Table S1 presents the results of the initial quality control of records following URL canonicalization. In the source corpus, 425 duplicate source URLs and 1,107 duplicate canonical identities were identified. One canonical URL contained inconsistent markup, which led to the exclusion of two associated records. After removing the identified duplicates, the source cohort was reduced from 235,795 to 234,687 observations. In the target corpus, no exact duplicates were found; however, 23 duplicate canonical identities were identified. These were resolved according to a pre-established rule retaining the first identified record resulting in a final count of 168,037 URLs.

Table S1**.** Results of the check for duplicates and conflicting URL identities

Dataset	Invalid URLs	Exact duplicates	Canonical duplicates	Conflicting canonical URLs	Conflicting lines removed	Redundant repetitions removed	Saved	Mixed-label domains
Source	0	425	1 107	1	2	1 106	234 687	105
Target	0	0	23	0	0	23	168 037	123

It is crucial that mixed-label domains were not interpreted as erroneous entries. In the source and target corpora, 105 and 123 domains, respectively, were identified as containing URLs of different classes. These domains were retained because a single registered domain can simultaneously host both legitimate and phishing pages; removing such instances would have artificially simplified the task. Conversely, a conflict was considered critical only when the exact same identity or canonical identity was assigned different labels. Thus, the cleaning procedure is aimed at resolving ambiguous object identities, rather than creating artificially homogeneous domains. The results indicate that the primary loss of entries is not due to URL errors or labeling conflicts, but rather to the management of recurring identities. This reduces the risk of the same object appearing in different experimental roles and ensures a more accurate assessment of the models' generalization capability.

Table S2 illustrates how the initial chronological distribution of the target data changed following forward domain purging. In the adaptation phase, all 29,016 observations were retained, as this role occurs first in the temporal sequence and cannot contain domains that appeared earlier. For the selection phase, 10,162 URLs were initially available; however, 4,318 of these entries belonged to domains that had already appeared in the adaptation phase, resulting in a final count of 5,844 URLs. Similarly, 11,876 out of 33,606 entries were excluded from the calibration phase, and 12,861 out of 59,803 were excluded from the final test phase.

Table S2**.** The impact of forward domain purging on the chronological roles of target data

Role	Before cleaning	Deleted	Saved	Domains
Adapt	29 016	0	29 016	16 662
Selection	10 162	4 318	5 844	5 279
Calibration	33 606	11 876	21 730	17 216
Test	59 803	12 861	46 942	33 428

The result highlights a key distinction between group-aware temporal validation and a simple chronological data split. If domains that reappeared over time were retained, later-stage evaluation would partially measure the model's ability to recognize previously encountered domain groups, rather than its capacity to generalize to new entities. In this protocol, each subsequent role contains only domains that were not observed in any preceding target roles. The highest absolute number of exclusions occurs in the test set, confirming a significant degree of domain recurrence over time. Consequently, the application of purging substantially increases the rigor of the final evaluation and prevents the inflation of results due to temporal domain recurrence.

Table S3 summarizes 37 individual leakage checks performed after the final assignment of roles. These checks covered three independent identifiers for potential overlap: the registered domain, row_id, and the canonical URL. For the source corpus, the fit, selection, calibration, and test sets were compared pairwise, resulting in six role combinations. An identical, comprehensive set of checks was performed for the target-domain fit, selection, calibration, and test sets. Additionally, the initial overlap between the source and target domains was examined separately.

Table S3**.** Comprehensive audit of the absence of informational overlap between experimental roles

Level being checked	Number of pairwise comparisons	Maximum intersection found
Source–target registrable domains	1	0
Source roles: domain	6	0
Source roles: row ID	6	0
Source roles: canonical URL	6	0
Target roles: domain	6	0
Target roles: row ID	6	0
Target roles: canonical URL	6	0
Total checks	37	0

In all cases, the number of shared objects is zero. This confirms that domain independence is not merely a stated rule for creating the split, but is verified directly following the formation of the datasets. Furthermore, the zero intersection of row_id values precludes the reuse of any single record, while the absence of overlapping canonical URLs protects against hidden textual duplicates that might have persisted even after URL normalization. Such multi-layered verification is crucial, as relying on a single form of identity is insufficient; for instance, multiple rows may correspond to the same canonical URL, and various URLs may belong to the same domain. Verifying all three levels simultaneously significantly reduces the likelihood of hidden leakage between the development and evaluation populations.

Table S4 illustrates the further partitioning of the target calibration population following chronological and domain-based filtering. Rather than using the same dataset for both probability calibration and the selection of operating thresholds, two independent sets of domains were established. The probability calibration set comprises 5,509 observations from 4,304 domains, including 3,048 legitimate URLs associated with 2,857 benign domains. The threshold calibration set is substantially larger, containing 16,221 observations from 12,912 domains, including 9,087 legitimate URLs across 8,572 benign domains. This separation is necessary due to the distinct statistical functions of the two procedures: probability calibration estimates the mapping of a model score to a probability, whereas threshold calibration defines an operating rule based on a specified, acceptable false-positive rate. Using the same sample for both tasks would introduce an unintended dependency between the probability estimation and the selection of the operating threshold.

Table S4**.** Division of the calibration population into independent functional roles

Role	**Observations**	**Domains**	**Legitimate URL**	**Benign domains**
Probability calibration	5 509	4 304	3 048	2 857
Threshold calibration	16 221	12 912	9 087	8 572

Distinguishing "benign domains" is crucial for low-FPR (false-positive rate) analysis, as one cannot automatically assume that entries within the same domain are independent. Consequently, the assessment of statistical support for rare false-positive events is based on the number of independent domain groups, rather than the total count of URLs. Table S5 evaluates whether the threshold-calibration population is sufficiently large to handle rare false-positive occurrences. For a nominal FPR of 1%, the required number of independent benign domains based on the zero-exceedance criterion employed—is 299, whereas the actual calibration cohort contains 8,572 such domains. For a significantly stricter FPR of 0.1%, the required support increases to

2,995 independent groups; however, the available number of domains still exceeds this threshold by nearly a factor of three. A status of Supported=True in the table indicates only that the calibration sample is sufficiently large according to the adopted statistical criterion. It should not be interpreted as a guarantee that the actual FPR will remain at its nominal value when applied to subsequent data. Such a guarantee would further require the distribution of benign scores to be stable an assumption that does not hold in the presence of a temporal shift.

Table S5**.** Statistical support for specified low-FPR operating levels

Номинальный FPR	**Независимые benign domains**	**Минимально необходимое число доменов**	**Поддержка**
0,010	8 572	299	*Д*а
0,001	8 572	2 995	*Д*а

Thus, the table confirms that the calibration cohort is quantitatively sufficient for establishing rare operating thresholds, while simultaneously maintaining a fundamental distinction between sample support and future operating guarantee. This distinction must be taken into account when comparing ROC-derived low-FPR metrics with results obtained using fixed thresholds on a locked test set. Table S6 concludes the description of the research design; it specifies not the numerical size of the samples, but rather the purpose of each role and the appropriate stage at which it may be used. For the source corpus, 70% of the domain groups are allocated for training, with 10% each assigned to development, calibration, and source testing. For the target data, the roles are linked to sequential time intervals: the first 45% of unique dates are used for adaptation, the next 10% for selection, the following 20% for calibration, and the final 25% for the locked test. The calibration interval is further subdivided based on domain groups into segments for probability calibration and threshold calibration. Consequently, the function of each sample is defined in advance and remains unchanged following the observation of model performance.

Table S6**.** Fixed protocol for access to experimental roles

Dataset	**Role**	**Formation rule**	**Access to freeze**
Source	Fit	70% domain groups	Training
Source	Selection	10% domain groups	Development
Source	Calibration	10% domain groups	Calibration
Source	Test	10% domain groups	Notebook 06 only
Target	Adapt	First 45% unique dates	Training
Target	Selection	Next 10% unique dates	Development
Target	Probability calibration	25% domains of next 20% dates	Probability only
Target	Threshold calibration	75% domains of next 20% dates	Threshold only
Target	Locked test	Last 25% unique dates	Notebook 06 only

Figure S1 illustrates the daily dynamics in the number of legitimate and phishing URLs within the development period of the target corpus. The timeframe spans from September 8 to October 23, 2025 covering the adaptation, selection, and calibration phases prior to the locked test. The daily number of observations ranges from 254 to 4,414, indicating an uneven intensity in data collection. This variation affects not only the total number of URLs but also the ratio between the classes. Such temporal heterogeneity carries significant methodological implications. A random row-level split would intermix observations from different dates across the training and validation sets; consequently, the resulting evaluation would reflect the model's

average interpolation capability while partially obscuring any chronological shift. In contrast, the temporal scheme employed here preserves the natural chronological order in which the data were collected.

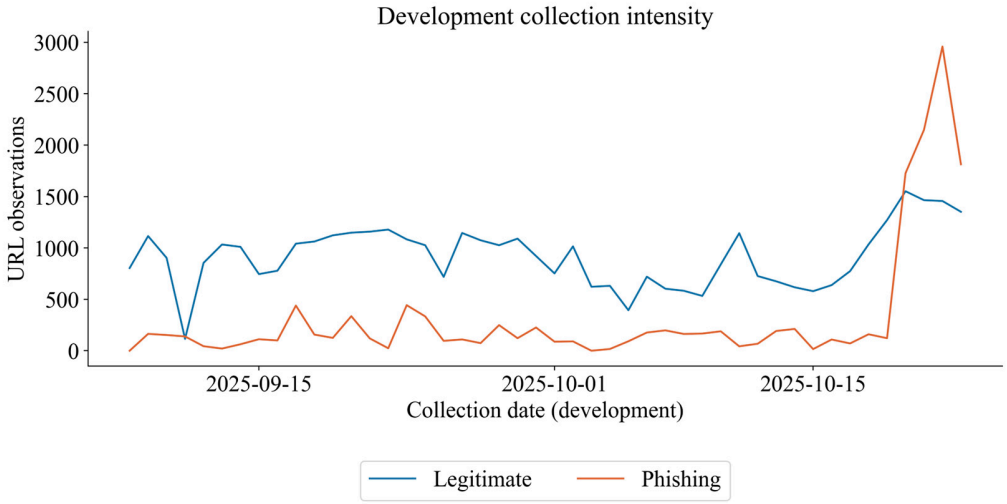


Figure S1. Development Collection Intensity

The chart further explains why the boundaries for the target roles were defined based on ordered, unique dates rather than a specific number of rows. Given varying data collection intensities, an identical number of rows could correspond to significantly different time intervals, thereby altering the nature of the experimental scenario. Ultimately, Figure S1 demonstrates that the temporal factor cannot be dismissed as a mere technical detail. Fluctuations in the volume and composition of observations over time constitute a distribution shift that the model must be robust enough to handle. Consequently, the subsequent locked-test period serves as a distinct evaluation of the model's transferability to new domains and evolving target distributions, rather than as a standard random test split.

Table S7 presents an initial audit of the fixed feature space, constructed using the target-adaptation population prior to model training. This assessment covers all 55 deterministic URL-based features and is primarily designed to identify missing values, constant or near-constant variables, and to evaluate the actual variance of each input parameter. No missing values were detected for any feature during the adaptation phase, confirming that the general URL extractor functioned correctly on the prepared dataset. However, the features base_dots and host_nonascii were found to be constant, while several other infrequent indicators showed a dominant value in at least 99% of the observations. These include, specifically, host_punycode, query_urls, path_double_slash, path_dotdot, host_ip, and explicit_port.

Table S7. Audit of the accessibility and variability of the shared space of URL attributes

Features	**Unique values**	**Constant**	**Almost constant**
host_length	66	Нет	Нет
host_digits	155	Нет	Нет
host_entropy	1393	Нет	Нет
host_hyphens	8	Нет	Нет
host_dots	7	Нет	Нет
host_digit_run	17	Нет	Нет
host_letters	234	Нет	Нет
host_unique	236	Нет	Нет
base_length	49	Нет	Нет
base_digits	89	Нет	Нет

base_entropy	680	Нет	Нет
base_hyphens	7	Нет	Нет
base_dots	1	Да	Да
base_digit_run	15	Нет	Нет
base_letters	114	Нет	Нет
base_unique	149	Нет	Нет
path_length	201	Нет	Нет
path_digits	1353	Нет	Нет
path_entropy	10928	Нет	Нет
path_hyphens	37	Нет	Нет
path_dots	7	Нет	Нет
path_digit_run	25	Нет	Нет
path_letters	1654	Нет	Нет
path_unique	1184	Нет	Нет
query_length	89	Нет	Нет
query_digits	115	Нет	Да
query_entropy	200	Нет	Нет
query_hyphens	17	Нет	Да
query_dots	4	Нет	Да
query_digit_run	17	Нет	Да
query_letters	138	Нет	Нет
query_unique	137	Нет	Нет
host_labels	7	Нет	Нет
host_subdomains	5	Нет	Нет
host_longest_label	53	Нет	Нет
host_nonascii	1	Да	Да
host_punycode	2	Нет	Да
host_vowels	194	Нет	Нет
host_consonant_run	12	Нет	Нет
path_depth	15	Нет	Нет
path_at	3	Нет	Да
path_percent	26	Нет	Нет
query_parameters	11	Нет	Нет
query_equals	9	Нет	Нет
query_urls	2	Нет	Да
path_double_slash	2	Нет	Да
path_dotdot	2	Нет	Да
url_nonascii	2	Нет	Да
url_length	223	Нет	Нет
url_at	3	Нет	Да
host_ip	2	Нет	Да
explicit_port	2	Нет	Да
host_sensitive_words	3	Нет	Нет
path_sensitive_words	4	Нет	Нет
query_sensitive_words	4	Нет	Да

The presence of features with low variance was not used as a criterion for their removal based on the final test results. The feature schema was defined in advance, and constant components were stabilized by a

numerical preprocessor. This approach prevents outcome-driven feature selection and ensures that the input space remains identical across all models being compared. At the same time, the table demonstrates that the majority of URL-based features exhibit sufficient variability; specifically, entropy-based, proportional, and structural characteristics of the path, host, and base components show the greatest diversity in terms of unique values. Consequently, this fixed feature space encompasses both common, general properties of URLs and rare indicators that may point to specific, distinct patterns.

Table S8 summarizes the distributional characteristics of 12 representative URL-derived features from the target adaptation dataset. The selected features cover host, path, query, and overall URL characteristics, including continuous structural measures and rare indicators. This reduced summary highlights distributional asymmetry and extreme values relevant to numerical preprocessing, while the complete set of 55 features is retained in the modeling pipeline.

Table S8**.** Descriptive statistics and tail properties of URL features on the target-adaptation population

Feature	Mean	Median	P95	P99	Max	Skewness	Kurtosis
host_length	15.5861	15.0000	26.0000	34.0000	110.0000	1.9024	10.0905
host_entropy	3.3253	3.3788	3.8212	4.0204	4.7175	−0.4167	0.4516
host_digit_run	0.1352	0.0000	1.0000	3.0000	31.0000	19.4980	617.6742
path_length	33.5499	28.0000	85.0000	118.0000	322.0000	1.8465	8.7972
path_entropy	3.3076	3.7623	4.4188	4.6162	5.1213	−1.7812	1.9452
path_percent	0.0508	0.0000	0.0000	1.0000	48.0000	30.5224	1317.2651
path_dotdot	0.0001	0.0000	0.0000	0.0000	1.0000	120.4429	14505.5000
query_length	0.5128	0.0000	0.0000	11.0000	426.0000	31.2133	1288.0720
query_entropy	0.0533	0.0000	0.0000	2.9207	5.2117	8.3984	72.8511
query_urls	0.0002	0.0000	0.0000	0.0000	1.0000	64.3628	4140.8569
url_length	49.6661	44.0000	101.0000	136.0000	456.0000	2.3485	14.9706
explicit_port	0.0000	0.0000	0.0000	0.0000	1.0000	170.3408	29016.0000

The results indicate substantial heterogeneity in the distributions of URL-derived features. URL length exhibits a median of 44 characters and a maximum of 456, while path length ranges up to 322 characters. Particularly pronounced right-skewed distributions are observed for query_length, path_percent, and host_digit_run. Rare structural indicators, including path_dotdot, query_urls, and explicit_port, exhibit extremely high skewness and kurtosis, reflecting their highly concentrated distributions and infrequent nonzero observations. These characteristics support the use of scale-aware numerical preprocessing, including log1p transformation for nonnegative features, while preserving rare observations rather than removing them as outliers. The statistics are descriptive and do not independently establish the predictive contribution of individual features.

Table S9 evaluates representational duplication that is, the recurrence not of the URLs themselves, but of their 55-dimensional numerical representations. In the source development population, 107,245 out of 212,143 rows have a feature vector that has appeared previously. For the target development set, this figure is significantly lower: 4,290 recurring rows out of 56,590. Furthermore, some of these identical aggregated representations appear across multiple experimental roles: 9,579 in the source set and 523 in the target set. More importantly, some identical feature vectors correspond to different classes—6,680 instances in the source set and 161 in the target set.

Table S9**.** Repeatability of numerical URL representations in development populations

Dataset	Development rows	Repeated vector rows	Vectors in multiple roles	Conflicting vector labels
Source	212 143	107 245	9 579	6 680
Target	56 590	4 290	523	161

These results should not be interpreted as evidence of identity leakage. Two different URLs can have identical aggregated features, since the fixed space describes structural properties rather than storing the URL

string itself. On the contrary, the presence of conflicting vector labels demonstrates a limitation of the representation: the same set of aggregated features does not always uniquely define a class. For this reason, such vectors were not removed post hoc and were not used to adjust the labels. Statistically, this feature implies that the resulting classification task is not trivially separable in the given space. This repetition is particularly pronounced in the source corpus, while the target data exhibits a higher diversity of aggregated combinations. Control over precise URL and domain identities remains a separate mechanism for preventing leakage; feature-vector collisions are considered a limitation of the chosen representation.

Table S10 presents a multicriteria assessment of 15 representative URL-derived features from the development data. Feature informativeness was evaluated using Pearson and Spearman correlations, mutual information, tree-based impurity importance, and permutation-based ROC-AUC reduction. The selected features include those with the largest observed permutation importance and representative variables with strong statistical associations. This comparison enables the identification of complementary feature contributions while highlighting differences between marginal association and model-dependent predictive importance.

Table S10**. Multi-criteria evaluation of the informativeness of URL features**

Feature	Pearson log	Spearman	Mutual information	Tree impurity	Permutation AUC drop
path_sensitive_words	0.159587	0.156736	0.006388	0.034297	0.028223
path_letters	−0.250210	−0.241507	0.068176	0.043797	0.018936
path_entropy	−0.187411	−0.030728	0.070671	0.036447	0.017165
path_digits	0.149454	0.137290	0.050511	0.036922	0.010782
path_depth	−0.205750	−0.183953	0.024572	0.042069	0.008352
base_unique	−0.004698	0.003182	0.017983	0.027660	0.007123
path_unique	0.052615	0.073069	0.075230	0.044752	0.007059
path_length	−0.158750	−0.088270	0.066645	0.052945	0.006749
query_letters	0.320213	0.329963	0.030669	0.039111	0.006075
query_entropy	0.334269	0.339370	0.033649	0.035304	0.005508
query_unique	0.325438	0.339370	0.033571	0.029709	0.005237
host_length	0.146497	0.121608	0.029039	0.029945	0.004672
query_length	0.314974	0.339370	0.035451	0.019784	0.004565
host_digit_run	0.207654	0.185560	0.017020	0.012579	0.001850
host_entropy	0.159725	0.159469	0.081720	0.033376	0.000119

The results reveal differences between statistical associations and model-dependent feature importance. The largest permutation-based ROC-AUC reductions are observed for path_sensitive_words (0.028223), path_letters (0.018936), and path_entropy (0.017165), indicating the relevance of path-related characteristics for the evaluated classifier. In contrast, host_entropy exhibits the highest mutual information among the selected features (0.081720), but its permutation AUC reduction is only 0.000119. Query-related features demonstrate relatively strong positive correlations, although their individual permutation contributions are more moderate. These findings suggest that marginal association does not necessarily correspond to predictive importance when features are evaluated jointly. The analysis is descriptive and model-dependent; it does not establish causal relationships or justify removing features from the predefined 55-dimensional representation.

Table S11 presents a validation of leakage-safe preprocessing for a target-adaptation population using three domain-grouped out-of-fold (OOF) folds. The sizes of the held-out sets are nearly equal: 9,672, 9,673, and 9,671 observations, respectively. For each fold, the numerical preprocessor was re-trained exclusively on the two remaining training sets, which together contain approximately 19.3 thousand observations. Consequently, the medians, means, and standard deviations used to process a given held-out fold were not estimated based on the held-out samples themselves.

Table S11. Fold-local validation of numerical preprocessing for target-adaptation data

Fold	Fit rows	Held-out rows	Input features	Output values and masks	Finite
0	19 344	9 672	55	110	Yes
1	19 343	9 673	55	110	Yes
2	19 345	9 671	55	110	Yes

In all cases, the 55 initial numerical features are transformed into a 110-dimensional neural representation: the first 55 components contain standardized values, while the second 55 consist of explicit binary indicators of availability. All generated values are finite; specific checks confirm that an artificially missing value results in a standardized component of zero and a corresponding mask value of zero. This outcome is technically significant for subsequent out-of-fold (OOF) predictions and hard-negative mining. If a unified preprocessing pipeline were trained on the entire adaptation population, information from a held-out fold could implicitly leak into the intermediate representation via scaling or imputation parameters. Training on a fold-local basis eliminates this potential for data leakage. Furthermore, maintaining a consistent output dimensionality across all folds ensures compatibility for stored OOF predictions and downstream models. Figure S2 displays the class-conditioned distributions of four representative features host_length, path_length, query_length, and host_entropy following a logarithmic transformation.

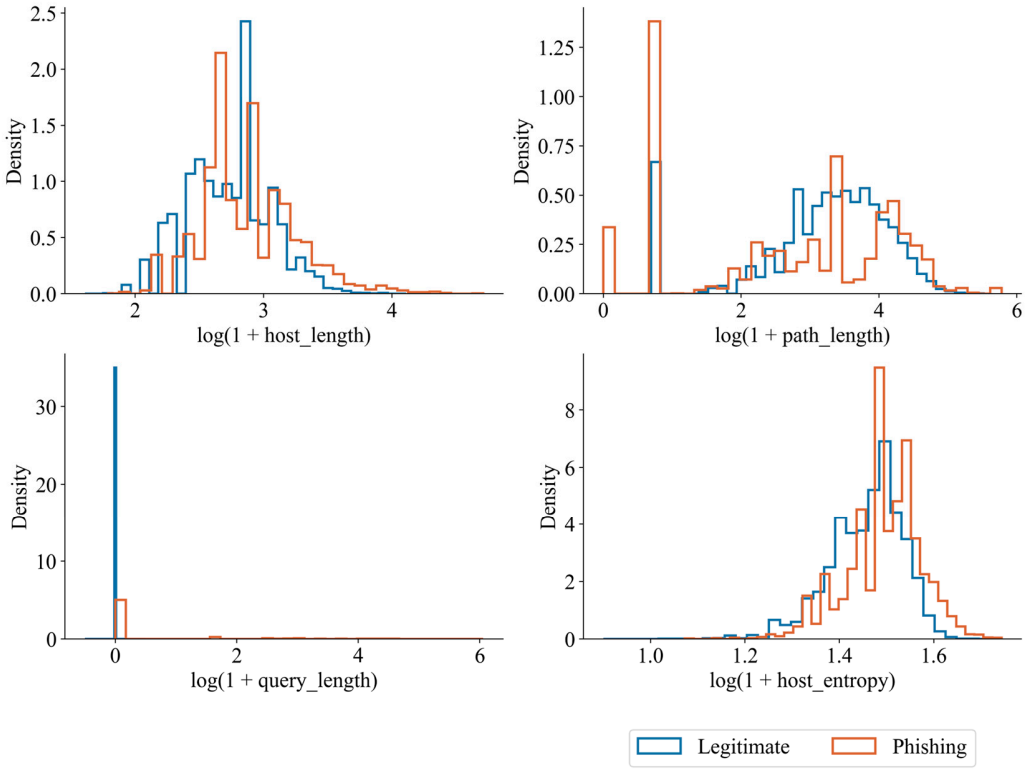


Figure S2. Class-Conditioned Feature Distributions in Target Adaptation

The comparison was performed on a target-adaptation population consisting of 25,169 legitimate URLs and 3,847 phishing URLs. The distributions indicate that while class differences exist, they do not form a simple, one-dimensional threshold. For host_length and host_entropy, the density plots overlap significantly, though phishing observations exhibit noticeable local shifts. The structure for `path_length` is more heterogeneous: the median path length is 28 for legitimate URLs and 19 for phishing URLs, yet both groups feature long, extended tails. The `query_length` metric is particularly revealing: a large portion of both classes is concentrated near zero, while a small fraction of URLs forms a long right-hand tail. Consequently, the absence of a query component is not a sufficiently discriminative signal, and the infrequent, long queries necessitate a robust numerical transformation. A logarithmic scale allows these distributions to be compared visually without the need to exclude extreme outliers. The results shown in Figure S2 highlight two key considerations for subsequent modeling. First, individual marginal distributions do not provide a clear linear separation between classes, which justifies the investigation of non-linear interactions. Second, the significant heterogeneity in scales validates the use of a fixed log-compression step prior to neural standardization. It should be noted that this visual analysis served exclusively as a diagnostic tool during the development phase and did not dictate the final feature set or the ultimate model architecture.

Table S12 illustrates the process of constructing a specialized TailTree branch, utilizing a domain-grouped, out-of-fold "hard-negative mining" approach. The primary objective of this procedure was to identify challenging legitimate URLs without relying on predictions generated by a model trained on the same domain as the instance in question. To achieve this, the target-adaptation population was partitioned into three grouped folds. In each instance, there was zero overlap between the domains present in the training set and those in the held-out set, confirming that the procedure adheres to the principle of being leakage-safe. The sizes of the held-out portions are nearly identical - 9,672, 9,673, and 9,671 observations so no single fold dominates in terms of volume.

Table S12. Domain-grouped OOF-процедура формирования hard-negative specialist

Fold	Fit observations	Held-out observations	Common domains	Training duration
0	19 344	9 672	0	5.730850
1	19 343	9 673	0	5.736077
2	19 345	9 671	0	6.032064

After obtaining OOF (out-of-fold) scores for all adaptation objects, the 99th percentile of the scores for legitimate URLs was calculated, yielding a value of −0.397864. A total of 252 out of 25,169 legitimate URLs fell into the top 1% of the benign score distribution. During the final training of the specialist branch, these specific observations were assigned a weight of 5, while all other observations retained a weight of 1. Thus, the specialization was not directed at the entire negative class, but rather at a small subset of legitimate URLs that, according to the independent OOF predictions, most closely resembled phishing URLs. The total computational cost for the hard-negative mining stage was approximately 17.50 seconds, after which three final CatBoost specialist models were trained using all 29,016 adaptation URLs. Table S13 presents the development-set ablation analysis of TARA, comparing the complete framework with its individual TailTree and FT context branches, as well as an unrestricted constant-fusion configuration. The experiment evaluates the contribution of the tail-aware ranking mechanism in terms of both global discrimination and low-FPR (false positive rate) metrics, using a chronologically selected dataset.

Table S13**. Developmental ablation of individual components and the TARA integration mechanism**

Model configuration	**ROC-AUC**	**AP**	**pAUC@1%**	**pAUC@0.1%**	**TPR@0.1%**	**TPR@1%**
TARA	0.928064	0.829987	0.777125	0.739640	0.515063	0.605442
TailTree only	0.923542	0.822780	0.776080	0.739640	0.515063	0.592809
FT context branch only	0.924072	0.825671	0.771317	0.694594	0.479106	0.601555
Constant fusion without guard	0.928129	0.830483	0.779775	0.742572	0.504373	0.614189

The results show that unrestricted constant fusion achieved slightly higher ROC-AUC (0.928129), AP (0.830483), and pAUC@1% (0.779775) than the complete TARA framework. However, TARA and the isolated TailTree branch achieved identical pAUC@0.1% (0.739640) and TPR@0.1% (0.515063), consistent with the intended preservation of specialist ranking within the protected region. The FT context branch demonstrated competitive global discrimination but lower performance in the most restrictive low-FPR region. These findings illustrate the trade-off between unrestricted score fusion and controlled tail-aware ranking. The development results were used to characterize component behavior rather than to modify the predefined architecture based on final-test performance.

Table S14 presents the affine normalization parameters required to correctly combine the Tree and Neural branches. Since the raw margins from CatBoost and the logits from the FT-Transformer are generated by different algorithms and exist on different numerical scales, averaging them directly would lack a consistent interpretation. Therefore, normalization was performed exclusively on a population of legitimate observations selected chronologically. For each branch, the 95th percentile of the benign score was used as the center, while

the distance between the 99.5th and 95th percentiles served as the scale. In both cases, the calculations were based on the same 4,815 legitimate URLs, but were performed separately for each score distribution.

Table S14. Normalization parameters for TARA score components

Component	Center $Q_{0.95}$	Scale $Q_{0.995} - Q_{0.95}$	Legitimate selection observations
TailTree	−1.364494	1.210672	4 815
FT-Transformer	−0.890407	1.571178	4 815

The resulting parameters differ noticeably. For TailTree, the center is −1.364494 and the scale is 1.210672; for FT-Transformer, the corresponding values are −0.890407 and 1.571178. This confirms that the raw outputs of the two components cannot, in fact, be considered as sharing a common scale. Following normalization, a value of =1 for a tree branch corresponds to its predetermined 99.5th-percentile benign tail boundary. It is precisely this level that serves as the boundary for the protected region. Crucially, once calculated, the normalization parameters were frozen and not subsequently re-estimated using calibration or final-test observations. Thus, the table does not document an additional model optimization, but rather a fixed transformation from heterogeneous component scores to a common scale, upon which a conditional-ranking mechanism becomes mathematically well-defined. To verify the implementation consistency of TARA, two complementary checks were performed before the final evaluation. First, the preservation of the frozen source encoder was assessed by comparing its output scores before and after target-domain adaptation. Second, checkpoint reproducibility was evaluated by comparing the predictions of the TailTree and Context FT components before and after serialization and reloading. Both procedures used identical observations to detect unintended changes in model behavior (Table S15).

Table S15**. Verification of Frozen Source Preservation and Checkpoint Reload Consistency in TARA**

Verification procedure	Evaluated component	Evaluation data	Maximum absolute score difference
Source preservation after target adaptation	Frozen source route	Source selection (23,338 URLs)	0.000000
Checkpoint reload consistency	TailTree	Target selection (identical observation IDs)	0.000000
Checkpoint reload consistency	Context FT	Target selection (identical observation IDs)	0.000000

As shown in Table S15, all three verification procedures yielded zero maximum score differences at the reported numerical precision. The unchanged source checkpoint hash and identical source-selection predictions confirm that target adaptation did not modify the frozen source route in the evaluated implementation. Similarly, the TailTree and Context FT components produced identical predictions after checkpoint reloading, supporting the reproducibility of the saved model states. These findings verify implementation consistency and architectural source isolation but do not constitute additional evidence of predictive superiority or guarantee the absence of catastrophic forgetting in other adaptation settings. To assess the statistical reliability of the observed ablation effects, paired domain-level bootstrap analysis was performed on the locked test set. The complete TARA framework was compared with constant fusion without the tail-aware guard, the isolated protected tree head, and the neural body across three metrics: ROC-AUC, pAUC@1%, and TPR@0.1%. The resulting nine comparisons were evaluated using Holm-adjusted p-values to account for multiple hypothesis testing (Table S16).

Table S16**. Pairwise statistical testing of TARA ablation effects**

Ablation configuration	Metric	TARA − Ablation	95% CI	Holm-adjusted p-value
Constant fusion	ROC-AUC	+0.000562	[0.000063, 0.001257]	0.338831
Constant fusion	pAUC@1%	+0.021810	[−0.002900, 0.059056]	0.637681
Constant fusion	TPR@0.1%	+0.230433	[−0.019843, 0.253739]	0.338831
Protected tree head	ROC-AUC	+0.013996	[0.012326, 0.015918]	0.004498
Protected tree head	pAUC@1%	+0.001716	[0.000000, 0.006719]	0.921539
Protected tree head	TPR@0.1%	0.000000	[0.000000, 0.000000]	1.000000
Neural body	ROC-AUC	+0.004062	[0.002315, 0.005886]	0.004498
Neural body	pAUC@1%	+0.040510	[−0.007954, 0.112326]	0.637681
Neural body	TPR@0.1%	+0.418446	[0.021849, 0.480623]	0.132934

As shown in Table S16, TARA demonstrated statistically significant improvements in ROC-AUC over the protected tree head (Δ = 0.013996, adjusted p = 0.004498) and the neural body (Δ = 0.004062, adjusted p = 0.004498). In contrast, none of the comparisons with unrestricted constant fusion remained statistically significant after Holm correction, despite positive observed differences. The identical TPR@0.1% values for TARA and the protected tree head are consistent with the deterministic preservation of specialist ranking within the protected region. Although TARA showed a substantial observed TPR@0.1% improvement over the neural body (Δ = 0.418446), this difference was not statistically significant after correction (adjusted p = 0.132934). Overall, the results support the contribution of neural refinement to global discrimination while indicating that the predictive advantage of the tail-aware guard over unrestricted fusion remains statistically inconclusive.

Table S17 presents the deterministic verification of two fundamental invariants of the proposed tail-aware ranking mechanism. First, the final TARA scores must preserve the specialist scores within the predefined protected region. Second, observations assigned to the lower region must remain strictly below the protected boundary after neural refinement. These properties were evaluated on the locked test set without modifying the predefined ranking mechanism.

Table S17. Verification of Deterministic Invariants of the TARA Tail-Aware Ranking Mechanism

Verified invariant	Number of observations	Verification statistic	Observed value
Preservation of protected specialist scores	27,392	Maximum absolute difference between TARA and TailTree scores	0.000000

Separation of lower-region scores from the protected boundary	19,550	Maximum lower-region score minus the boundary value (1.0)	−0.200492

The results confirm that both ranking invariants were satisfied. For all 27,392 observations in the protected region, the maximum absolute difference between TARA and TailTree scores was 0.000000, demonstrating exact preservation at the reported numerical precision. For the remaining 19,550 observations, the maximum score relative to the protected boundary was -0.200492, confirming that no lower-region observation crossed the threshold of 1.0. These findings verify the intended deterministic behavior of the tail-aware guard and demonstrate consistency between its mathematical formulation and implementation. However, satisfying these invariants does not independently establish predictive superiority over unrestricted fusion or other competing models.

Table S18 examines the impact of the assumed operational prevalence of phishing on the interpretation of TARA's fixed class-conditional rates. In contrast to the locked-test corpus, which features a high proportion of phishing, prevalence rates ranging from 0.1% to 10% were considered for the applied scenario. Sensitivity and specificity remained constant, based on the "frozen" nominal 1% operating rule; consequently, the results represent an analytical projection rather than new empirical measurements.

Table S18**. Sensitivity of TARA's projected operational performance metrics to phishing prevalence**

Phishing prevalence	Expected PPV	Expected NPV	False alerts / 100k	Missed phishing / 100k
0.001	0.069699	0.999731	979.02	26.65
0.005	0.273314	0.998649	975.10	133.25
0.010	0.430530	0.997289	970.20	266.51
0.050	0.797541	0.986032	931.00	1 332.54
0.100	0.892660	0.970963	882.00	2 665.08

The most significant change is observed in the positive predictive value (PPV). At a prevalence of 0.1%, the expected PPV is only 0.069699; even a fairly robust detector will, in such an environment, generate a substantial proportion of false alarms among all alerts. Given these assumptions, for every 100,000 input URLs, one would expect approximately 979 false alerts and 26.65 missed phishing URLs. At a prevalence of 1%, the PPV rises to 0.430530, and at 10%, it reaches 0.892660. Meanwhile, the negative predictive value (NPV) remains high across all scenarios, though it gradually decreases as the proportion of attacks increases.

This result highlights an important limitation when using precision and average precision (AP) metrics derived from a research corpus. Predictive values depend not only on a model's discriminative power but also on the base rate. Therefore, the operational interpretation of TARA must account for the actual prevalence within a specific data stream. The table provided does not imply that sensitivity and specificity will necessarily remain constant under a different deployment distribution; rather, it serves as a conditional estimate, assuming no additional distribution shift occurs. Figure S3 shows the dynamics of binary cross-entropy and the discrimination of the frozen source encoder during the pre-training process.

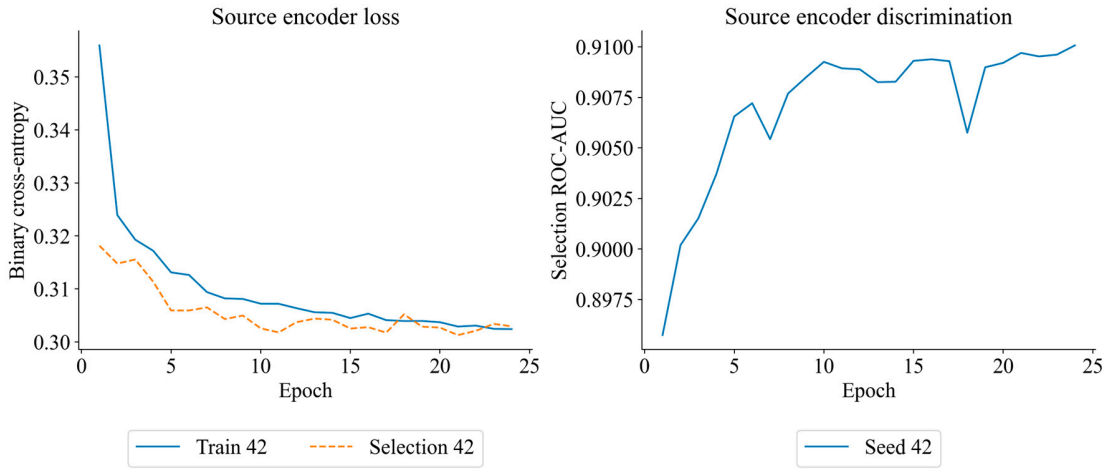


Figure S3**. Source encoder loss / Source encoder discrimination**

The left panel shows the training loss and selection LogLoss, while the right panel displays the ROC-AUC on a separate source-selection population. During the initial epochs, the training loss decreases rapidly from approximately 0.356 to a range around 0.32, after which the decline becomes more gradual. The selection loss exhibits a similar decrease, with no significant divergence from the training trajectory in the later stages. The selection ROC-AUC rises from approximately 0.896 at the start of training to about 0.909–0.910 in the later epochs. The curve contains local fluctuations—including a notable drop around the 18th epoch which underscores the necessity of employing a predefined early-stopping rule rather than simply selecting the final epoch. According to the saved training record, epoch 17 was chosen as the final model based on a development criterion; the source calibration and source test sets were not used for the early-stopping decision.

Figure S4 illustrates the training dynamics of three Context FT members, initialized with seeds 42, 43, and 44. Unlike the source encoder, all three neural members are trained on the target-adaptation population and receive a frozen source representation as an additional context token. In the left panel, the training binary cross-entropy for all seeds decreases steadily from approximately 0.30-0.32 to a level around 0.15. Meanwhile, the chronological-selection LogLoss, following an initial improvement, fluctuates primarily within the 0.23–0.26 range and does not mirror the monotonic decline observed in the training loss.

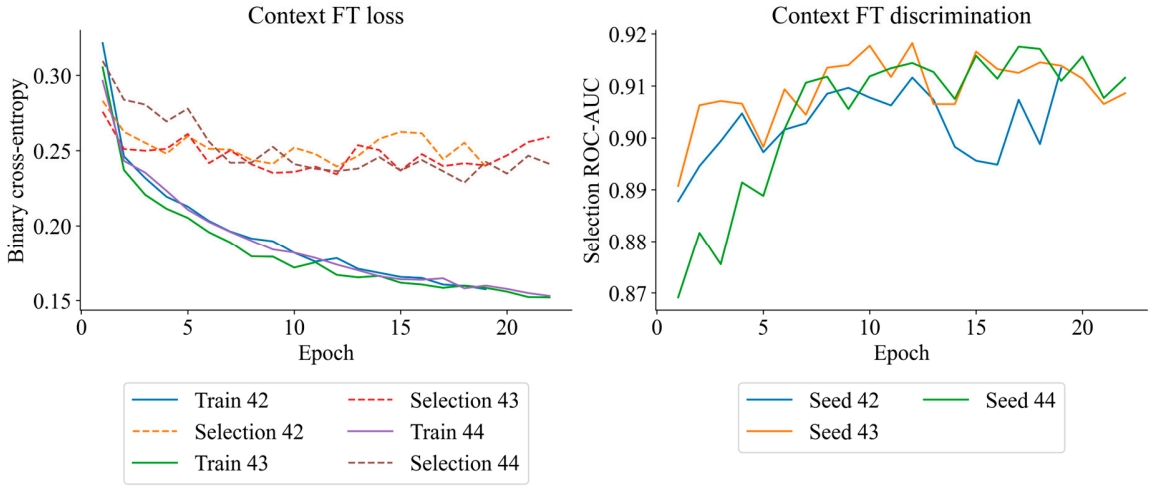


Figure S4**. Context FT loss / Context FT discrimination**

This discrepancy is significant: continuing to minimize the training objective does not guarantee improved performance on a subsequent temporal selection set. Consequently, the number of retained epochs

varies by seed: 12 for seed 42, 15 for seed 43, and 18 for seed 44. The right-hand panel confirms the stochastic variability of the discrimination task: the selection ROC-AUC generally increases, but local maxima are reached at different epochs, after which temporary declines may occur. Thus, the ensemble of three members is not constructed by artificially enforcing a uniform stopping epoch; rather, each member retains a checkpoint corresponding to its own specific development trajectory. In this context, the different seeds reflect variability in initialization and training, rather than independent dataset replications. The figure also demonstrates an absence of numerical divergence: all loss and ROC-AUC values remain finite throughout the training process. While this confirms the technical stability of the Context FT fitting procedure, it does not, in itself, constitute proof of its superiority.

Figure S5 compares two distinct mechanisms for interpreting the internal components of TARA. For the TailTree specialist, mean absolute SHAP values were used, whereas for Context FT, mean absolute input gradients were employed. This analysis was conducted on 512 randomly selected observations from the frozen test set, with the model parameters held constant. Because these attribution methods possess different semantics and scales, the bar lengths across the two panels cannot be directly compared as absolute importance values; the comparison focuses primarily on the ranking of the features.

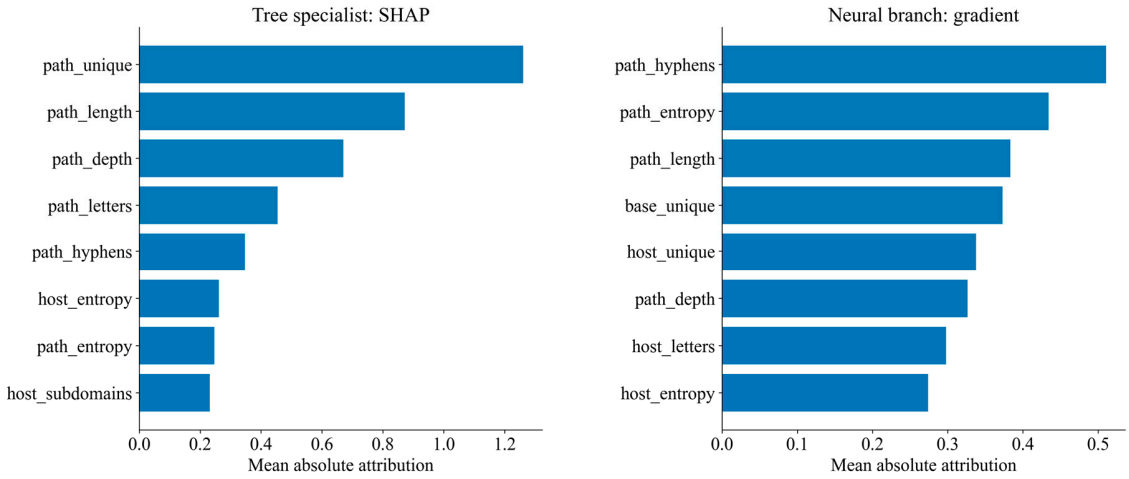


Figure S5**. Tree specialist: SHAP / Neural branch: gradient**

For the tree specialist, the most significant contribution comes from path_unique, followed by path_length, path_depth, and path_letters. For the "neural branch, the highest local sensitivity is observed for path_hyphens, path_entropy, and path_length; other significant features include base_unique, host_unique, and path_depth. Thus, while both branches place considerable emphasis on path-related structure, they assign different priorities. The Spearman correlation between the full rankings of all 55 features is 0.716. This indicates a notable, though not complete, agreement between the components, which aligns with the architectural concept of functional complementarity. Neural gradients are calculated with a fixed (frozen) source token and, therefore, do not represent the full derivative through the source preprocessing and encoder stages. Similarly, SHAP values describe the raw specialist score rather than providing a causal explanation for phishing. Consequently, this figure should be interpreted as a component-level interpretability analysis, rather than as evidence of the causal importance of individual URL characteristics.

Figure S6 examines the sensitivity of low-FPR ranking performance to the protected-region threshold τ and the lower-region neural weight w on the target-calibration partition comprising 21,730 URLs. The analysis evaluates multiple guard and fusion configurations without modifying the locked TARA model. Performance remains relatively stable around the selected configuration ($\tau = 1$, $w = 0.5$), whereas unrestricted fusion with a high neural weight produces more substantial degradation.

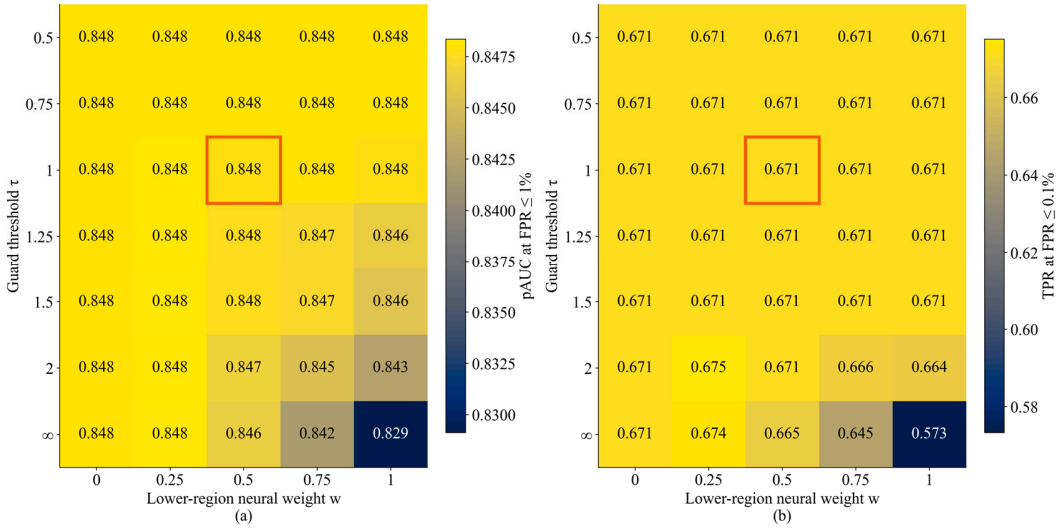


Figure S6. Sensitivity of ranking performance to the protected-region threshold τ and lower-region neural weight w on the target-calibration partition (21,730 URLs): (a) standardized pAUC at FPR ≤ 1%; (b) TPR at FPR ≤ 0.1%. The row marked ∞ denotes unrestricted fusion without ranking protection. The outlined cell indicates the fixed TARA configuration (τ = 1, w = 0.5).

Figure S7 illustrates the effects of hard-negative selection quantiles and sample-weighting strength on the low-FPR performance of TARA. The experiment compares four selection quantiles and three weighting levels against a uniformly weighted reference. The results show that the selected configuration (quantile 0.99, weight 5) achieves favorable performance on the target-calibration partition, although differences among nearby configurations are relatively small.

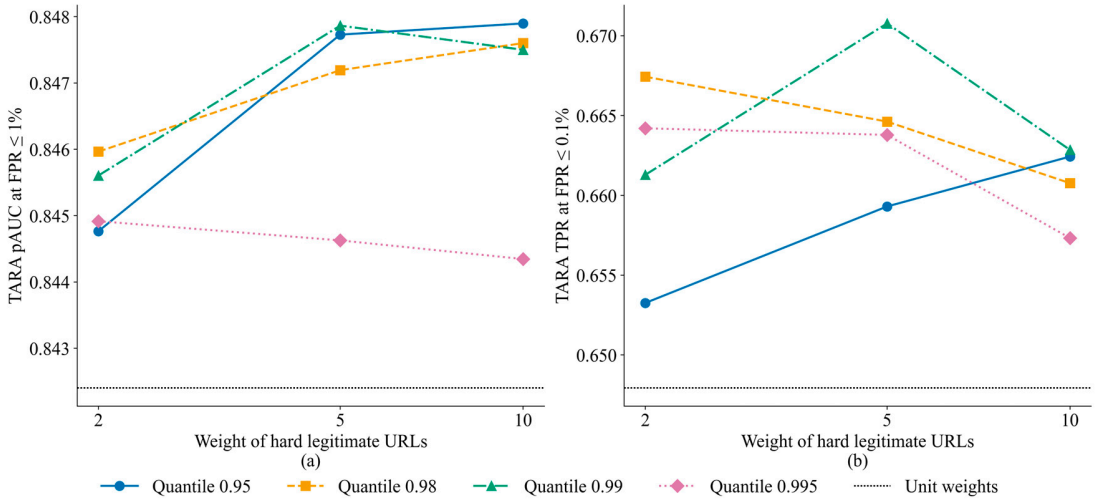


Figure S7. Sensitivity of the complete TARA framework to hard-negative selection parameters on the target-calibration partition: (a) standardized pAUC at FPR ≤ 1%; (b) TPR at FPR ≤ 0.1%. Curves represent different hard-negative selection quantiles, and the horizontal dotted line indicates the unweighted specialist reference.

Figure S8 illustrates the uncertainty in low-FPR performance estimates across the evaluated models using 2,000 paired domain-level bootstrap replicates. The results show substantial variability in TPR at FPR ≤ 0.1%, while standardized pAUC at FPR ≤ 1% exhibits more concentrated distributions, indicating greater stability of the integrated low-FPR metric.

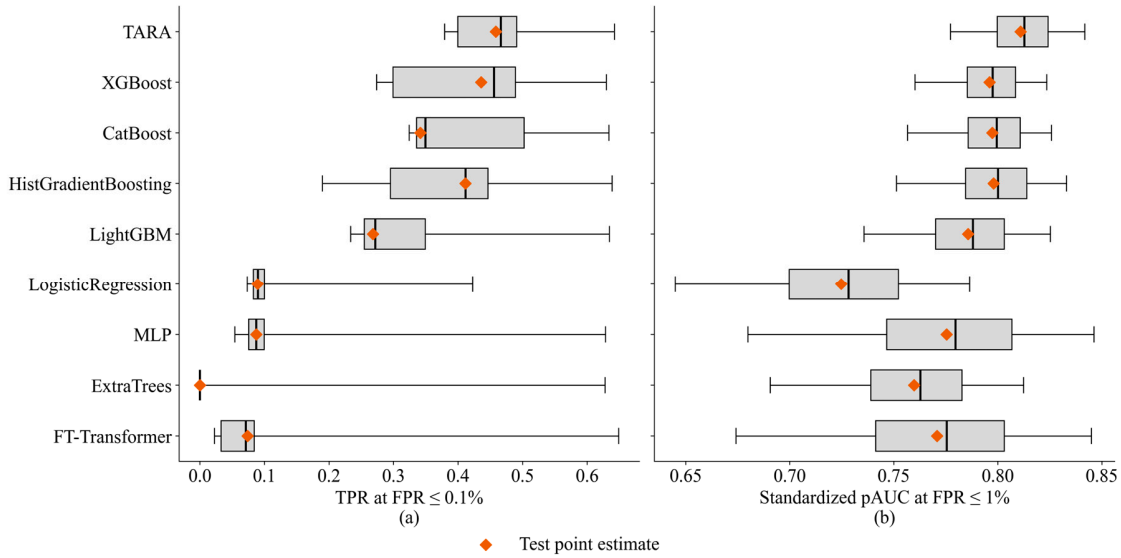


Figure S8. **Distributions of 2,000 paired domain-level bootstrap replicates on the locked test set: (a) TPR at FPR ≤ 0.1%; (b) standardized pAUC at FPR ≤ 1%**

Table S21 summarizes the configurations of six TARA variants and six additional comparator models used to evaluate the contribution of source-domain information, individual architectural components, and model capacity. The table specifies the modifications, ensemble composition, and target-training time for each configuration, providing a reproducible basis for assessing supervised cross-dataset adaptation and the effectiveness of source-aware learning strategies.

Table S21. Configurations of TARA variants and additional comparator models.

Model	**Modification or configuration**	**Members**	**Target training time (s)**
Without source token	Contextual branch without the source token (input: 55 feature values and 55 masks); TailTree and guard unchanged from TARA	3 × branches	203.2
Random frozen encoder	Encoder with the same architecture, randomly initialized (seed 42) and frozen	3 × branches	243.8
Target-trained encoder	Encoder with the same architecture, trained on the target-adaptation set (up to 24 epochs, batch size 1,024, stopping on selection data), then frozen	1 encoder + 3 branches	137.6 + 19.1
Fine-tunable source encoder	Source encoder updated jointly with the contextual branch on the target-adaptation set	3 × (encoder + branch)	285.8
TailTree without OOF weights	CatBoost depth 8; 700 iterations; learning rate 0.04; L2 = 8; uniform sample weights	3 × CatBoost	21.8
Without source token and OOF weights	Combination of the preceding two changes with the same guard	6 members	225.0
Source-only encoder	Source-trained network applied to the target without target-domain training	1 network	0
Pooled source–target CatBoost	Trained on 165,557 source-fit URLs and 29,016 target-adaptation URLs; depth 8; ≤700 iterations; learning rate 0.04; L2 = 8; patience 50 on selection data	3 × CatBoost	90.5

Model	Modification or configuration	Members	Target training time (s)
CatBoost + source representation	55 URL-derived features and 64 values from the frozen source representation; otherwise configured as above	3 × CatBoost	36.9
Fine-tuned source network	Initialized with source-pretrained weights; AdamW; learning rate 10^{-3}; up to 22 epochs; patience 7	3 × networks	155.6
CatBoost×3 + FT-Transformer×3	Locked comparator models; normalization using legitimate URLs in the selection set; equal weights	6 members	231.9
Character-level CNN	URL string ≤256 characters; embedding size 16; convolution widths 3/5/7 with 32 filters each; AdamW; learning rate 10^{-3}; up to 22 epochs; patience 7	3 × CNNs	286.4

Table S23 evaluates the sensitivity of TARA to the quantiles used for normalizing the specialist and contextual branch scores. Nine combinations of center and upper quantiles were examined on the target-calibration partition while retaining the fixed guard threshold and neural fusion weight. The analysis assesses whether moderate changes in normalization affect the protected region and the resulting ranking performance.

Table S23. Sensitivity of TARA to branch-normalization quantiles on the target-calibration partition.

Center quantile	Upper quantile	Proportion of the protected area	ROC-AUC	pAUC at FPR ≤ 1%	TPR at FPR ≤ 0.1%
0,900	0,990	32,52%	0,9394	0,8480	0,6708
0,900	0,995	31,45%	0,9393	0,8479	0,6708
0,900	0,999	29,88%	0,9393	0,8474	0,6708
0,950	0,990	32,52%	0,9393	0,8480	0,6708
0,950	0,995	31,45%	0,9391	0,8479	0,6708
0,950	0,999	29,88%	0,9393	0,8473	0,6708
0,975	0,990	32,52%	0,9385	0,8480	0,6708
0,975	0,995	31,45%	0,9386	0,8479	0,6708
0,975	0,999	29,88%	0,9391	0,8473	0,6708

Across the nine normalization configurations, standardized pAUC@1% ranged from 0.8473 to 0.8480, while TPR@0.1% remained constant at 0.6708. These results indicate limited sensitivity to the selected normalization quantiles within the evaluated parameter range. Table S24 examines the sensitivity of hard-negative specialist training to the selection quantile and sample weight assigned to difficult legitimate URLs. The analysis compares an unweighted reference with twelve weighted configurations. For each configuration, the specialist and complete TARA model were evaluated on the target-calibration partition using the same predefined ranking procedure.

Table S24. Sensitivity of TailTree and TARA to hard-negative selection quantiles and weights on the target-calibration partition.

Quantile	Weight	Weighted URLs	TailTree pAUC@1%	TailTree TPR@0.1%	TARA ROC-AUC	TARA pAUC@1%	TARA TPR@0.1%
No weighting	1	0	0.8421	0.6479	0.9358	0.8424	0.6479
0.950	2	1,259	0.8446	0.6533	0.9372	0.8448	0.6533

0.950	5	1,259	0.8483	0.6593	0.9379	0.8477	0.6593
0.950	10	1,259	0.8473	0.6624	0.9369	0.8479	0.6624
0.980	2	504	0.8458	0.6674	0.9364	0.8460	0.6674
0.980	5	504	0.8465	0.6646	0.9364	0.8472	0.6646
0.980	10	504	0.8469	0.6608	0.9362	0.8476	0.6608
0.990	2	252	0.8453	0.6613	0.9362	0.8456	0.6613
0.990	5	252	0.8480	0.6708	0.9391	0.8479	0.6708
0.990	10	252	0.8477	0.6628	0.9375	0.8475	0.6628
0.995	2	126	0.8444	0.6642	0.9362	0.8449	0.6642
0.995	5	126	0.8448	0.6638	0.9364	0.8446	0.6638
0.995	10	126	0.8447	0.6573	0.9378	0.8443	0.6573

The results show that all evaluated hard-negative weighting configurations improved pAUC@1% and TPR@0.1% relative to uniform weighting on the calibration partition. The fixed configuration achieved a pAUC@1% of 0.8479 and TPR@0.1% of 0.6708. Nevertheless, the differences among weighted configurations were modest, and the sensitivity analysis should not be interpreted as independent confirmation of parameter optimality.

Table S28 compares class-specific URL structural characteristics between the PhiUSIIL source-fit and PhreshPhish target-adaptation partitions. The results reveal substantial differences in path structure, particularly the frequency of empty or root paths. In PhiUSIIL, such paths are more common among legitimate URLs (100.00%) than phishing URLs (74.04%), whereas the opposite pattern is observed in PhreshPhish (9.21% versus 28.36%). These differences indicate a shift in class-dependent URL characteristics and provide a plausible explanation for the negative transfer observed in source-aware models.

Table S28. URL structure by class in the source and target development partitions.

Dataset and role	Class	URLs	Empty or root path	Median path length	Median URL length	Query present
PhiUSIIL, source fit	Legitimate	94,451	100.00%	0	15	0.00%
PhiUSIIL, source fit	Phishing	71,106	74.04%	1	26	6.03%
PhreshPhish, target adaptation	Legitimate	25,169	9.21%	28	44	0.00%
PhreshPhish, target adaptation	Phishing	3,847	28.36%	19	43	13.05%

To further quantify the incremental contribution of the hybrid architecture, Table S29 presents a detailed comparison between TARA and the standalone TailTree-only detector on the original locked test set. The comparison includes global discrimination, low-FPR ranking performance, probability calibration, and fixed-threshold operating characteristics. Statistical significance was assessed using 2,000 paired domain-level bootstrap replicates with Holm correction.

Table S29. Performance Comparison of TARA and the Standalone TailTree-Only Detector on the Original Locked Test Set.

Metric	TARA	TailTree-only	Holm-adjusted p
ROC-AUC	0.9443	0.9303	0.009
AP	0.9891	0.9862	0.009
pAUC@0.1%	0.6959	0.6959	1.000
pAUC@0.5%	0.7754	0.7754	1.000
pAUC@1%	0.8109	0.8092	1.000

pAUC@2%	0.8431	0.8325	0.009
pAUC@5%	0.8731	0.8542	0.009
TPR@0.1%	0.4582	0.4582	1.000
TPR@0.2%	0.5027	0.5027	1.000
TPR@0.5%	0.6744	0.6744	1.000
TPR@1%	0.7360	0.7030	1.000
TPR@2%	0.7710	0.7309	0.009
TPR@5%	0.8117	0.7597	0.009
Brier score	0.1074	0.1288	0.009
LogLoss	0.3232	0.3701	0.009
TPR, nominal FPR 1%	0.7335	0.7054	0.009
FPR, nominal FPR 1%	0.0098	0.0110	0.530
TPR, nominal FPR 0.1%	0.6384	0.6384	1.000
FPR, nominal FPR 0.1%	0.0032	0.0032	1.000

To assess the statistical significance of source-domain pretraining, paired comparisons were conducted between pretrained, randomly initialized, and non-pretrained frozen representations. Table S30 reports differences in ROC-AUC, standardized pAUC at FPR ≤ 1%, and TPR at FPR ≤ 1%, together with 95% domain-level bootstrap confidence intervals. Holm correction was applied separately to full-model and contextual-branch comparisons.

Table S30. Paired ROC-AUC Differences Between Pretrained and Randomly Initialized Frozen Encoders on the Locked Test Set.

Comparison	Δ ROC-AUC [95% CI]
PhiUSIIL − Random 42	+0.0019 [0.0012, 0.0025]*
PhiUSIIL − Random 43	+0.0040 [0.0031, 0.0050]*
PhiUSIIL − Random 44	+0.0011 [0.0006, 0.0016]*
PhreshPhish − Random 42	+0.0046 [0.0040, 0.0053]*
PhreshPhish − Random 43	+0.0067 [0.0059, 0.0076]*
PhreshPhish − Random 44	+0.0039 [0.0032, 0.0046]*
PhreshPhish − PhiUSIIL	+0.0028 [0.0020, 0.0034]*

The influence of target-label availability was examined by training contextual branches using different proportions of the adaptation domains. Table S31 summarizes the mean performance and standard deviation across three random seeds for each frozen representation configuration.

Table S31. Branch Performance under Reduced Target-Label Availability.

Target domains	Representation	ROC-AUC	pAUC@1%
10%	No token	0.8846 ± 0.0192	0.7398 ± 0.0103
10%	Random encoder	0.8821 ± 0.0128	0.7560 ± 0.0117
10%	PhiUSIIL	0.8789 ± 0.0140	0.7284 ± 0.0180
10%	PhreshPhish	0.9109 ± 0.0035	0.7499 ± 0.0170
25%	No token	0.9269 ± 0.0034	0.7551 ± 0.0141
25%	Random encoder	0.9268 ± 0.0092	0.7568 ± 0.0212
25%	PhiUSIIL	0.9237 ± 0.0167	0.7121 ± 0.0664
25%	PhreshPhish	0.9415 ± 0.0031	0.7795 ± 0.0051
100%	No token	0.9217 ± 0.0061	0.7570 ± 0.0066
100%	Random encoder	0.9311 ± 0.0046	0.7673 ± 0.0072
100%	PhiUSIIL	0.9330 ± 0.0041	0.7480 ± 0.0202
100%	PhreshPhish	0.9445 ± 0.0046	0.7676 ± 0.0049

To independently evaluate the informativeness of frozen representations, logistic-regression linear probes were trained using different fractions of target-adaptation domains. Table S32 presents the corresponding ROC-AUC and standardized pAUC at FPR ≤ 1%, averaged across the experimental fits.

Target domains	Representation	ROC-AUC	pAUC@1%
1%	No token	0.6960	0.6114
1%	Random encoder	0.7357	0.5944
1%	PhiUSIIL	0.6511	0.6422
1%	PhreshPhish	0.9009	0.6098
5%	No token	0.7989	0.6877
5%	Random encoder	0.8267	0.6997
5%	PhiUSIIL	0.8218	0.7509
5%	PhreshPhish	0.9089	0.7058
25%	No token	0.8711	0.7266
25%	Random encoder	0.8925	0.7261
25%	PhiUSIIL	0.8282	0.7533
25%	PhreshPhish	0.9360	0.7430
100%	No token	0.9204	0.7386
100%	Random encoder	0.9193	0.7393
100%	PhiUSIIL	0.8579	0.7521
100%	PhreshPhish	0.9529	0.7499

To further assess the contribution of source-domain pretraining, linear-probe experiments were conducted using different proportions of labeled target-adaptation domains (1%, 5%, 25%, and 100%). Logistic regression was trained on frozen representations without updating the encoder parameters. Table S32 reports the mean ROC-AUC and standardized pAUC at FPR ≤ 1%, together with standard deviations and the number of model fittings. These results allow the informativeness of pretrained and randomly initialized representations to be compared under varying levels of target-label availability.

Table S32. Linear probes of frozen representations on a held-out test set.

Domain share	Representation	Number of fittings	ROC-AUC	pAUC at FPR ≤ 1%
1%	Without a representation token	5	0,6960 ± 0,1926	0,6114 ± 0,0586
1%	Random encoders (3 initializations)	15	0,7357 ± 0,0710	0,5944 ± 0,0398
1%	PhiUSIIL encoder	5	0,6511 ± 0,1318	0,6422 ± 0,0589
1%	PhreshPhish encoder	5	0,9009 ± 0,0368	0,6098 ± 0,0625
5%	Without a representation token	5	0,7989 ± 0,0255	0,6877 ± 0,0524
5%	Random encoders (3 initializations)	15	0,8267 ± 0,0437	0,6997 ± 0,0569
5%	PhiUSIIL encoder	5	0,8218 ± 0,0550	0,7509 ± 0,0272
5%	PhreshPhish encoder	5	0,9089 ± 0,0383	0,7058 ± 0,0687
25%	Without a representation token	5	0,8711 ± 0,0572	0,7266 ± 0,0085
25%	Random encoders (3 initializations)	15	0,8925 ± 0,0344	0,7261 ± 0,0317
25%	PhiUSIIL encoder	5	0,8282 ± 0,0895	0,7533 ± 0,0480
25%	PhreshPhish encoder	5	0,9360 ± 0,0192	0,7430 ± 0,0312
100%	Without a representation token	1	0,9204 ± 0,0000	0,7386 ± 0,0000
100%	Random encoders (3 initializations)	3	0,9193 ± 0,0048	0,7393 ± 0,0286
100%	PhiUSIIL encoder	1	0,8579 ± 0,0000	0,7521 ± 0,0000
100%	PhreshPhish encoder	1	0,9529 ± 0,0000	0,7499 ± 0,0000

To evaluate the generalizability of TARA beyond the original target period, an additional source-target configuration (Configuration A) was constructed using PhiUSIIL as the source dataset and an earlier PhreshPhish collection as the target population. Table S33 summarizes the chronological data partitions, including their observation periods, sample sizes, registered domains, and class distributions. Domain-disjoint splitting was applied to prevent information leakage between experimental roles.

Table S33. Dataset Composition and Experimental Partitions of the Additional Source-Target Configuration.

Population and Role	Collection dates	URL	Domains	Phishing	Legitimate
Configuration A: Adaptation	02.07.2024–15.08.2024	66 503	42 950	34 699	31 804
Configuration A: Selection	16.08.2024–26.08.2024	14 723	13 645	2 458	12 265
Configuration A: Probability Calibration	27.08.2024–15.09.2024	6 503	5 840	1 625	4 878
Configuration A: Threshold Calibration	27.08.2024–15.09.2024	19 822	17 520	5 180	14 642
Configuration A: Locked Test	16.09.2024–11.10.2024	29 187	23 557	15 768	13 419
Configuration B (Source): Training	02.07.2024–07.09.2025	234 246	125 275	126 331	107 915
Configuration B (Source): Selection	02.07.2024–07.09.2025	42 979	17 896	28 324	14 655
Configuration B (Source): Calibration	02.07.2024–07.09.2025	41 192	17 897	26 328	14 864
Configuration B (Source): Source Test	02.07.2024–07.09.2025	50 971	17 897	36 140	14 831
Extended Population of Legitimate URLs	31.07.2024–08.09.2025	94 752	72 560	0	94 752

To assess the statistical reliability of TARA's performance in Configuration A, paired comparisons with baseline models were conducted using 2,000 domain-level bootstrap replicates. Table S34 reports performance differences in low-FPR metrics, corresponding 95% confidence intervals, and Holm-adjusted p-values. These comparisons provide additional evidence regarding the reproducibility of TARA's ranking performance across different target periods.

Table S34. Paired Statistical Comparisons of TARA and Baselines in Configuration A.

The system being compared	Indicator	TARA - system [95% CI]	p	Conclusion
Logistic Regression	ROC-AUC	+0,0330 [0,0272; 0,0405]	0,016	in favor of TARA
Logistic Regression	pAUC at FPR ≤ 1%	+0,0874 [0,0750; 0,1008]	0,016	in favor of TARA
Logistic Regression	TPR at FPR ≤ 0.1%	+0,2055 [0,1159; 0,2722]	0,016	in favor of TARA
Logistic Regression	Brier score	−0,0412 [−0,0482; −0,0355]	0,016	in favor of TARA
Logistic Regression	pAUC at FPR ≤ 0.1%	+0,0887 [0,0543; 0,1254]	0,004	in favor of TARA
Extra Trees	ROC-AUC	+0,0090 [0,0072; 0,0112]	0,016	in favor of TARA
Extra Trees	pAUC at FPR ≤ 1%	+0,0434 [0,0314; 0,0553]	0,016	in favor of TARA
Extra Trees	TPR at FPR ≤ 0.1%	+0,0736 [0,0023; 0,1628]	0,924	not permitted
Extra Trees	Brier score	−0,0168 [−0,0204; −0,0135]	0,016	in favor of TARA
Extra Trees	pAUC at FPR ≤ 0.1%	+0,0517 [0,0031; 0,0997]	0,075	not permitted
HistGradient-Boosting	ROC-AUC	+0,0007 [−0,0002; 0,0018]	1,000	not permitted

The system being compared	Indicator	TARA - system [95% CI]	p	Conclusion
HistGradient-Boosting	pAUC at FPR ≤ 1%	+0,0037 [−0,0036; 0,0117]	1,000	not permitted
HistGradient-Boosting	TPR at FPR ≤ 0.1%	+0,0732 [−0,0176; 0,1481]	0,962	not permitted
HistGradient-Boosting	Brier score	+0,0009 [−0,0006; 0,0023]	1,000	not permitted
HistGradient-Boosting	pAUC at FPR ≤ 0.1%	+0,0386 [0,0054; 0,0696]	0,062	not permitted
CatBoost	ROC-AUC	+0,0017 [0,0008; 0,0028]	0,057	not permitted
CatBoost	pAUC at FPR ≤ 1%	+0,0184 [0,0136; 0,0237]	0,016	in favor of TARA
CatBoost	TPR at FPR ≤ 0.1%	+0,0275 [−0,0323; 0,0845]	1,000	not permitted
CatBoost	Brier score	−0,0031 [−0,0055; −0,0013]	0,170	not permitted
CatBoost	pAUC at FPR ≤ 0.1%	+0,0221 [0,0034; 0,0405]	0,069	not permitted
XGBoost	ROC-AUC	+0,0009 [0,0003; 0,0016]	0,168	not permitted
XGBoost	pAUC at FPR ≤ 1%	+0,0116 [0,0056; 0,0177]	0,016	in favor of TARA
XGBoost	TPR at FPR ≤ 0.1%	+0,0607 [−0,0110; 0,1075]	0,728	not permitted
XGBoost	Brier score	+0,0003 [−0,0009; 0,0015]	1,000	not permitted
XGBoost	pAUC at FPR ≤ 0.1%	+0,0280 [0,0069; 0,0490]	0,062	not permitted
LightGBM	ROC-AUC	+0,0007 [−0,0002; 0,0018]	1,000	not permitted
LightGBM	pAUC at FPR ≤ 1%	+0,0056 [−0,0018; 0,0135]	1,000	not permitted
LightGBM	TPR at FPR ≤ 0.1%	+0,0845 [0,0135; 0,1592]	0,382	not permitted
LightGBM	Brier score	+0,0020 [0,0003; 0,0035]	0,280	not permitted
LightGBM	pAUC at FPR ≤ 0.1%	+0,0459 [0,0128; 0,0759]	0,024	in favor of TARA
MLP	ROC-AUC	+0,0044 [0,0015; 0,0087]	0,382	not permitted
MLP	pAUC at FPR ≤ 1%	+0,0112 [0,0018; 0,0206]	0,304	not permitted
MLP	TPR at FPR ≤ 0.1%	+0,0596 [−0,0147; 0,1373]	1,000	not permitted
MLP	Brier score	−0,0057 [−0,0105; −0,0019]	0,270	not permitted
MLP	pAUC at FPR ≤ 0.1%	+0,0213 [−0,0139; 0,0542]	0,223	not permitted
FT-Transformer	ROC-AUC	+0,0015 [0,0005; 0,0028]	0,228	not permitted
FT-Transformer	pAUC at FPR ≤ 1%	+0,0124 [0,0045; 0,0204]	0,099	not permitted
FT-Transformer	TPR at FPR ≤ 0.1%	+0,0228 [−0,0286; 0,1738]	1,000	not permitted
FT-Transformer	Brier score	−0,0003 [−0,0028; 0,0017]	1,000	not permitted
FT-Transformer	pAUC at FPR ≤ 0.1%	+0,0509 [0,0139; 0,0781]	0,036	in favor of TARA
TARA without a representation token	ROC-AUC	+0,0013 [0,0007; 0,0019]	0,003	in favor of TARA
TARA without a representation token	pAUC at FPR ≤ 1%	+0,0014 [0,0005; 0,0031]	0,204	not permitted
TARA without a representation token	TPR at FPR ≤ 1%	+0,0053 [−0,0007; 0,0236]	0,754	not permitted
TARA with a random frozen encoder	ROC-AUC	+0,0009 [0,0003; 0,0016]	0,045	in favor of TARA
TARA with a random frozen encoder	pAUC at FPR ≤ 1%	−0,0006 [−0,0015; 0,0004]	0,703	not permitted

The system being compared	Indicator	TARA - system [95% CI]	p	Conclusion
TARA with a random frozen encoder	TPR at FPR ≤ 1%	−0,0055 [−0,0124; 0,0115]	0,754	not permitted

To evaluate the effectiveness of temporal threshold recalibration, three strategies were compared under a simulated weekly deployment protocol. Table S35 reports the observed false-positive rates and detection sensitivity at a nominal FPR of 0.1% for the original configuration and Configuration A.

Table S35. Comparison of Fixed and Temporally Recalibrated Thresholds at a Nominal FPR of 0.1%.

Configuration	Threshold strategy	Observed FPR (%)	Maximum weekly FPR (%)	TPR
Original	Fixed	0.318	3.727	0.6384
Original	Cumulative	0.318	3.727	0.6405
Original	Rolling, 14 dates	0.318	3.727	0.6432
Configuration A	Fixed	0.089	0.126	0.4355
Configuration A	Cumulative	0.075	0.126	0.4153
Configuration A	Rolling, 14 dates	0.089	0.158	0.3977