**Supplementary Materials**

**S1. Analytical flow**

The analytical cohort comprises 176,470 procedures, including 44,549 unsuccessful outcomes. Cross-source reconciliation aligned all 176,470 tender IDs and binary labels, with no split-membership changes; the raw extraction contains 443,178 source records. The checks below distinguish (a) agreement between cohort files and predictions, (b) independent operations on fitted checkpoints, and (c) a separately specified reconstruction from raw administrative events. These validation levels do not imply identical feature derivation or public observability.

The analytical unit is a unique tender procedure. The consolidated 2016–2025 cohort contains 176,470 finalized tenders, including 44,549 unsuccessful and 131,921 awarded procedures. The 2024 and 2025 forward tests contain 29,606 and 30,113 procedures, respectively. The source-data reconciliation matched all 176,470 tender IDs and found no target or temporal-split-membership changes.

**S2. Announcement-date diagnostics**

Among 204,632 raw unique procedures, 420 have multiple announcement dates; 370 occur in the analytical cohort. Within these 370 cohort cases, 197 cohort endpoints agree with the earliest observed date and 173 with the latest. The observed endpoints do not follow a single uniform earliest- or latest-date rule. Alternative endpoint choices can alter calendar-year assignment and historical windows; administrative placeholder dates must not be treated as public information timestamps. The 27-procedure 2016/2017 reassignment under an alternative annual-count convention is not identical to the number of multi-date cases.

Multiple announcement records per tender were consolidated chronologically. Administrative dates do not establish the first public availability of historical fields; historical feature analyses are therefore conditional rather than verified announcement-time deployment claims.

**S3. Feature admissibility**

Historical features are classified as (i) focal announcement-observable, (ii) historical announcement-only, (iii) historical outcome/competition/supplier-dependent with unverified first-publication time, or (iv) prohibited focal post-announcement. The 82 historical feature columns were computationally covered in the retained pipeline, and a separate exact file-to-file comparison of 176,470 rows × 82 columns found no differences against the archived historical input file (numeric tolerance $10^{-8}$, categorical equality, and separate missingness). This is agreement between those retained files, not equivalence of their construction to a fresh independent derivation from raw Excel. Independently specified raw-window diagnostics produced discrepancies dependent on source population, administrative announcement dates and buyer mapping, and public first-availability of historical results remains unresolved. A historical administrative date alone is insufficient to establish first-publication observability.

Primary predictors are focal title, announced budget, procurement type and procedure. Focal bids, bidder participation, winners and outcomes are excluded. Historical outcome/supplier/winner feature admissibility cannot be verified without field-level first-publication timestamps.

**S4. Prior buyer-announcement availability**

Under strict prior-event canonical dates, 6,712 tenders have zero reconstructed buyer-announcement history; earliest-source-date sensitivity yields 6,659. Exactly 53 focal tenders differ, explained by the dated eligibility of a small set of prior events. These are administrative-event timing proxies, not verified public-first-appearance times.

**Table S1. Announcement-history availability by year.**

| YEAR | n | announcement_history_eligib | mean_sequence_length | zero_history | verified_historical_outcome |
|---|---|---|---|---|---|
| 2016 | 336 | 160 | 3.238095238095238 | 176 | 0 |
| 2017 | 3817 | 3119 | 13.39795651034844 | 698 | 0 |
| 2018 | 9335 | 8437 | 15.573540439207283 | 898 | 0 |
| 2019 | 16256 | 15360 | 16.744032972440944 | 896 | 0 |
| 2020 | 21549 | 20833 | 17.66392872059028 | 716 | 0 |
| 2021 | 19673 | 19080 | 17.83652722004778 | 593 | 0 |
| 2022 | 21499 | 20968 | 18.100562816875204 | 531 | 0 |
| 2023 | 24286 | 23472 | 17.851889977764966 | 814 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| 2024 | 29606 | 28887 | 17.91238262514355 | 719 | 0 |
| 2025 | 30113 | 29442 | 18.120180652874176 | 671 | 0 |

**S5. Model specifications and available hyperparameters**

Model specifications are reported separately for the retained forward comparisons and additional sensitivity analyses. The evaluation design uses chronological training and testing, with model-specific parameters and software details described below.

Checkpoints and replay programs were available for the GRU and Transformer sequence branches. Independent evaluation of six checkpoint/sequence combinations for 2023, 2024 and 2025 reproduced 168,010 branch-level probability scores, using aligned tender keys and labels; the maximum absolute prediction discrepancy was $1.0543 \times 10^{-7}$ (tolerance $10^{-6}$). The sequence inputs encode up to 20 historical events. This verifies fitted-weight inference on stored sequence tensors, not fresh sequence formation from the raw Excel extract or refitting of every neural weight.

The announcement-only specifications use CatBoost for structured attributes and character n-gram TF–IDF for tender titles. The historical and sequence specifications are evaluated separately. The reported settings describe the evaluated specifications; complete tuning histories are not available.

For the distinct buyer-announcement-history CatBoost experiment, the verified 16-input training/testing matrices were verified against their recorded hashes. Training-year median imputation was applied separately within the chronological pre-test data for each forward year, and six .cbm checkpoints (seeds 2026–2028 × 2024/2025) reproduced all 179,157 seed-level probabilities from the corresponding matched model outputs to maximum absolute error $9.71 \times 10^{-17}$. These checkpoints belong to the distinct buyer-announcement-history CatBoost experiment and are not treated as the full-history Structured branch in Table 4. The checkpoint configuration used CatBoost 1.2.8 for the verified inference replay.

**S6. Development-period fusion coefficients**

Fusion coefficients fitted on the 2023 development period are distinguished from the retrospective 2024 coefficient-refit sensitivity evaluated on 2025 outcomes.

Three logistic fusion estimators were independently refitted on 24,286 matched 2023 development predictions using C = 0.1, solver = lbfgs and max_iter = 1000. All coefficient/intercept sets matched the stored models to numerical precision (maximum absolute coefficient difference $1.64 \times 10^{-13}$). To reproduce the GRU and Transformer coefficients it was necessary to use the architecture-specific matched 2023 development branch files, not similarly named columns from a different aligned archive. Forward inference with the recorded coefficients on clipped/logit-transformed probabilities reproduced the three fused model probabilities in each of 2024 and 2025: 6/6 model–year comparisons, 179,157 scores, maximum absolute error $2.22 \times 10^{-16}$. This refit establishes the matched retrospective development operation but does not constitute prospective model selection.

Structured + Text, Structured + Text + GRU, and Structured + Text + Transformer combine branch predictions using development-period logistic fusion. Historical-input specifications are conditional on the availability of the relevant fields before the focal announcement. The 2024 coefficient-refit analysis is a retrospective sensitivity evaluation.

**S7. Feature ablation**

The historical CatBoost ablation compares current-announcement A, buyer-announcement B and richer history families C–G. The B–A PR-AUC changes are +0.026353 (2024) and +0.022359 (2025); these use administrative-date prior announcements and remain conditional on historical publication-time eligibility. Outcomes, supplier/winner relations and exact-title recurrence are not promoted to strictly verified primary predictors.

**S8. Feature ablation paired buyer-cluster bootstrap (2,000 draws per comparison/year)**

Paired buyer-cluster bootstrap used 2,000 common cluster resamples per year/contrast. For B–A, 95% intervals for PR-AUC differences are [+0.004595,+0.046615] in 2024 and [+0.010562,+0.033092] in 2025. Historical richer-family intervals frequently include zero. Paired intervals quantify uncertainty; intervals containing zero do not prove equivalence.

For the retained Structured + Text comparator against GRU-augmented and Transformer-augmented fusion, a separate 2,000-resample buyer-cluster bootstrap per year/model/metric was recomputed from aligned saved probabilities. All twelve paired point differences agreed within $5 \times 10^{-6}$; all twelve independent 95% percentile intervals overlapped their reported

counterparts and all twelve produced the same zero-inclusion interpretation. The independent resampling seeds differ, so exact quantile endpoints are not expected to coincide. The 2024 Transformer PR-AUC difference was +0.002541 (independent 95% CI [−0.001576, +0.006838]); the 2025 difference was +0.001403 (independent CI [−0.004966, +0.006441]). The metric/interval results are conditional on saved fitted predictions, not an uncertainty distribution over retrained model weights.

**S9. Domain shift**

The expanded classifier includes procurement_type and therefore may exploit procedure coding. The expanded specification is therefore coding-sensitive. The fold-mean ROC-AUC values were 0.895707 and 0.669704 under the specified fold-local preprocessing protocol. No observed domain-classifier separation alone identifies a causal regulatory effect.

Domain classifiers use fold-local preprocessing. Fold-mean ROC-AUC for the expanded categorical versus three-fold numeric-only specifications was 0.895707 versus 0.616162 (2023–2024), and 0.669704 versus 0.605836 (2024–2025). "Coding-robust" does not demonstrate complete exclusion of coding proxies.

**S9A. Feature-level temporal distribution differences**

The full feature-distance table is exploratory: features must be screened for semantic coding and announcement-time admissibility before describing any particular distance as coding-independent. Jensen–Shannon and Wasserstein diagnostics are distributional summaries, not causal estimators.

Feature-level temporal differences are summarized using squared Jensen–Shannon divergence (base 2, in bits). Categorical variables use pooled empirical frequencies and numeric variables use pooled deciles. Numeric Wasserstein distances are standardized by the pooled standard deviation. These descriptive distances complement, but are distinct from, domain-classification performance.

**Table S2. Feature-level distribution-shift distances.**

| Years | Field | Type | JS²(bits) | W/pooled SD |
|---|---|---|---|---|
| 2023-2024 | log_budget | numeric | 0.0042 | 0.1085 |
| 2023-2024 | log_current_budget | numeric | 0.0044 | 0.1075 |
| 2023-2024 | current_year_budget_ratio | numeric | 0.0035 | 0.1464 |
| 2023-2024 | lot_flag | categorical | 0.0000 | NA |
| 2023-2024 | procurement_type | categorical | 0.0690 | NA |
| 2023-2024 | funding_source | categorical | 0.3603 | NA |
| 2023-2024 | national_investment | categorical | 0.0172 | NA |
| 2024-2025 | log_budget | numeric | 0.0025 | 0.0465 |
| 2024-2025 | log_current_budget | numeric | 0.0025 | 0.0628 |
| 2024-2025 | current_year_budget_ratio | numeric | 0.0000 | 0.1313 |
| 2024-2025 | lot_flag | categorical | 0.0028 | NA |
| 2024-2025 | procurement_type | categorical | 0.0069 | NA |
| 2024-2025 | funding_source | categorical | 0.0539 | NA |
| 2024-2025 | national_investment | categorical | 0.0069 | NA |

**S9B. Domain-classifier validation**

Domain classifiers were evaluated using three stratified folds with preprocessing fitted within each training fold. Reported fold means and sample standard deviations summarize the held-out fold metrics. The categorical specification includes procurement_type, which may retain information about procurement-method coding; the numeric-only specification provides a narrower sensitivity analysis.

A separate, stricter coding-excluded check used five stratified, fold-safe logistic-classifier fits per year transition, with log_budget, log_current_budget, current_year_budget_ratio and national_investment as numeric inputs; procedure type, funding source and categorical procedure codes were omitted. The five-fold mean ROC-AUC (fold SD, ddof = 1) was 0.540689 (0.002952) for 2023–2024 (n = 53,892), and 0.539297 (0.004414) for 2024–2025 (n = 59,719). These restricted-input values are not estimates for the expanded three-fold CatBoost diagnostic and substantially narrow the scope of a coding-independent shift claim.

The domain classifier used CatBoostClassifier (120 iterations, depth 5, learning rate 0.08, L2 regularization 4, Logloss, four threads). Three shuffled stratified folds used random state 2026 and model seeds 2027–2029. Numeric inputs were log_budget, log_current_budget, and current_year_budget_ratio; nonfinite values were handled using training-fold median imputation, with zero fallback. Categorical inputs were lot_flag, procurement_type, funding_source, and national_investment, with missing values encoded as __MISSING__. Numeric-only models excluded categorical inputs. Fold summaries use unweighted means and sample SD (ddof=1); pooled out-of-fold AUC is reported separately.

**Table S3. Domain-classifier fold-level validation.**

| Transition | Specification | Fold | n | ROC-AUC | PR-AUC | Model seed |
|---|---|---|---|---|---|---|
| 2023->2024 | coding_robust | 1 | 17964 | 0.894858 | 0.920523 | 2027 |
| 2023->2024 | coding_robust | 2 | 17964 | 0.896411 | 0.921678 | 2028 |
| 2023->2024 | coding_robust | 3 | 17964 | 0.895853 | 0.921861 | 2029 |
| 2023->2024 | numeric_only | 1 | 17964 | 0.615279 | 0.664097 | 2027 |
| 2023->2024 | numeric_only | 2 | 17964 | 0.616307 | 0.672011 | 2028 |
| 2023->2024 | numeric_only | 3 | 17964 | 0.616901 | 0.665551 | 2029 |
| 2024->2025 | coding_robust | 1 | 19907 | 0.668758 | 0.679829 | 2027 |
| 2024->2025 | coding_robust | 2 | 19906 | 0.669498 | 0.673459 | 2028 |
| 2024->2025 | coding_robust | 3 | 19906 | 0.670856 | 0.679043 | 2029 |
| 2024->2025 | numeric_only | 1 | 19907 | 0.611293 | 0.624466 | 2027 |
| 2024->2025 | numeric_only | 2 | 19906 | 0.610732 | 0.623080 | 2028 |
| 2024->2025 | numeric_only | 3 | 19906 | 0.595483 | 0.614314 | 2029 |

**Table S4. Domain-classifier fold-summary metrics.**

| Transition | Specification | Mean ROC | SD ROC | Pooled ROC | Mean PR | Pooled PR |
|---|---|---|---|---|---|---|
| 2023->2024 | coding_robust | 0.895707 | 0.000787 | 0.895552 | 0.921354 | 0.921718 |
| 2023->2024 | numeric_only | 0.616162 | 0.000821 | 0.615426 | 0.667220 | 0.666970 |
| 2024->2025 | coding_robust | 0.669704 | 0.001064 | 0.669448 | 0.677444 | 0.677219 |
| 2024->2025 | numeric_only | 0.605836 | 0.008970 | 0.604911 | 0.620620 | 0.620685 |

**S10. Yearwise models**

**Table S5. Conditional multiview performance (2024–2025).**

| Year | Model | PR-AUC | ROC-AUC | Brier |
|---|---|---|---|---|
| 2024 | GRU | 0.2929 | 0.5670 | 0.1868 |
| 2024 | Structured_Text | 0.3269 | 0.5789 | 0.1849 |
| 2024 | Structured_Text_GRU | 0.3292 | 0.5843 | 0.1844 |
| 2024 | Structured_Text_Transformer | 0.3294 | 0.5864 | 0.1843 |
| 2024 | TabM_style | 0.2898 | 0.5455 | 0.1874 |
| 2024 | Transformer | 0.2991 | 0.5725 | 0.1863 |
| 2024 | structured | 0.3111 | 0.5647 | 0.1861 |
| 2024 | text | 0.3036 | 0.5624 | 0.1886 |
| 2025 | GRU | 0.3000 | 0.5858 | 0.1798 |
| 2025 | Structured_Text | 0.3236 | 0.6165 | 0.1768 |
| 2025 | Structured_Text_GRU | 0.3269 | 0.6189 | 0.1764 |
| 2025 | Structured_Text_Transformer | 0.3250 | 0.6158 | 0.1767 |
| 2025 | TabM_style | 0.3068 | 0.6022 | 0.1783 |
| 2025 | Transformer | 0.2974 | 0.5811 | 0.1799 |
| 2025 | structured | 0.3111 | 0.6053 | 0.1784 |
| 2025 | text | 0.3007 | 0.5826 | 0.1805 |

**S10A. Retrospective 2024 coefficient-refit sensitivity for 2025**

The 2024 coefficient-refit comparison is a retrospective analysis of the matched branch predictions. The separate neural sensitivity analysis is reported in S19 using distinct row-aligned predictions; the results are interpreted according to their distinct training and fusion protocols.

This retrospective sensitivity compares penalized logistic fusion coefficients fitted on 2024 out-of-sample branch predictions with independently fitted 2023 development comparators. The comparator refits are distinct from the coefficients used to produce the main forward-evaluation probabilities. The results are summarized in Table S6.

For each matched branch combination, logistic fusion with C=0.1 was fitted on 2024 out-of-sample predictions and applied without updating coefficients to 2025 predictions. The corresponding 2023 development comparator was fitted independently. No 2025 outcomes were used to estimate coefficients. These comparisons are descriptive and do not replace the primary forward evaluation protocol.

**Table S6. Retrospective fusion coefficient-refit sensitivity.**

| Fusion | Fit year | Test n | PR-AUC | ROC-AUC | Brier |
|---|---|---|---|---|---|
| Structured_Text | 2023 | 30113 | 0.3236 | 0.6165 | 0.1768 |
| Structured_Text | 2024 | 30113 | 0.3249 | 0.6154 | 0.1773 |
| Structured_Text_GRU | 2023 | 30113 | 0.3269 | 0.6190 | 0.1764 |
| Structured_Text_GRU | 2024 | 30113 | 0.3301 | 0.6178 | 0.1764 |
| Structured_Text_Transformer | 2023 | 30113 | 0.3238 | 0.6164 | 0.1768 |
| Structured_Text_Transformer | 2024 | 30113 | 0.3166 | 0.6119 | 0.1773 |

**S11. Forward-prediction subgroup analyses**

Subgroup tests use saved forward predictions, not buyer-disjoint retraining. Separate buyer-disjoint tests give 2024 identity-bearing/free PR-AUC of 0.2808/0.2795 and 2025 of 0.2707/0.2739; paired intervals include zero.

**Table S7. Selected executed subgroup counts.**

| year | model | group_type | group_value | n | positives |
|---|---|---|---|---|---|
| 2024 | structured | seq_group | zero | 719 | 142 |
| 2024 | structured | seq_group | 1-2 | 553 | 126 |
| 2024 | structured | seq_group | 3-5 | 914 | 166 |
| 2024 | structured | seq_group | 6-10 | 1192 | 222 |
| 2024 | structured | seq_group | 11-20 | 26228 | 6773 |
| 2024 | structured | buyer_group | has_prior_buyer_announcement | 28887 | 7287 |
| 2024 | structured | buyer_group | no_prior_buyer_announcement | 719 | 142 |
| 2024 | structured | tender_type | Ажил | 8679 | 2451 |
| 2024 | structured | tender_type | Бараа | 18646 | 4308 |
| 2024 | structured | tender_type | Зөвлөх үйлчилгээ | 22 | 2 |
| 2024 | structured | tender_type | Зөвлөхөөс бусад үйлчилгээ | 13 | 2 |
| 2024 | structured | tender_type | Үйлчилгээ | 2246 | 666 |
| 2024 | structured | budget_group | Q1 | 5515 | 1223 |
| 2024 | structured | budget_group | Q2 | 7913 | 1904 |
| 2024 | structured | budget_group | Q3 | 8285 | 2172 |
| 2024 | structured | budget_group | Q4 | 7893 | 2130 |
| 2025 | structured | seq_group | zero | 671 | 158 |

**S11B. Transformer-fusion subgroup analyses for 2024**

Table S8 reports the 2024 subgroup decomposition. It is descriptive unless buyer-cluster confidence intervals from the same aligned model predictions are verified. The additional sensitivity subgroup estimates in S20 use different fitted models and cannot be merged into Table S8.

The 2024 subgroup comparisons use aligned forward predictions and define ΔPR-AUC as Transformer-augmented fusion minus Structured + Text. Groups are defined by prior buyer-announcement availability and sequence length. Subgroup estimates are descriptive because subgroup-specific clustered confidence intervals were not available. A consistent project-novelty grouping could not be established, so no project-novelty estimates are reported.

**Table S8. Transformer-fusion 2024 subgroup comparisons.**

| Family | Subgroup | n | Positive | Base PR | Transformer PR | Delta |
|---|---|---|---|---|---|---|
| seq_group | 1-2 | 553 | 126 | 0.318240 | 0.295340 | -0.022900 |
| seq_group | 11-20 | 26228 | 6773 | 0.332449 | 0.335433 | +0.002984 |
| seq_group | 3-5 | 914 | 166 | 0.259320 | 0.245507 | -0.013813 |
| seq_group | 6-10 | 1192 | 222 | 0.268038 | 0.271554 | +0.003516 |
| seq_group | zero | 719 | 142 | 0.203240 | 0.201207 | -0.002034 |
| buyer_group | has_prior_buyer_announcement | 28887 | 7287 | 0.328363 | 0.330915 | +0.002552 |
| buyer_group | no_prior_buyer_announcement | 719 | 142 | 0.203240 | 0.201207 | -0.002034 |

**S12. Calibration**

Chronological native and Platt-scaled calibration estimates were obtained in an auxiliary evaluation protocol separate from the primary fusion and selective-prediction analyses. The auxiliary estimates should not be interpreted as calibration of the primary fusion probabilities.

Auxiliary chronological Platt scaling was fitted on a prior-year calibration slice, separate from the forward evaluation year. In 2024 native versus Platt Brier was 0.1848868 versus 0.1851381 and ECE was 0.0181075 versus 0.0238118. Calibration improvement is therefore not claimed. This auxiliary procedure is distinct from the native fusion protocol.

**Table S9. Auxiliary calibration sample sizes.**

| evaluation_year | model | train_fusion_n | platt_n | conformal_n | n |
|---|---|---|---|---|---|
| 2024 | AuxiliaryNative | 21312 | 2002 | 972 | 29606 |
| 2024 | AuxiliaryPlatt | 21312 | 2002 | 972 | 29606 |
| 2025 | AuxiliaryNative | 25704 | 3002 | 900 | 30113 |
| 2025 | AuxiliaryPlatt | 25704 | 3002 | 900 | 30113 |

**S13. Selective prediction**

Selective risk–coverage curves were computed retrospectively from native conditional Structured + Text probabilities. Lower retained coverage was associated with lower observed classification error, but the curves do not establish an optimal prospective deferral threshold or review-capacity policy.

The selective-prediction probabilities differ from those used in the auxiliary calibration and conformal analyses; their reliability metrics are therefore reported separately.

**S14. Conformal**

The auxiliary chronological split-conformal evaluation yielded 2024 empirical coverage of 86.44% and 90.09% at nominal levels of 90% and 95%, respectively. Temporal distribution shift may invalidate the exchangeability conditions needed for conventional marginal coverage guarantees.

The conformal results refer to a separate prior-year calibration protocol and do not provide coverage guarantees for the primary rolling-model probabilities.

**Table S10. Auxiliary conformal coverage.**

| evaluation_year | nominal_coverage | empirical_coverage | mean_set_size | empty_set_rate | two_class_set_rate |
|---|---|---|---|---|---|
| 2024 | 0.9 | 0.8644193744511248 | 1.3766466256839829 | 0.0 | 0.376646625683983 |
| 2024 | 0.95 | 0.9008984665270553 | 1.5259406876984396 | 0.0 | 0.5259406876984395 |
| 2025 | 0.9 | 0.9137913857802278 | 1.5069903363995616 | 0.0 | 0.5069903363995617 |
| 2025 | 0.95 | 0.9495898781257265 | 1.6830936804702288 | 0.0 | 0.6830936804702288 |

**S15. Buyer-identity and buyer-disjoint comparisons**

Direct buyer identifiers excluded from predictors do not imply the absence of residual identity signals. Buyer-ID ablation, held-out buyer retraining, and size-matched controls address different estimands; buyer-name-normalized cluster identifiers are proxies and may imperfectly represent legal procuring entities.

**Table S11. Buyer-ID and buyer-disjoint comparisons.**

| Setting | Year | Identity-free PR-AUC | Buyer-ID PR-AUC |
|---|---|---|---|
| Ordinary forward | 2024 | 0.304409 | 0.327065 |
| Ordinary forward | 2025 | 0.309070 | 0.329119 |
| Buyer-disjoint | 2024 | 0.2795 | 0.2808 |
| Buyer-disjoint | 2025 | 0.2739 | 0.2707 |

**S16. Implementation protocol and information-timing constraints**

For each focal tender i at announcement $a_i$, eligible historical event j must satisfy $a_i - W \leq a_j < a_i$ for $W \in \{365,730,1095\}$ calendar days, subject to the additional requirement that the source information was publicly observable before $a_i$. This operational interval includes the lower boundary and excludes the focal boundary. Without verified publication timestamps, supplier, winner, bid-price and outcome histories remain conditional sensitivity features. Deduplication and sample-versus-population dispersion choices must be specified by feature family; undocumented choices are not reconstructed by matching target values.

The unit of analysis is the tender procedure. The raw extract contains 443,178 records and the finalized cohort contains 176,470 tenders. The outcome equals one for an unsuccessful tender and zero for an awarded tender. Training covers 2016–2022, development covers 2023, and forward evaluations cover 2024 and 2025. Historical event eligibility is based on administrative dates; first-publication times of outcome-, bid-, supplier-, and winner-dependent fields were not established. Same-day events are excluded from prior histories, and sequences retain up to 20 events.

Let i index a tender, b(i) its buyer, a_i its announcement date, and y_i its finalized outcome. Current-announcement predictors are drawn from attributes and title available at a_i. A prior buyer announcement h is eligible when b(h)=b(i), h≠i, and a_h<a_i. For a window of w days, the intended administrative-date condition is a_i−w≤a_h<a_i. The inclusion of outcome-, supplier-, winner-, or bid-dependent historical information additionally requires public availability before a_i; without first-publication timestamps, these inputs remain conditional. The exact boundary convention for some historical windows, bid-price coefficient-of-variation treatment, and multiple-winner deduplication are not fully established for every historical specification.

For each tender, eligible prior announcements are ordered chronologically, with up to 20 retained for sequence inputs. Announcement-count and budget aggregates use full-history and 365-, 730-, and 1095-day windows where applicable. Focal bids and outcomes are excluded from predictors. Preprocessing and text vocabulary are fitted on training data only; fusion is fitted using development-period predictions.

**S17. Hypothesis-linked experimental design**

**Table S12. Hypothesis-linked design and provenance.**

| Hypothesis / comparison | Model and features | Period / selection | Reported evidence | Reconstructed settings / provenance |
|---|---|---|---|---|
| H1 Historical announcement contribution | CatBoost A: current announced fields vs B: prior buyer announcements | 2016–22 train; 2023 dev; 2024/25 tests | ΔPR-AUC 2024 +0.02635 [0.00460,0.04662]; 2025 +0.02236 [0.01056,0.03309] | Historical 82-field dictionary; 2,000 buyer-cluster draws; historical observability conditional |
| H2 Temporal distribution shift | Fold-safe coding-robust vs numeric-only domain classifiers | 2023–24 and 2024–25, fold-local preprocessing | Coding-robust ROC-AUC 0.8957 / 0.6697; numeric-only 0.6162 / 0.6058 (fold-mean) | Archived execute_fold_safe_domain.py, Domain_OOF.csv; S9 JS/Wasserstein |
| H3 Buyer ID and cold start | Fixed CatBoost buyer ID present/absent; buyer-disjoint; cold-start | 2016–23 train→2024; 2016–24 train→2025 | Matched PR-AUC 2024 0.327065 vs 0.304409; 2025 0.329119 vs 0.309070 | CatBoost 220 trees, depth 6, lr .045, L2 6, random_strength 1, seed 2026, threads 2; disjoint separate |
| H4 Incremental sequence value | Structured+Text vs +GRU / +Transformer | 2023 development coefficients; 2024/25 tests | Transformer ΔPR-AUC 2024 +0.00254 CI [-0.00148,0.00687]; 2025 +0.00140 CI [-0.00491,0.00655] | Transformer d_model 24, 4 heads, FF48, 1 layer, 4 epochs; GRU exact original hyperparameters unavailable |
| H5 Reliability | Calibration, conformal, selective risk | Auxiliary chronological evaluation | 90% conformal 2024 86.44%, 2025 91.38%; calibration not improved | S12–S14 auxiliary predictions and uncertainty; not strict primary claim |

The numerical and inference analyses used Python 3.13.5, CatBoost 1.2.8, scikit-learn 1.8.0, NumPy 2.3.5, pandas 2.2.3 and PyTorch 2.10.0+cpu; model-specific artifact hashes and matched-input checks are summarized in S21. Neural checkpoint replay was executed with the available CPU runtime, without inferring training GPU configuration, wall times or hyperparameter-search histories beyond the available experiment records.

**S18. Feature specifications and generalization analyses**

Historical feature definitions and their sensitivity to announcement-date conventions are summarized in this section. Administrative recording dates do not always establish the first date at which outcome or competition information became public; these fields are therefore treated separately from announcement-only predictors.

Among the 82 engineered historical fields in the conservative admissibility registry, 11 announcement-history candidates require additional date/source verification, six project-family candidates require title/identity and timestamp validation, four opening-event timing variables have conditional availability, and 61 remaining historical variables are restricted pending first-publication evidence. No historical field in this 82-feature registry was formally certified as observable at announcement solely from the available administrative extract. This statement does not apply to current-announcement structured attributes or title text. The raw-history alternatives also differ from the retained feature values, even though the two archived 82-column files agree exactly; neither comparison should be substituted for the other.

The current-tender structured variables include announced budget, announcement calendar fields, tender type, and procurement procedure. The text branch uses title character n-gram TF–IDF with training-only vocabulary. The matched buyer-identity comparison additionally uses prior buyer-announcement availability, elapsed time since prior announcements, and historical announcement counts and budget summaries across full-history and 365-, 730-, and 1095-day windows. An explicit buyer ID is included only in the identity-bearing comparator. Other historical feature families may require additional publication-time eligibility checks.

In the 2024 cold-start subgroup without prior buyer announcements (n=719; unsuccessful n=142), buyer-ID and identifier-free PR-AUC values were 0.227611 and 0.245365. Among tenders with prior buyer announcements (n=28,887; unsuccessful n=7,287), the corresponding values were 0.328284 and 0.305332. These are descriptive subgroup comparisons and do not establish causal attribution to buyer identity.

The 2025 retrospective fusion comparison evaluates independently fitted 2023 and 2024 coefficient sets on the same 2025 outcomes. The 2025 outcomes were used for evaluation only.

**S19. Additional neural-fusion sensitivity analysis**

The following PR-AUC values were recalculated from a separately trained, row-aligned neural sensitivity analysis. Its 2024-refit sensitivity is a retrospective experiment. These sensitivity-model estimates are reported separately from the main forward-evaluation models.

**Table S13. Retrospective neural-fusion sensitivity: PR-AUC by evaluation year and model specification. The sensitivity-model fits are distinct from the primary forward-evaluation fits.**

| Evaluation | Model/protocol | PR-AUC |
|---|---|---|
| 2025 | Structured + Text pilot | 0.259601 |
| 2025 | Transformer fusion, 2023-frozen pilot | 0.263621 |
| 2025 | Transformer fusion, 2024-refit pilot | 0.26418 (prior executed pilot; precision limited) |

**S20. Additional subgroup sensitivity analysis**

Delta PR-AUC is Transformer-augmented sensitivity fusion minus Structured + Text sensitivity. Intervals are 95% percentile intervals from 300 paired resamples of normalized buyer-name proxy clusters, without multiplicity adjustment. Buyer-history-depth threshold ≤159 was derived from the development/training reference; project-family novelty is a proxy and its public-time observability is not fully verified. These estimates are not uncertainty intervals for retained Table 6 or Table S8.

**Table S14. Exploratory Transformer augmentation by buyer-history depth and project-family novelty proxy, 2024 and 2025. Entries are paired PR-AUC differences with 95% percentile intervals from 300 buyer-name-proxy cluster resamples.**

| Stratum | 2024 ΔPR-AUC [95% CI] | 2025 ΔPR-AUC [95% CI] |
|---|---|---|
| Buyer history ≤159 | +0.00768 [0.00099, 0.01424] | +0.01163 [0.00405, 0.01968] |
| Buyer history >159 | −0.00096 [−0.00685, 0.00488] | −0.00324 [−0.00819, 0.00260] |
| New project-family proxy | +0.00550 [0.00116, 0.00986] | +0.00549 [0.00005, 0.01130] |
| Previously seen project-family proxy | +0.00323 [−0.00411, 0.01016] | +0.00115 [−0.00595, 0.00848] |

**S21. Computational verification of model outputs and uncertainty**

Probability files, model checkpoints, development branch predictions, feature files and evaluation labels were checked using prediction-level calculations. Probabilities were joined one-to-one using tender ID, outcome and evaluation year

where applicable, and PR-AUC was implemented as average precision. Inference replay and metric recalculation use saved model inputs, not a new fit of each originally reported model. Table S15 records the operation, its observed agreement, and the boundary of the inference it supports.

For the CatBoost buyer-announcement-history experiment, training-year imputation and 16-column input order were applied before inference with fitted checkpoints. For the neural branches, checkpoint state dictionaries were executed on the 20-event sequence tensors. Development fusion weights were refitted using matched 2023 branch predictions and were replayed through the logit-link transform using probabilities clipped to $[10^{-7}, 1 - 10^{-7}]$. Paired bootstrap results used separate independent resampling streams (2,000 cluster replicates); equality of percentile endpoints is not expected when seeds differ.

**Table S15. Independent computational verification and the scope of inference for the retained evaluation artifacts.**

| Component | Independent executed check | Interpretation / boundary |
|---|---|---|
| Cohort identity and outcome | 176,470 tender IDs; 44,549 positive labels; 0 cohort-ID / label discrepancies | Confirms compared tender-level cohorts, not field publication time |
| 82 historical feature columns | 176,470 × 82 retained-versus-archived matrix cells: no value or missingness differences at specified tolerances | Exact stored-file alignment; independent raw-window derivation is distinct |
| Table 4 metrics | 48/48 PR-AUC, ROC-AUC and Brier metric cells reproduced at displayed precision from aligned probabilities | Probability-to-metric calculation, not model retraining |
| GRU and Transformer branches | 6/6 checkpoint/year tests; 168,010 prediction scores; maximum absolute error $1.0543 \times 10^{-7}$ | Archived checkpoint and sequence tensor inference |
| Historical CatBoost B experiment | 6/6 seed/year tests; 179,157 prediction scores; maximum error $9.71 \times 10^{-17}$ | Distinct 16-input buyer-history experiment, not Table 4 full-history Structured branch |
| 2023 fusion estimators | 3/3 fitted coefficient/intercept sets reproduced; n = 24,286; C = 0.1 | Matched retrospective development prediction specification |
| 2024/2025 fused probabilities | 6/6 model/year comparisons; 179,157 scores; maximum difference $2.22 \times 10^{-16}$ | Exact probability replay, not prospective choice of architecture |
| Table 5 paired differences | 12/12 point contrasts within $5 \times 10^{-6}$; 12/12 independent 95% intervals overlap and agree on zero inclusion | 2,000-draw buyer-cluster bootstrap conditional on saved probabilities |
| Procedure-coding-restricted shift | 5-fold numeric logistic ROC-AUC: 0.540689 (2023–24), 0.539297 (2024–25) | Distinct from 3-fold expanded CatBoost and original archived AUCs |
| Exploratory neural pilot subgroup | 59,719 forward tender IDs, years and labels aligned; 300 buyer-name-proxy clustered resamples | Distinct new sensitivity model; not intervals for Table 6 retained model |

The checks above establish numerical reproducibility of the specified evaluation components. Full raw-extract-to-all-82-feature equality, field-specific first-publication admissibility, complete historical model-tuning trajectories, and independent retraining of every original model are distinct from checkpoint inference and metric replay. The sensitivity analyses in S3, S9B, S19 and S20 are interpreted as sensitivity analyses under their specified protocols. The intended use is comparative decision support with uncertainty-aware human review, not fully automated procurement adjudication.