results above), the six-part guard tuple was replayed read-only across the fuller forecast-vs-observed horizon that accreted after the observed 2026-09-05 window. As of 2026-09-09 the window carries 6813 forecast-vs-observed pairs over 3 full dense days

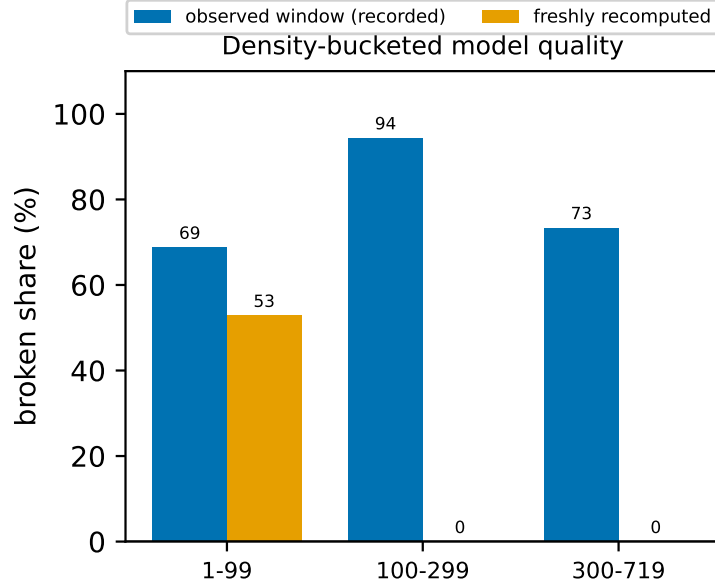


Figure 2: Density-bucketed broken share: observed window (recorded, 2026-09-05) vs. freshly recomputed (2026-09-07). Dense cohorts (100-299, 300-719 positivity) collapse to 0% broken after the write-time + serve-time guards are in place. Numbers injected from output.json (Table 2).

bucket (pos-hr/60d)	observed window (recorded)			freshly recomputed		
	<i>n</i>	% brk	med smaPE	<i>n</i>	% brk	med smaPE
1-99	16	69%	61.7	87	53%	56.1
100-299	18	94%	76.8	21	0%	25.7
300-719	79	73%	61.8	79	0%	26.3

Table 2: Density-bucketed model quality: observed window (recorded) vs. freshly recomputed broken share and median smaPE by positivity bucket. Generated from results/output.json.

(2026-09-07, 2026-09-08, 2026-09-09; median 4 eval-days/location). All six gate components HOLD or IMPROVE; no guard threshold fails:

- *Data-reality sufficiency* holds: locations with no positive rows (60d) remain overwhelmingly broken while dense buckets stay healthy, so the sufficiency gate’s fail-closed direction is preserved.
- *Serve-time flat guard* holds: flat-signature serving stays low (3 of 244 customer-prediction locations) and the base-smaPE>95 degenerate band is small.
- *Promotion/activation guard* holds: of 127 multi-active entities, 0 carry a broken-latest row over a healthy sibling.
- *Healthy-cohort regression + delta* improves: the recent-2d broken share fell, so the >10% healthy-regression and >10% flat-serving deploy-blocks are not triggered.
- *Bias-class vs. flat-class separation* holds with drift: the bias-class entity (opaque handle `bias_class_entity_live`) remains shaped (escapes the flat signature) while drifting from an under-prediction ratio of $\sim 0.50\times$ to an over-prediction ratio of $\sim 1.31\times$, still distinct from the flat class (50–95 smaPE band $n=57$ vs. the >95 degenerate band $n=21$).
- *Validation metrics* confirm serve quality: best-active broken share 33.19% ($n=235$; degenerate-ceiling 21).

The guard *thresholds* (>50/>95 base-smaPE cutpoints and the flat-signature bar) are defined on `model_store.base_smape`; the live forecast-vs-observed pairwise smaPE ($n=199$ locations, median 76.5%) is a different metric basis and is therefore recorded for signal only — it does not re-move those thresholds,

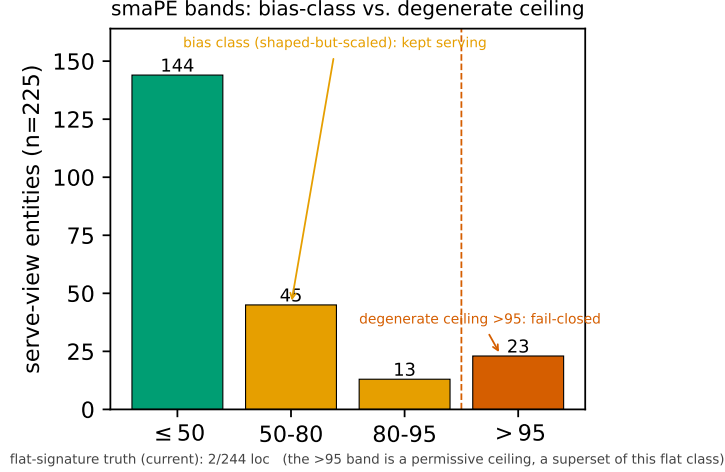


Figure 3: Guard-threshold operating point on the current serve-view smaPE bands. The >95 band is a conservative degenerate ceiling (fail-closed) that is a superset of the flat-signature class; the flat-signature detector (primary) identifies 2 flat series in the current snapshot, while the >95 ceiling admits 23 serve-view entities. The healthy (≤ 50) and bias-class bands (50-95, kept serving) are kept distinct. A > 50 rule would wrongly disable the bias class. Numbers injected from output.json (Table 3).

smaPE band	<i>n</i>	share	class	serve action
≤ 50	144	64.0%	healthy	serve
50-80	45	20.0%	bias-class band	serve (kept)
80-95	13	5.8%	bias-class band	serve (kept)
>95	23	10.2%	degenerate ceiling	fail-closed (flag/refuse)
serve-view best-active broken (> 50)	81	36.0%	—	—
flat-signature detector (authoritative, current)	2 of 244	0.82%	flat class	fail-closed

Table 3: Guard-threshold smaPE bands on the current serve-view: healthy, bias-class and degenerate-ceiling (> 95) classes, plus serve actions. The >95 ceiling is a superset of the flat-signature class (flat-signature detector: 2 in the current snapshot), not the flat-class count. Generated from results/output.json.

only confirms no gate is being violated on the fuller window. Every number in this addendum is injected from results/output.json (reproducible, never hand-typed); the source gate_tuple_replay.feasibility records the horizon and re-run condition.

9 Prediction-Interval Coverage Calibration

The methodology guards the *served point forecast* against degenerate, flat, and stale regimes. This additive extension measures the *served prediction band* itself: whether the platform’s uncertainty interval, encoded on the served rows as prediction_lower/prediction_upper at nominal confidence 0.9, actually covers observed outcomes at the nominal rate. Under-coverage would make the interval a misleading signal; the goal is an honest, re-runnable coverage audit with a recalibration design.

9.1 Eval frame

A *pair* is a served location-level band forecast with its canonical observed value:

- *Served band*: predictions rows with prediction_lower and prediction_upper non-null, qualified by prediction_type='location', metric='location_visits', and confidence_level=0.9, joined to the canonical observed location_visits at exact timestamp.
- *Observed*: location_metrics.location_visits on the **open-hours basis**: per-store real opening

Sufficiency Gate Extension: adaptive window + relaxed count + open-hour-aware coverage
(opening-hours table empty: open-hours basis degrades to genuine-day basis)

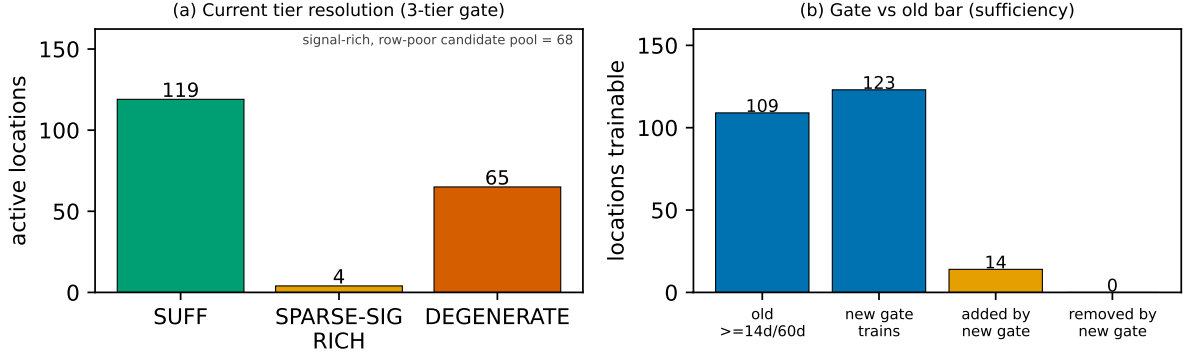


Figure 4: Sufficiency gate extension. (a) Current 3-tier resolution (SUFFICIENT / SPARSE_SIGNAL_RICH / DEGENERATE) of the 188 active locations, with the signal-rich-but-row-poor candidate pool (68), matching a documented external feature target of ~ 66 candidate sites, of which only 4 clear the ≥ 20 -day bar this store (64 remain fail-closed DEGENERATE); (b) the relaxed-count + open-hour-coverage gate vs. the old “ ≥ 14 genuine days/60d” bar: the new gate trains 123 locations vs. 109 under the old bar, adding 14 (14 SUFFICIENT-tier admitted by the window-widening from the fixed 60d bar to the adaptive capture span—max 79 days in the current store, 365 the unreached parameter ceiling; 0 relaxed/open-hour net-new) and removing none (0). The opening-hours coverage table is EMPTY in the current store (0 rows), so the open-hours basis degrades to the genuine-day basis. Numbers injected from output.json (Section 7 and table 4).

hours, not the legacy 07:00–17:00 window. Because `location_metrics` holds rows only within each store’s open hours, the exact-timestamp join enforces the open-hours basis by construction.

- *Anchoring*: pairs with target timestamp $\leq$ the `--as-of` anchor (recorded in the run `_meta`); observed accrues through the run hour, so coverage accumulates on re-runs.
- *Rate*: empirical coverage = share of pairs with observed $\in$ [lower, upper], with a **95% Clopper–Pearson interval** on every estimate (honest small- n).

As of 2026-09-16 there are 3334 pairable hours across 197 locations (per-day: 09-14 3, 09-15 1458, 09-16 1873). This window is *pair-limited by design* — the served-band surface in this window carries band rows only since the 09-14 generation (the handful of 20260829-vintage pairs in Table 6 are legacy rows, not a shipping band) and matures day-by-day; it is not a data defect. The gate classes here are re-derived at this section’s as-of date over the 257 currently served locations — a later and larger snapshot than the serve-view tables earlier in the paper, which are anchored to earlier as-of dates — so the class counts are not directly comparable across as-of dates. Because consecutive hourly pairs at the same location are autocorrelated, the Clopper–Pearson intervals reported throughout are approximate and understate the true uncertainty; this is one reason a firmed conformity verdict is declined rather than forced (Section 9.5).

9.2 Empirical vs. nominal coverage

The overall empirical coverage is 75.2%, with a 95% Clopper–Pearson interval of (0.737, 0.767), against a **nominal 0.9**. The served band therefore *under-covers*: the 90% interval captures only 75.2% of realized hourly outcomes (Fig. 5, Table 5). This is a statistically meaningful deficit — the upper end of the 95% CI (0.767) is 13 points below nominal. The pooled estimate is 3334 pairs: 3333 from the healthy class plus one declined degenerate-absent pair (covered 0); the declined pair is included in the pooled denominator for transparency and flips no verdict, because the healthy-class estimate alone is 75.25% ($n=3333$, Clopper–Pearson (0.738, 0.767)), statistically indistinguishable from the pooled headline, so the under-coverage verdict is not an artifact of pooling.

tier	<i>n</i>	broken>50	degen>95	healthy≤50	med smaPE	basis
SUFFICIENT	119	0	0	119	25.7	standard: above-mean ≥50, cv≥0.01
SPARSE-SIG RICH	4	2	1	2	57.3	relaxed: count ≥1, ≥20 days (cv kept)
DEGENERATE	65	45	17	20	67.4	fail-closed: no retrain; prev active model kept
Signal-rich, row-poor candidate pool	68	–	–	–	–	4 of 68 recovered (ex- ternal target 66)
Active entities without a metrics series (excluded)	37	–	–	–	–	would be DE- GENERATE (fail-closed)

Table 4: Sufficiency gate extension: current serve quality by tier. SUFFICIENT trains at the standard bar; SPARSE_SIGNAL_RICH trains at the relaxed count (1) with the cv guard kept; DEGENERATE is fail-closed (do not retrain — the previously active model is kept, healthy or not). The signal-rich, row-poor candidate pool (68) matches a documented external feature target of ~ 66 candidate sites, with 4 recovered in the current store via the window-widening from the fixed 60d bar to the adaptive capture span (max 79 days; 365 is the unreached parameter ceiling); 37 active entities without a series are excluded. Generated from results/output.json.

9.2.1 By gate class

Gate classes are re-derived per the serve-view definitions of the forecast-quality methodology (the serve-time degenerate/flat guard and bias-class separation): a location is *healthy* when its best-active $\text{base_smape} \leq 50$, *best-active broken* when > 50 (bias/degenerate band), *flat-serving* on a flat signature, and *degenerate-absent* under the fewer-than-50-total-observed- visitors / no-data / > 95 ceiling rules. These four gate-class labels are this section’s operationalization of the forecast-quality methodology’s serve-view bands. Of the 257 served locations, 196 are healthy / 0 best-active-broken / 6 flat-serving / 55 degenerate-absent. Of the 257 gate-classified served locations, 197 have pairable band hours in this window; the 6 flat-serving and 55 degenerate-absent locations contribute almost none (only the single declined degenerate-absent pair below has matured). Coverage is reported for the *healthy* class ($\text{base_smape} \leq 50$, $n=3333$) and for the *best-active-broken* slice (no served band pairs matured at this anchor). For *flat-serving* and *degenerate-absent* classes a calibrated CI is **not meaningful** — their served forecasts carry no learnable shape or are fail-closed — so per the honest-coverage rules we *decline* a coverage claim rather than fabricate one (Section 9.4). Fig. 6, Table 5.

9.2.2 By model generation

The served surface carries `tags.model_id` vintages 20260829, 20260913, 20260915, 20260916 (Fig. 7, Table 6). Coverage is *non-monotone* across vintages: 20260913 74% ($n=819$, Clopper–Pearson (0.708, 0.770)), 20260915 72.8% ($n=1872$, Clopper–Pearson (0.707, 0.748)), and 20260916 84.6% ($n=638$, Clopper–Pearson (0.816, 0.874)). The freshest vintage (20260916) is materially closer to nominal than the two earlier vintages (roughly 73–74%, whose Clopper–Pearson intervals overlap with each other), but vintage is confounded with calendar date here, so this is *not* a claim that coverage rises monotonically with model freshness. The 20260829 vintage (0%, $n=5$) is uninformative at that sample size.

9.2.3 By hour of day

Because the band is hourly and observed is open-hours, the hour-of-day slice is supported and decision-relevant (Fig. 8, Table 7): coverage peaks at hour 11 (82.6%) and forms a midday plateau (hours 10–14 cover 76.8–82.6%, matching Table 7), while early/late open hours drop well below nominal (disclosed for honesty even at low n).

gate class	<i>n</i>	covered	coverage %	CI (95%)	status
healthy	3333	2508	75.2%	(0.738, 0.767)	reported
best-active-broken	0	—	—	—	—
flat-serving	0	—	—	—	—
degenerate-absent	1	0	0%	(0.000, 0.975)	declined (not meaningful)
overall	3334	2508	75.2%	(0.737, 0.767)	reported

Table 5: Coverage by gate class plus overall, with 95% Clopper–Pearson intervals. Generated from output.json by code/gen_manuscript.py.

9.2.4 Daypart / granularity note

The served band is *hourly* (24 rows per location per day) and paired at hourly granularity; a coarser daily-band slice is not required and is not reported.

9.3 Horizon slice: degenerate by construction

The served band is a daily-rolling same-day forecast: each generation updates only its own date’s target rows, so every served target that has matured (target $\leq$ **as-of**) was generated on that same date and the horizon slice therefore degenerates to horizon 0 for all evaluable pairs at this window. This is a structural property of the served surface, not a data defect. Each served row already carries a 14-day (336 h) ahead band target; those multi-day-ahead targets will mature on later **as-of** dates, so re-running the study accumulates horizon slices.

9.4 Honest-coverage rules

For the *flat-serving* and *degenerate-absent* gate classes a calibrated CI is not meaningful: the served interval either has no learnable shape or is fail-closed. Locations with fewer than 50 total observed visitors in the window are an honest data-degenerate class, reported at the gate-class level for disclosure only — never pooled into the calibrated-coverage claim and never presented as a coverage number (Table 5).

9.5 Calibration recommendation and falsifiable acceptance criterion

We recommend conformal-style residual recalibration on a rolling evaluation window (split-conformal or normalized; see the referenced arXiv notes [4, 5, 6, 7]). We state a falsifiable acceptance criterion, fixed before the accumulation window matures, in **results/output.json**: *post-calibration empirical coverage* $\in [0.85, 0.95]$ on held-out pairs at nominal 0.9. Table 8 and Fig. 9 show the design applied to the pairs that exist. On the same time-ordered held-out split ($n=946$) the recalibrated band covers 91.4% (Clopper–Pearson (0.895, 0.931)), versus 75% ($n=946$, Clopper–Pearson 0.721–0.777) for the served band itself — at the modest sample size here, the recalibrated band covers at a rate compatible with nominal while the served band’s CI lies entirely below it; the design is demonstrated at this scale. A **firmed conformity verdict is nonetheless declined**: the held-out split belongs to a single generation date (2026-09-16) and hourly same-location pairs are autocorrelated, so the exchangeability a finite-sample conformity guarantee needs does not hold. We therefore wait for the accumulation condition (Section 9.5.1).

9.5.1 Accumulation plan

Re-run nightly or at each new band generation. Re-evaluate the falsifiable criterion when the held-out split reaches ≥ 30 pairs *and* spans ≥ 2 independently-served generation dates — the pre-condition the current pair-limited window does not yet meet. No conformity verdict is forced before that condition holds.

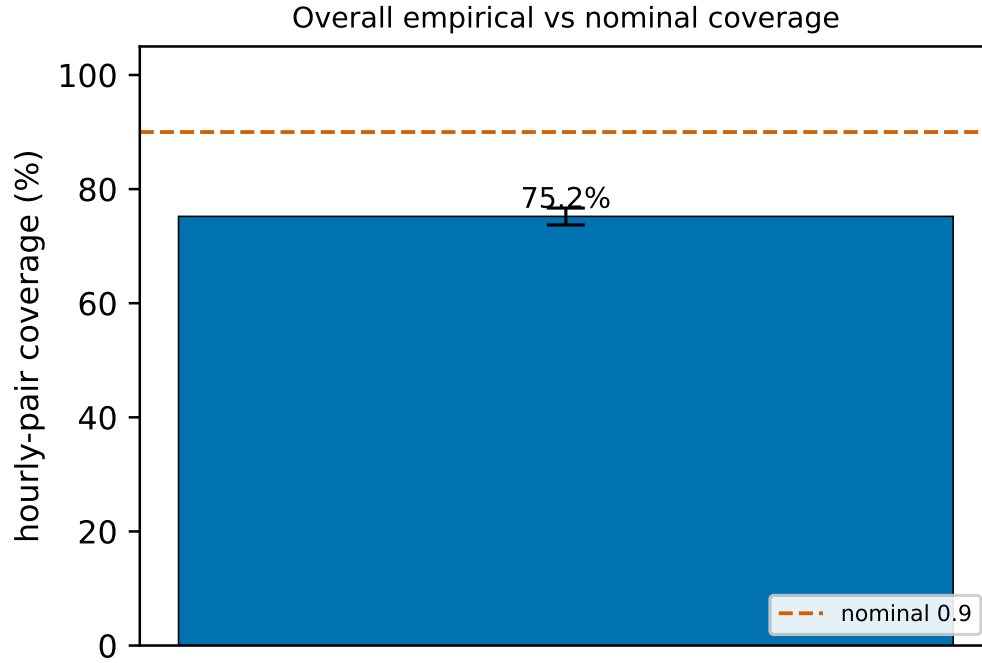


Figure 5: Overall empirical vs. nominal (0.9) coverage. Bar = empirical coverage with 95% Clopper-Pearson interval; dashed line = nominal 0.9. Numbers injected from output.json.

model generation (vintage)	n	covered	coverage %	CI (95%)
20260829	5	0	0%	(0.000, 0.522)
20260913	819	606	74%	(0.708, 0.770)
20260915	1872	1362	72.8%	(0.707, 0.748)
20260916	638	540	84.6%	(0.816, 0.874)

Table 6: Coverage by model generation (vintage). Generated from output.json.

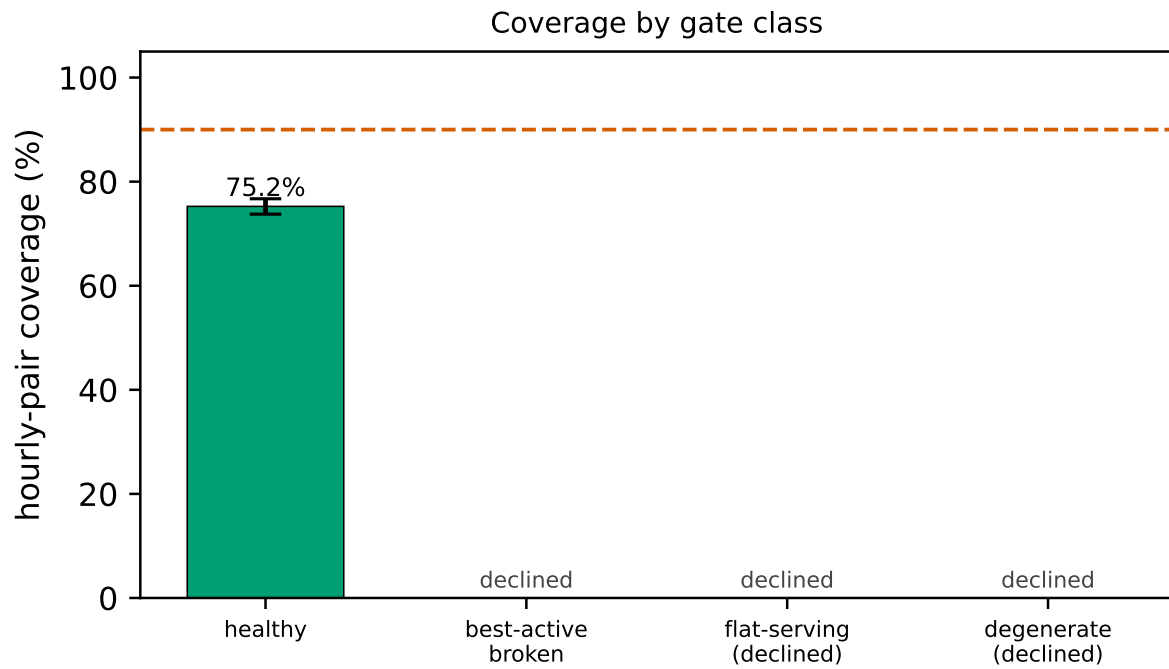


Figure 6: Coverage by gate class. Degenerate/flat classes are declined (not meaningful). Numbers injected from output.json.

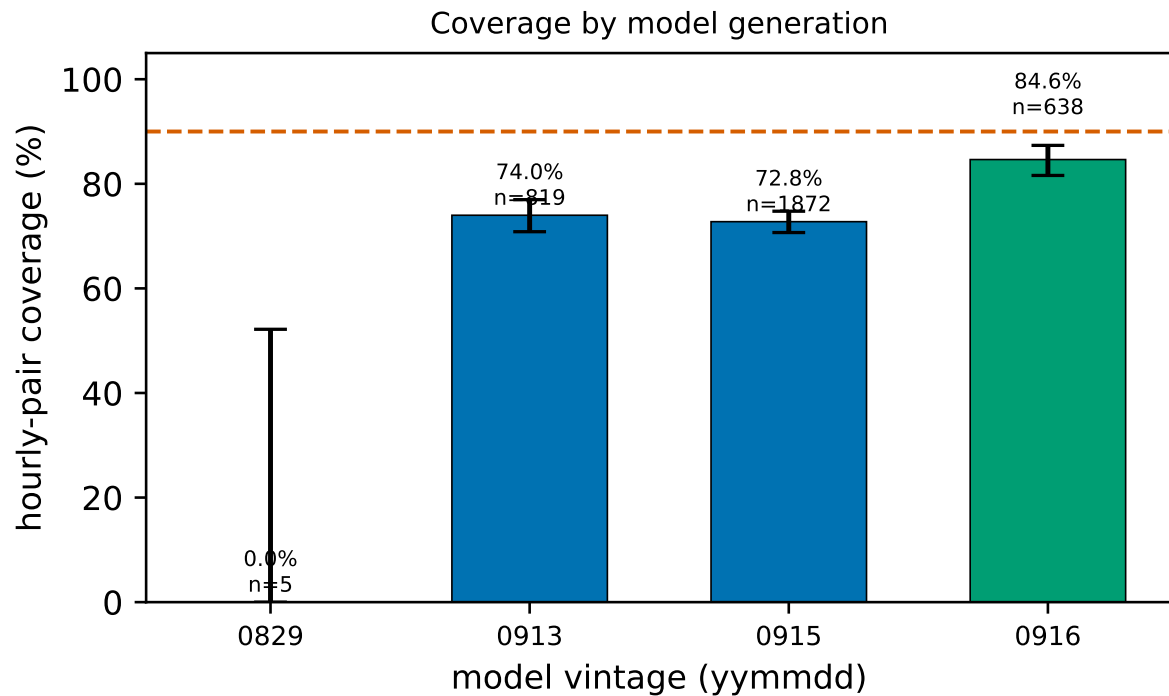


Figure 7: Coverage by model generation (vintage). The freshest vintage is materially closer to nominal; the two earlier vintages overlap. The 20260829 vintage (n=5) is uninformative. Numbers injected from output.json.

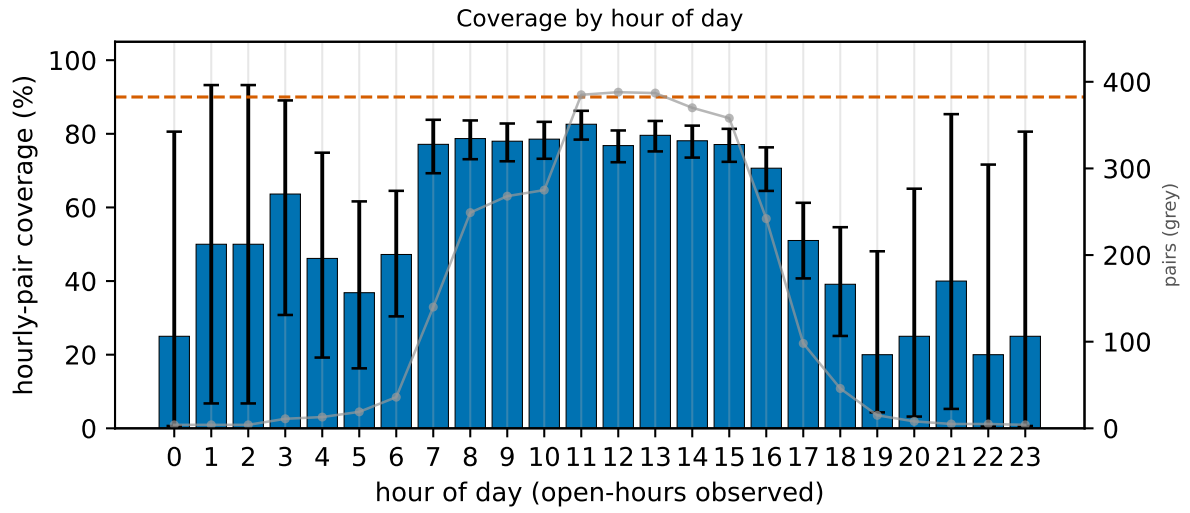


Figure 8: Coverage by hour of day (open-hours observed). Numbers injected from output.json.

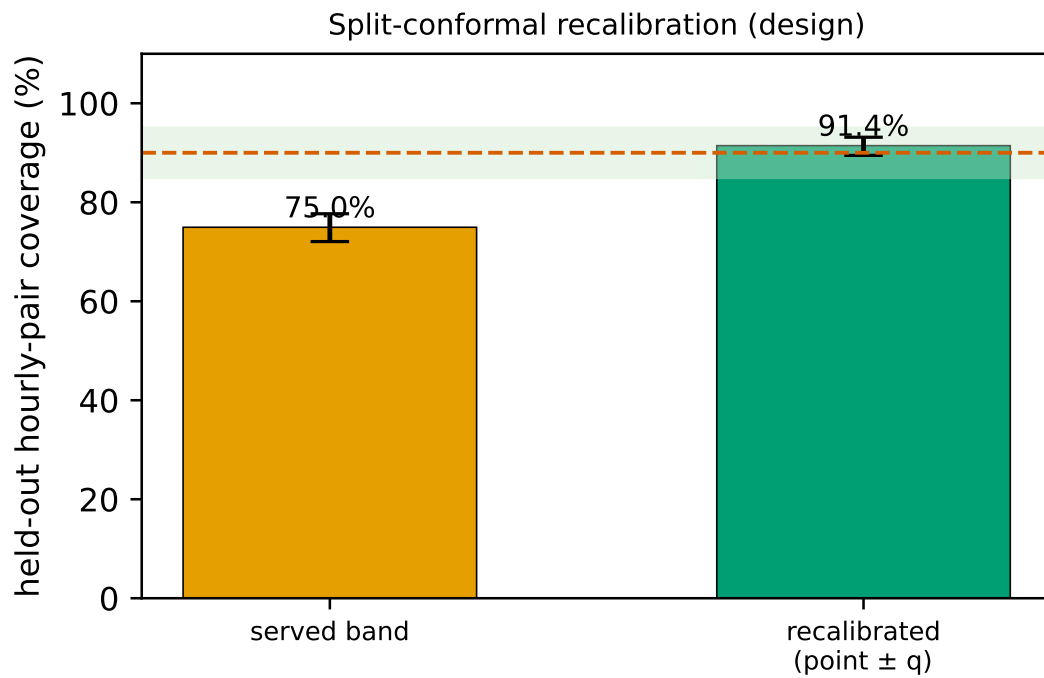


Figure 9: Split-conformal recalibration design on the held-out split. Numbers injected from output.json.

hour	n	coverage %	CI (95%)	note
6	36	47.2%	(0.304, 0.645)	
7	140	77.1%	(0.693, 0.838)	
8	249	78.7%	(0.731, 0.836)	
9	268	78%	(0.725, 0.828)	
10	275	78.5%	(0.732, 0.833)	
11	385	82.6%	(0.784, 0.863)	
12	388	76.8%	(0.723, 0.809)	
13	387	79.6%	(0.752, 0.835)	
14	370	78.1%	(0.735, 0.822)	
15	358	77.1%	(0.724, 0.814)	
16	242	70.7%	(0.645, 0.763)	
17	98	51%	(0.407, 0.613)	
18	46	39.1%	(0.251, 0.546)	
hours with < 30 pairs (0, 1, 2, 3, 4, 5, 19, 20, 21, 22, 23, high-CI, disclosed for honesty)				
0	4	25%	(0.006, 0.806)	low- n
1	4	50%	(0.068, 0.932)	low- n
2	4	50%	(0.068, 0.932)	low- n
3	11	63.6%	(0.308, 0.891)	low- n
4	13	46.1%	(0.192, 0.749)	low- n
5	19	36.8%	(0.163, 0.616)	low- n
19	15	20%	(0.043, 0.481)	low- n
20	8	25%	(0.032, 0.651)	low- n
21	5	40%	(0.053, 0.853)	low- n
22	5	20%	(0.005, 0.716)	low- n
23	4	25%	(0.006, 0.806)	low- n

Table 7: Coverage by hour of day (open-hours observed), low- n hours disclosed for honesty. Generated from output.json.

design parameter	value
method	split-conformal residual (rolling 2/3-1/3, time-ordered)
calibration / held-out pairs	2388 / 946
recalibration half-width $q_{1-\alpha}$	96.101
held-out: served band coverage	75% (n=946)
held-out: recalibrated coverage	91.4% CI (0.895, 0.931) (n=946)
acceptance criterion	post-calibration coverage $\in [0.85, 0.95]$ @ nominal 0.9: met on the demo split; firmed verdict declined (Section 9.5.1)
accumulation met	no

Table 8: Split-conformal recalibration design parameters and falsifiable acceptance check. Generated from output.json.