Supplementary Materials

Targeted Allocation and Caps for Food Delivery Promise Extensions Under Protective Operating Requirements

S1. Numerical verification

Numerical checks test whether promise adjustments use a consistent feasibility calculation. In Figure S1(a), independent enumeration of 48 instances gives a maximum upper-bound excess of 0.003034 min (about 0.18 s), below the 0.02-min classification threshold. Figure S1(b) recomputes 100 states after uniform extensions of 0, 1, 3 and 5 min. All 400 interval checks pass against the 0.005-min reference tolerance. The first test checks the bracket around G; the second checks its response to a known deadline shift. These tests verify the declared calculation before and after allocation; retained intervals identify decisions close to the classification threshold.

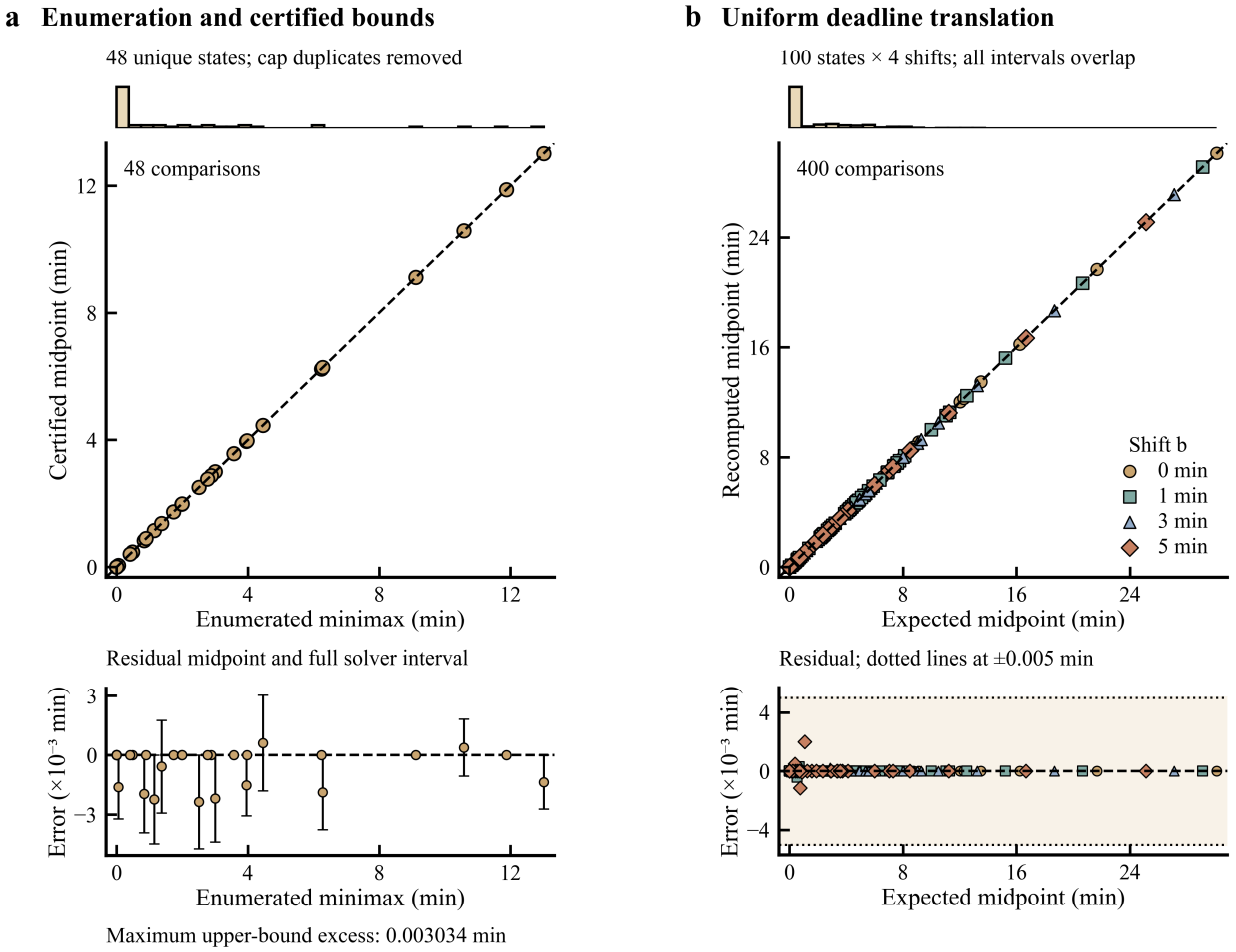

Figure S1. Numerical checks by independent enumeration and uniform promise translation. (a) Enumerated G versus the midpoint of the solver interval for 48 states. The identity line marks agreement. Whiskers show numerical bounds. (b) Uniform deadline translations of 0, 1, 3 and 5 min for 100 states. The tolerance band is ±0.005 min and zeros remain in the distribution.

S2. Travel, preparation and service inputs

Figure S2(a) displays the travel-time inputs for 179,119 remaining nodes relative to their riders, using the same kilometre scale in both directions. Figure S2(b) compares release times with recorded pickups for 57,440 pending-pickup references. Of these, 24,735 releases are clamped or stored as zero, and 813

references have a zero raw forecast timestamp. A zero release treats the order as already available for service. Among zero waits, 23,922 have a nonzero forecast at or before the snapshot.

Figure S2(*c*) examines service accounting at shared merchants. All 48,167 merchant–state groups have zero coordinate span; counts for one to seven pickup nodes are 41,018, 5,602, 1,121, 307, 92, 22 and 5. Repeated pickups add little travel at the shared location, while the model charges service per node. Figure S2(d) pairs reference and no-service *G* values to isolate the response to this assumption. The checks identify three input-review tasks: distance conversion, forecast waiting and service accounting at shared locations.

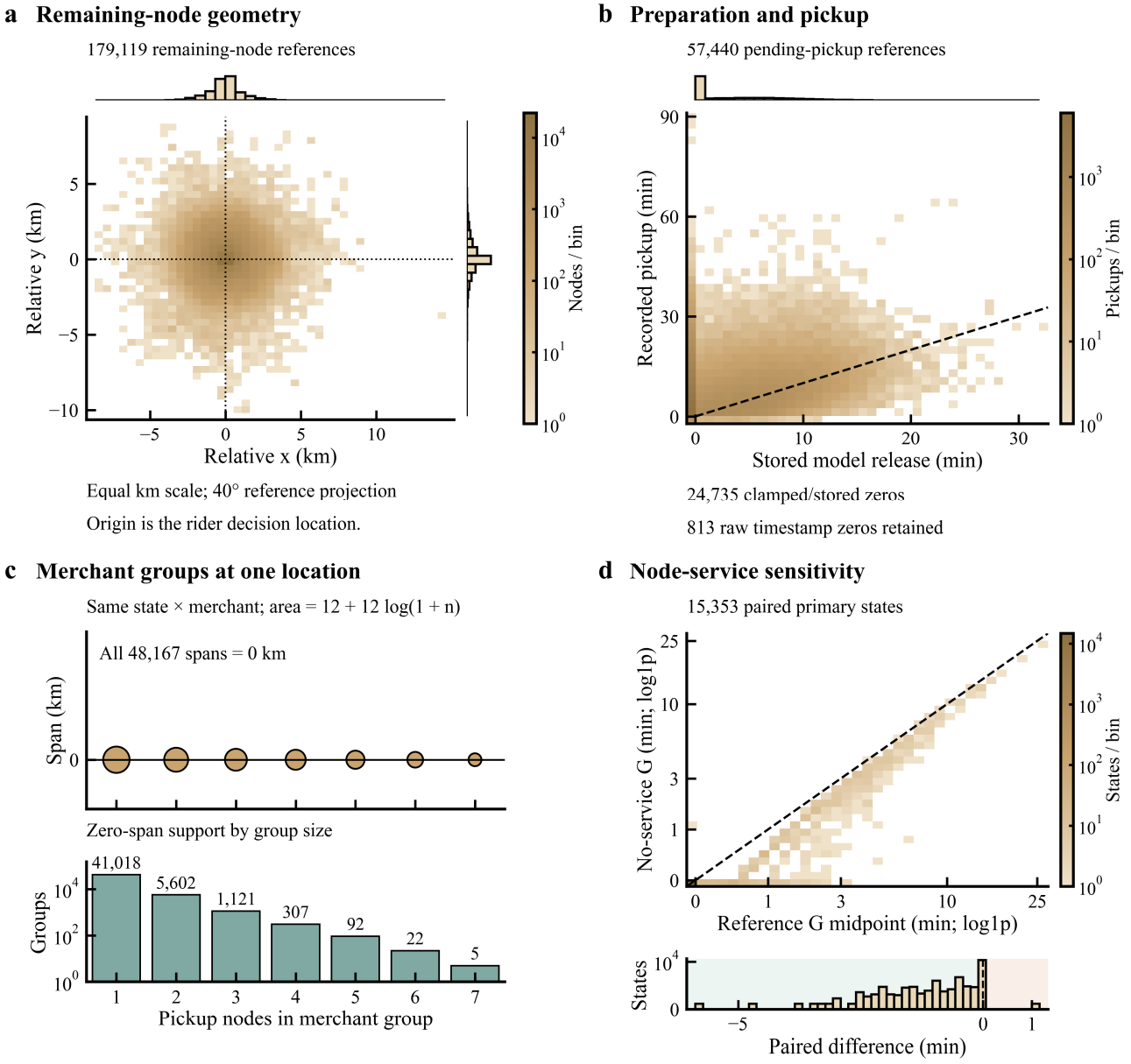

Figure S2. Relative geometry, preparation releases and node service inputs. (a) Relative positions of 179,119 nodes in kilometres. (b) Forecast releases and subsequent pickups for 57,440 pending-pickup references. (c) Within-merchant coordinate spans for 48,167 merchant–state groups. (d) Reference versus zero-service deficits in minutes. The diagonal marks equality.

S3. Comparison with recorded fulfilment

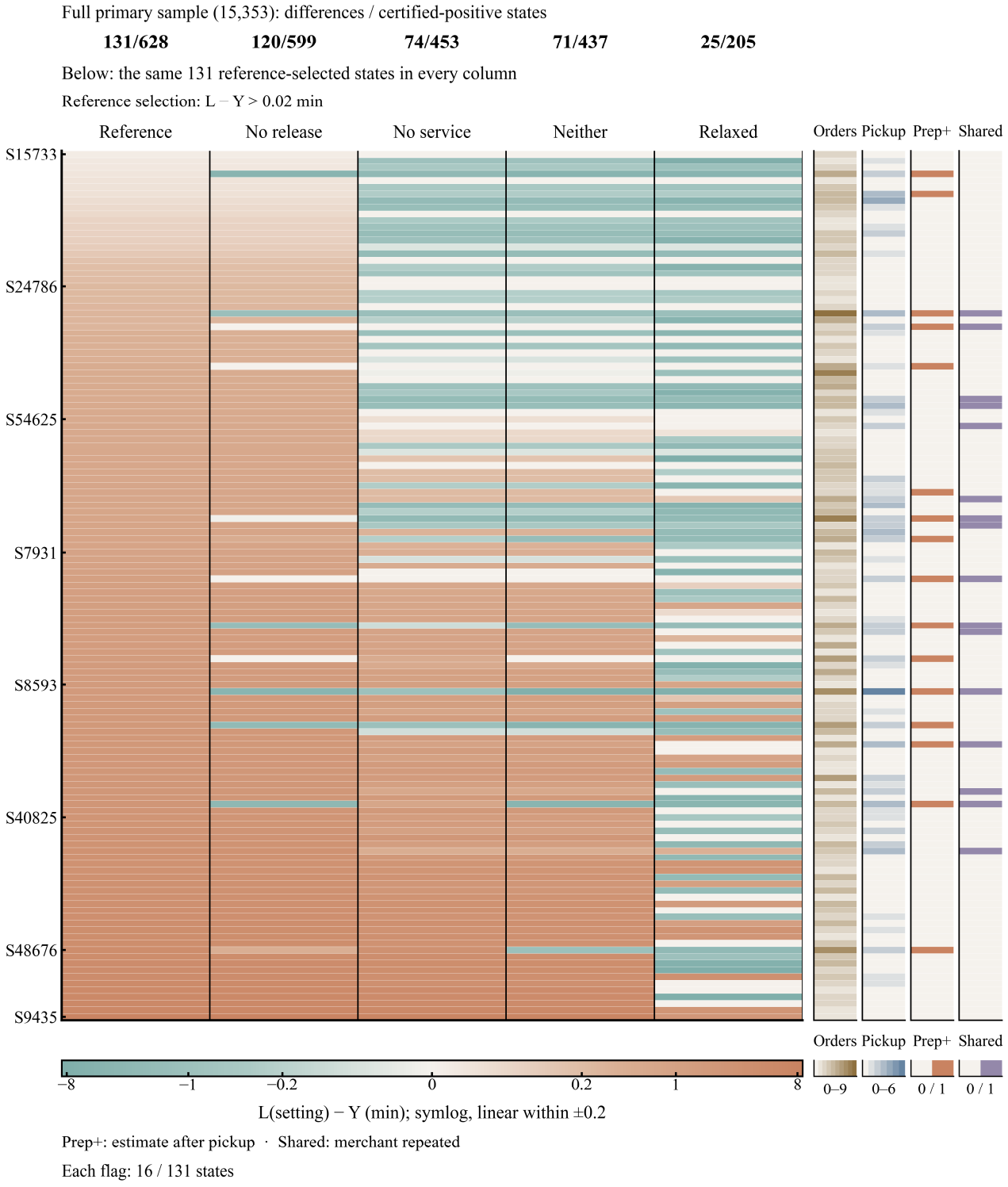

Figure S3. Complete fingerprints of 131 model–record compatibility differences. Each row is one of the 131 reference-selected states. Columns compare five operating settings. Terracotta denotes a lower bound above the recorded maximum lateness, teal a lower bound below it. The signed-log scale is linear within ±0.2 min. Side annotations identify workload, preparation forecasts later than pickup, and shared merchants.

The 131 reference compatibility differences form an order-linked list for reviewing operating inputs (Figure S3). Each row retains the same cohort across five settings. Colour represents $D=L(\text{setting})-Y_{\max}$: terracotta marks a lower bound above recorded lateness and teal the reverse. The scale is linear within ±0.2 min, with signed-log tails. Orders and Pickup give task counts. Prep+ marks estimates

later than recorded pickup, and Shared marks repeated merchants; each flag identifies 16 states. These attributes direct review to the preparation and service inputs associated with each discrepancy.

The setting-specific differences and certified-positive counts in the full primary sample are 131/628, 120/599, 74/453, 71/437 and 25/205. The matrix retains the same 131 reference-selected rows. Reference differences represent 0.85% of all 15,353 primary states but 20.86% of the 628 certified-positive states; the relaxed-setting ratio is 25/205 (12.20%). Compatibility checks are therefore particularly relevant to states targeted for promise review. A usable deficit record should carry its operating settings, solution interval and linked orders, allowing the platform to examine the assumptions behind a proposed extension.

S4. Internal prediction comparisons

Figure S4 compares alternative time quantities for screening tasks before detailed promise adjustment. Lin+*G*, Spl+*G* and Tree+*G* augment the linear, spline and tree models with *G*; Lin+*L* and Lin+*U* use the individual analytic and inexpensive route bounds. Lin+G_0 removes release constraints from *G*. Points and intervals use the same test states. Higher AUC and AP indicate better ranking; lower Brier score and log loss indicate smaller probability errors. The comparison tests whether the information gained for screening justifies full joint solving.

AUC increments from *G* are 0.01037, 0.00192 and 0.00066 for linear, spline and tree models. The latter intervals are [−0.00008, 0.00404] and [−0.00190, 0.00333]. Spline AP improves by 0.01490 (95% interval: [0.00384, 0.02484]), while its probability-error intervals and all four tree-model intervals cross zero. Linear models with analytic bounds, route bounds and no-release *G* reach AUCs of 0.7645, 0.7649 and 0.7653, close to 0.7655 with full *G*. Lower-cost structural quantities are plausible screening candidates. Full *G* provides the joint feasibility calculation when a selected cohort requires a quantified promise adjustment.

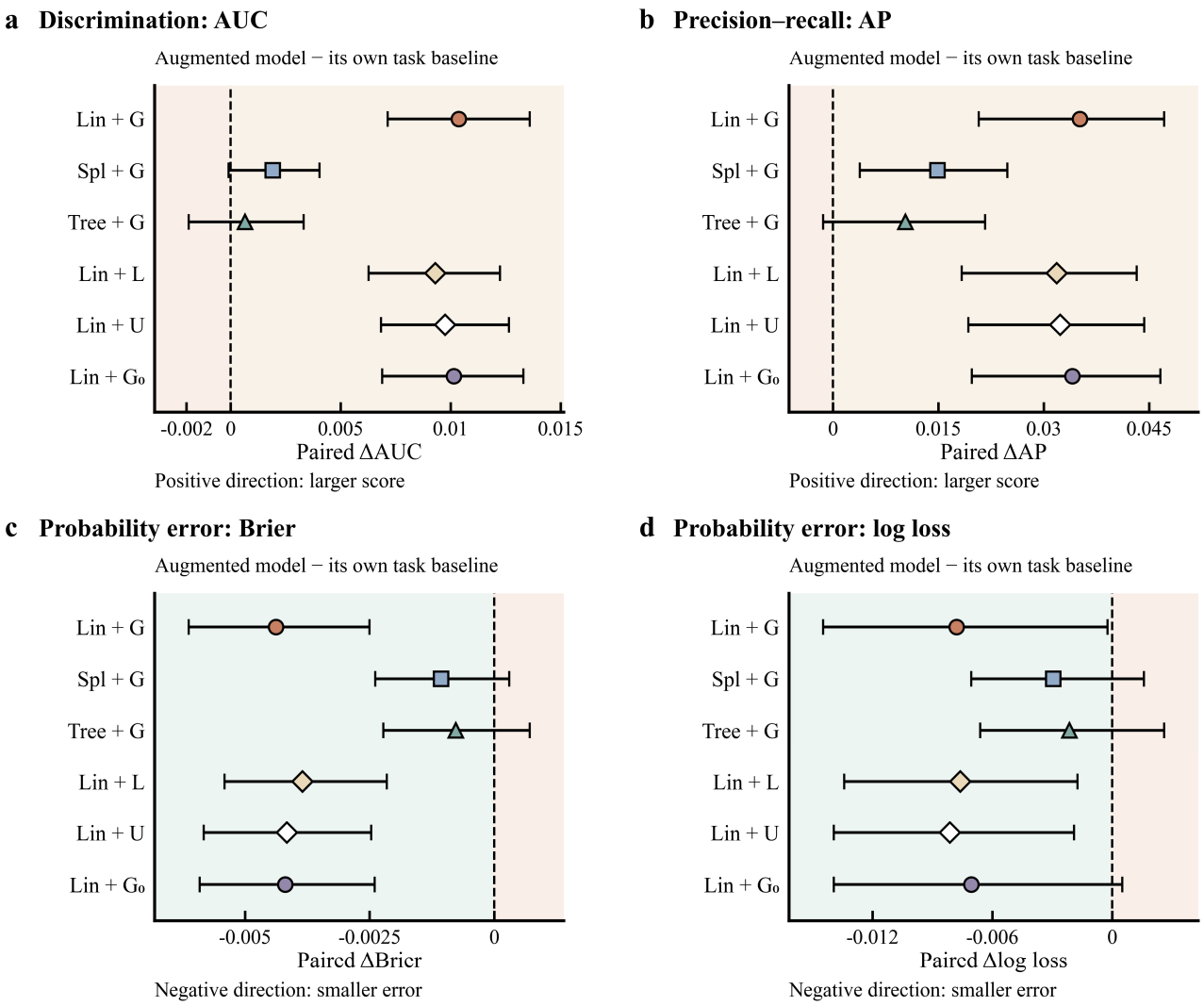

Figure S4. Paired predictive increments from structural time quantities across task baselines. (a–d) Paired changes in area under the ROC curve (AUC), average precision (AP), Brier score and log loss across six contrasts based on nine fixed models. Horizontal bars show pointwise 95% rider-bootstrap intervals on the two historical test dates (3,761 states). Zero is the no-change reference. Smaller values are preferred for Brier score and log loss.

S5. Prediction across riders and dates

Figure S5(a) separates riders unseen in training from those unseen throughout development: 295 riders with 354 states and 218 riders with 256 states, respectively. The latter excludes both training and validation riders. Figure S5(b) compares the fixed models on 1,887 states from 23 October and 1,874 from 24 October. Linear point increments are positive on both days, while nonlinear increments remain near zero. Directions and interval support vary across models in the development-unseen subset. These internal comparisons distinguish date variation from performance on new riders within the same release and identify where the screening rule needs further evaluation.

Screening and promise adjustment call for different levels of computation. A linear screen can test G or cheaper structural bounds; a flexible model needs evidence of an incremental benefit. Calibration and sample support govern use of

the resulting probabilities. Once a cohort is selected for promise adjustment, joint solving returns the required minutes and their bounds. This separation reserves detailed computation for task sets requiring an explicit promise decision.

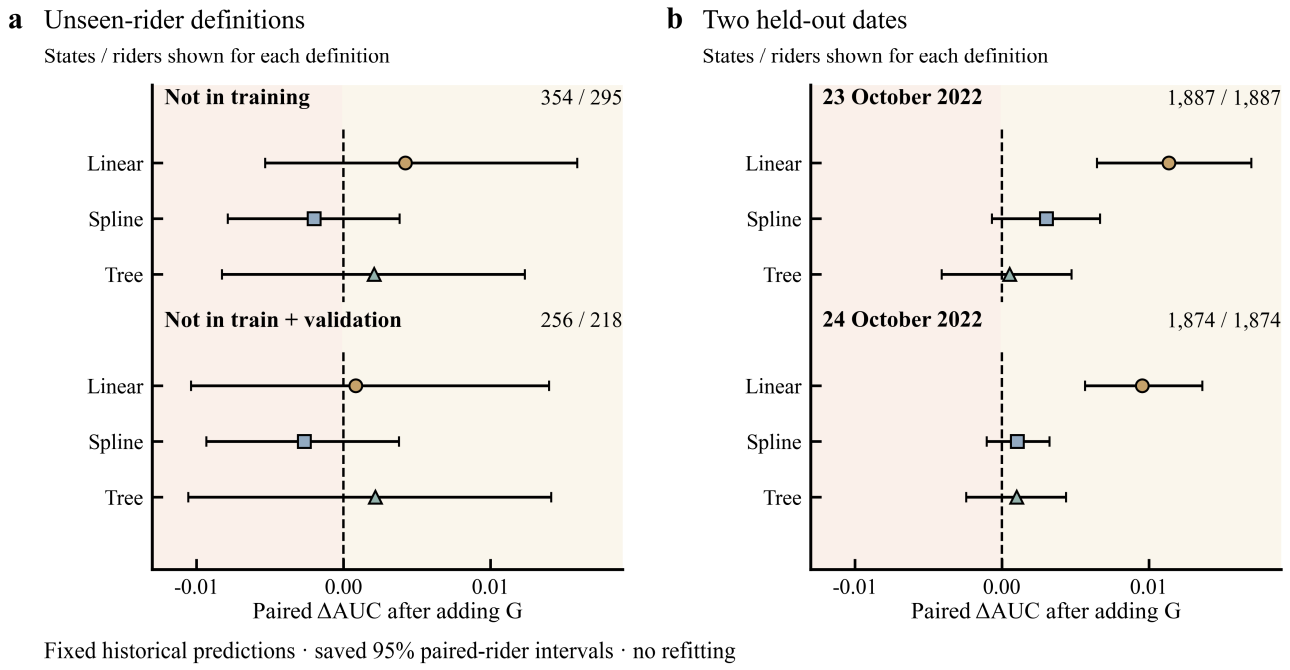

Figure S5. Predictive increments for unseen riders and two historical test dates. (a) AUC increments on 354 states from riders absent from training and 256 from riders absent throughout development. (b) Increments on the two test dates. Symbols identify the model family and whiskers show historical pointwise 95% paired rider intervals. The vertical zero line marks equal AUC.

S6. Within-cohort budget accounting

Figure S6(a) verifies equal resource use across 13,772 cohorts and 16 implementable requests, giving 220,352 cohort–strategy records. Requested and assigned budgets differ by at most about 2.84×10^{-14} order-minutes. Reconstructing 618,016 order allocations from q and cohort-specific η gives a maximum difference of 5.684×10^{-14} min. The feasibility differences thus arise at effectively identical cohort budgets. An operational implementation can use the same checks in sequence: verify individual caps, verify the cohort sum, then evaluate residual feasibility.

Figure S6(*b*) shows why priority alone does not determine an order's extension. The 36,896 orders at q=0.1 min account for about 95.5% of the sample; the remaining orders form four groups of 433, 432, 432 and 433. Most individual earliest completions fit their promises, but shared routing can still create a joint shortage. Even floor-priority orders receive different minutes because cohort scaling differs. Quantile strips and the 6-min reference show the resulting allocation range and cap use. Order identity, q, cohort η and b_j make the allocation traceable when the heuristic leaves a joint requirement unresolved.

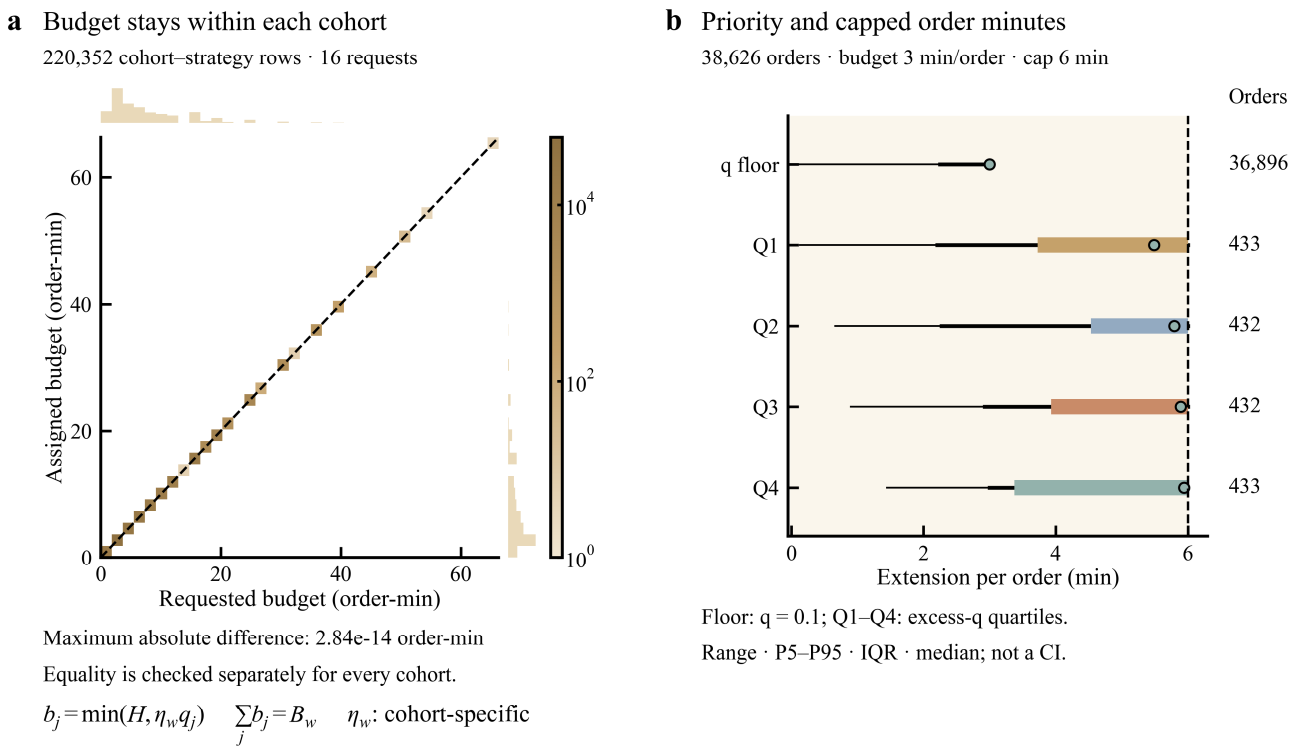

Figure S6. Cohort budget conservation and capped order-level extensions. (a) Requested and assigned minutes in 220,352 cohort–policy combinations. The diagonal denotes full use of the budget, and cell colour gives frequency on a log scale. (b) Priorities and allocated minutes for 38,626 orders at a 6-min cap. Strips show the range, 5th–95th percentiles and interquartile range; dots mark medians. The common multiplier depends on the other orders in the cohort.

S7. Route timing and secondary selection

Figure S7(a) examines the computational demand of 512 fixed states by remaining-node count. Each method uses the median of three warmed calls per state. Overall medians are 0.0387 ms for structural bounds and 0.0422 ms for certification; certification has a longer upper tail. The marginals retain 512 states and 1,024 method medians, with compilation excluded. The stored timings support using node count to identify demanding cohorts and retaining bounds when solving is incomplete. Scheduling must account for the long upper tail as well as the short median call.

Figure S7(*b*) isolates the effect of route selection on reported service. Across 27,544 matched pairs from uniform and uncapped burden-based allocation, the returned upper bound U is fixed while the original route is compared with the earliest-end route feasible at that bound. Points below the identity line indicate shorter completion at the same feasibility target. The paired strip retains the full variation. Using this common secondary criterion separates service differences caused by allocation from those caused by which feasible route is returned. Each policy record therefore needs both its promise minutes and the completion time under the stated route criterion.

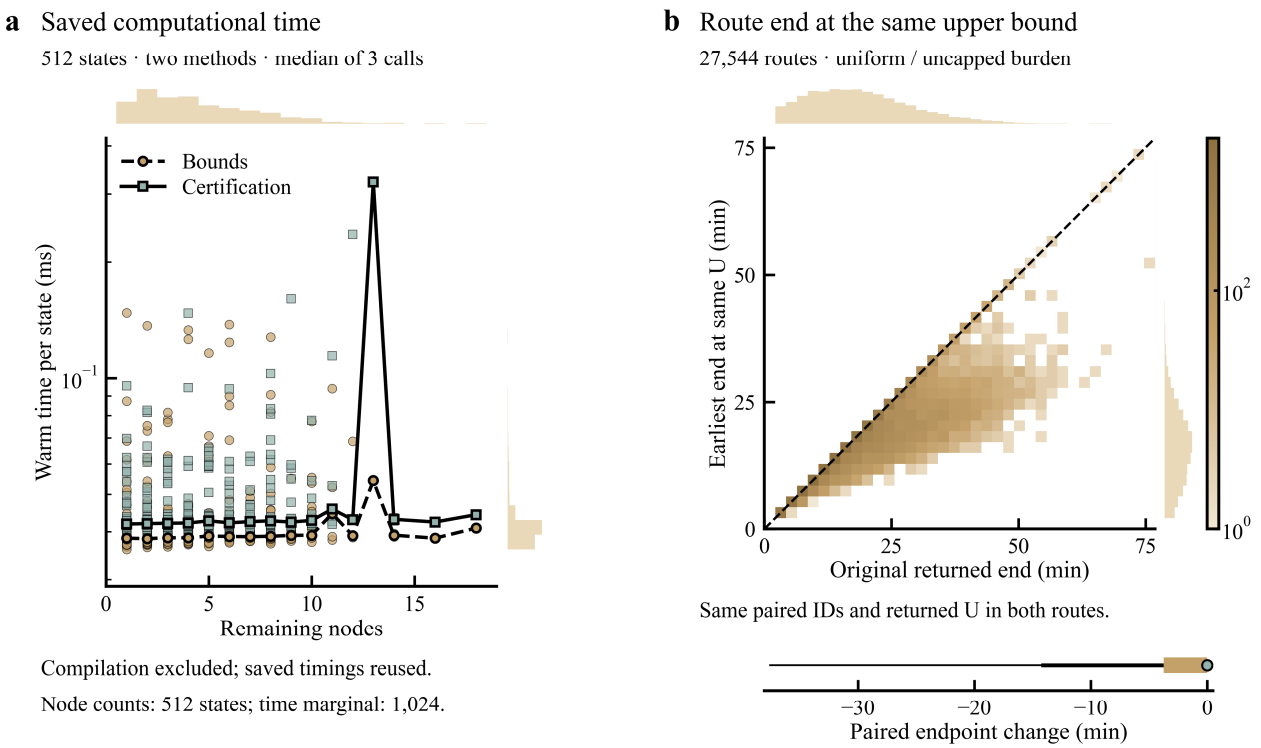

Figure S7. Stored computation times and route end times at a common upper bound. (a) Warmed computation times for bounds and certification on 512 states, grouped by node count. Each state contributes a median of three calls. (b) Route-end times for 27,544 paired original and earliest-finish routes. The diagonal marks equal completion times at the same feasible upper bound.

S8. Variation across anonymous zones

Figure S8 compares the common allocation rule across 23 anonymous zones. The left panel gives the 6-min-cap-minus-uniform change in positive-cohort share with rider-cluster 95% intervals; the centre gives cohort counts. The right reports the original positive share, mean order count, P95 capped extension and mean route end time, with a separate colour scale for each attribute. Negative changes indicate fewer positive cohorts. Reading the outcome with task composition and sample support identifies zones where the same rule warrants different levels of review.

Zone point estimates range from −4.31 to 0.00 percentage points. Support varies from 3,061 cohorts in zone 00 to 20 in zone 22, so intervals accompany each local comparison. P95 capped extensions are 4.2 min in zone 07 and 5.5 min in zone 17; approximately 95% of each zone's orders fall below these values. The latter distribution lies closer to the common 6-min cap. These zones describe heterogeneity within the same data release. Task composition, original conflict levels and extension tails inform local parameter review, with unequal sample support explicit in each comparison.

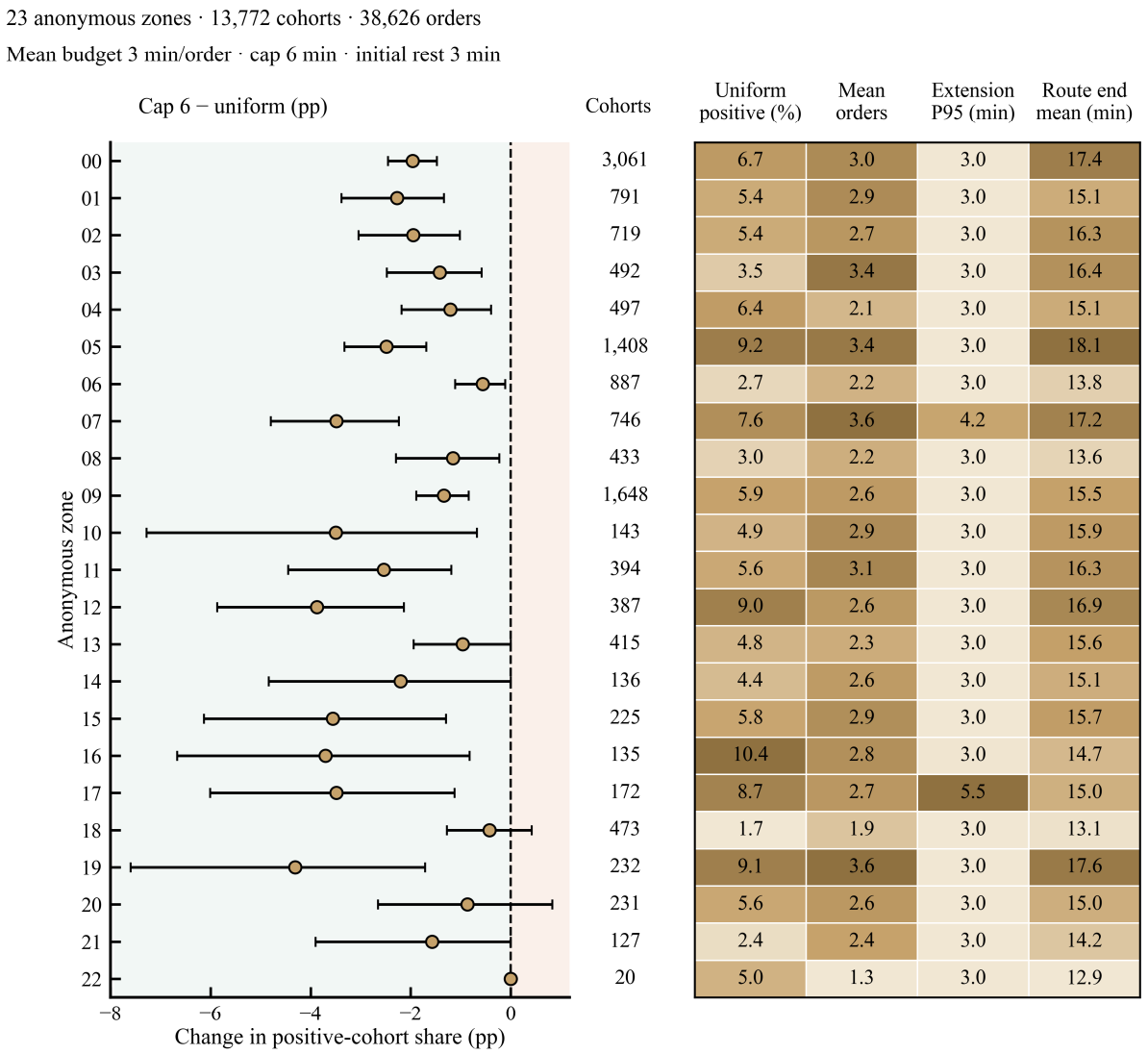

Figure S8. Allocation responses and task composition across 23 anonymous zones. Capped-minus-uniform differences in residual-positive share across 23 anonymous zones are paired within cohort. Points give percentage-point differences and whiskers pointwise 95% rider-bootstrap intervals. The adjoining workload summaries and sample counts describe local support.

S9. Independent survey context

Figure S9 provides independent behavioural context for the management decision. The public survey contains 423 respondents. PAM measures perceived algorithmic management, and WDB4 records self-reported traffic violations to deliver on time. Figure S9(a) gives joint counts and marginals; Figure S9(b) groups PAM by five WDB4 response levels with 34, 65, 102, 148 and 74 respondents. Spearman's correlation is 0.1231 (95% interval: [0.0297, 0.2164]), a weak positive association. This weak association provides behavioural context from respondents unlinked to the delivery records; allocation performance is assessed on the operational cohorts.

Research on platform health and safety and the job demands–resources model connects work organization with riding behaviour [73, 74]. Comparisons of delivery and private riding and studies of working conditions further establish this occupational context [75, 76]. The procedure specifies a corresponding operational action: reserve stated operating and initial-rest time, allocate promise

extensions, and recheck feasibility and service. The order-level experiments quantify feasibility and simulated service under those specified conditions. Linking implemented promise changes and actual rest to independent health observations would allow their effects on rider wellbeing to be assessed.

Reference numbers [73–76] correspond to the reference list in the main article.

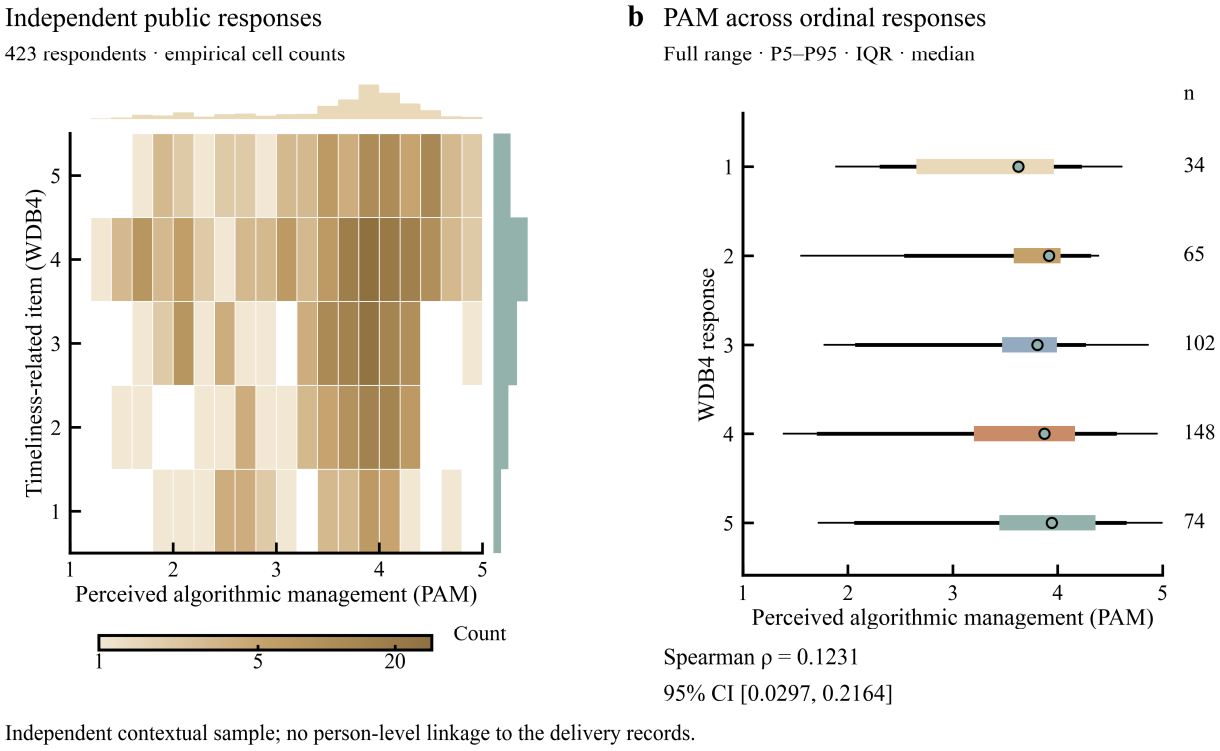

Figure S9. Perceived algorithmic management and self-reported behaviour related to timely delivery. (a) Binned perceived algorithmic management (PAM) and the five levels of a timeliness-related self-reported behaviour item (WDB4) among 423 independent respondents. Colour is the log count. (b) PAM within each response level. The dot marks the median, the box the interquartile range, and outer segments the tails. Spearman ϱ = 0.1231. Survey respondents are unlinked to the delivery records.

S10. Sample coverage and decision thresholds

Table S1. The first common snapshot occurs near 11:25 local time on each date. All counts use consistent active states.

States	n	Mean orders	Mean pickups	Mean nodes	Median tightest time (min)	Mean G bounds (min)
Later snapshots	28,199	2.93	1.37	4.30	18.48	0.306–0.311
Over 14 nodes	47	8.98	7.38	16.36	20.40	0.000–1.145
Policy sample	13,772	2.80	1.35	4.15	18.83	0.284–0.284

The release supplies the three dispatch epochs on each day. The first falls between 11:25:06 and 11:25:19, the second exactly two minutes later and the third five minutes after the first. One common epoch per day avoids allocating two budgets to the same order. Coverage bounds in Table S1 use the reference input before initial rest. In the larger-cohort comparison, the policy input includes three minutes of rest. For 15–18 nodes we raise the DP limit to 18. For 19–23 nodes the individual lower bound and constructive route upper bound remain available. All 47 states enter each comparison.

Table S2. Positive / undetermined counts above each operational threshold. All rates use 13,772 cohorts. If all undetermined cases are positive, the upper rate is the sum of both counts divided by 13,772.

Threshold (min)	Uniform	Capped	Uncapped	Net fewer	Relative (%)	Retained (%)
0.02	835 / 2	565 / 0	470 / 2	270	32.3	74.0
1	652 / 0	441 / 0	361 / 1	211	32.4	72.5
2	526 / 1	347 / 0	289 / 0	179	34.0	75.5
5	247 / 0	167 / 0	148 / 0	80	32.4	80.8

Table S3. Capped-minus-uniform differences in the positive-cohort percentage. Each row uses 1,000 paired rider draws (seed 20261111), all defined.

Threshold (min)	Difference (pp)	Pointwise 95% interval
0.02	-1.96	-2.19 to -1.74
1.0	-1.53	-1.75 to -1.32
2.0	-1.30	-1.50 to -1.10
5.0	-0.58	-0.72 to -0.45

Table S4. Fixed-prediction rider resampling on the two historical test dates. Values are pointwise 95% intervals for the increment over each baseline. Each comparison had 1,000 attempted and valid draws. The analysis scripts in the repository cited in the main article generate prediction_intervals.csv with the full metric-specific results.

Baseline + G	AUC increment	AP increment	Brier difference	Log loss difference
linear	0.00703 to 0.01358	0.02131 to 0.04813	-0.00609 to -0.00267	-0.01469 to -0.00057
spline	-0.00022 to 0.00401	0.00361 to 0.02541	-0.00233 to 0.00027	-0.00696 to 0.00162
tree	-0.00236 to 0.00345	-0.00219 to 0.02159	-0.00227 to 0.00090	-0.00663 to 0.00290

The survey interval in Figure S9 resamples respondents. Existing exploratory figure intervals use the original rider draws. The independently drawn intervals in Tables S3 and S4 keep policy outcomes and test predictions fixed. Within the two test dates and each date separately, all metric-specific draws were defined. The side panels in Figure S5 retain the historical subgroup analysis.

S11. Allocation rules and input scenarios

Table S5. Recomputed positive counts and differences under each set of inputs. Unless indicated, n = 13,772. Counts in parentheses straddle the 0.02-min threshold. The difference is uniform minus capped in percentage points.

Input	Uniform	Capped	Uncapped	Difference (pp)	Retained (%)
Reference	835 (2)	565 (0)	470 (2)	1.96	74.0
Longitude at 30 deg	924 (2)	642 (1)	548 (3)	2.05	75.0
Longitude at 50 deg	751 (0)	499 (2)	412 (1)	1.83	74.3
Speed 12 km/h	1272 (2)	932 (1)	854 (1)	2.47	81.3
Speed 18 km/h	666 (1)	421 (0)	342 (1)	1.78	75.6
Service 0.25 min	760 (0)	498 (1)	409 (2)	1.90	74.6
Service 1 min	1077 (1)	788 (2)	717 (1)	2.10	80.3
No readiness wait	829 (2)	556 (0)	462 (2)	1.98	74.4

Input	Uniform	Capped	Uncapped	Difference (pp)	Retained (%)
Shared pickup stop	830 (1)	560 (0)	466 (2)	1.96	74.2
Initial rest 0 min	422 (0)	279 (0)	241 (0)	1.04	79.0
Initial rest 5 min	1354 (2)	914 (1)	792 (1)	3.19	78.3
Recorded readiness available (n = 13529)	813 (2)	546 (0)	455 (2)	1.97	74.6
Over 14 nodes (n = 47)	3 (1)	2 (1)	3 (1)	2.13	Undefined

The 30° and 50° tests change the longitude conversion while retaining both spatial coordinates. The readiness-subset row applies the reference inputs to 13,529 common cohorts after removing 243 cohorts containing a pending pickup with raw forecast zero. In the larger-cohort row the uncapped and uniform counts coincide, so their difference cannot be used as a denominator.

Table S6. Allocation alternatives on the 13,772 policy cohorts. Direction compares separated bounds with uniform allocation. The mean residual interval reports the cohort means of the lower and upper bounds (minutes).

Rule	Positive / undetermined	Mean bounds (min)	Improved / worsened	Extension P95 / max (min)
Capped	565 / 0	0.1960–0.1963	617 / 38	3.00 / 6.00
Route lateness	512 / 0	0.1875–0.1877	688 / 0	3.00 / 6.00
Slack rank	565 / 1	0.1941–0.1943	634 / 63	6.00 / 6.00
Uncapped	470 / 2	0.1685–0.1687	615 / 46	3.00 / 24.51
Uniform	835 / 2	0.2790–0.2793	0 / 0	3.00 / 3.00

A full-budget comparison isolates the placement of the minutes. In the optional-budget variant, an already feasible unextended route receives zero. Otherwise retain the capped route and its upper bound U, then assign each order the positive part of its completion minus its recorded deadline minus U. Tightening the deadlines retains the previous lower bound and this route attains the previous upper bound within numerical tolerance. For the zero-allocation shortcut, the original route certifies feasibility within 0.02 min. Across all 13,772 cohorts the resulting allocation used 6,227.01 order-minutes and left 109,650.99 unused. There were 36,483 zero-extension orders and 12,090 zero-extension cohorts. The construction gives one feasible use of the allowance at the retained bounds.

S12. Exact resource gaps and latency

Among the 42 feasible small instances, 24 had an exact minimum of zero and 18 had a positive minimum. Relative excess on the latter group averaged 21.5%, with median zero and maximum 155.7%. The analysis scripts generate exact_comparison.csv, which includes the residual bounds for the six infeasible instances. Across all 48, the largest capped upper bound was 9.15 min. The per-instance excess uses the positive part of completion minus the recorded deadline minus 0.02 min on the selected route, summed over orders, less the exact minimum under that same tolerance and cap.

Table S7. End-to-end calculation from one in-memory state through distances, individual completion, allocation, revised promises, joint bounds and earliest ending route. One serial call per cohort on an AMD Ryzen 7 9700X, Python 3.13.5. The first call, including

initialization of cached compiled functions, took 364.2 ms. Values below are warm elapsed times in milliseconds.

Remaining nodes	n	P50	P90	P95	P99	Max
All warm	13771	0.028	0.065	0.095	0.281	8.475
1	1841	0.023	0.042	0.057	0.089	0.275
2	2831	0.024	0.045	0.059	0.098	0.235
3	1987	0.025	0.044	0.057	0.111	0.190
4	2147	0.026	0.048	0.070	0.107	0.220
5	1207	0.028	0.048	0.065	0.105	0.163
6	1305	0.029	0.053	0.069	0.115	0.221
7	718	0.036	0.063	0.076	0.120	0.194
8	747	0.043	0.076	0.093	0.151	0.216
9	340	0.056	0.099	0.121	0.210	0.261
10	307	0.081	0.132	0.161	0.282	0.489
11	124	0.149	0.268	0.343	1.190	1.464
12	121	0.207	0.397	0.500	1.228	1.661
13	41	0.451	1.017	1.308	2.731	3.628
14	55	0.979	2.026	2.712	6.010	8.475

All 13,772 states met the numerical width target. Of them, 55 (0.40%) had 14 nodes. The separate large-state test retains bounds where the node limit prevents further search. Wall-clock values exclude disk reads, network communication and order ingestion.

S13. Decision comparisons

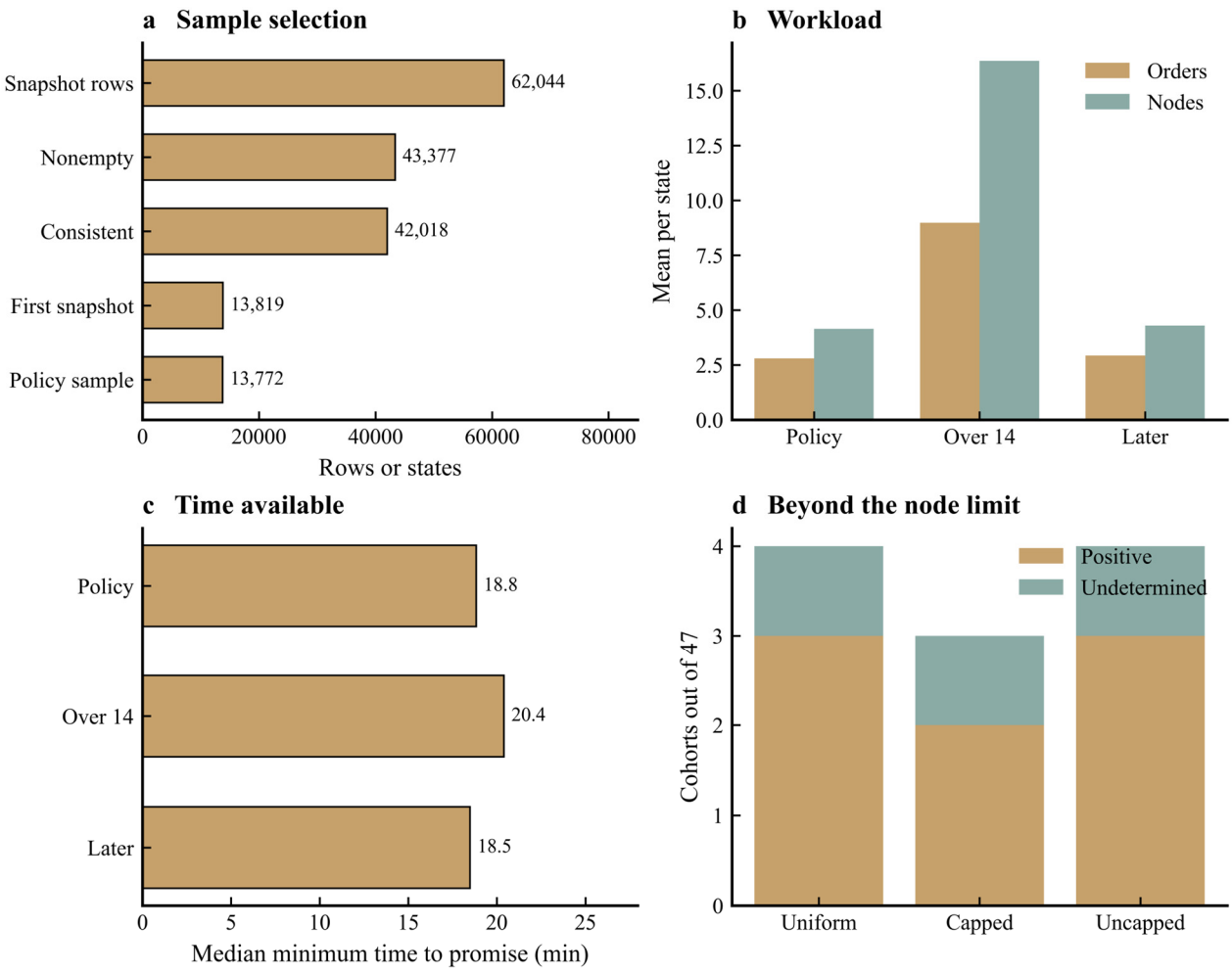

Figure S10. Sample selection and coverage. (a) Counts through the first daily snapshot and node screen. (b) Orders and nodes per state. (c) Median time to the tightest recorded promise. (d) Positive and threshold-undetermined counts for all 47 larger cohorts. The main sample contains 13,772 cohorts.

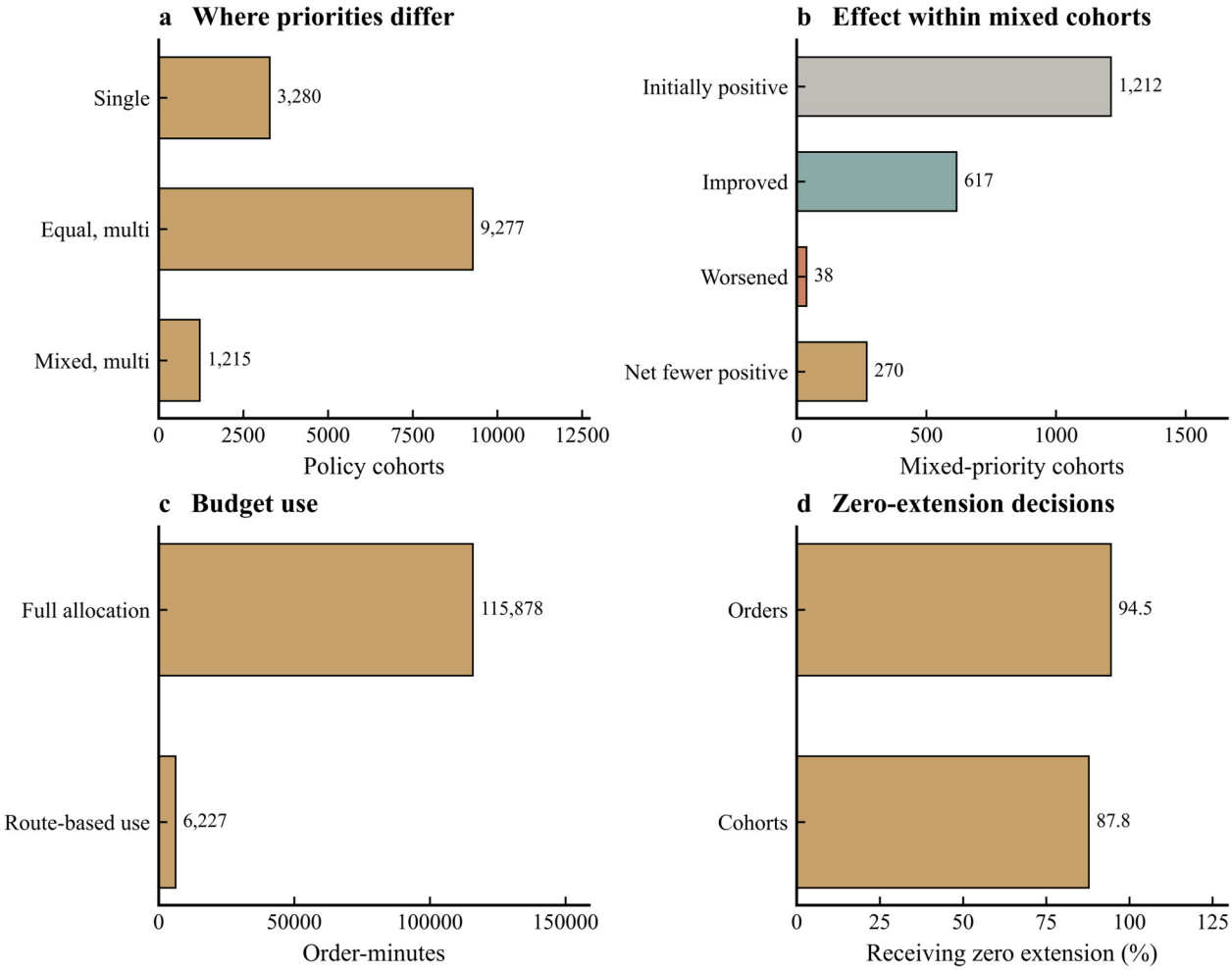

Figure S11. Priority groups and optional budget use. (a) Policy cohorts by priority composition. (b) Joint conflicts before allocation, paired changes in the deficit and net fewer positive cohorts within the mixed group. (c) Full budget versus minutes needed by the retained routes. (d) Shares receiving no extension, counted over 38,626 orders and 13,772 cohorts.

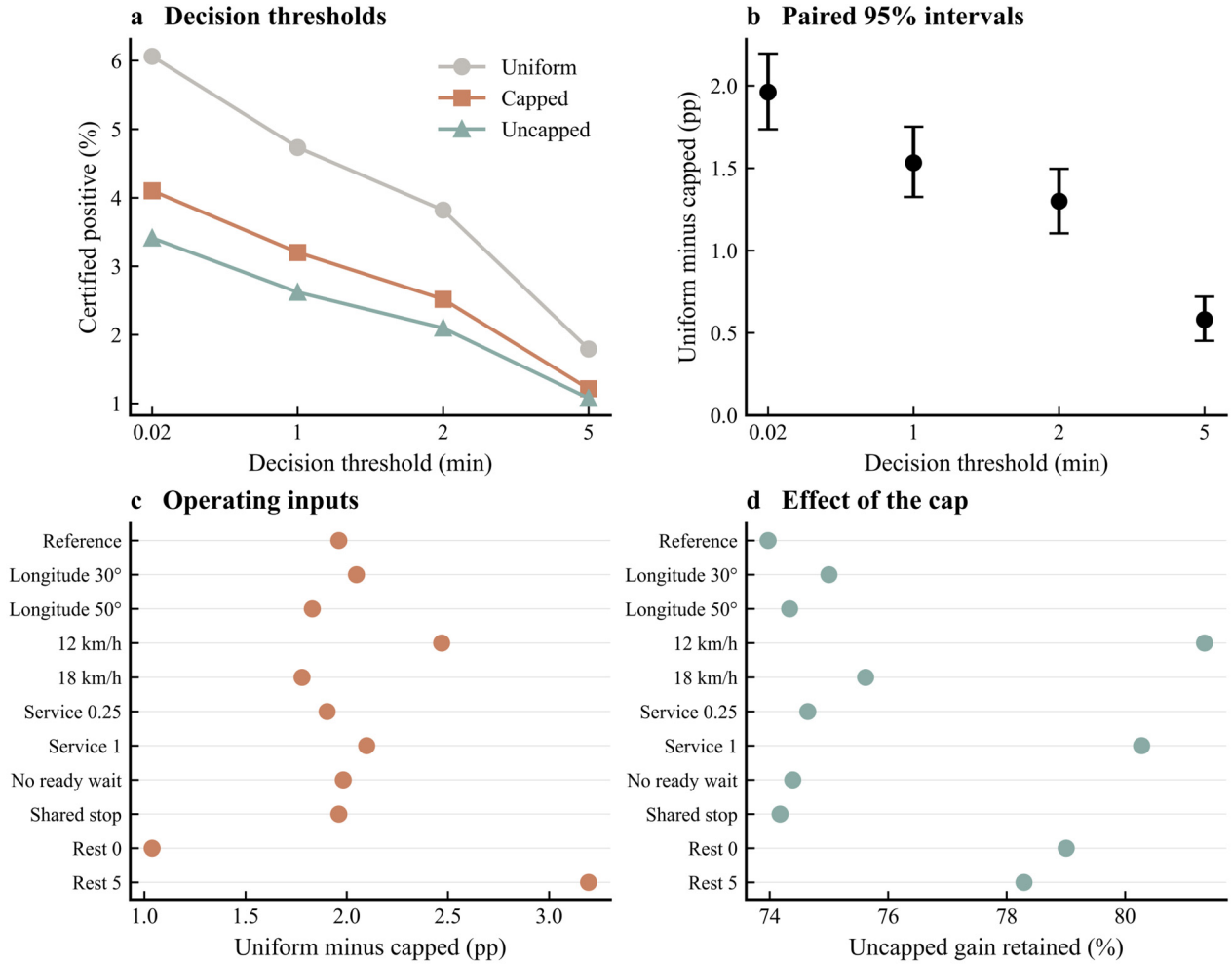

Figure S12. Decision thresholds and operating inputs. (a) Lower-bound-positive shares at four thresholds. (b) Paired percentage-point differences with pointwise 95% rider-bootstrap intervals (1,000 draws). (c,d) Capped reduction relative to uniform allocation and fraction of the uncapped reduction retained. Each row recalculates the policies on the same 13,772 cohorts.

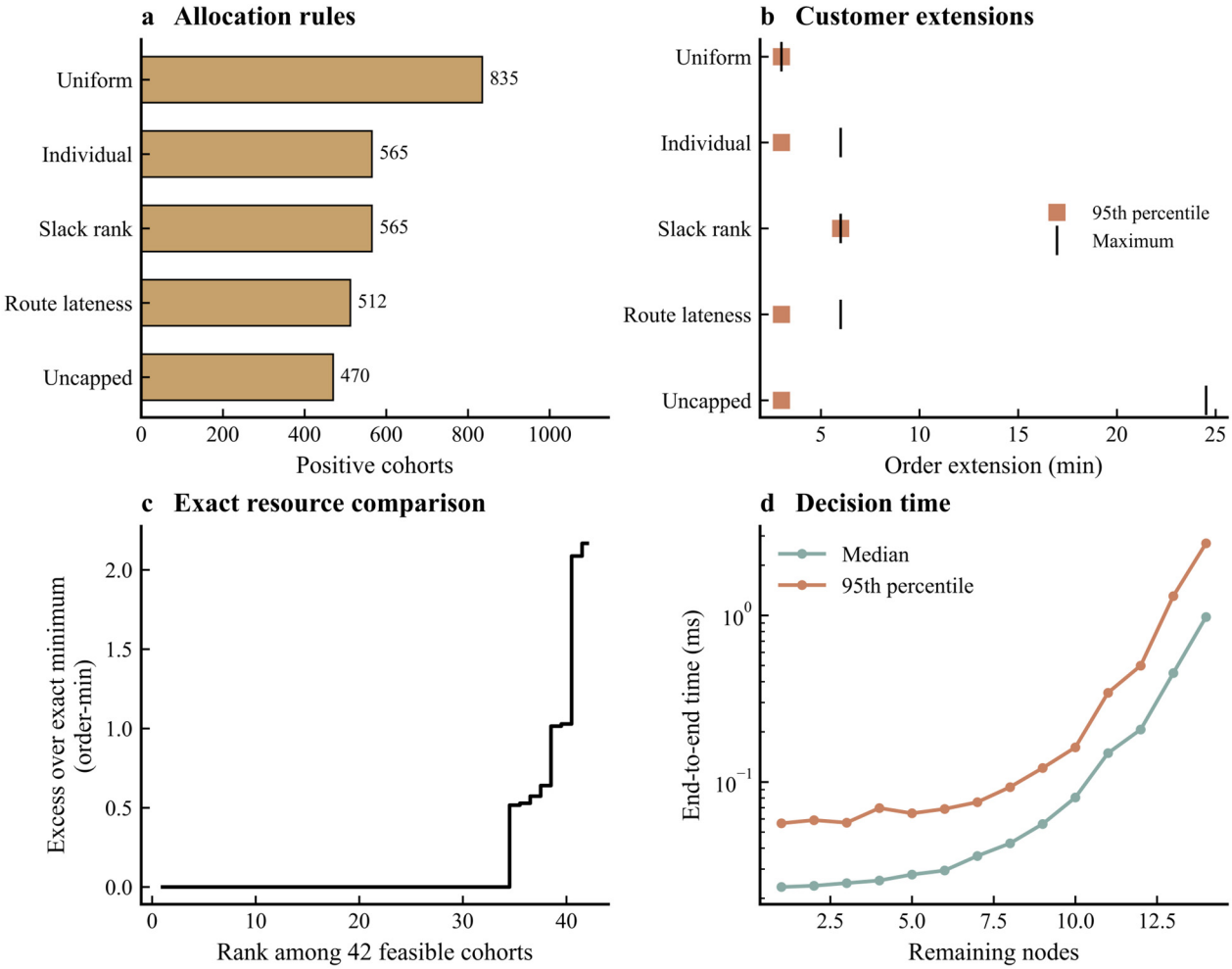

Figure S13. Allocation alternatives, exact resource use and computation. (a) Positive counts under five rules. (b) The 95th percentile and maximum extension per order. (c) Additional order-minutes required by the selected route over the exact minimum on 42 feasible instances. (d) Median and 95th percentile of warm end-to-end time by node count, on a logarithmic scale.

S14. Route certificates, extension distributions and algorithms

Table S8. Per-order extensions (minutes) for the 125 orders in the 42 feasible exact instances at the six-minute cap and 0.02-min tolerance. Required extensions are the positive excess beyond each recorded deadline and tolerance. The full allocation spends the entire allowance.

Order extension	Mean	Median	P95	Maximum
Assigned by capped individual rule	3.000	3.000	6.000	6.000
Required on its selected route	0.404	0.000	2.742	5.148
Required on minimum-resource route	0.336	0.000	2.697	4.441

For a fixed common shift, the following recurrence tests feasibility. The start time includes the initial rest. Subsets are visited in increasing mask order. A node with an outstanding pickup is eligible only after the pickup belongs to the visited set.

1. Initialize one label for each eligible first node. Add travel, wait until its release, then add service. Store the time if it meets the shifted deadline.

2. From each reachable (visited set, endpoint) label, try each unvisited eligible node. Apply the same travel, release, service and deadline rule. Retain the earliest new time and its predecessor for the enlarged set and endpoint.

3. A reachable full-set label certifies a feasible sequence. Choose the earliest full-set completion and follow its predecessors back to the start. A failed feasibility test raises the lower bound, and a successful test supplies an upper bound and a route. Stop the common-shift search when the bracket is at most 0.02 min wide.

4. To allocate the budget, divide unassigned minutes in proportion to the current positive weights. Fix all provisional shares exceeding the cap simultaneously, remove those orders and repeat. Write the remaining shares once they meet the cap, then rerun the feasibility test using the revised deadlines.

The analysis scripts available in the repository cited in the main article generate all_rule_intervals.csv for the common-rider resamples across rules and thresholds, rule_timing.csv for rule-specific processing times after distance construction, and feasible_exact_orders.csv for the aligned extension values underlying Table S8. These measurements supplement the full in-memory decision timings in Table S7. The scripts also generate the complete budget and cap grid and the 400 translation checks.