WITHIN-TOP RANKING QUALITY: A MASKED DIMENSION IN REGRESSION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Regression models are widely used to score candidates and select the best for downstream action, from drug compounds to financial assets. In many such settings, model scores do more than determine which candidates are selected: they also determine how resources are allocated among them, often in proportion to predicted score. A reversed ranking within the top group can therefore direct larger resource commitments toward poorer-performing candidates within that group. Yet existing metrics either evaluate prediction accuracy across the entire candidate pool, or measure how many good candidates enter the model's predicted top group. Neither directly isolates whether the candidates with the highest realised outcomes are correctly ordered relative to one another. We formalise this missing evaluation dimension with W_T, the rank correlation within the true top-outcome group rather than the model's predicted top group. To characterise when within-top reversal can occur, we prove a safety theorem: under independent additive noise with a log-concave density, the Bayes-optimal score cannot reverse; extending this guarantee to a fitted model requires a separate rank-alignment condition. Under feature-dependent noise structure, reversal is possible and standard metrics are largely insensitive to it. Across 17 regression domains, the dominant reversal pattern is domain-structured and persists across model classes. In the reversing case, standard model selection tends to favour models with poorer within-top ranking, and W_T-aware constrained selection raises point-estimate returns in chronological backtests. Beyond the reversing case, the degree to which W_T provides information beyond standard metrics varies widely, and the two criteria can prefer different models even when all W_T values are positive. This suggests that W_T monitoring is warranted whenever regression models drive costly selection decisions.

1 INTRODUCTION

Regression models are widely deployed as gatekeepers: they score a large pool of candidates, and only the top-scoring group is selected for downstream action. In many practical settings, model scores do more than determine who is selected. They also determine how resources are allocated among the selected candidates, often in proportion to predicted score. A drug screening pipeline prioritises synthesis in score order (Vamathevan et al., 2019), committing resources that cannot be recovered; a portfolio allocates capital by predicted alpha (Yang et al., 2020). A reversed ranking within the realised top group means the model assigns its lowest scores to the candidates that turn out best; under score-proportional allocation, these candidates would then receive the fewest resources. Aggregate accuracy does not capture this within-top ordering.

The standard toolkit includes aggregate metrics (R^2, Spearman ρ_S, RMSE) that assess overall prediction accuracy, and top-K metrics (NDCG@K (Järvelin and Kekäläinen, 2002), precision@K) that evaluate which candidates enter the model's predicted top group. Together, these metrics have driven substantial progress in model development across domains (Gu et al., 2020). Neither class, however, directly isolates ordering within the true top-outcome group, whose members are defined by realised outcomes rather than model predictions. In score-driven resource allocation, this leaves an important question unanswered: among the candidates that ultimately achieve the best outcomes, did the model assign higher scores to the better ones?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

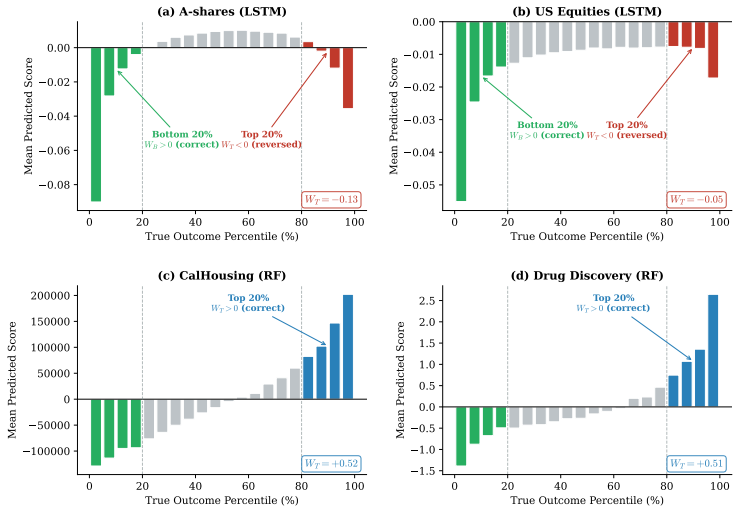


Figure 1: Mean predicted score by true outcome percentile. **Top (a,b):** Financial return prediction—the overall trend is positive ($\rho_S > 0$), but within the top quintile (red bars) predicted scores decrease as outcomes improve ($W_T < 0$): the model assigns lower scores to better candidates within the top group. **Bottom:** (c) CalHousing and (d) Drug discovery—predicted scores continue to increase within the top quintile ($W_T > 0$). W_B denotes the same statistic within the bottom quintile. The reversal in (a,b) is masked by ρ_S: with the true-top group in consecutive prediction ranks, a complete internal reversal changes ρ_S by only 0.016 (Proposition 1).

We find that within-top ranking can be reversed while all standard metrics remain positive. Figure 1 shows this concretely. In the financial model (top panels), the curve rises from left to right, confirming a positive overall rank association ($\rho_S > 0$). But among the truly best candidates (top quintile, red shading), predicted scores *decrease* as true outcomes increase: the model assigns lower scores to better candidates within the top group. In the drug model (bottom panels), the curve continues to rise within the top quintile. Both models have positive ρ_S, yet they differ completely in within-top ordering. As we show in Section 3, this is not a coincidence: with the true-top group in consecutive prediction ranks, a complete within-top reversal can change ρ_S by as little as 0.016 (Proposition 1), so aggregate metrics are largely insensitive to within-top ordering quality.

Masking explains why the problem has been overlooked; why it arises depends on data structure. In Figure 1(a,b), the model learns enough about outcome magnitude to rank the pool correctly, but the largest realised outcomes require predicting direction, which the model does not capture (Section 4.4).

To capture this dimension, we introduce W_T, the Spearman rank correlation between model predictions and actual outcomes, computed only among the candidates whose true outcomes place them in the top 20% (Definition 1). In other words, W_T asks: among the candidates who actually turned out to be the best, did the model assign them scores in the right order? Because this group is defined by realised outcomes, W_T is a retrospective diagnostic, not a real-time serving metric; it complements retrieval metrics (NDCG@K, precision@K), which evaluate the model's *predicted* top group. A model can achieve high NDCG@K while having $W_T < 0$. Whether this diagnostic dimension translates to deployment value depends on whether the difference from standard metrics aligns with outcomes and carries forward to future selection, which we test in Section 5.

More broadly, W_T can be informative even when it remains positive. Across the 9 domains in Table 2, the correlation between W_T and standard metrics ranges from $+0.94$ (strongly coupled) to -0.78 (conflicting), with financial reversal as the extreme case. Even in non-financial domains, however, coupling can be incomplete, meaning that standard evaluation alone cannot guarantee within-top quality even where no reversal is present.

We investigate this through theory and experiments across 17 regression domains and 684 models. Our main contributions are:

- **Cross-domain evidence and complementary evaluation.** Across 17 domains and 684 models, coupling between W_T and standard metrics ranges from strong to conflicting. Systematic reversal is concentrated in finance, where all 161 standard-loss models across 14 architectures produce $W_T < 0$; all 15 non-financial domains have positive median W_T.
- **Structural characterisation and safety conditions.** Under independent additive noise with a log-concave density, the Bayes-optimal score cannot reverse (Theorem 1); a separate rank-alignment condition extends the result to fitted models. Feature-dependent noise permits reversal, and a feature-level directional statistic separates the reversing and non-reversing domains in the tested sample; a matched residual-only interpolation reproduces reversal at fixed signal and SNR. An exact score-cut identity characterises the sign of W_T without requiring a cross-domain threshold (Appendix N.1).
- **Decision relevance and constrained model selection.** On fixed splits, ρ_S and W_T prefer different objectives even when all W_T values are positive; historical W_T correlates with retrospective capture across nine domains and provides incremental model-selection information in finance and Traffic.

2 Related Work

Financial ML and direction–magnitude asymmetry. Cross-sectional stock prediction has advanced rapidly (Li et al., 2024; Xu et al., 2021; Duan et al., 2022; Gu and Dao, 2024; Liu et al., 2024), but is typically evaluated by ρ_S or portfolio Sharpe—aggregate metrics blind to within-quantile quality. Christoffersen and Diebold (2006) documented that return direction is linked to volatility dynamics, establishing a fundamental asymmetry between magnitude and direction components; we find that this asymmetry has consequences for within-top ranking that aggregate metrics mask. Gu et al. (2020) use R^2 for return prediction.

Selection bias, collider effects, and subgroup evaluation. Conditioning on $Y \in \mathcal{T}$ induces Berkson's paradox (Berkson, 1946), studied in epidemiology (Hernán et al., 2004) and causal inference (Bareinboim and Pearl, 2012). The broader principle that aggregate statistics can misrepresent subgroup behaviour is well established in fairness (Barocas and Selbst, 2016) and medical AI (Oakden-Rayner et al., 2020). In information retrieval and recommendation, NDCG@K and MRR@K evaluate ranking quality within the model's *predicted* top group; re-ranking pipelines reorder this group. W_T addresses a different question, namely ranking quality within the *true* top group, as detailed in Section 3. Slice-based methods (Slice finder (Chung et al., 2019), Domino (Eyuboglu et al., 2022)) partition data by features but lack theoretical bounds on masking severity. Our subgroup is defined by outcome; Proposition 1 establishes a structural masking effect; and Theorem 1 gives sufficient conditions under which the Bayes score cannot reverse.

Ranking metrics and molecular prediction. Ranking losses such as ListMLE (Xia et al., 2008) and LambdaRank (Burges, 2010) optimise whole-list ordering but share ρ_S's insensitivity to within-group quality. Virtual screening benchmarks in drug discovery (Wu et al., 2018) evaluate by RMSE, R^2, or enrichment factor (Truchon and Bayly, 2007), none of which detects within-top reversal. W_T is designed to complement these metrics.

3 Problem Formulation

We now formalise this quantity as a metric and show that aggregate Spearman can mask the reversal, a finding we extend empirically to other standard metrics.

3.1 Within-top ranking quality: W_T

Definition 1 (Within-Top Ranking Quality)**.** *Let $\hat{y} = f(X)$ be predictions, Y the outcome, and $\mathcal{T}_q = \{i : Y_i \geq Q_{1-q}(Y)\}$ the top q-fraction by true outcome ($q = 0.2$ throughout; tested at $q \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$, Appendix B; we drop the subscript when the context is clear). The within-top ranking quality is*

$$W_T \coloneqq \rho_S(\hat{y}, Y)_{\mathcal{T}},$$

the Spearman rank correlation between predictions and outcomes, computed only among samples in $\mathcal{T}$. The standard aggregate $\rho_S \coloneqq \rho_S(\hat{y}, Y)$ (RankIC) computes the same over all samples.

Interpretation. $W_T > 0$ means the model assigns higher scores to the truly better candidates within the top group; $W_T < 0$ means the ranking is reversed; $W_T = 0$ indicates zero monotone rank association. As a Spearman correlation, $W_T \in [-1, +1]$. Random predictions produce $W_T \approx 0 \pm O(1/\sqrt{qn})$, so small top groups require explicit uncertainty reporting (Section 5).

Interpretation boundary. $W_T < 0$ does not by itself imply Bayes-suboptimality or that a higher-W_T model improves deployment utility. Because the conditioning set is defined by realised outcomes, W_T diagnoses how scores order realised winners; its decision relevance depends on retrieval, allocation rule and temporal stability (Section 5).

Distinction from retrieval metrics. Retrieval metrics (NDCG@K, MRR@K) evaluate the model's *predicted* top group; W_T evaluates the true-top group $\mathcal{T}$. A model can achieve high NDCG@K while having $W_T < 0$. Because $\mathcal{T}$ is outcome-defined, W_T is a retrospective diagnostic for model selection, not a real-time serving metric (Appendix A).

3.2 Structural masking of W_T

Proposition 1 (Structural Masking)**.** *Assume $N \geq m \geq 2$ and no ties. Let the m true-top observations occupy consecutive global prediction ranks among N observations. Holding all outcomes and all other predictions fixed, changing the group's ordering from $W_T = +1$ to $W_T = -1$ changes overall Spearman by $2m(m^2-1)/[N(N^2-1)] \approx 2q^3$, where $q = m/N$—only 0.016 for $q = 0.2$. Proof: Appendix C.*

A complete within-top reversal can therefore accompany only a small change in aggregate Spearman. Precision and binary-relevance NDCG can remain unchanged under internal reorderings of true-top candidates; graded NDCG can respond, but it does not directly isolate this target. The empirical metric comparisons are reported in Appendices Q.2 and B; Appendix A distinguishes true-top from predicted-top evaluation. A natural follow-up is whether structural conditions can rule out within-top reversal; Section 4.2 addresses this after the cross-domain survey.

4 Where and Why Reversal Occurs

We begin with a cross-domain survey to establish the empirical landscape (Section 4.1), then identify sufficient conditions that rule out reversal (Section 4.2), analyse the financial case in detail (Section 4.3), and test a candidate mechanism behind the reversal (Section 4.4).

4.1 Survey: where does reversal occur?

We first establish *where* reversal occurs. Table 1 reports W_T for each of 17 domains (median across models); the key column is W_T, which is positive for all non-financial domains and negative only for finance. The 17 domains span drug discovery, materials science, real estate, engineering, and finance, assembled from publicly available regression benchmarks to span a range of sample sizes (n from 442 to 139K), feature types, and application areas without pre-screening for W_T behaviour. All non-financial datasets are publicly available (UCI, MoleculeNet); financial data uses the open-source Qlib platform (Yang et al., 2020). Full experimental details are in Appendix G. Two lines of evidence indicate that this split reflects domain structure rather than model class: tabular models (Ridge, RF, XGBoost, MLP) yield positive domain-level median W_T across all 15 non-financial domains, and the same LSTM architecture produces $W_T < 0$ on financial data but $W_T > 0$ on the two non-financial time-series domains tested (Electricity, Traffic; Appendix G).

Robustness. Of 492 non-financial models, 96% produce $W_T > 0$; the 21 exceptions are low-accuracy models with $\rho_S < 0.10$. The pattern persists across label transformations (Appendix H).

This pattern raises two questions: what structural mechanism produces reversal in the financial case, and why do the other 15 domains not exhibit it (Section 4.4)?

4.2 Safety theorem: when is reversal ruled out?

We study noise structure through the signal $G = g(X) = E[Y \mid X]$ and its learned approximations.

Table 1: Cross-domain results (W_T: median across models). Financial domains (bold) are the only ones with $W_T < 0$. $\max|\rho_{\text{dir}}|$: strongest feature-level correlation with outcome direction (Section 4.4). †: primary-configuration models only; $W_T < 0$ counts standard-loss only. See Appendix G for the full experimental breakdown and per-model results.

Domain	n	**Mod.**	ρ_S	W_T	$\max\|\rho_{\text{dir}}\|$	$W_T<0$
MiamiHousing	13.9K	15	+0.91	+0.75	0.58	0/15
FreeSolv	642	15	+0.85	+0.67	0.19	0/15
Drug (AChE)	6842	51	+0.80	+0.51	0.27	0/51
ESOL	1128	15	+0.90	+0.31	0.18	0/15
Lipophilicity	4200	15	+0.67	+0.20	0.16	0/15
Superconduct.	21.3K	7	+0.89	+0.55	0.68	0/7
CalHousing	20.6K	15	+0.86	+0.52	0.56	0/15
EnergyEff.	768	51	+0.92	+0.49	0.32	11/51
Abalone	4177	51	+0.77	+0.33	0.59	0/51
Wine	1599	50	+0.62	+0.54	0.44	0/50
Electricity	26K×321	51	+0.55	+0.24	0.40	0/51
NBA	1069	3	+0.65	+0.38	0.46	0/3
Traffic	17.5K×862	51	+0.58	+0.23	0.52	0/51
Weather	139K	51	+0.32	+0.21	0.21	0/51
Diabetes	442	51	+0.66	+0.08	0.50	10/51
A-shares	**4.8K/d**	**149**†	**+0.06**	**−0.13**	**0.04**	**115/115**†
US eq.	**2K/d**	**14**†	**+0.05**	**−0.05**	**0.01**	**12/12**†

Theorem 1 (Safety Floor)**.** *For $E|Y| < \infty$, let $\mathcal{T}_q = \{Y \geq c_q\}$ be the true-top event, where $c_q = Q_{1-q}(Y)$, with $P(Y = c_q) = 0$, and write $W_T(\tilde{F}) = \rho_S(\tilde{F}, Y \mid \mathcal{T}_q)$ for any score $\tilde{F}$. The conditioning notation $\rho_S(\cdot,\cdot \mid \mathcal{T}_q)$ is equivalent to the subscript $\rho_S(\cdot,\cdot)_{\mathcal{T}_q}$ of Definition 1.*

Assume nonzero conditional rank variances. If, for almost every selected pair $g_2 > g_1$ and every $u \geq 0$,

$$P(Y - c_q > u \mid G = g_2, \mathcal{T}_q) \geq P(Y - c_q > u \mid G = g_1, \mathcal{T}_q), \tag{1}$$

then $W_T(G) \geq 0$.

Intuition: Within the true-top group, higher-signal candidates are at least as likely to exceed any given outcome threshold. For $Y = G + \sigma\varepsilon$ with constant $\sigma > 0$, $\varepsilon \perp X$, $E[\varepsilon] = 0$, and a log-concave noise density, this ordering holds (Appendix D).

Varying scale. Under the same noise assumptions, $Y = G + s(G)\varepsilon$ preserves the selected ordering if $s(g)$ is non-decreasing and $t_q(g) = (c_q - g)/s(g)$ non-increasing. When the noise is also symmetric and log-concave, the threshold condition $t_q(g) \downarrow$ can be dropped: non-decreasing $s(g)$ alone guarantees $W_T(G) \geq 0$ (Theorem 2). The proof does not establish selected stochastic dominance (1); instead it shows that every pairwise winning probability $K(g,h) \geq 1/2$ via measure-preserving reflections, a weaker route that suffices for nonnegative rank correlation. Three boundary results delineate the theorem's scope: asymmetric log-concave noise can reverse (Proposition 2), symmetric non-log-concave noise can reverse (Proposition 3), and independent random scale can reverse even with log-concave elementary noise (Proposition 4). Symmetry and log-concavity are each needed: dropping either one, while keeping the other and a non-decreasing deterministic scale, permits reversal in the stated constructions. Under symmetric log-concave noise with non-decreasing deterministic scale, reversal cannot occur; the financial results are consistent with a change in residual shape, as formalised by the direction–magnitude mechanism of Section 4.4.

Fitted predictors. Theorem 1 is invariant to feature composition. A fitted score $F = f(X)$ (the $\hat{y}$ of Definition 1), however, does not automatically inherit the signal's guarantee. Let $r = \rho_S(F, G \mid \mathcal{T}_q)$ and $w = W_T(G)$, using the same selected midranks. The Gram bound gives

$$W_T(F) \geq rw - \sqrt{(1 - r^2)(1 - w^2)}. \tag{2}$$

Thus $r, w \geq 0$ and $r^2 + w^2 \geq 1$ ensure non-reversal; the proof (Theorem 3) and bounded-approximation corollary (Corollary 4) are in Appendix F. The theorem identifies when reversal is ruled out; we next examine the case where it is not.

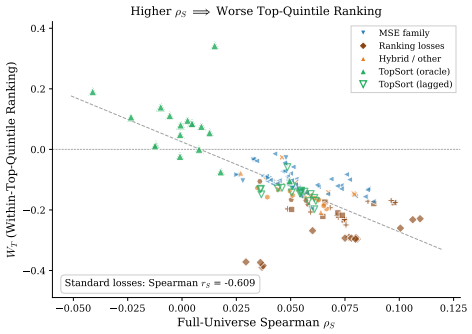


Figure 2: ρ_S vs. W_T across 192 financial models. Standard-loss models (blue/orange) cluster in the $W_T < 0$ region with $\rho = -0.883$ among 53 loss-sweep models ($\rho = -0.609$ over all standard-loss variants). TopSort with completed training-label supervision (green) achieves $W_T > 0$ by sacrificing ρ_S. The negative slope means that higher aggregate ranking *tends to coincide with* lower within-top ranking—the two quality dimensions are opposed.

4.3 Financial analysis: how broad and persistent?

Financial return prediction exhibits the deepest reversal in our survey: the only setting where W_T is consistently negative. We analyse it in detail to establish how robust the reversal is; Section 4.4 then tests a candidate structural mechanism.

The reversal spans all tested architectures and standard losses. Using Alpha360 features (Yang et al., 2020) with six price/volume fields (open, high, low, close, VWAP, volume) over a 60-day lookback, all 161 standard-loss models produce $W_T < 0$, spanning 14 architectures: MASTER (Li et al., 2024), HIST (Xu et al., 2021), FactorVAE (Duan et al., 2022), AlphaForge (Shi et al., 2025), HRFT (Xu et al., 2025), LSTM (Hochreiter and Schmidhuber, 1997), GRU, ALSTM (Qin et al., 2017), iTransformer (Liu et al., 2024), Mamba (Gu and Dao, 2024), Transformer, GATs (Veličković et al., 2018), XGBoost, and Linear (Appendix G). TopSort and TopListMLE (losses that rank observations within the realised training-set top group) achieve $W_T > 0$ in all 13 direct-supervision runs, but lagged or random group masks do not (0/18).

The reversal is not a test-set or regime artefact: it is present on training data and persists across four calendar years spanning both bull and bear markets (Appendix I).

Higher ρ_S correlates with deeper reversal. Among 53 financial models, $\rho(\rho_S, W_T) = -0.883$ (Figure 2). Higher aggregate ranking tends to coincide with deeper reversal in this domain, the conflicting end of the coupling spectrum examined in Section 5.1.

Feature invariance. Neither feature selection nor feature source eliminates the reversal: training on individual price/volume field families, applying three input transformations, and replacing price-derived features with fundamentals all produce $W_T < 0$ (Appendices G, J, K, L). This empirical pattern is consistent with Corollary 2, which shows that the Bayes-level safety result does not depend on feature composition. The breadth and persistence of the reversal motivate a structural explanation; we turn to this next.

4.4 A sufficient mechanism: residual structure, not signal strength

Direction–magnitude structure and reversal. The pattern in Figure 1 is consistent with the asymmetry between return magnitude and direction (Christoffersen and Diebold, 2006). In the model $Y = D\widetilde{M}$, features carry magnitude information but the probability of a positive return is constant: $D \perp (X, \widetilde{M})$ and $P(D = +1) = p < 1/2$. The signal $G = (2p - 1)E[\widetilde{M} \mid X]$ decreases with predictable magnitude, whereas the true top consists of large positive outcomes when $q < p$. Appendix N gives conditions under which this produces positive overall Spearman and negative W_T.

Matched controls. Because fitted R^2 does not isolate latent signal-to-noise ratio (SNR), we vary known signal fractions, scale coupling and architecture while holding the features, split, components

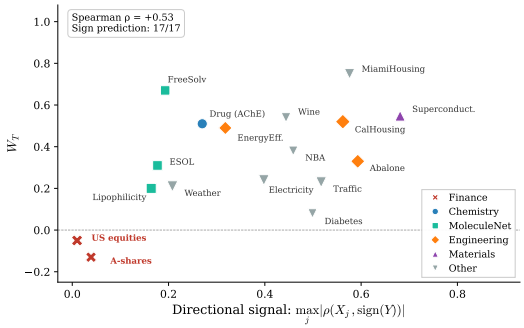


Figure 3: Directional signal strength vs. W_T across 17 domains. Red: financial domains—the only ones in the reversing regime. Both reversing domains have markedly lower directional signal than all non-reversing ones.

and experimental pipeline fixed. All 27 runs retain $W_T > 0$, including at a 1% signal fraction. A separate residual-only interpolation fixes the signal and SNR: at $\lambda = 0, 0.5, 1$, learned W_T is $+0.0130, -0.2607, -0.8935$, respectively, while overall Spearman stays positive (Appendix O). Thus low signal and the tested scale changes do not reproduce reversal, whereas the residual-structure change does within this controlled family. The two endpoints bracket the safety theorem: at $\lambda = 0$ the residual is homoscedastic Gaussian, so Corollary 1 guarantees $W_T(G) \geq 0$ (Bayes value is $+0.014$); at $\lambda = 1$ the direction–magnitude residual violates the selected ordering, and $W_T(G) = -0.968$. The interpolation thus traces the boundary of the theorem in a single controlled experiment.

Supporting evidence. In retrospective seed-42 interventions on both markets, learned magnitude paired with realised direction yields $W_T = +0.310/+0.440$; shuffling magnitude reduces these values to $0.000/0.002$ (Appendix O). The same-feature additive control remains positive (Appendix P), while the residual diagnostic gives $\rho(|\hat{\varepsilon}|, |\hat{y}|) = +0.133$ in finance (Appendix M.1). Injecting a synthetic inverse-magnitude feature into drug data lowers W_T monotonically (Appendix P.1). Together, these results establish a sufficient mechanism in the constructed family and are consistent with both markets, without identifying a unique cause of natural-market reversal.

The non-financial counterpart. The direction–magnitude mechanism predicts three signatures in reversing domains: weak feature-level directional signal, sign flips between upper- and lower-tail regression coefficients, and, in non-reversing domains, positivity that survives label noise. The tested evidence separates the reversing financial domains from their non-financial counterparts: $\max_j |\rho(X_j, \text{sign}(Y))| \leq 0.04$ in both financial domains versus above 0.16 in all 15 others (Table 1); quantile-regression coefficients flip sign in 7/42 financial cases but none of nine non-financial domains tested; and, in Drug and MiamiHousing, label noise up to $R^2 = 0$ drives W_T toward zero but never below it (Appendix G). These patterns do not verify the safety conditions of Theorem 1, but they show that the mechanism's tested fingerprints are absent in the corresponding non-reversing comparisons.

5 When Does W_T Matter for Decisions?

5.1 Does W_T differ from standard metrics?

We now turn to the central question: across the tested domains, does W_T carry information that standard metrics miss, and does this information improve decisions? If ρ_S and W_T always agree, W_T is redundant; if they disagree, standard model selection may favour a model with poorer within-top ordering. Table 2 reveals the coupling spectrum: $\rho(\rho_S, W_T)$ ranges from $+0.94$ (Superconduct.) to -0.78 (finance). Diabetes illustrates the resulting instability (-0.28 pooled, $+0.24$ within-split mean; Appendix G.6).

W_T need not be negative to be informative; it only needs to be not fully coupled. Wine and Abalone ($+0.31$ and $+0.28$) have substantially weaker coupling than the top five domains, yet all W_T values remain positive. In the conflicting financial panel, higher-ρ_S models tend to have lower W_T, though the correlation alone does not show that the highest-ρ_S model has the worst W_T.

Table 2: Cross-domain coupling and retrospective capture. Outer ρ: Spearman across 90 model–split combinations (9 configs $\times$ 10 splits) for non-financial domains, 13 default-loss architectures for finance ($\star$). Figure 2 extends this to 53 loss-sweep models. R^2 and $-\mathrm{RMSE}$ can differ across splits due to varying outcome variance. Best std.$= \max_m |\rho(m, W_T)|$ over R^2, $-$RMSE, $-$MAE, ρ_S, EF@20% (Appendix Q.2). $\rho(W_T$,dec.): correlation with retrospective capture (Appendix Q.3). Wine: red+white ($n = 6{,}497$; Table 1 red-only). Superconduct.: 90 models here, 7 in Table 1. ρ^2 uses unrounded values.

Domain	$\rho(\rho_S, W_T)$	$\rho^2(\rho_S, W_T)$	**Best** $\lvert\rho(m, W_T)\rvert$	$\rho(W_T$**,dec.)**
Superconduct.	$+0.94$	0.88	0.94	$+0.91$
CalHousing	$+0.91$	0.83	0.94	$+0.91$
Drug	$+0.92$	0.85	0.92	$+0.84$
EnergyEff.	$+0.83$	0.69	0.85	$+0.78$
MiamiHousing	$+0.76$	0.58	0.86	$+0.89$
Wine	$+0.31$	0.10	0.45	$+0.94$
Abalone	$+0.28$	0.08	0.65	$+0.49$
Diabetes	-0.28	0.08	0.28	$+0.63$
Financial*	$\mathbf{-0.78}$	**0.61**	**0.85**	$\mathbf{+0.96}$

Objective comparison on a fixed split. A more direct test holds the split, architecture, optimiser and training budget fixed and varies only the objective among MSE, pairwise ranking, ListMLE and TopSort (CalHousing and Diabetes, three seeds each, 24 runs; Appendix R). All W_T values are positive, yet the two criteria prefer different objectives: on CalHousing, ρ_S favours pairwise ranking while W_T favours TopSort; on Diabetes, they favour ListMLE and pairwise ranking, respectively. This confirms that W_T reveals within-top ordering differences even without reversal.

5.2 Does the difference align with outcomes?

The introduction motivates W_T through resource allocation: among the candidates that turn out best, does the model rank them correctly? Fixed-membership allocation tests this directly. Across nine financial objectives, W_T correlates $+0.983$ with the score-rank return gain over equal allocation when realised true-top membership is held fixed (Appendix S.4). This is retrospective, not a deployable rule, but it confirms that within-top ordering quality affects how resources are distributed among the best candidates.

Beyond the financial case, W_T correlates with retrospective within-true-top capture in all 9 domains ($\rho > +0.49$; mean $+0.82$), from moderate (Abalone $+0.49$) to very strong (finance $+0.96$). In strongly coupled domains such as CalHousing, W_T adds little beyond ρ_S on this grid. In Diabetes, models with higher W_T produce better top-group outcomes in the majority of splits, though the effect does not reach significance given only 18 true-top observations (Appendix G.6). Predicted-top allocation gain is a separate target that can move in the opposite direction (Appendix G).

5.3 Can historical W_T guide future selection?

The objective comparison in Section 5.1 shows that W_T and standard metrics can prefer different models. A natural use is to exploit this: enforce an aggregate-quality floor, then select the model with the best historical W_T among survivors. Direct training-time optimisation of W_T is an alternative route but lowers ρ_S in every tested configuration (Appendix G).

Financial backtest. The rule applies a 0.045 overall-Spearman floor to three-block histories, then selects the largest W_T. A maturity-purged 92-model reconstruction improves gross annualised return by 6.7 pp over historical-Spearman selection (Sharpe $+0.02$ equal-weight, $+0.11$ score-rank; the portfolio holds the predicted top 5%, while W_T is evaluated at $q = 0.2$; Appendix T). Historical-to-next-block rankings are positively associated in all 33 audited windows.

Non-financial selection. Across 122 Traffic future blocks, adding historical W_T to a predicted-top capture selector improves allocation gain by 4.2% over the baseline (39-comparison-adjusted $p = 0.015$; Appendix U).

Scope and conditions. A comparison with aggregate and retrieval-based selectors shows that W_T selects different models in the majority of windows and produces higher point-estimate returns than ρ_S-based selection under both allocation schemes (Appendix T.4). In non-financial domains, Traffic shows the strongest improvement; Electricity and Weather show smaller or no gains (Appendix U). All of these comparisons use matured historical outcomes; extending them to new cohorts requires that the relevant feedback is available and that model quality is stable. In practice, three regimes arise: complete feedback permits direct evaluation; selective feedback requires consistency, ignorability and overlap; and future transfer requires estimation accuracy and stability, with regret bounded by $2(\epsilon + \Delta)$ (Appendix V).

6 Discussion

Contributions. The central finding is that within-top ranking quality constitutes a dimension of model performance not fully captured by standard metrics. This conclusion rests on three pillars: an exact masking identity (Proposition 1) explains why this dimension has been overlooked; a safety theorem gives sufficient conditions under which the Bayes score cannot reverse, with a transfer certificate for fitted models; and experiments across 17 domains confirm that W_T provides additional information wherever coupling is incomplete.

Relationship to broader principles. This finding parallels a principle established in fairness (Barocas and Selbst, 2016) and medical AI (Oakden-Rayner et al., 2020): aggregate metrics can mask consequential subgroup-level structure. Our contribution is to show that in top-K regression this masking can be extreme.

Implications for practice. Because the coupling regime cannot be inferred from aggregate metrics alone, we recommend reporting W_T as a low-cost retrospective complement when matured outcomes are available in top-K selection settings; whether it should influence model selection depends on feedback availability and temporal stability (Section 5.3).

Limitations and future work. The safety theorem gives sufficient conditions, not necessary ones: a domain may have $W_T > 0$ without satisfying the theorem's assumptions, and the theorem does not cover all noise structures. In particular, the empirical signal-noise relationship is positive but does not establish a non-decreasing functional form (Appendix M); scale–signal independence alone is insufficient (Appendix E), and the directional statistic $\max_j |\rho(X_j, \operatorname{sign} Y)|$ is neither necessary nor sufficient for the sign of W_T over unrestricted data-generating processes (DGPs) (Proposition 5). Matched controls establish a sufficient residual-structure mechanism within a constructed family, not a unique attribution of natural-market reversal. The observed domain-level separation by $\max |\rho_{\text{dir}}|$ is descriptive and has not been validated on held-out domains. Top-aware losses improve within-top ordering using completed training outcomes, but no tested configuration simultaneously improves W_T and aggregate quality. Selective feedback settings require the identification conditions stated in Section 5.3. Why strong directional signal does not guarantee high coupling (e.g., Abalone: $\max |\rho_{\text{dir}}| = 0.59$ but $\rho = +0.28$) remains open.

7 Conclusion

In score-driven resource allocation, within-top ranking quality describes how well the model orders the candidates that turn out best; this affects, but does not solely determine, score-based allocation among selected candidates, yet standard evaluation metrics are largely insensitive to it. This paper introduces W_T to capture this masked dimension. A structural masking result explains why it has been overlooked; a safety theorem gives sufficient conditions under which the Bayes score cannot reverse, and a transfer certificate extends these conditions to fitted models; and experiments across 17 domains show that coupling between W_T and standard metrics ranges from strong to conflicting, and W_T can provide additional information where coupling is incomplete. The practical consequence is direct: in the conflicting case, standard model selection tends to favour models with poorer within-top ranking, and W_T-aware constrained selection raises point-estimate returns in chronological backtests. More broadly, this work illustrates that model evaluation metrics are not interchangeable—different metrics capture different quality dimensions, and which dimension matters depends on how the model's output is used. We expect W_T to be easy to report as a low-cost diagnostic complement to existing evaluation practices.

AI USE STATEMENT

In this work, we used generative AI tools to assist in exploring and checking some auxiliary counterexample constructions and related arguments in the appendix. We did not use generative AI tools to develop the main theoretical framework or main mathematical results, including the safety theorem and its proof, nor to design or conduct the experiments, implement the evaluated methods, process the experimental data, or generate the reported numerical results; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for manuscript editing and readability improvement, literature and information search, and feedback on manuscript presentation and structure. We independently reviewed and verified all AI-assisted mathematical arguments against our own derivations and checked relevant literature and sources where applicable. All experiments and reported results were produced and validated by the authors. We take responsibility for the final content of this work, including all text, claims, proofs, experimental results, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Theoretical proofs and derivations are in Appendices D–F (safety conditions, scale counterexamples, transfer certificate). Dataset descriptions, preprocessing, model configurations and evaluation protocols are in Appendix G. The chronological model-selection procedures, label-maturity handling and statistical tests for the financial backtest are in Appendix T; for non-financial domains in Appendix U. The matched-control experimental design is in Appendix O. Sensitivity analyses include top-fraction variation (Appendix B), label noise (Appendix G), feature and label transformations (Appendices L, P, H), and bootstrap uncertainty (Appendix G.6). Identification and transport conditions for selective feedback are in Appendix V. All non-financial datasets are publicly available (UCI, MoleculeNet); financial data uses the open-source Qlib platform.

REFERENCES

A. Ang, R. J. Hodrick, Y. Xing, and X. Zhang. The cross-section of volatility and expected returns. *J. Finance*, 61(1):259–299, 2006.

M. Bagnoli and T. Bergstrom. Log-concave probability and its applications. *Econ. Theory*, 26(2):445–469, 2005.

E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *AISTATS*, 2012.

S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. Law Rev.*, 104(3):671–732, 2016.

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biom. Bull.*, 2(3):47–53, 1946.

C. J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

P. Christoffersen and F. X. Diebold. Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Manag. Sci.*, 52(8):1273–1287, 2006.

Y. Chung, T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang. Slice finder: Automated data slicing for model validation. In *ICDE*, 2019.

Y. Duan et al. FactorVAE: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *AAAI*, 2022.

S. Eyuboglu et al. Domino: Discovering systematic errors with cross-modal embeddings. In *ICLR*, 2022.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024.

S. Gu, B. Kelly, and D. Xiu. Empirical asset pricing via machine learning. *Rev. Financ. Stud.*, 33(5):2223–2273, 2020.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Univ. Press, 1934.

M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

W. Hoeffding. Masstabinvariante Korrelationstheorie. *Schriften Math. Inst. Univ. Berlin*, 5:181–233, 1940.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

E. L. Lehmann. Some concepts of dependence. *Ann. Math. Statist.*, 37(5):1137–1153, 1966.

T. Li et al. MASTER: Market-guided stock transformer for stock price forecasting. In *AAAI*, 2024.

Y. Liu et al. iTransformer: Inverted transformers are effective for time series forecasting. In *ICLR*, 2024.

L. Oakden-Rayner et al. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *CHIL*, 2020.

M. Parkinson. The extreme value method for estimating the variance of the rate of return. *J. Bus.*, 53(1):61–65, 1980.

Y. Qin et al. A dual-stage attention-based recurrent neural network for time series prediction. In *IJCAI*, 2017.

H. Shi et al. AlphaForge: A framework to mine and dynamically combine formulaic alpha factors. In *AAAI*, 2025.

J.-F. Truchon and C. I. Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.*, 47(2):488–508, 2007.

J. Vamathevan et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.*, 18(6):463–477, 2019.

P. Veličković et al. Graph attention networks. In *ICLR*, 2018.

Z. Wu et al. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.*, 9(2):513–530, 2018.

F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: Theory and algorithm. In *ICML*, 2008.

W. Xu et al. HIST: A graph-based framework for stock trend forecasting via mining concept-oriented shared information. arXiv preprint arXiv:2110.13716, 2021.

W. Xu, R. Wang, C. Li, Y. Hu, and Z. Lu. HRFT: Mining high-frequency risk factor collections end-to-end via transformer. In *WWW Companion*, 2025.

X. Yang, W. Liu, D. Zhou, J. Bian, and T.-Y. Liu. Qlib: An AI-oriented quantitative investment platform. arXiv preprint arXiv:2009.11189, 2020.

A Definition: True Top vs Predicted Top

Let $F = f(X)$. We distinguish two complementary conditional statistics:

$$W_q^Y(F) = \rho_S(F, Y \mid Y \geq Q_{1-q}(Y)), \qquad W_q^F(F) = \rho_S(F, Y \mid F \geq Q_{1-q}(F)).$$

Ranks are recomputed within the respective groups using the declared tie convention. The first is the paper's W_T: it evaluates ordering among candidates that ultimately realise the best outcomes. The second evaluates outcome ordering within the model's predicted shortlist. Although predicted membership is available before outcomes mature, its outcome-based rank correlation is still retrospective.

The original comparisons report $(W_q^Y, W_q^F) = (+0.53, +0.65)$ for CalHousing and $(+0.38, +0.37)$ for Abalone. The financial model-averaged values are $(-0.146, -0.002)$ across the reported 18-model comparison. These describe reversed association within the realised top group and little net monotone association within the predicted group; they do not identify random ordering, the overlap of the two sets, or whether high raw scores correspond to high volatility.

Retrieval is measured separately. In the concentrated-portfolio comparison, precision of the predicted top 5% relative to the realised top quintile is 25.1% for LSTM–MSE and 11.3% for LSTM–ListMLE, against a 20% random-selection benchmark. The two fractions here are different: the predicted shortlist is 5%, and the target relevant set is 20%. These direct retrieval results, rather than a difference of conditional correlations, quantify the overlap.

We retain the true-outcome definition because it answers the paper's within-top ranking question. Predicted-top ordering, retrieval and allocation utility remain complementary evaluation targets, not alternative names for W_T or consequences of its sign.

B Quantile Sensitivity and Alternative Metrics

B.1 Persistence over the tested top fractions

We hold predictions fixed for the same 20-model China-market panel. The evaluated fractions are $q = 0.05, 0.10, 0.20, 0.30$ and 0.50. Every model has negative W_T at each tested fraction; Table B1 gives the across-model means and sign counts. These are sensitivity results for the same predictions, not independent training replications or additional domains.

Table B1: Top-fraction sensitivity in the fixed 20-model China-market panel. The $2q^3$ column is an analytic illustration for consecutive global prediction ranks, not an empirical bound for these models. Percentage entries express absolute correlation changes: 1.6% denotes 0.016, not a relative decrease; entries are rounded.

		Consecutive ranks		
q	**Top group**	$2q^3$	**Mean** W_T	**%** $W_T < 0$
0.05	Top 5%	0.03%	-0.084	100%
0.10	Top 10%	0.2%	-0.124	100%
0.20	Top 20%	1.6%	-0.149	100%
0.30	Top 30%	5.4%	-0.149	100%
0.50	Top 50%	25%	-0.127	100%

The mean ranges from -0.084 to -0.149, with the most negative reported means at $q = 0.20$ and 0.30. Changing q changes both the selected group and its internal ranks. Persistence over this finite grid does not establish sign invariance at untested fractions or certify the corresponding population signs.

The analytic column refers only to the following rank configuration: with no ties, the m true-top observations occupy consecutive global prediction ranks. Holding all other predictions fixed, changing their internal ordering from fully concordant to fully reversed changes overall Spearman by $2m(m^2-1)/[N(N^2-1)] \simeq 2q^3$ for $m/N = q$ (Proposition 1; Appendix C). The table does not assert that this configuration holds for the panel.

B.2 Retrieval metrics and within-top ordering

In the reported panel, all 20 models have positive NDCG@20% and precision@20% despite negative W_T. These observations do not make positive retrieval scores a certificate of within-top non-reversal. Table B2 reports the cross-model associations, not a structural blindness result.

Table B2: Reported model-level Spearman correlations in the financial retrieval comparison. NDCG@20% uses the linear-gain convention described below.

Metric pair	**Spearman ρ**
ρ_S vs W_T	-0.893
NDCG@20% vs W_T	-0.497
Precision@20% vs W_T	$+0.931$

The NDCG calculation uses raw standardised labels as linear gains with logarithmic rank discounts; these values are not binary NDCG or a non-negative graded-gain convention. The five models with the highest NDCG@20% have mean $W_T = -0.211$; the five with the highest overall Spearman have mean $W_T = -0.286$. Higher precision is associated with less-negative W_T in this panel ($\rho = +0.931$), although the highest-precision models still have negative W_T.

Precision measures overlap, not the order of the true-top identities. Binary-relevance NDCG also records where relevant observations occur, but exchanging their identities can leave it unchanged while reversing their internal order. Graded-gain NDCG can respond to such changes. The reported associations therefore describe co-movement in this panel, not identification of the complete within-true-top ordering by a retrieval metric.

C Structural Masking: Complete Proof

Proof of Proposition 1. Assume $N \geq m \geq 2$ and no ties. Order the true-top observations by outcome, with global outcome ranks $b_k = N - m + k$, $k = 1, \ldots, m$. Let $a_1 < \cdots < a_m$ be their occupied global prediction ranks. The set of these ranks stays fixed when prediction values are permuted within the group. For an internal permutation π, write

$$D(\pi) = \sum_{k=1}^{m} (a_{\pi(k)} - b_k)^2 = \sum_{k=1}^{m} a_k^2 + \sum_{k=1}^{m} b_k^2 - 2\sum_{k=1}^{m} a_{\pi(k)} b_k.$$

The first two sums are fixed. The rearrangement inequality (Hardy et al., 1934) therefore makes D smallest under concordant ordering and largest under reverse ordering. Because all other squared rank differences are unchanged, these are also the maximum and minimum attainable aggregate Spearman values under this within-group operation. Moreover,

$$\begin{aligned} D_{\max} - D_{\min} &= 2\sum_{k=1}^{m} a_k (b_k - b_{m+1-k}) \\ &= 2\sum_{k=1}^{m} (2k - m - 1) a_k = 2 \sum_{1 \leq i < j \leq m} (a_j - a_i). \end{aligned}$$

Using the no-tie Spearman formula gives the exact general range

$$\rho_{S,+} - \rho_{S,-} = \frac{12}{N(N^2-1)} \sum_{1 \leq i < j \leq m} (a_j - a_i), \tag{3}$$

where $+$ and $-$ denote groupwise perfect agreement and complete reversal, respectively. If $a_k = a_1 + k - 1$, then

$$\sum_{i<j} (a_j - a_i) = \sum_{d=1}^{m-1} d(m-d) = \frac{m(m^2-1)}{6}.$$

Consequently,

$$\rho_{S,+} - \rho_{S,-} = \frac{2m(m^2-1)}{N(N^2-1)} \longrightarrow 2q^3 \qquad (N \to \infty,\ m/N \to q). \tag{4}$$

The same two assignments give groupwise $W_T = +1$ and $W_T = -1$. This proves the stated consecutive-rank calculation.

Equation (3) describes the best-to-worst range for any fixed set of occupied prediction ranks; it is not the change obtained by reversing an arbitrary starting order. Nonconsecutive ranks can give a larger range. Indeed, $a_j - a_i \geq j - i$ for distinct integer ranks, so the consecutive case gives the smallest such range over occupied rank sets. The main proposition therefore demonstrates structural masking without asserting an unconditional $2q^3$ upper bound. Its 0.016 is an absolute correlation change, not a percentage decrease from a model's initial correlation. These no-tie formulas are not applied to tied data without the corresponding rank calculation.

D Safety Theorem: Conditions and Proofs

D.1 Population, selection and ranks

All results in this appendix refer to one fixed population law. Assume $E|Y| < \infty$ and define $G = E[Y \mid X]$, where X denotes the information available to the predictor before the outcome is observed. Conditioning on the true-top event changes the evaluation law, not the definition of G. The term Bayes score below refers to this conditional mean; its interpretation as a squared-loss optimum additionally requires $E[Y^2] < \infty$.

For $q \in (0, 1)$, choose a cutoff c_q such that $P(Y > c_q) \leq q \leq P(Y \geq c_q)$. When $P(Y = c_q) = 0$, put $\mathcal{T}_q = \{Y \geq c_q\}$, so that $P(\mathcal{T}_q) = q$. For a cutoff atom, an exact probability-q population target can be defined using $U_q \sim \mathrm{Unif}(0, 1)$ independent of (X, Y):

$$\lambda_q = \frac{q - P(Y > c_q)}{P(Y = c_q)}, \tag{5}$$

$$\mathcal{T}_q = \{Y > c_q\} \cup \{Y = c_q,\ U_q \leq \lambda_q\}. \tag{6}$$

Here (5) is used only when the denominator is positive; otherwise set $\lambda_q = 0$. This population construction does not change the empirical include-all-boundary-ties rule in Definition 1. If the population target instead includes every cutoff tie, take $\lambda_q = 1$ and use its actual selection probability, which may exceed q. The proof below applies to either specified convention; one statement must use one common selected law throughout.

For any variable A, define its conditional mid-distribution by

$$M_{A|\mathcal{T}_q}(a) = P(A < a \mid \mathcal{T}_q) + \tfrac{1}{2} P(A = a \mid \mathcal{T}_q). \tag{7}$$

We use $\rho_S(A, B \mid \mathcal{T}_q) = \mathrm{Corr}(M_{A|\mathcal{T}_q}(A), M_{B|\mathcal{T}_q}(B) \mid \mathcal{T}_q)$. Every correlation below requires positive variances of both conditional rank transforms. A zero variance makes the correlation undefined, rather than zero. Selection ties determine membership; midranks are computed from the original values after selection.

Use jointly measurable versions of the conditional probabilities

$$\bar{F}_Y(g, y) = P(Y > y \mid G = g), \qquad p_c(g) = P(Y = c_q \mid G = g),$$

and put

$$\pi_q(g) = \bar{F}_Y(g, c_q) + \lambda_q p_c(g), \qquad \mathcal{G}_q = \{g : \pi_q(g) > 0\}.$$

Statements concerning g are restricted to this selected support. The pairwise inequalities below hold outside one common $P_{G|\mathcal{T}_q}^{\otimes 2}$-null set, for every stated threshold.

D.2 Proof of the selected-ordering theorem

With a possible atom at the cutoff, the sufficient condition is

$$\pi_q(g_1)\bar{F}_Y(g_2, y) \geq \pi_q(g_2)\bar{F}_Y(g_1, y), \qquad g_2 > g_1, \quad y \geq c_q. \tag{8}$$

Without a cutoff atom, this is exactly (1) after setting $y = c_q + u$.

Proof of Theorem 1. For $y \geq c_q$, an outcome exceeding y is selected under either cutoff convention. Hence

$$P(Y > y \mid G = g, \mathcal{T}_q) = \frac{\bar{F}_Y(g, y)}{\pi_q(g)}.$$

Condition (8) makes this probability non-decreasing in g. For $y < c_q$ it is identically one. Consequently, the selected outcome distribution is stochastically non-decreasing in g.

Set $h(g) = M_{G|\mathcal{T}_q}(g)$, $V = M_{Y|\mathcal{T}_q}(Y)$, and $b(g) = E[V \mid G = g, \mathcal{T}_q]$. Both h and b are non-decreasing: the first by definition, the second because the outcome rank is an increasing function evaluated under the ordered conditional laws. Let G_1, G_2 be independent copies drawn from $G \mid \mathcal{T}_q$. Then

$$\begin{aligned}\mathrm{Cov}(h(G), V \mid \mathcal{T}_q) &= \mathrm{Cov}(h(G), b(G) \mid \mathcal{T}_q) \\ &= \tfrac{1}{2} E\big[\{h(G_1) - h(G_2)\}\{b(G_1) - b(G_2)\}\big] \\ &\geq 0.\end{aligned}$$

Division by the positive conditional rank standard deviations gives $W_T(G) \geq 0$. The independent copies are used only in this covariance identity; they are not an assumption that observed financial rows are independent. Equivalently, the selected stochastic ordering gives positive quadrant dependence, and the Hoeffding covariance identity (Hoeffding, 1940; Lehmann, 1966) yields the same nonnegative rank covariance.

The condition is sufficient, not necessary. Strict positivity follows if the product in the last expectation is positive with positive probability. This alone does not provide a uniform positive margin across distributions.

D.3 Fixed and deterministic varying scales

Consider $Y = G + S\varepsilon$, where G, S are measurable functions of X, $S > 0$, $\varepsilon \perp X$, $E[\varepsilon] = 0$, and $E|G| + E|S\varepsilon| < \infty$. These assumptions ensure $G = E[Y \mid X]$. We use S as a positive scale, not necessarily a conditional standard deviation; the latter interpretation needs $E[\varepsilon^2] = 1$. Assume that ε is atomless.

For a residual Z, write $\bar{F}_Z(z) = P(Z > z)$ and $H_Z(z) = -\log \bar{F}_Z(z)$, taking $H_Z = +\infty$ when the survival probability is zero. Relevant-tail increasing-failure-rate (IFR) means that H_Z is extended-real convex on an upper interval containing all entry thresholds and subsequent arguments used below. For any entry threshold with $\bar{F}_Z(t) > 0$, let

$$Q_Z(t, x) = \frac{\bar{F}_Z(t + x)}{\bar{F}_Z(t)}, \qquad x \geq 0.$$

Lemma 1 (Ordering of residual excesses)**.** *If H_Z is convex on the relevant upper interval, then $Q_Z(t_2, x) \geq Q_Z(t_1, x)$ for $t_2 \leq t_1$, $x \geq 0$, provided both entry thresholds have positive survival probability.*

Proof. For a finite cumulative hazard, $Q_Z(t, x) = \exp\{-H_Z(t + x) + H_Z(t)\}$. The increment $H_Z(t + x) - H_Z(t)$ of a convex function over a fixed step is non-decreasing in t. The claimed ordering follows. If an exit argument is beyond the support, its survival is zero; the same inequality follows by extension at that endpoint.

Corollary 1 (Sufficient scale conditions)**.** *Under the location–scale assumptions above, Theorem 1 applies in either of the following cases:*

(i) $S = \sigma > 0$ is constant and H_ε is convex on $\{(c_q - g)/\sigma + x : g \in \mathcal{G}_q,\ x \geq 0\}$.

(ii) $S = s(G)$, where $s(g)$ is non-decreasing and $t_q(g) = (c_q - g)/s(g)$ is non-increasing on $\mathcal{G}_q$, and H_ε is convex on $\{t_q(g) + x : g \in \mathcal{G}_q,\ x \geq 0\}$.

All entry thresholds are required to have positive survival probability. In each case, $W_T(G) \geq 0$.

Proof. For constant scale,

$$P(Y - c_q > u \mid G = g, \mathcal{T}_q) = Q_\varepsilon((c_q - g)/\sigma, u/\sigma).$$

The entry threshold decreases with g, so the lemma proves the ordering in (1).

For varying scale, take $g_2 > g_1$ and let $s_i = s(g_i)$, $t_i = (c_q - g_i)/s_i$. For every $u \geq 0$,

$$\begin{aligned} Q_\varepsilon(t_2, u/s_2) &\geq Q_\varepsilon(t_2, u/s_1) \\ &\geq Q_\varepsilon(t_1, u/s_1). \end{aligned}$$

The first step uses $s_2 \geq s_1$ and monotonicity of survival in its second argument. The second uses $t_2 \leq t_1$ and the lemma. The resulting selected ordering gives $W_T(G) \geq 0$.

A log-concave residual density is a familiar sufficient condition for IFR (Bagnoli and Bergstrom, 2005); the proofs require only the relevant-tail condition. Increasing $s(g)$ alone does not imply decreasing $t_q(g)$. If s is differentiable, the latter condition is equivalent to

$$s(g) + (c_q - g)s'(g) \geq 0.$$

For $g > c_q$ and $s'(g) \geq 0$, this requires $s'(g) \leq s(g)/(g - c_q)$. A positive empirical correlation between estimated signal and scale does not establish either pointwise ordering or the requirement that the scale be a deterministic function of G.

D.4 Symmetric noise with non-decreasing scale

The two orderings in Corollary 1(ii) ensure selected stochastic dominance. Symmetry gives another route to non-reversal: it is enough that a higher-signal selected observation beats a lower-signal selected observation with probability at least one half. This does not require the selected survival curves to be ordered.

Theorem 2 (Symmetric-noise safety under non-decreasing scale)**.** *Let $Y = G + s(G)\varepsilon$, where $\varepsilon \perp G$, $s(g) > 0$ is non-decreasing on the relevant support, and ε has a non-degenerate density that is symmetric about zero and log-concave. For any fixed c with $0 < P(\mathcal{T}_c) < 1$, where $\mathcal{T}_c = \{Y \geq c\}$, assume that the conditional midranks of G and Y have positive variances. Then*

$$\rho_S(G, Y \mid \mathcal{T}_c) \geq 0.$$

In particular, $W_T(G) \geq 0$ at every population top-q cutoff for which these ranks are non-degenerate. No monotonicity of $(c - g)/s(g)$ is required.

The rank assertion uses only independence from G. With $G = g(X)$, $E|G| + E|s(G)\varepsilon| < \infty$ and $E[\varepsilon \mid X] = 0$ (for example, $\varepsilon \perp X$), it is a result for the population conditional mean $G = E[Y \mid X]$. It does not by itself certify a fitted score or a finite-sample estimate.

Lemma 2 (Selected pairwise precedence and rank covariance)**.** *Fix a selected event $\mathcal{T}_c$ and write $\Pi = \mathcal{L}(G \mid \mathcal{T}_c)$, $U(g) = M_{G|\mathcal{T}_c}(g)$ and $V = M_{Y|\mathcal{T}_c}(Y)$. For independent draws Y_g, Y_h from their respective conditional laws given $(G = g, \mathcal{T}_c)$ and $(G = h, \mathcal{T}_c)$, let*

$$K(g, h) = P(Y_g > Y_h) + \tfrac{1}{2}P(Y_g = Y_h).$$

Using jointly measurable conditional versions,

$$\mathrm{Cov}(U(G), V \mid \mathcal{T}_c) = \tfrac{1}{2} \iint [U(g) - U(h)]\,[K(g, h) - \tfrac{1}{2}]\,\Pi(dg)\Pi(dh). \tag{9}$$

Consequently, $K(g, h) \geq 1/2$ for $\Pi^{\otimes 2}$-almost every pair $g > h$ implies nonnegative conditional rank covariance.

Proof. Put $m(g) = E[V \mid G = g, \mathcal{T}_c]$ and $d(g, h) = K(g, h) - 1/2$. An independent reference draw from the selected outcome distribution gives $m(g) = \int K(g, h)\Pi(dh)$. Also, $E[U(G) \mid \mathcal{T}_c] = E[V \mid \mathcal{T}_c] = 1/2$ and $d(g, h) = -d(h, g)$. Hence

$$\begin{aligned} \mathrm{Cov}(U(G), V \mid \mathcal{T}_c) &= \iint U(g)d(g, h)\,\Pi(dg)\Pi(dh) \\ &= \tfrac{1}{2} \iint [U(g) - U(h)]d(g, h)\,\Pi(dg)\Pi(dh). \end{aligned}$$

All integrands are bounded, so the interchange is valid. The last integrand is nonnegative under the stated pairwise condition, since U is non-decreasing. This proof does not assume that $m(g)$ is non-decreasing, nor that selected outcome distributions are ordered.

Proof of Theorem 2. **Step 1: reduce to two selected signal levels.** Fix selected-relevant $h < g$, put $s_L = s(h) \le s_H = s(g)$, and define

$$\theta = s_L/s_H \in (0,1], \qquad \delta = (g-h)/s_H > 0, \qquad a = (c-h)/s_L, \quad b = (c-g)/s_H.$$

Then $b = \theta a - \delta < \theta a$. Let R_t have the residual-life law $(\varepsilon - t) \mid \varepsilon \ge t$. With independent residual draws,

$$Y_h - c \overset{d}{=} s_L R_a, \qquad Y_g - c \overset{d}{=} s_H R_b, \qquad K(g,h) = P(R_b > \theta R_a).$$

Log-concavity gives the IFR ordering from Lemma 1, so $R_b \succeq_{\mathrm{st}} R_{\theta a}$ and

$$K(g,h) \ge P(R_{\theta a} > \theta R_a). \tag{10}$$

The auxiliary law at θa is defined: for $a \ge 0$ its survival probability is at least that at a, and for $a < 0$ symmetry gives $P(\varepsilon \ge \theta a) \ge 1/2$.

Step 2: a nonnegative entry threshold. If $a \ge 0$, then $\theta a \le a$ and

$$R_{\theta a} \succeq_{\mathrm{st}} R_a \succeq_{\mathrm{st}} \theta R_a.$$

For independent atomless $Z_1 \succeq_{\mathrm{st}} Z_2$, $P(Z_1 > Z_2) = E[F_{Z_2}(Z_1)] \ge E[F_{Z_2}(Z_2)] = 1/2$. Thus the right-hand side of (10) is at least $1/2$.

Step 3: a negative entry threshold. For $a < 0$, let $A = -a > 0$ and take independent noise draws x, z with density f. We must show that $z > \theta x$ is at least as probable as $z < \theta x$, conditional on

$$E_A = \{x \ge -A,\ z \ge -\theta A\}.$$

Partition the losing region into

$$L_1 = E_A \cap \{-A \le x \le A,\ z < \theta x\},$$
$$L_2 = E_A \cap \{x > A,\ z < \theta x\}.$$

Reflection $T_1(x,z) = (-x,-z)$ sends L_1 injectively into the winning region inside E_A. It has unit absolute Jacobian and preserves the density by symmetry. Its image satisfies $z' \le \theta A$.

For L_2, use $T_2(x,z) = (z/\theta, \theta x)$, also with unit absolute Jacobian. Here $z/\theta \ge -A$, $\theta x > \theta A$ and $\theta x > z$, so its image is winning and lies inside E_A. Its image has $z' > \theta A$ and is disjoint from $T_1(L_1)$.

It remains to check density. Write $f = e^{-\phi}$ with ϕ even and convex, allowing $+\infty$ off the support. On the interval where both terms are finite, the function

$$D_\theta(v) = \phi(v/\theta) - \phi(v), \qquad v \ge 0,$$

is non-decreasing: at almost every differentiability point, $D'_\theta(v) = \theta^{-1}\phi'(v/\theta) - \phi'(v) \ge 0$. Convexity gives local absolute continuity on the support interior, so this yields the same monotonicity without a smoothness assumption. For positive-density points of L_2, $|z| < \theta x$, whence

$$\begin{aligned}\phi(z/\theta) + \phi(\theta x) - \phi(z) - \phi(x) \\ = D_\theta(|z|) - D_\theta(\theta x) \le 0.\end{aligned}$$

Thus $f(z/\theta)f(\theta x) \ge f(z)f(x)$. If the symmetric support has finite endpoint r_+, then $|z|/\theta < x < r_+$ and $\theta x \le x < r_+$ at almost every positive-density source point. All displayed differences are therefore finite; the argument also covers bounded-support densities.

The two injections map all losing mass into disjoint winning regions without reducing mass. Since the boundary has zero probability, $P(R_{\theta a} > \theta R_a) \ge 1/2$ also when $a < 0$.

Step 4: combine the cells. We have proved $K(g,h) \ge 1/2$ for every relevant $g > h$. Lemma 2 now gives nonnegative conditional rank covariance. Division by the positive conditional rank standard deviations proves the theorem.

This proof establishes selected pairwise precedence, not selected first-order stochastic dominance. The selected survival curves may cross. The theorem is therefore an additional route to $W_T(G) \geq 0$, not another case claimed to satisfy (1). It supplies no uniform positive margin for the learned-score certificate.

Corollary 2 (Feature Invariance)**.** *The conclusion $W_T(G) \geq 0$ under Theorem 1 does not depend on whether the features forming g are associated with outcome magnitude, direction, or both.*

Proof. The theorem uses the selected joint law of (G, Y) and its stochastic ordering. It does not use a particular decomposition of g into features. This assertion is about the signal G, not an arbitrary score fitted to the features.

D.5 Limits of scale and noise assumptions

The following elementary calculation will be used in both this appendix and Appendix E.

Lemma 3 (Two-level signal)**.** *Suppose that G has two values $g_H > g_L$ on $\mathcal{T}_q$, with $\pi_H = P(G = g_H \mid \mathcal{T}_q) \in (0, 1)$. Let Y_H, Y_L be independent draws from their respective selected conditional outcome distributions, assumed continuous. If $\mathrm{AUC} = P(Y_H > Y_L)$, then*

$$W_T(G) = \sqrt{12\pi_H(1-\pi_H)}\,(\mathrm{AUC} - \tfrac{1}{2}). \tag{11}$$

Proof. Let V be the CDF rank of Y under the selected mixture. It is uniform on $[0, 1]$, with variance $1/12$. Conditioning on the two signal levels gives

$$E[V \mid H] = \pi_H/2 + (1-\pi_H)\mathrm{AUC}, \qquad E[V \mid L] = \pi_H(1-\mathrm{AUC}) + (1-\pi_H)/2.$$

Their difference is $\mathrm{AUC} - 1/2$. The signal midrank equals $(1-\pi_H)/2 + \mathbf{1}_H/2$, so its variance is $\pi_H(1-\pi_H)/4$ and its covariance with V is $\pi_H(1-\pi_H)(\mathrm{AUC} - 1/2)/2$. Normalizing yields the result.

Decreasing scale can produce reversal. The low-volatility anomaly (Ang et al., 2006) motivates considering a decreasing signal–scale relation. Take a standard normal ε independent of G, with $(g_H, s_H) = (0.01, 0.01)$ and $(g_L, s_L) = (-100, 100)$. Set $q = 0.2$, $a = \Phi(1)$, $b = 1 - \Phi(1)$, and $P(G = g_H) = (q-b)/(a-b)$. Then $P(Y \geq 0) = q$, making zero the top-quintile cutoff. After selection, $Y_H = 0.01(\varepsilon_H + 1)$ with $\varepsilon_H \geq -1$, and $Y_L = 100(\varepsilon_L - 1)$ with $\varepsilon_L \geq 1$. The event $Y_H > Y_L$ implies $Y_H > 0.1$ or $Y_L < 0.1$, hence

$$P(Y_H > Y_L) \leq \frac{1-\Phi(9)}{\Phi(1)} + \frac{\Phi(1.001) - \Phi(1)}{1-\Phi(1)} < 0.001525 < \tfrac{1}{2}.$$

Lemma 3 proves reversal. This is a possibility result; it does not make non-decreasing scale a necessary condition for safety.

Proposition 2 (Increasing scale with asymmetric log-concave noise)**.** *A strictly increasing deterministic scale and independent, zero-mean asymmetric log-concave noise need not give $W_T(G) \geq 0$.*

Proof. Let $B \sim \mathrm{Beta}(2, 10)$ and $\varepsilon = B - 1/6$. Its density is a translated version of $110x(1-x)^9$ on $[0, 1]$; it is continuous and log-concave, and $E[\varepsilon] = 0$. Take $\varepsilon \perp G$ and

g	-100	0	0.01
$s(g)$	0.5	1	2
$P(G = g)$	0.8	0.1	0.1

with $Y = G + s(G)\varepsilon$ and $X = G$. The lowest-signal component ends at $-100 + 5/12$, strictly below the supports of both remaining components; those two supports overlap with each other. Thus $c_q = -100 + 5/12$ is the 80% quantile, and the selected law places probability $1/2$ on each of $G = 0$ and $G = 0.01$.

In this group, $Y_L = B_L - 1/6$ and $Y_H = 2B_H - 97/300$ for independent beta draws. Writing $C_B(x) = 1 - (1-x)^{10}(1+10x)$ on $[0,1]$ gives

$$\begin{aligned}\mathrm{AUC} &= P(Y_H > Y_L) \\ &= \int_{47/600}^{347/600} C_B(2x - 47/300)\, 110x(1-x)^9\, dx \\ &\quad + \int_{347/600}^{1} 110x(1-x)^9\, dx \\ &\simeq 0.469959215 < \tfrac{1}{2}.\end{aligned}$$

Therefore $W_T(G) = \sqrt{3}(\mathrm{AUC} - 1/2) \simeq -0.0520322$. On the selected support, $t_q(0) < t_q(0.01)$, so the sufficient threshold-ordering test fails even though s increases. The noise is asymmetric; this example is outside Theorem 2.

Proposition 3 (Symmetry without log-concavity is insufficient)**.** *Symmetric, smooth, zero-mean noise and a strictly increasing deterministic scale can give $W_T(G) < 0$.*

Proof. Let k_η be a symmetric smooth density with support closure $[-\eta, \eta]$, $\eta = 0.01$, and set

$$\begin{aligned}f_\varepsilon(x) = 0.45k_\eta(x-1) + 0.45k_\eta(x+1) \\ + 0.05k_\eta(x-10) + 0.05k_\eta(x+10).\end{aligned}$$

This density is symmetric and smooth but not log-concave. Take $X = G$, $\varepsilon \perp G$, $P(G = 0) = 2/3$, $P(G = 0.1) = 1/3$, $s(0) = 1$, $s(0.1) = 2$, and $Y = G + s(G)\varepsilon$. The conditional mean is G. At $c = 1.01$, the lower-signal group is selected only from the $+10$ component, with probability 0.05, and has $Y \in [9.99, 10.01]$. The higher-signal group is selected from its $+1$ and $+10$ components, with total probability 0.50. Thus

$$P(Y \geq c) = \tfrac{2}{3}(0.05) + \tfrac{1}{3}(0.50) = 0.20.$$

Continuity and the component endpoints make c the population 80% quantile. In the selected higher-signal group, Y lies in $[2.08, 2.12]$ with probability 0.9 and in $[20.08, 20.12]$ with probability 0.1. The winning probability against the selected lower-signal group is therefore 0.1. Its selected weight is $\pi = 5/6$, so Lemma 3 gives

$$W_T(G) = \sqrt{12\pi(1-\pi)}\,(0.1 - 0.5) \simeq -0.516398.$$

This example shows that symmetry alone is insufficient. It does not make log-concavity a necessary condition for every safe noise law.

D.6 Summary of structural conditions

Table D1 summarises the sufficient conditions and counterexamples.

The reported U-shaped-scale experiment, $s(g) = 1 + \alpha(g - \bar{g})^2$ with $\alpha \in \{0, 0.5, 1, 2, 5, 10, 20\}$, gave positive W_T for the signal and the random forest. It remains a distribution-specific experimental observation, not a sign guarantee for every U-shaped scale or a proof that only decreasing scale can cause reversal.

E Independent Random Scale: Safety Conditions and a Counterexample

For $Y = G + \Sigma\varepsilon$, independence of Σ and G allows a useful reduction, but does not by itself preserve within-top ordering. The relevant noise law is that of the product $Z = \Sigma\varepsilon$. Theorem 2 instead assumes a scale fixed by G. It does not cover a non-degenerate scale mixture conditional on G, even when ε is symmetric and log-concave.

Table D1: Sufficient conditions and counterexamples for the population signal. Independence and rank non-degeneracy assumptions are those of the cited statements.

Setting	Conclusion
Constant scale, relevant-tail IFR	Non-reversal: Corollary 1(i).
Non-decreasing $s(g)$, non-increasing $t_q(g)$, relevant-tail IFR	Non-reversal: Corollary 1(ii).
Non-decreasing deterministic $s(g)$; symmetric log-concave noise	Non-reversal without threshold monotonicity: Theorem 2.
Non-decreasing $s(g)$; asymmetric log-concave noise	Reversal is possible: Proposition 2.
Non-decreasing $s(g)$; symmetry alone	Reversal is possible: Proposition 3.
Decreasing $s(g)$	Reversal is possible; see the Gaussian construction.
Independent random scale with an IFR effective residual	Non-reversal: Corollary 3.
Independent random scale alone	Reversal is possible even at ratio 2: Proposition 4.
General scale dependence	Selected ordering (8) is one sufficient test, not an exhaustive characterisation.

Corollary 3 (Safety through the effective residual)**.** *Let $G = E[Y \mid X]$, $Y = G+Z$, and assume that $Z \perp G$ is atomless. Suppose that $H_Z = -\log P(Z > \cdot)$ is convex on $\{c_q - g + u : g \in \mathcal{G}_q,\ u \geq 0\}$, with positive survival probability at all selected entry thresholds $c_q - g$. If the conditional rank variances are positive, then $W_T(G) \geq 0$. One instance is $Z = \Sigma\varepsilon$ with $\Sigma \perp G$ and $\varepsilon \perp (G, \Sigma)$, provided this effective residual satisfies the stated tail condition and $G = E[Y \mid X]$.*

Proof. For every $u \geq 0$,

$$P(Y - c_q > u \mid G = g, \mathcal{T}_q) = Q_Z(c_q - g, u).$$

The entry threshold $c_q - g$ decreases as g increases. Lemma 1 therefore gives the selected ordering, and Theorem 1 proves the claim.

The condition concerns H_Z, not just H_ε. The next example has a log-concave elementary noise density and an independent scale taking only the values 1 and 2, yet its within-top correlation is negative.

Proposition 4 (An independent-scale counterexample)**.** *There is a location–scale model with $\Sigma \perp G$, independent continuous symmetric zero-mean log-concave noise, and scale ratio 2, for which the conditional-mean signal has $W_T(G) < 0$ at $q = 0.2$.*

Proof. Let G, Σ, ε be mutually independent, with

$$P(G = 1/2) = 169/185, \qquad P(G = 0) = 16/185,$$
$$P(\Sigma = 1) = 3/4, \qquad P(\Sigma = 2) = 1/4,$$
$$f_\varepsilon(e) = (1 - |e|)\mathbf{1}\{|e| \leq 1\}.$$

Set $X = (G, \Sigma)$ and $Y = G + \Sigma\varepsilon$. The residual has mean zero, hence $E[Y \mid X] = G$. Write L for $G = 0$ and H for $G = 1/2$. For $0 \leq t \leq 1$, the residual survival is $\bar{C}(t) = (1 - t)^2/2$. At $c = 9/10$, the four entry probabilities are

	$\Sigma = 1$	$\Sigma = 2$
$G = 0$	$1/200$	$121/800$
$G = 1/2$	$9/50$	$8/25$

so that

$$p_L = P(Y > c \mid L) = 133/3200, \qquad p_H = P(Y > c \mid H) = 43/200.$$

Consequently,

$$P(Y > c) = \frac{16}{185}\frac{133}{3200} + \frac{169}{185}\frac{43}{200} = \frac{1}{5}.$$

The continuous mixture has $c_q = 0.9$ at its 80% quantile.

Selection changes the scale weights to

$$(\lambda_{H,1}, \lambda_{H,2}) = (27/43, 16/43), \qquad (\lambda_{L,1}, \lambda_{L,2}) = (12/133, 121/133).$$

For independent selected draws, define $A_{rs} = P(Y_H > Y_L \mid \Sigma_H = r, \Sigma_L = s)$. With $t_{H,r} = (c - 1/2)/r$, $t_{L,s} = c/s$, and $C(x) = P(\varepsilon \le x)$, these probabilities are obtained from

$$A_{rs} = \frac{1}{\bar{C}(t_{H,r})} \int_{t_{H,r}}^{1} (1 - e) \frac{[C((1/2 + re)/s) - C(t_{L,s})]_{[0,\bar{C}(t_{L,s})]}}{\bar{C}(t_{L,s})}\, de,$$

where $[z]_{[0,a]} = \min\{\max\{z, 0\}, a\}$. Integrating the polynomial pieces gives

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} 193/216 & 38/121 \\ 491/512 & 953/1536 \end{pmatrix}.$$

Using the selected scale weights,

$$A = \sum_{r=1}^{2} \sum_{s=1}^{2} \lambda_{H,r} \lambda_{L,s} A_{rs} \simeq 0.47225076 < \frac{1}{2}.$$

The selected high-signal share is $\pi = P(H \mid \mathcal{T}_q) = 7267/7400$. By Lemma 3,

$$W_T(G) = \sqrt{12\pi(1 - \pi)}\,(A - 1/2) \simeq -0.0127707.$$

Here the low-signal candidates that reach the top group come predominantly from the larger-scale component. Before selection, the scale was independent of the signal; after selection, its composition differs sharply between the two signal levels. This explains why scale–signal independence is not enough.

The example also defeats an entry-threshold-only alignment argument. For the high and low signals, $(c_q - g_H)/(c_q - g_L) = 4/9$, while every possible scale ratio Σ_H/Σ_L is at least $1/2$. Thus the event $\Sigma_H/\Sigma_L < (c_q - g_H)/(c_q - g_L)$ has probability zero, although $A < 1/2$. Entry-threshold alignment alone does not order the scaled excess outcomes.

Positive results in a particular bounded-scale simulation remain empirical evidence for that simulated family. They do not supply a universal bound on the admissible scale ratio, nor replace the effective-residual condition above.

F Transfer to Learned Scores

The safety theorem concerns the signal $G = E[Y \mid X]$. We now give conditions under which a fitted score $F = f(X)$ also has nonnegative within-top correlation. Throughout this appendix, the conditional midranks of F, G, Y are non-degenerate and use the same cutoff and selection convention.

Theorem 3 (Conditional Gram bound)**.** *Let $r = \rho_S(F, G \mid \mathcal{T}_q)$, $w = \rho_S(G, Y \mid \mathcal{T}_q)$ and $z = \rho_S(F, Y \mid \mathcal{T}_q)$. Then*

$$rw - \sqrt{(1 - r^2)(1 - w^2)} \le z \le rw + \sqrt{(1 - r^2)(1 - w^2)}. \tag{12}$$

If $r, w \ge 0$ and $r^2 + w^2 \ge 1$, then $z \ge 0$. The inequality is strict if $r^2 + w^2 > 1$.

Proof. Center and standardize the three conditional midrank variables. Their correlation matrix is

$$\begin{pmatrix} 1 & r & z \\ r & 1 & w \\ z & w & 1 \end{pmatrix}.$$

It is positive semidefinite, so its determinant is nonnegative:

$$1 + 2rwz - r^2 - w^2 - z^2 \ge 0.$$

Rearranging gives $(z - rw)^2 \le (1 - r^2)(1 - w^2)$, which proves (12), including boundary values of r, w. For $r, w \ge 0$, its lower endpoint is nonnegative exactly when $r^2 w^2 \ge (1 - r^2)(1 - w^2)$, or $r^2 + w^2 \ge 1$.

The three correlations are constrained by their joint rank distribution; independence between the two ranking errors is not assumed. For an observed sample, the same matrix argument applies to three midrank vectors computed on the same rows. Pairwise deletion with separate re-ranking does not provide this common matrix, and a sample identity is not a population confidence statement.

Corollary 4 (A bound from approximation error)**.** *Let a be strictly increasing, and assume that G and $a(F)$ have atomless conditional distributions on $\mathcal{T}_q$. Suppose*

$$P(|a(F) - G| \le \epsilon \mid \mathcal{T}_q) = 1, \qquad \sup_x p_{G|\mathcal{T}_q}(x) \le L.$$

Then $r \ge r_0 = 1 - 24L^2\epsilon^2$. If also $0 \le r_0 \le 1$ and $W_T(G) \ge w_0 > 0$, then

$$W_T(F) \ge r_0 w_0 - \sqrt{(1 - r_0^2)(1 - w_0^2)}. \tag{13}$$

In particular, $r_0^2 + w_0^2 \ge 1$ certifies $W_T(F) \ge 0$.

Proof. Work under $P(\cdot \mid \mathcal{T}_q)$, and put $\tilde{F} = a(F)$. Let $C_G, C_{\tilde{F}}$ be the corresponding conditional CDFs. The almost-sure error bound implies, for every x,

$$C_G(x - \epsilon) \le C_{\tilde{F}}(x) \le C_G(x + \epsilon).$$

Since $G - \epsilon \le \tilde{F} \le G + \epsilon$, evaluating at $x = \tilde{F}$ gives

$$C_G(G - 2\epsilon) \le C_{\tilde{F}}(\tilde{F}) \le C_G(G + 2\epsilon).$$

The selected density bound makes C_G L-Lipschitz, hence

$$|C_{\tilde{F}}(\tilde{F}) - C_G(G)| \le 2L\epsilon.$$

Both conditional CDF ranks are uniform on $[0, 1]$, and a does not change the ordering of F. Therefore

$$r = 1 - 6E[(C_{\tilde{F}}(\tilde{F}) - C_G(G))^2 \mid \mathcal{T}_q] \ge 1 - 24L^2\epsilon^2.$$

On $[0, 1]^2$ the Gram lower endpoint $\phi(r, w) = rw - \sqrt{(1 - r^2)(1 - w^2)}$ is non-decreasing in each argument. Substituting $r \ge r_0 \ge 0$ and $w \ge w_0 > 0$ proves (13).

Taking a to be the identity recovers the bounded-error condition on $F - G$. The density bound must hold after true-top selection; a bound for the unselected population does not substitute for it. The bound $r \ge 1 - 24L^2\epsilon^2$ is valid for any such ϵ, but its substitution into (13) requires $r_0 \in [0, 1]$. If the available signal guarantee is only $w_0 = 0$, this worst-case bound alone cannot certify non-reversal for $0 \le r_0 < 1$. This is a limitation of the certificate, not a claim that every imperfectly aligned score must reverse.

Connection to the multi-architecture results. The multi-architecture results and negative training-set W_T show that the observed pattern is not confined to one learner or to test data, but do not establish $W_T(G) < 0$ or verify the transfer conditions for every tested model. The certificate therefore provides a conditional bridge rather than a retrospective verification of the financial experiments.

G Experimental Details

G.1 Notation and computation

Spearman correlation is the Pearson correlation of the two midrank vectors. With no ties, it reduces to $\rho_S(a, b) = 1 - 6\sum_{i=1}^{n} d_i^2/[n(n^2 - 1)]$, where $d_i = \mathrm{rank}(a_i) - \mathrm{rank}(b_i)$. For W_T, both vectors are ranked within the selected true-top group. For financial data this is computed date by date and then averaged; for non-financial tabular data it is computed on the selected test group. The primary fraction is $q = 0.2$. Correlation is undefined if either rank variance vanishes.

Finite-sample selection conventions. The original non-financial evaluations include all observations tied at the empirical outcome cutoff. The financial W_T implementation selects exactly $m_d = \lfloor qN_d \rfloor$ observations per date using ordinal rank; the semi-synthetic controls use a percentile-threshold rule. Ranks within the selected group are midranks of the original values. The population boundary rule in Appendix D is a separate theoretical convention and does not replace these empirical selectors. The retrospective capture diagnostic, including its financial rows, uses an inclusive outcome cutoff (Appendix Q.3); its boundary membership can therefore differ from financial W_T. New experiment-specific masks and shortlist rules are stated with their respective experiments.

G.2 FINANCIAL MODELS

Qlib framework (Yang et al., 2020), Alpha360 features (360 dimensions). Stock universe: full Chinese A-shares (5,624 unique stocks; mean 4,836 stocks per day; 801 test days). All financial data is downloaded via Qlib's public data pipeline (Yahoo Finance source, freely available). Adam, lr 3×10^{-4} with 5-epoch linear warmup, batch 4096, max 200 epochs, patience 20 on validation-set IC. Train 2007–2020, val 2021–2022, test 2023–2026. LSTM (Hochreiter and Schmidhuber, 1997): 2 layers, hidden 64, dropout 0, 201,857 parameters; best epochs range 10–25. Labels are next-day returns, cross-sectionally standardised per day. For the 13 non-LSTM architectures, we use each architecture's default Qlib configuration with 16 loss functions (MSE, MSE+IC, Cosine, ListMLE, SoftRank, TopSort, etc.). Financial RF and XGB (tabular models) also produce $W_T < 0$, confirming the reversal is not specific to deep architectures.

Model-set summary. The paper uses nested financial model sets: 13 default-loss architectures (Table 2); 53 primary models including loss variants (Figure 2, $\rho = -0.883$); 62 single-horizon models (training-set W_T analysis); 161 standard-loss models excluding TopSort/TopListMLE; 192 total including direct-supervision and alternative-mask runs. Table 1 reports 149 A-share and 14 US primary-configuration models.

G.3 NON-FINANCIAL MODELS

For the 8 primary domains (Drug, Wine, EnergyEff., Abalone, Diabetes, Electricity, Traffic, Weather), each is tested with 51 models: Ridge ($\alpha \in \{0.01, 1, 100\}$), Lasso ($\alpha = 0.01$), ElasticNet ($\alpha = 0.1$), RF (depth $\in \{5, 10, 20\}$, $n_{\text{estimators}} \in \{100, 200\}$, seeds $\{42, 123\}$), XGBoost (depth $\in \{3, 5, 8\}$, $n_{\text{estimators}} \in \{100, 200\}$, seeds $\{42, 123\}$), MLP (hidden $(64, 32)$ and $(128, 64)$, early stopping). Wine has 50 models (one MLP failed to converge). For the 6 benchmark domains (MiamiHousing, FreeSolv, ESOL, Lipophilicity, CalHousing), each uses 15 models: Ridge $\times 3$, Lasso, ElasticNet, RF $\times 4$, XGB $\times 4$, MLP $\times 2$. Superconduct. uses 7 models (Ridge $\times 3$, RF $\times 2$, XGB $\times 2$). NBA uses 3 models. All use scikit-learn with random 80/20 split (seed 42), no validation set or hyperparameter tuning—diversity achieved by running multiple fixed configurations.

Model complexity cross-check. To rule out model class as a confound, we trained MLPs of increasing depth (1–4 hidden layers, up to 512 units) on 4 non-financial domains. All produce $W_T > 0$: CalHousing $W_T \in [+0.47, +0.59]$, Abalone $[+0.39, +0.43]$, Wine $[+0.30, +0.40]$, Diabetes $[+0.03, +0.06]$ (10 trials each). We also trained the same LSTM architecture used for financial prediction on two non-financial time series domains: Electricity ($\rho_S = +0.508$, $W_T = +0.247$) and Traffic ($\rho_S = +0.563$, $W_T = +0.040$)—both $W_T > 0$. The complete 2×2 design is shown in Table G1.

Table G1: 2×2 design: domain type $\times$ model class. Tabular models (RF, XGB) cover all 15 non-financial domains; sequential models (LSTM) cover the two non-financial time-series domains (Electricity, Traffic).

	Tabular (RF/XGB)	**Sequential (LSTM)**
Financial	$W_T < 0$	$W_T < 0$
Non-financial	$W_T > 0$	$W_T > 0$

Model class is not a confound; reversal is associated with domain structure rather than model class.

G.4 NOISE EXPERIMENTS

$\mathcal{N}(0, k\hat{\sigma}_Y)$ added to labels. Drug: 14 levels ($k \in \{0, 0.5, \ldots, 50\}$). MiamiHousing: 9 levels ($k \in \{0, 1, \ldots, 100\}$). RF retrained at each level; W_T, R^2 on uncorrupted test set. At the highest noise levels, R^2 approaches zero in both domains, yet W_T remains positive throughout (Pearson $\rho(R^2, W_T) > +0.96$ across noise levels in both domains). Drug and MiamiHousing approach $W_T = 0$ asymptotically from above, confirming that near-zero predictive signal alone does not produce reversal.

G.5 QUANTILE REGRESSION

`statsmodels.QuantReg` at $\tau = 0.1, 0.9$. Sign flip: $\text{sign}(\hat{\beta}_{0.1}) \neq \text{sign}(\hat{\beta}_{0.9})$, both $p < 0.05$. Finance: Corr $= -0.747$, 7/42 flips. All 9 non-financial domains: zero flips.

G.6 DIABETES PANEL-MEAN UNCERTAINTY

We assess the Diabetes panel mean using 10,000 joint paired-bootstrap draws over the shared held-out rows. Each draw recomputes the true-top cutoff, membership, ranks, model-level W_T values and their panel mean, using the same resampled rows for all models. The reported result is

$$\widehat{W_{T\,\text{panel}}} = +0.0820, \qquad 95\%\ \text{CI} = [-0.2783, +0.7435].$$

The interval crosses zero, leaving the panel-mean sign unresolved. It is not an interval for the coupling between overall Spearman and W_T, nor for the across-model median. The original panel also contains 10 negative-W_T models out of 51; uncertainty in the mean does not eliminate that model-level disagreement.

Across-split coupling stability. Over 50 random 80/20 splits, the Spearman correlation between ρ_S and W_T has mean $+0.24$ and ranges from -0.83 to $+0.88$. No single standard metric correlates with W_T above $|\rho| = 0.3$ in the primary split. Models selected by higher W_T produce better top-group outcomes in 80% of splits (mean gap $+0.15\sigma$), though the effect does not reach significance given 18 true-top observations per split.

The 50-split coupling summary and the three-seed objective comparison concern different evaluation quantities and should not be read as estimates of this same uncertainty.

H DE-VOLATILISATION ACROSS DOMAINS

Table H1 reports the effect of de-volatilising labels across six non-financial domains.

Table H1: De-volatilisation across six non-financial domains: $|\Delta W_T| \leq 0.09$ in five of six domains (Abalone $+0.147$); all six remain positive.

Domain	$W_{T,\text{std}}$	$W_{T,\text{devol}}$	Δ
CalHousing	$+0.537$	$+0.519$	-0.018
EnergyEff.	$+0.846$	$+0.761$	-0.085
Abalone	$+0.316$	$+0.463$	$+0.147$
Diabetes	$+0.151$	$+0.125$	-0.026
Electricity	$+0.508$	$+0.504$	-0.004
Traffic	$+0.365$	$+0.318$	-0.047

All six maintain $W_T > 0$ after de-volatilisation. In finance, $X/\hat{\sigma}$ worsens W_T from -0.117 to -0.196.

I TIME STABILITY OF WITHIN-TOP REVERSAL

To rule out market-regime dependence, we compute yearly W_T across 20 primary financial models (Table I1).

Table I1: Yearly W_T across 20 primary financial models (A-shares, 2023–2026). Reversal is present in every year.

Year	n_{models}	**Mean** W_T	**%** $W_T < 0$	**Mean** ρ_S
2023	20	-0.146	95%	$+0.040$
2024	20	-0.121	90%	$+0.053$
2025	20	-0.137	90%	$+0.062$
2026	20	-0.130	95%	$+0.053$

The reversal is present in every year (90–95% of models), spanning both rising and falling markets. Mean W_T varies by only 0.025 across four years (-0.146 to -0.121), supporting that the reversal persists across the tested years and is not specific to a single market period

J Feature-Level Signal Decomposition

Drug (AChE, 208 descriptors): mean magnitude/direction ratio $\bar{r} = 1.8$; 34% features have $r < 1$ (direction-dominant). Finance (Alpha360, 42 features at lag 1): $\bar{r} = 4.2$; zero features have $r < 1$. The two strongest financial features (`high/close`, `low/close`) have magnitude signal 3–6$\times$ direction signal (Parkinson, 1980).

J.1 Directional strength does not determine the sign

Fix the outcome origin and feature dictionary used to define $S = \max_j |\rho(X_j, \text{sign}(Y))|$. The constructions below use Pearson correlation for S; they give the same values with midrank Spearman correlation because direction is independent of the continuous feature and the additional direction feature is balanced binary. They concern the population signal $G = E[Y \mid X]$, not the empirical sign of an arbitrary fitted model.

Proposition 5 (Limits of a directional-signal threshold)**.** *For any $0 < q < 1/2$, there are two atomless outcome laws with the same joint law of features and outcome direction, and with $S = 0$, whose conditional-mean signals have $W_T = +1$ and $W_T = -1$, respectively. Moreover, for every $t \in [0, 1)$, there is an atomless outcome law with $S > t$ and $W_T(G) = -1$.*

Proof. Let $X \sim \text{Unif}(0, 1)$ and let $D \in \{-1, +1\}$ be independent of X, with both signs having probability $1/2$. Define

$$Y_+ = \begin{cases} 2 + X, & D = +1, \\ -(2 - X), & D = -1, \end{cases} \qquad Y_- = \begin{cases} 2 + X, & D = +1, \\ -(2 + 3X), & D = -1. \end{cases}$$

Both outcome laws are atomless and have finite second moments. In both, $\text{sign}(Y) = D$, so the joint feature–direction law is identical and $S = 0$. Their conditional means are

$$G_+(X) = X, \qquad G_-(X) = -X.$$

For either law, the true-top event is $\mathcal{T}_q = \{D = +1,\ X > 1 - 2q\}$, up to a null boundary. On this event $Y_+ = Y_- = 2 + X$, hence $W_T(G_+) = +1$ and $W_T(G_-) = -1$.

For the second claim, choose $s \in (t, 1)$ and independently draw $B \in \{-1, +1\}$ with $P(B = +1) = (1 + s)/2$. Add the feature $Z = DB$ and use the feature vector (X, Z). Then $\rho(X, D) = 0$, $\rho(Z, D) = s$, and $p_z = P(D = +1 \mid Z = z) = (1 + zs)/2$. Put

$$m_z(x) = \frac{p_z(2 + x) + x}{1 - p_z} > 0, \qquad Y_s = \begin{cases} 2 + X, & D = +1, \\ -m_Z(X), & D = -1. \end{cases}$$

Since $s < 1$, these outcomes have finite second moments and an atomless law. Their direction is again D, while

$$G_s(X, Z) = p_Z(2 + X) - (1 - p_Z)m_Z(X) = -X.$$

The true-top event is still $\{D = +1,\ X > 1 - 2q\}$, where the outcome increases in X and the signal decreases. Thus $S = s > t$ but $W_T(G_s) = -1$.

The first construction preserves even the full joint feature–direction law, not just one correlation. The second shows that high directional strength is not a sufficient safety certificate. Conversely, the first law is non-reversed with $S = 0$, so high strength is not necessary either. These facts do not dispute the observed domain gap; they rule out interpreting that gap as a universal sign rule.

The empirical statistic also depends on the feature dictionary, outcome origin and correlation convention. These must be specified before comparisons. If direction or every eligible feature is constant, S is undefined, not zero. The cross-domain unit of comparison is a domain, not each of its model runs.

K Drug Feature Stratification

Stratifying 208 descriptors by ratio $r_j = |\rho(X_j, |Y|)|/|\rho(X_j, \text{sign}(Y))|$ (Table K1).

Table K1: Drug feature stratification: W_T by molecular feature family.

Group	**Ratio**	n_{feat}	ρ_S	W_T
Extreme magnitude	$r > 5$	17	$+0.738$	$+0.399$
High ratio	$r > 2$	29	$+0.763$	$+0.487$
Mixed	0.5–2	76	$+0.800$	$+0.528$
Direction-dom.	$r < 0.5$	103	$+0.792$	$+0.539$

Strictly monotonic: $+0.40 \rightarrow +0.54$. Even extreme-magnitude features produce $W_T = +0.40$—magnitude and direction are coupled in molecular descriptors, so models that learn magnitude implicitly capture direction.

L Encoding and Stacking Experiments

L.1 Targeted encoding strategies

Six encoding strategies designed to strengthen directional signal were tested across 5 non-financial domains (5 trials each). Values are ΔW_T vs baseline (Table L1).

Table L1: Targeted encoding strategies: ΔW_T across five non-financial domains under negative-only and absolute-value transformations.

Strategy	**Drug**	**CalH.**	**Abal.**	**Diab.**	**Wine**
Dir-weighted	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
Coupling D×M	-0.001	-0.001	-0.007	-0.050	$+0.009$
Orthogonalise ($-$mag)	-0.084	-0.031	$+0.049$	-0.116	—
Rank-encoded	-0.031	-0.017	≈ 0	-0.013	—
Dir-boosted	$+0.001$	≈ 0	$+0.002$	-0.004	—

No strategy achieves consistent dual improvement ($\Delta\rho_S > 0$ and $\Delta W_T > 0$) across domains. Random Forest already captures nonlinear relationships via tree splits; manually adding these features provides information the model already has.

L.2 Stacking with coupled-feature anchor

Stacking uses coupled-feature predictions as an additional input. Tested across 5 domains, 30 trials each (Table L2).

Stacking achieves significant dual improvement only in CalHousing (3/3 model families, 100% win rate). The method requires sufficient coupled features as anchor signal—a condition not universally met.

Table L2: Stacking with coupled-feature anchor: W_T when combining financial-style and non-financial-style features across domains.

Domain	Coupled %	Model	$\Delta\rho_S$ [CI]	ΔW_T [CI]	Win%
CalHousing	62%	RF	$+0.010$ $[+0.010, +0.011]$	$+0.061$ $[+0.057, +0.066]$	100%
CalHousing	62%	XGB	$+0.009$ $[+0.008, +0.009]$	$+0.020$ $[+0.017, +0.024]$	100%
CalHousing	62%	MLP	$+0.049$ $[+0.048, +0.051]$	$+0.051$ $[+0.045, +0.058]$	100%
Abalone	86%	RF	-0.003	$+0.007$	23%
Drug	22%	RF	-0.033	-0.038	0%
Wine	45%	RF	-0.007	-0.014	7%
Diabetes	30%	RF	-0.007	-0.011	23%

M Financial Residual Diagnostics

Per-stock $\hat{g}_s = \bar{Y}_s$ and $\hat{\sigma}_s = \mathrm{std}(Y_s)$ are computed over the test period ($n = 5{,}525$ stocks, at least 30 observations each). Spearman $\rho = +0.266$ ($p < 10^{-89}$) indicates a positive association (Figure M1). It does not establish a non-decreasing deterministic scale $s(G)$, the relevant noise assumptions, or the selected-ordering condition. The observed reversal is analysed in Section 4.3.

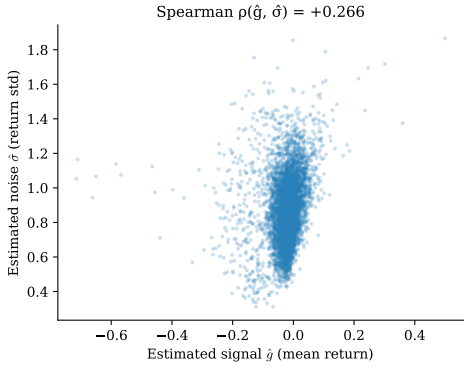


Figure M1: Per-stock $\hat{g}$ vs. $\hat{\sigma}$ across 5,525 stocks. Spearman $\rho = +0.266$.

M.1 Residual independence test

Under homoscedastic additive structure ($\sigma(X) = \text{const}$), residual magnitude $|\hat{\varepsilon}|$ should not depend on prediction magnitude $|\hat{y}|$. Under heteroscedastic additive structure, $|\hat{\varepsilon}| \approx \sigma(X)|\varepsilon|$ can depend on X through σ, but the pattern of dependence differs from the non-additive case. We compute $\rho(|\hat{\varepsilon}|, |\hat{y}|)$ where $\hat{\varepsilon} = Y - \hat{y}$ (Table M1).

Table M1: Residual independence test: $\rho(|\hat{\varepsilon}|, |\hat{y}|)$ across domains.

Domain	$\rho(\lvert\hat{\varepsilon}\rvert, \lvert\hat{y}\rvert)$	Interpretation
Financial (daily mean)	$+0.133 \pm 0.041$	Residual structure depends on $\lvert\hat{y}\rvert$
% days > 0	100%	Consistent across all trading days
CalHousing	-0.088	≈ 0: consistent with additive
Drug (AChE)	-0.021	≈ 0: consistent with additive

The financial result ($\rho = +0.133$, positive on 100% of trading days) indicates that residual variance systematically increases with prediction magnitude—a pattern consistent with non-additive structure but not conclusively ruling out heteroscedastic additive structure. The near-zero values in non-financial domains indicate no systematic relationship between residual and prediction magnitude, consistent with additive structure. Combined with the feature-invariance result (Section 4.3), the synthetic DGP experiment (Appendix P), and quantile regression sign flips (Section 4.4), these

patterns are consistent with the direction–magnitude mechanism of Section 4.4, without uniquely identifying the data-generating process.

N Within-Top Reversal: Sign Characterisation and a Sufficient Mechanism

N.1 Score-cut characterisation

Fix a score $F = f(X)$ and an evaluation event A of positive probability. Use the same conditional law to define midranks $U_A = M_{F|A}(F)$ and $V_A = M_{Y|A}(Y)$, where $M_{Z|A}(z) = P(Z < z \mid A) + \frac{1}{2}P(Z = z \mid A)$. Assume both rank variances are positive. For $t \in [0, 1]$, split this population at $U_A > t$ and put $p_A(t) = P(U_A > t \mid A)$. Let Y_H, Y_L be independent draws from the high- and low-score groups, and set

$$\mathrm{AUC}_A(t) = P(Y_H > Y_L) + \tfrac{1}{2}P(Y_H = Y_L).$$

When a group has zero mass, its contribution below is zero.

Proposition 6 (Score-cut sign identity)**.** *Under these conventions,*

$$\begin{aligned} J_A(F) &= \int_0^1 p_A(t)\{1 - p_A(t)\}[\mathrm{AUC}_A(t) - \tfrac{1}{2}]\, dt \\ &= \mathrm{Cov}(U_A, V_A \mid A), \\ \operatorname{sign} \rho_S(F, Y \mid A) &= \operatorname{sign} J_A(F). \end{aligned} \tag{14}$$

Proof. The identity $U_A = \int_0^1 \mathbf{1}\{U_A > t\}\, dt$ gives $\mathrm{Cov}(U_A, V_A \mid A) = \int_0^1 \mathrm{Cov}(\mathbf{1}\{U_A > t\}, V_A \mid A)\, dt$. For a fixed cut, write $p = p_A(t)$ and $\theta = \mathrm{AUC}_A(t)$. The two groups' mean pooled outcome midranks are $p/2 + (1 - p)\theta$ and $p(1 - \theta) + (1 - p)/2$, respectively. Their difference is $\theta - 1/2$, so the integrand is $p(1 - p)(\theta - 1/2)$. Division by positive rank standard deviations preserves the sign.

Let $\mathcal{B}_q$ be the population bottom-q event corresponding to $\mathcal{B}$ in Appendix P, and write $W_B(F) = \rho_S(F, Y \mid \mathcal{B}_q)$. Apply the identity separately to the full population Ω, the true-top event $\mathcal{T}_q$, and the true-bottom event $\mathcal{B}_q$. The pattern $(\rho_S > 0, W_T < 0, W_B > 0)$ is equivalent to $(J_\Omega > 0, J_{\mathcal{T}_q} < 0, J_{\mathcal{B}_q} > 0)$. These are exact sign conditions, not a unique explanation of the DGP. For date-wise summaries, apply the identity within each date before averaging its normalised correlations; a pooled calculation is a different target. This identity is also the mathematical basis for retrospective score-cut audit, not independent mechanism evidence.

N.2 A direction–magnitude mechanism

Consider $Y = D\widetilde{M}$, $\widetilde{M} > 0$, $E[\widetilde{M}^2] < \infty$, and $D \perp (X, M)$ with $P(D = +1) = p < 1/2$. Write $\mu(X) = E[\widetilde{M} \mid X]$ and assume that $\widetilde{M}$ and $\mu(X)$ are atomless, with the rank variances below nonzero. Then

$$G = E[Y \mid X] = (2p - 1)\mu(X).$$

For $0 < q < p$, define $m_T = Q_{1-q/p}(M)$ and $m_B = Q_{1-q/(1-p)}(M)$, and write

$$\begin{aligned} \kappa &= \rho_S(\mu(X), M), \\ \kappa_T &= \rho_S(\mu(X), \widetilde{M} \mid \widetilde{M} \geq m_T), \\ \kappa_B &= \rho_S(\mu(X), \widetilde{M} \mid \widetilde{M} \geq m_B). \end{aligned}$$

Proposition 7 (Direction–magnitude reversal)**.** *Under the conditions above,*

$$\rho_S(G, Y) = (1 - 2p)\kappa, \qquad W_T(G) = -\kappa_T, \qquad W_B(G) = \kappa_B. \tag{15}$$

Thus $\kappa, \kappa_T, \kappa_B > 0$ give positive aggregate correlation, reversed true-top ordering, and positive true-bottom ordering.

Proof. Let $U = F_\mu(\mu(X))$ and $V = F_M(M)$, both uniform ranks. The rank of Y is $(1-p)(1-V)$ on $D=-1$ and $(1-p)+pV$ on $D=+1$. Independence yields $E[F_Y(Y) \mid U, V] = 1-p+(2p-1)V$. As $2p-1<0$, the rank of G is $1-U$, giving $\rho_S(G,Y) = (1-2p)\kappa$. The selected events are $\mathcal{T}_q = \{D=+1, \widetilde{M} \geq m_T\}$ and $\mathcal{B}_q = \{D=-1, \widetilde{M} \geq m_B\}$. Within the first, $Y = \widetilde{M}$ while G decreases with μ; within the second, $Y = -M$ and G both have reversed magnitude order. Rank invariance therefore gives $W_T(G) = -\kappa_T$ and $W_B(G) = \kappa_B$.

Here higher predictable magnitude lowers the signed mean score, yet realised winners have large positive magnitude. This explains the ordering conflict without a signal-estimation error. The residual interpolation in Appendix O connects a Gaussian residual to this family at fixed signal and SNR. Low SNR, learnable magnitude and weak covariate-dependent direction information motivate the diagnostics; they are not individually necessary conditions for every reversal. Other noise structures can also yield $W_T < 0$, so the sign alone does not identify this mechanism.

O Matched Controls and Market Interventions

O.1 Known signal fraction and scale coupling

Cross-domain comparisons vary several factors at once. We therefore use a fixed standardised signal $g(X)$ and a fixed positive scale component $\sigma(X)$, with Gaussian draws $\varepsilon \sim N(0,1)$ independent of X. Let ν be the signal-variance fraction and η the scale-coupling parameter:

$$\begin{aligned} Y_{\eta,\nu} &= \sqrt{\nu}\, g(X) + \sqrt{1-\nu}\, s_\eta(X)\varepsilon, \\ s_\eta^2(X) &\propto (1-\eta) + \eta\sigma(X)^2, \qquad E[s_\eta^2(X)] = 1. \end{aligned} \tag{16}$$

The corresponding SNR is $\nu/(1-\nu)$, not ν itself. The comparisons share the feature inputs, split, components and noise draws. The grid evaluates $\nu \in \{0.01, 0.05, 0.20\}$ and $\eta \in \{0, 0.5, 1\}$ with LSTM, MLP and Transformer models. All 27 runs have positive W_T. The 27 runs comprise 3 signal fractions ($\nu \in \{0.01, 0.05, 0.20\}$) $\times$ 3 scale couplings ($\eta \in \{0, 0.5, 1\}$) $\times$ 3 architectures (LSTM, MLP, Transformer), each with a single seed; Table O1 reports the LSTM means.

Table O1: Matched SNR/scale control: mean W_T over three LSTM seeds. The reported training-seed SD is at most 0.002; it is not a sampling-confidence interval. SNR values are rounded.

ν	SNR	$\eta=0$	$\eta=0.5$	$\eta=1$
0.01	0.010	0.039	0.101	0.128
0.05	0.053	0.099	0.189	0.228
0.20	0.250	0.247	0.352	0.411

These are results for the tested family. In particular, the positive learned-score values in the feature-dependent-scale cells are not certificates supplied by the signal-level safety theorem.

O.2 Residual-only interpolation

Let $D \in \{-1,+1\}$, $P(D=+1) = p < 1/2$, and $\widetilde{M} = \mu(X)V > 0$, where $\mu(X) > 0$, $V \perp X$ and $E[V] = 1$. Assume $D \perp (X,V)$ and $\varepsilon \sim N(0,1) \perp (X,D,V)$. With E_{tr} denoting expectation under the training distribution, define the signal and direction–magnitude residual by

$$\begin{aligned} \delta &= 2p-1, & \bar{\mu}_{\mathrm{tr}} &= E_{\mathrm{tr}}[\mu(X)], \\ G_{\mathrm{raw}} &= \delta(\mu(X) - \bar{\mu}_{\mathrm{tr}}), & R_{\mathrm{DM}} &= D\widetilde{M} - \delta\mu(X), \\ a^2 &= E_{\mathrm{tr}}[G_{\mathrm{raw}}^2], & b^2 &= E_{\mathrm{tr}}[R_{\mathrm{DM}}^2], \end{aligned} \tag{17}$$

$$G_0 = G_{\mathrm{raw}}/a, \qquad \widetilde{R} = R_{\mathrm{DM}}/b, \qquad \gamma = \frac{a^2}{a^2+b^2}. \tag{18}$$

The normalising moments are finite and positive. We fix G_0, $\gamma \simeq 0.01$, features, split and training pipeline, and vary only the residual law through $\lambda \in [0, 1]$:

$$Y_\lambda = \sqrt{\gamma}\, G_0 + \sqrt{1-\gamma}\left[\cos\left(\frac{\pi\lambda}{2}\right)\varepsilon + \sin\left(\frac{\pi\lambda}{2}\right)\widetilde{R}\right]. \tag{19}$$

Since $E[\widetilde{M} \mid X] = \mu(X)$, $E[R_{\mathrm{DM}} \mid X] = 0$. Consequently,

$$E[Y_\lambda \mid X] = \sqrt{\gamma}\, G_0 \quad \text{for every } \lambda. \tag{20}$$

The normalised residual components have zero mean, unit second moment under E_{tr}, and zero cross-moment by independence. Thus their sine–cosine combination also has unit second moment, preserving the training-law signal fraction γ and SNR $\gamma/(1-\gamma)$. At the endpoint,

$$Y_1 = \frac{D\widetilde{M} - \delta\bar{\mu}_{\mathrm{tr}}}{\sqrt{a^2 + b^2}},$$

which is an increasing affine standardisation of the original direction–magnitude outcome.

Table O2: Three residual-interpolation anchors. The learned scores use the seed-42 LSTM pipeline. The signal-score columns correspond to the known conditional mean. Values are the reported evaluation summaries.

λ	ρ_S^{Bayes}	W_T^{Bayes}	$\rho_S^{\mathrm{learned}}$	W_T^{learned}
0	+0.0471	+0.0135	+0.0467	+0.0130
0.5	+0.0398	-0.2634	+0.0384	-0.2607
1	+0.1100	-0.9677	+0.1007	-0.8935

The reported 21-point signal-score path retains positive overall Spearman while W_T changes sign; Table O2 records the three numerical anchors also evaluated with the learned model. This is a controlled change in the outcome-generating process, with a separate fitted model at each learned anchor, rather than a comparison of scores on one fixed realised top set. The results establish sufficiency within this family, not the unique mechanism of market data.

O.3 Retrospective direction–magnitude interventions

The seed-42 diagnostics in the two equity markets find weak direction prediction but learnable magnitude. Pairing learned magnitude with realised direction yields positive W_T in both markets; pairing realised direction with shuffled magnitude removes that recovery. Table O3 summarises these diagnostics.

Table O3: Retrospective direction–magnitude diagnostics in the two equity markets (seed 42).

Diagnostic	Two-market values
Direction-prediction AUC	$0.517/0.523$
Magnitude-prediction Spearman	$0.344/0.412$
W_T: learned magnitude with realised direction	$+0.310/+0.440$
W_T: shuffled magnitude with realised direction	$0.000/0.002$

Realised direction is used only for this diagnostic intervention, not as an input available at prediction time. These one-seed comparisons support the direction–magnitude interpretation without identifying it as the only possible source of reversal.

P Rotation Experiment and Forward Control

Drug Discovery (coupled): Rotating from pure direction ($\theta = 0°$) to pure magnitude ($\theta = 90°$) features: $W_T \in [+0.34, +0.44]$, no zero crossing. W_T is highest at pure magnitude ($+0.44$), indicating that in this coupled domain magnitude features also carry directional information.

Synthetic additive DGP (decoupled, $g \perp \sigma$): $W_T > 0$ for all θ. Peak at $\theta = 45°$ ($W_T = +0.18$), not at pure magnitude ($W_T = +0.04$). No tested additive configuration produces $W_T < 0$. Table P1 reports the full results.

Here $W_B \coloneqq \rho_S(\hat{y}, Y)_{\mathcal{B}}$ denotes the analogous metric on the bottom q-fraction $\mathcal{B}$.

Table P1: Rotation experiment: W_T and W_B across rotation angle θ in drug and synthetic DGPs.

θ	ρ_S	W_T	W_B
Drug Discovery			
$0°$ (pure direction)	$+0.539$	$+0.355$	$+0.176$
$45°$ (equal mix)	$+0.612$	$+0.353$	$+0.329$
$90°$ (pure magnitude)	$+0.680$	$+0.441$	$+0.418$
Synthetic additive DGP			
$0°$ (pure direction)	$+0.181$	$+0.019$	$+0.003$
$25°$	$+0.173$	$+0.115$	-0.112
$45°$ (equal mix)	$+0.162$	$+0.177$	-0.151
$65°$	$+0.129$	$+0.173$	-0.165
$90°$ (pure magnitude)	-0.011	$+0.035$	-0.081

P.1 Forward control

Injecting synthetic inverse-magnitude feature $X_{\text{fake}} = -|Y_{\text{train}}| \cdot s + \varepsilon$ into drug data at increasing strengths s. W_T decreases monotonically, providing evidence that direction–magnitude decoupling is associated with reversal.

Q Diagnostic Analysis: Metrics, Coupling, and Decision Impact

Q.1 Within-model-family ρ_S–W_T relationship

We examine coupling within model families using 21 configurations per domain: 7 Ridge, 8 Random Forest and 6 MLP configurations. The reported coefficients summarise cross-configuration association within each family and in the combined grid (Table Q1).

Table Q1: Descriptive ρ_S–W_T association across the tested configurations within each model family. Each coefficient is computed across model configurations, not across observations or loss functions.

Domain	**Overall ρ**	**Ridge**	**RF**	**MLP**
CalHousing	$+0.92$	$+0.99$	$+0.98$	$+0.94$
Abalone	$+0.51$	$+0.88$	$+0.57$	$+0.49$
Diabetes	$+0.07$	$+0.24$	-0.54	-0.09

CalHousing has strong positive point estimates in all three families. In Abalone, the positive association is strongest for Ridge and weaker for Random Forest and MLP. Diabetes is weakly associated overall in this grid ($+0.07$), with a negative Random Forest estimate (-0.54). These describe the tested configurations on shared data; they do not by themselves identify model flexibility or tree depth as the mechanism.

The observed ranges are 0.088 for ρ_S and 0.250 for W_T in Abalone, and 0.082 and 0.177 in Diabetes. On their common correlation scale, the W_T range is respectively $2.8\times$ and $2.2\times$ as wide. These are absolute range ratios, not variation relative to the means.

Q.2 Standard metrics vs. W_T: cross-domain comparison

Within each domain, we compare W_T across models with the higher-is-better scores R^2, $-\mathrm{RMSE}$, $-\mathrm{MAE}$, ρ_S and EF@20%. The outer coefficient is Spearman correlation across models; it is distinct from each model's observation-level Spearman score. The non-financial panels use 90 models per domain; the financial summary uses the reported primary-model panel (Table Q2).

Table Q2: Spearman associations with W_T across model–split combinations (9 configurations $\times$ 10 random splits per non-financial domain). RMSE and MAE are sign-reversed so that higher values mean better standard performance. Because each split has different outcome variance, R^2 and $-\mathrm{RMSE}$ need not coincide.

Domain	R^2	$-\mathrm{RMSE}$	$-\mathrm{MAE}$	ρ_S	EF@20%
CalHousing	$+0.94$	$+0.91$	$+0.88$	$+0.91$	$+0.90$
Drug	$+0.92$	$+0.90$	$+0.91$	$+0.92$	$+0.85$
Wine	$+0.32$	$+0.43$	$+0.45$	$+0.31$	$+0.32$
Abalone	$+0.65$	$+0.49$	$+0.58$	$+0.28$	$+0.58$
Diabetes	-0.13	$+0.24$	$+0.13$	-0.28	-0.28
Financial	$+0.30$	$+0.32$	$+0.43$	-0.78	$+0.85$

CalHousing and Drug show strong coupling on these model grids, not exact redundancy. No displayed absolute association exceeds 0.30 in Diabetes. In finance, EF@20% is positively aligned with W_T ($+0.85$), whereas overall Spearman is negatively aligned (-0.78). Thus, higher-ρ_S models tend to have poorer within-top ordering in this panel; neither coefficient alone identifies the worst selected model or proves that EF is a better future selection rule. Across the six displayed domains, no individual standard metric has a uniformly strong marginal association with W_T. This does not test what all five metrics could predict jointly. The held-out objective comparison in Appendix R separately shows disagreement even when every W_T estimate is positive.

Q.3 Decision impact of W_T

For this diagnostic, let $\mathcal{T} = \{i : Y_i \geq \widehat{Q}_{1-q}(Y)\}$ be the realised true-top quintile, including all cutoff ties. We select its $k = \min\{|\mathcal{T}|, \max\{3, \lfloor 0.25|\mathcal{T}| \rfloor\}\}$ highest-scored observations and compute retrospective capture as

$$\mathrm{Capture} = \frac{\bar{Y}_{\mathrm{scored}} - \bar{Y}_{\mathrm{random}}}{\bar{Y}_{\mathrm{oracle}} - \bar{Y}_{\mathrm{random}}}.$$

The oracle selects the k largest realised outcomes in $\mathcal{T}$, and $\bar{Y}_{\mathrm{random}} = \bar{Y}_{\mathcal{T}}$ is the expected mean of a uniformly random k-subset. The statistic is undefined if the denominator vanishes. It equals one for oracle allocation and zero for the expected-random baseline; negative values are possible. This is not a deployable predicted-top retrieval rule.

Financial capture also uses this inclusive cutoff, unlike the exact-m_d financial W_T selector (Appendix G); their boundary-tie memberships may differ. Table Q3 reports the results.

Table Q3: Reported model-level Spearman associations between W_T and retrospective within-true-top capture across nine domains.

Domain	$\rho_S(W_T, \mathrm{Capture})$
Wine	$+0.94$
Financial	$\mathbf{+0.96}$
CalHousing	$+0.91$
Superconduct.	$+0.91$
MiamiHousing	$+0.89$
Drug	$+0.84$
EnergyEff.	$+0.78$
Diabetes	$+0.63$
Abalone	$+0.49$

The reported associations range from $+0.49$ to $+0.96$ (mean $+0.82$): higher-W_T models tend to have higher retrospective capture. These associations do not establish a future selection advantage; the chronological experiments in Section 5.3 evaluate that separate question.

Q.4 Retrospective model-level audit

The audit evaluates frozen scores on matured outcomes rather than predicting a model's sign from the domain statistic S. Before inspecting the audit outcomes, specify the candidate scores, primary top fraction, any sensitivity grid, outcome definition, selection/tie rule and aggregation. Keep these choices common across models being compared. With selective feedback, use the identified target and weighting conditions in Appendix V; without full-population identification, name the restricted target.

Point estimates and evaluation support. For each model and top fraction, recompute the realised true-top group and both sets of ranks within it. Record $\widehat{W_T}(q)$ and the number of selected observations. In a panel, also report the evaluated dates and the distribution of date-wise top-group sizes; these counts are not a claim that all rows or dates are independent. Report constant ranks or other undefined cases explicitly, using a prespecified common rule rather than counting them as evidence of non-reversal.

Score-cut interpretation. Proposition 6 decomposes the selected rank covariance into weighted score-cut AUC contrasts. Within the true-top group, a cut with $\mathrm{AUC}_{\mathcal{T}_q}(t) < 1/2$ contributes negatively: the higher-scored side tends to have lower outcomes. Reversal occurs exactly when the weighted deficits exceed the surpluses. This is an exact description of the rank-sign calculation, not independent evidence identifying a data-generating mechanism. No cross-domain cutoff on S is required; the outcome fraction q and the score cuts still remain part of the calculation. For date-averaged correlations, apply the identity on each date and normalise by that date's rank standard deviations before averaging. The sign of an unnormalised pooled integral is not a substitute.

Uncertainty and model comparison. Use common resampled rows for models evaluated on a shared independent cohort, or a joint dependence-aware procedure for panel/date data. Where resampling is used, recompute selection and ranks in each replicate, retaining the same paired rows or date blocks across models. State whether the target is a model-specific value, a panel mean or median, or a date-averaged value; do not attach a mean interval to a median. An interval crossing zero leaves the corresponding sign unresolved, rather than establishing non-reversal. Also report how candidate-model ordering under W_T differs from ordering under aggregate metrics on the same candidate set. Claims across several models or top fractions require the corresponding joint or multiplicity-aware uncertainty, rather than treating each point estimate as a separate sign certificate.

The audit concerns the completed evaluation cohort. Carrying its model choice forward requires the feedback and stability conditions of Section 5.3; neither a point estimate nor the score-cut identity alone establishes future performance.

R Objective Comparison outside Finance

To test whether W_T distinguishes models when no reversal occurs, we compare MSE, pairwise ranking, ListMLE and TopSort on California Housing and Diabetes. The two datasets, four objectives and three prespecified seeds give 24 runs. Within each dataset, the split, MLP architecture, optimiser, training budget and validation-only checkpoint rule are fixed; only the objective changes. Completed training outcomes provide supervision, validation selects checkpoints, and test outcomes are used only for final evaluation at $q = 0.20$.

Table R1: California Housing: held-out objective comparison at $q = 0.20$. Values are means $\pm$ training-seed SD over three seeds; SD does not measure held-out sampling uncertainty.

Objective	n	Overall ρ_S	W_T	W_T signs $+/-$
MSE	3	$+0.8957 \pm 0.0020$	$+0.5860 \pm 0.0027$	3/0
Pairwise	3	$+0.8979 \pm 0.0018$	$+0.5592 \pm 0.0020$	3/0
ListMLE	3	$+0.8846 \pm 0.0028$	$+0.5254 \pm 0.0079$	3/0
TopSort	3	$+0.8850 \pm 0.0014$	$+0.6014 \pm 0.0072$	3/0

Table R2: Diabetes: held-out objective comparison at $q = 0.20$. Values are means ± training-seed SD over three seeds; SD does not measure held-out sampling uncertainty.

Objective	n	Overall ρ_S	W_T	W_T signs $+/-$
MSE	3	$+0.6725 \pm 0.0257$	$+0.1387 \pm 0.0310$	3/0
Pairwise	3	$+0.6666 \pm 0.0205$	$+0.1821 \pm 0.0455$	3/0
ListMLE	3	$+0.6749 \pm 0.0206$	$+0.1094 \pm 0.0176$	3/0
TopSort	3	$+0.6512 \pm 0.0424$	$+0.1194 \pm 0.0402$	3/0

All 24 runs have positive W_T. In Table R1, pairwise ranking has the highest mean overall Spearman, while TopSort has the highest mean W_T. In Table R2, ListMLE has the highest mean overall Spearman, while pairwise ranking has the highest mean W_T. The two criteria therefore prefer different objectives on both evaluated splits even though none of the runs reverses. This establishes a non-financial diagnostic role for W_T; it does not establish uniform utility gains or a universal best training objective. The reported SDs summarise training-seed variation, not uncertainty over held-out cohorts.

S Empirical Loss-Weight Sweeps and Allocation Diagnostics

S.1 MSE–ListMLE weight sweep

We evaluate 21 separately trained LSTM blends, $\mathcal{L}_\alpha = (1-\alpha)\mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{ListMLE}}$, for $\alpha \in \{0, 0.05, \ldots, 1\}$, using seed 42. Figure S1 shows a strong, approximately linear negative relationship between mean test ρ_S and W_T. The endpoint values are $\rho_S : 0.050 \to 0.081$ and $W_T : -0.107 \to -0.297$. No tested $\alpha > 0$ improves both metrics relative to $\alpha = 0$, but adjacent points are not strictly monotone: overall Spearman falls at four steps and W_T improves at two steps. This is a result on the evaluated single-seed grid, not a continuity or impossibility theorem for other objectives.

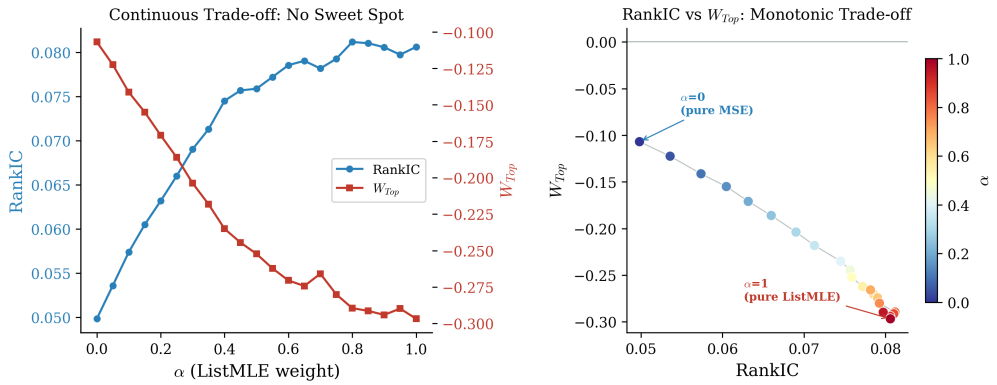


Figure S1: Twenty-one-point MSE–ListMLE weight sweep on LSTM. **Left:** both metrics versus α. **Right:** overall Spearman versus W_T, coloured by α. The configurations exhibit a near-linear trade-off with local deviations; no tested nonzero weight improves both metrics over pure MSE.("RankIC" on the vertical axis is the Spearman rank correlation ρ_S between predictions and outcomes, following quantitative-finance convention.)

Table S1 lists the seed-42 values at selected blend points.

Other ranking losses. LambdaRank (Burges, 2010) deepens reversal relative to the MSE baseline ($W_T = -0.169$ versus -0.107). De-volatilisation attenuates reversal by 66% but reduces ρ_S by 34% (Appendix H).

Table S1: Reported seed-42 sweep values used for the endpoint and adjacent-step checks. These are independently trained configurations, not repeated estimates of one model.

α	Overall ρ_S	W_T
0.00	0.049820	-0.106901
0.05	0.053579	-0.122458
0.10	0.057369	-0.141322
0.15	0.060481	-0.155002
0.20	0.063166	-0.170922
0.25	0.065989	-0.186055
0.30	0.069028	-0.203768
0.35	0.071294	-0.218236
0.40	0.074509	-0.235055
0.45	0.075684	-0.244492
0.50	0.075873	-0.252231
0.55	0.077198	-0.262211
0.60	0.078548	-0.270480
0.65	0.079036	-0.274333
0.70	0.078172	-0.265847
0.75	0.079266	-0.280067
0.80	0.081169	-0.289509
0.85	0.081029	-0.291381
0.90	0.080569	-0.294140
0.95	0.079732	-0.289773
1.00	0.080613	-0.296810

S.2 Completed-label top-aware supervision

TopSort adds a smooth pairwise term among observations whose completed training outcomes place them in the true-top group; TopListMLE is its listwise analogue. For completed labels y_i, y_j and fitted scores $s_i = f(x_i), s_j = f(x_j)$, the pairwise term is

$$\ell_{ij} = \log(1 + \exp\{-\operatorname{sign}(y_i - y_j)(s_i - s_j)\}) .$$

The top masks use training outcomes only. Validation selects the checkpoint, test outcomes are reserved for evaluation, and inference uses X alone. Thus these are completed training-label methods, not forward-looking supervision. The reported direct TopSort/TopListMLE runs give positive out-of-sample W_T; lagged or random masks do not, whereas whole-list ApproxNDCG is mixed.Specifically, 13 TopSort/TopListMLE configurations produce $W_T > 0$ (13/13); 18 lagged-label and random-mask configurations produce $W_T \leq 0$ (0/18); 5 ApproxNDCG configurations give mixed results (2/5 positive).

S.3 Separate retrieval and within-top weights

The additional date-wise experiment uses exact training-top masks and separates top-versus-rest retrieval from ordering within that group:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{cross}}\mathcal{L}_{\text{top-vs-rest}} + \lambda_{\text{within}}\mathcal{L}_{\text{within-top}}. \tag{21}$$

The two ranking terms apply to top-versus-rest and within-top pairs, respectively. All nine weight pairs and the two validation-preservation rules were fixed before training; model selection used validation only. Table S2 gives the four reported representative configurations. Tables S3 and S4 enumerate the complete weight grid and its predicted-top allocation results.

Every non-baseline configuration raises W_T and top overlap but lowers overall Spearman. None improves predicted-top-20% equal-weight Sharpe over the baseline or satisfies both validation-preservation constraints. The reported score decomposition shows that top-aware supervision raises realised winners but also raises large losers. Thus an improvement in the local ranking target need not improve the other components of retrospective within-true-top capture. This finding extends the observed trade-off beyond the MSE–ListMLE blend; it does not establish that all possible losses must have the same trade-off.

Table S2: Four reported representatives from the prespecified nine-point grid: baseline, balanced midpoint, within-only endpoint and highest-W_T configuration. Top overlap is reported under the same evaluation protocol.

λ_{cross}	λ_{within}	W_T	ρ_S	Top overlap
0	0	-0.1964	+0.0651	0.1789
0.25	0.25	-0.0547	+0.0329	0.2214
0	1	+0.1164	-0.0194	0.2356
0.50	1	+0.1967	-0.0342	0.2627

S.4 Fixed true-top membership: isolating allocation

For the nine trained objectives, hold membership fixed to the realised true-top group. Equal allocation is then identical across objectives. Within that same group, use score-rank weights

$$w_i = \frac{r_i}{\sum_j r_j}, \tag{22}$$

where r_i is the within-group score rank, with higher scores receiving higher ranks. All three negative-W_T configurations reduce score-rank return relative to equal allocation, the five clearly positive configurations increase it, and the near-zero case changes little. Across the nine objectives, the Spearman association between W_T and the annualised return gain is $+0.983$. Because membership is defined by realised outcomes, this is a retrospective, nondeployable mechanism diagnostic, not an investment rule.

S.5 Predicted-top allocation: the full nine-point results

We also compare equal and score-rank allocation within each model's predicted top-20% shortlist. Returns and Sharpe are annualised, and Δ is score-rank minus equal allocation. Unlike the fixed-membership test, this comparison retains differences in retrieval and in the composition of the predicted shortlist.

Table S3: Complete nine-point predicted-top-20% allocation comparison: annualised returns. Return levels are percentages; differences are percentage points. W_T is computed in the realised true-top group.

λ_{cross}	λ_{within}	W_T	Equal return (%)	Score-rank return (%)	Δ return (pp)
0	0	-0.196	34.23	37.55	+3.32
0	1	+0.116	15.79	15.87	+0.08
0.25	0.25	-0.055	34.24	38.06	+3.82
0.25	0.50	+0.0001	28.73	31.43	+2.70
0.50	0.25	-0.017	30.57	33.93	+3.36
0.50	0.50	+0.087	26.96	29.73	+2.77
0.50	1	+0.197	12.45	9.91	-2.54
1	0	+0.055	26.33	28.88	+2.55
1	0.50	+0.091	30.27	33.73	+3.46

Predicted-top outcomes combine retrieval, within-shortlist ordering and risk. Accordingly, the sign of true-top W_T does not determine these portfolio differences: the baseline has negative W_T but positive score-rank gain, whereas the highest-W_T configuration has a negative gain in its predicted shortlist. The fixed-membership test isolates the ordering channel; this predicted-top test does not. The historical selector in Section 5.3 therefore enforces aggregate quality before using W_T, rather than maximising W_T without constraints.

Table S4: The same nine predicted-top configurations: annualised Sharpe. The difference is score-rank minus equal allocation, not the difference from the baseline objective.

λ_{cross}	λ_{within}	Equal Sharpe	Score-rank Sharpe	Δ Sharpe
0	0	1.222	1.312	+0.090
0	1	0.544	0.547	+0.003
0.25	0.25	1.122	1.208	+0.086
0.25	0.50	0.962	1.040	+0.078
0.50	0.25	1.055	1.161	+0.106
0.50	0.50	0.894	0.977	+0.083
0.50	1	0.390	0.299	-0.091
1	0	0.852	0.916	+0.064
1	0.50	0.991	1.092	+0.101

T W_T-AWARE MODEL SELECTION: ROLLING BACKTEST

T.1 HISTORICAL RULE AND LABEL AVAILABILITY

In the original eight-loss LSTM comparison, W_T is strongly associated with gross top-5% portfolio return (Pearson $r = 0.97$). This cross-loss association is separate from the chronological model-selection experiment below; it does not imply that every pair of models has the same return and W_T ordering.

Let B_k denote an evaluation block and τ_k its selection time. For the preceding three blocks, retain only labels that have matured:

$$\mathcal{H}_k = \{d \in B_{k-3} \cup B_{k-2} \cup B_{k-1} : \text{avail}(Y_d) \le \tau_k\}.$$

The block length is 21 trading days. Write $\widehat{W}_{j,k}$ for the historical mean of the daily W_T values of model j, and $\widehat{R}_{j,k}$ for the historical overall-Spearman statistic under the frozen evaluation convention. With a prespecified floor κ,

$$\begin{aligned} &\mathcal{A}_k = \{j \in \mathcal{M} : \widehat{R}_{j,k} \ge \kappa\}, \\ &\widehat{j}_k^{\,W} \in \arg\max_{j \in \mathcal{A}_k} \widehat{W}_{j,k}, \qquad \widehat{j}_k^{\,R} \in \arg\max_{j \in \mathcal{M}} \widehat{R}_{j,k}. \end{aligned} \tag{23}$$

The submitted specification uses $\kappa = 0.045$. The floor is a historical eligibility constraint: an increase in W_T cannot compensate for failing that constraint. An empty-set fallback and a tie-breaking rule must also be fixed before future outcomes are examined. The future-best comparison in Appendix V is restricted to a common nonempty eligible set; it does not cover a fallback that ignores the floor.

When the candidate scores, eligibility rule, fallback and tie-breaking are fixed using information available at τ_k, both selectors in (23) are measurable with respect to that information. No outcome from B_k enters model selection. This establishes non-anticipation under the stated availability conditions. The frozen reconstruction removes two signal dates at each history boundary to account for label maturity.

T.2 SUBMITTED RESULTS AND THE FROZEN RECONSTRUCTION

The initial unpurged experiment reports improvements of 9.1 percentage points in gross annualised return, 0.19 in Sharpe, and 7.4 percentage points in maximum drawdown relative to historical-Spearman selection. These are the unpurged results. A separate 92-model maturity-purged reconstruction with the maturity purge retains a gross return advantage of approximately 6.7 percentage points. The result summary also reports the following allocation comparison using the same frozen, two-date-purged selection log: only weights inside the predicted top-5% shortlist change, over 713 days. For score-rank allocation, $w_i = r_i / \sum_{l \in \mathcal{P}} r_l$, where r_i is the model-score rank within that predicted shortlist $\mathcal{P}$. Table T1 keeps this reconstruction separate from the original headline figures. The selector comparison in Table T3 uses the same panel over 693 days; the shorter window reflects the bootstrap endpoint trimming required for paired block resampling.

Table T1: Frozen rolling reconstruction: equal versus score-rank allocation (predicted top 5% portfolio). Returns are annualised and gross of transaction costs. The difference row is the constrained W_T rule minus historical-Spearman selection.

	Equal weights		Score-rank weights	
Selector	Return	Sharpe	Return	Sharpe
Overall Spearman	28.46%	1.154	30.38%	1.250
Spearman floor $+W_T$	35.18%	1.176	41.01%	1.355
Difference	+6.72 pp	+0.022	+10.63 pp	+0.105

The return advantage is 3.91 percentage points larger under score-rank allocation, and the Sharpe advantage is larger by 0.083. These realised results indicate that the allocation rule changes the measured benefit of selection in this panel. They are not net-return estimates, a significance test of financial returns, or a guarantee that positive true-top W_T improves every predicted-shortlist portfolio.

T.3 Historical-to-next-block stability

For the frozen 92-model panel, each origin uses three complete 21-day history blocks, removes their last two signal dates for maturity, freezes the model ranking and evaluates the next complete block. Future W_T is used only for this audit, not for that origin's selection. Table T2 reports model-level Spearman associations.

Table T2: Stability audit on the frozen financial panel. Positive counts and the median/range summarise the reported window-level coefficients; windows and model comparisons are not independent replications.

Check	Model set	Positive	Median	Range
History to next block	All 92	33/33	0.984	[0.927,0.995]
History to next block	Eligible	33/33	0.986	[0.758,0.996]
Adjacent histories	All	32/32	0.996	[0.982,0.999]
Adjacent histories	Eligible intersection	32/32	0.996	[0.855,0.999]
Nonoverlapping histories	All	30/30	0.988	[0.966,0.997]

The nonoverlapping-history comparisons remain positive, so the reported stability is not limited to adjacent overlapping histories. The evidence supports historical use within this panel and tested period; it does not certify the drift bound of Proposition 9 or establish stability in arbitrary future periods or other domains.

T.4 Financial selector comparison

The three selectors represent structurally different information sources: overall ranking quality (ρ_S), predicted-top retrieval (top-K overlap, equivalent to Precision@K and EF@K up to scale), and true-top internal ordering (W_T). All use the same frozen 92-model panel, trailing three-block history, two-day maturity purge and historical RankIC floor 0.045 (predicted top 5% portfolio). Intervals use a paired circular moving-block bootstrap (63-day blocks) of the daily return vectors; they are not adjusted for the number of comparisons.

Table T3: W_T selector minus two baseline selectors (693 trading days): annualised gross return and Sharpe differences with pointwise 95% bootstrap intervals.

Allocation	**Comparison**	**Return Δ [95% CI]**	**Sharpe Δ [95% CI]**
Equal	W_T minus ρ_S	+6.40 [−6.76, +21.52]	−0.021 [−0.436, +0.385]
Equal	W_T minus overlap	+1.37 [−11.21, +17.42]	+0.048 [−0.350, +0.518]
Score-rank	W_T minus ρ_S	+10.43 [−5.39, +29.51]	+0.064 [−0.445, +0.543]
Score-rank	W_T minus overlap	+2.51 [−11.80, +19.86]	+0.099 [−0.342, +0.590]

The three selectors choose different models: W_T and ρ_S select the same model in 0 of 33 windows; W_T and overlap agree in 13 of 33 (Table T4). In point estimate, W_T selection produces higher gross returns than ρ_S under both allocation schemes; the difference relative to the retrieval-based overlap selector is small and not significant. Neither comparison reaches significance at the 5% level under the stated bootstrap. These results confirm that W_T captures information distinct from both aggregate ranking and predicted-top retrieval, but do not establish a return advantage over direct retrieval selection in this panel.

Table T4: Selector agreement: number of windows (out of 33) in which two selectors choose the same model.

Comparison	Same model
W_T vs ρ_S	0 / 33
W_T vs overlap	13 / 33

U Non-financial Historical Model Selection

U.1 Chronological comparisons and decision endpoints

Electricity, Traffic and Weather use a history-to-next-block protocol, with 184, 122 and 59 future blocks, respectively (365 in total). At each origin, model selection uses only completed historical blocks; the next block is opened for evaluation after the choice is frozen. For each historical criterion C, the same percentile-rank selector is compared with its $C + W_T$ version, keeping the candidate pool, history and selection convention fixed between the two versions.

Selector aggregation. At each decision origin t, let $\mathcal{H}_t$ denote the preceding three complete 21-day blocks and let $\mathcal{M}_t$ be the common candidate-model set. For each selection metric z, we first compute its trailing mean and then its average-tie percentile rank across models:

$$\bar{z}_t(m) = \frac{1}{|\mathcal{H}_t|} \sum_{\tau \in \mathcal{H}_t} z_\tau(m), \qquad R_{z,t}(m) = \frac{\mathrm{rank}^{\mathrm{avg}}_{\mathcal{M}_t}\{\bar{z}_t(m)\}}{|\mathcal{M}_t|},$$

where larger values receive larger ranks. The execution order is: trailing mean over historical blocks, then percentile rank across models, then equal-weight average across metrics. The two selector scores are

$$S_C(m;t) = \frac{1}{p} \sum_{j=1}^{p} R_{c_j,t}(m), \qquad S_{C+W_T}(m;t) = \frac{\sum_{j=1}^{p} R_{c_j,t}(m) + R_{W_T,t}(m)}{p+1}.$$

Each rule selects the model with the largest score. W_T contributes one equally weighted percentile-rank vote; the metrics are not combined lexicographically. Ties within a metric receive their average rank; exact ties in the final score are resolved by the lexicographically smallest model identifier.

The financial floor rule and training-objective comparison are separate experiments.

Allocation uses score-rank weights inside each model's predicted Top-K shortlist. Allocation gain is score-rank utility minus equal-allocation utility on that same selected set. Raw total utility is the end-to-end utility of the allocation rule. Neither is the retrospective within-true-top capture used in the earlier diagnostic analysis. All deltas below are the $C + W_T$ selector minus the C selector, evaluated on the subsequent block.

The all-direct-raw criterion includes overall Spearman, NDCG, true-top Recall, predicted-top rank correlation, predicted-top capture and raw historical utility. The normalised historical utility row uses random-to-oracle normalisation when aggregating historical blocks; its future endpoint is still expressed in raw units.

U.2 Traffic: allocation-specific incremental information

Tables U1 and U2 report the six highlighted baseline comparisons. For the within-predicted-top capture baseline, adding historical W_T increases next-block allocation gain by 0.000105, with a

95% moving-block bootstrap (MBB) interval of $[0.000049, 0.000161]$. Relative to the baseline mean of 0.002527, the reported increase is 4.2%. A model-identity permutation test with max-statistic correction across 13 contexts in each of three domains (39 comparisons) gives $p = 0.015$ for this allocation endpoint.

Table U1: Traffic: changes in next-block retrieval metrics for the six highlighted historical criteria. These point estimates describe retrieval, not allocation gain.

Historical criterion C	Δ Recall	Δ NDCG
All-direct-raw	-0.0080	-0.0177
NDCG@K	$+0.0015$	-0.0233
True-top Recall@K	-0.0105	-0.0199
Predicted-top rank correlation	-0.0221	-0.0545
Within-predicted-top capture	-0.0470	-0.0760
Normalised historical score-dependent utility	-0.0042	-0.0262

Table U2: Traffic: next-block allocation and total-utility changes, with reported 95% MBB intervals in brackets. These intervals are not 39-comparison familywise intervals; the capture-based allocation result has a separate adjusted permutation $p = 0.015$.

Historical criterion C	Δ allocation gain	Δ raw total
	$+0.0000529$	$+0.000094$
All-direct-raw	$[+0.0000087, +0.0000990]$	$[-0.000030, +0.000229]$
	$+0.000102$	$+0.000847$
NDCG@K	$[+0.000047, +0.000157]$	$[+0.000572, +0.001128]$
	$+0.000104$	$+0.000135$
True-top Recall@K	$[+0.000049, +0.000161]$	$[+0.000002, +0.000267]$
	$+0.000112$	$+0.000714$
Predicted-top rank correlation	$[+0.000049, +0.000176]$	$[+0.000300, +0.001138]$
	$+0.000105$	$+0.000031$
Within-predicted-top capture	$[+0.000049, +0.000161]$	$[-0.000182, +0.000256]$
	$+0.000102$	$+0.000607$
Normalised historical score-dependent utility	$[+0.000054, +0.000153]$	$[+0.000376, +0.000846]$

The confirmed familywise conclusion concerns capture-based allocation gain, not a general improvement in raw total utility. Several retrieval point estimates decline, and the capture-row raw-total interval includes zero. The raw-total intervals in the table provide supporting evidence, not a multiplicity-adjusted utility claim. In the stated exact-K setting, Precision and true-top Recall are equal and enrichment is a fixed rescaling of Recall; they should not be counted as independent sources of evidence.

Table U3 lists the full set of 13 Traffic allocation comparisons. Every point estimate is positive. A shared model-identity permutation that preserves dependence across contexts gives a mean standardised-effect statistic of $T = 2.113$ with one-sided $p = 0.001$, and a 13-of-13 positive-sign result with one-sided $p = 0.027$. This is a post-hoc joint sensitivity analysis. It does not replace the 39-comparison-adjusted capture result, and the 13 signs are not treated as independent Bernoulli observations.

U.3 All three chronological domains

Tables U4 and U5 separate allocation gain from raw total utility for the full three-domain, 13-context matrix. Raw units differ across domains and are not pooled. Allocation deltas are positive in 4/13 Electricity, 13/13 Traffic and 2/13 Weather contexts; raw-total deltas are positive in 3/13, 12/13 and 9/13, respectively. Electricity and Weather delimit the Traffic finding. These signs do not establish statistical significance for individual rows.

Table U3: Traffic: allocation-gain deltas across all 13 historical contexts. The last column gives the reported single-context identity-permutation p-value, not a familywise-adjusted value.

Historical context	Allocation delta	Single-context p
All-direct	$+0.00005675$	0.020979
All-direct-raw	$+0.00005293$	0.025974
NDCG@K	$+0.00010244$	0.000999
Within-predicted-top capture	$+0.00010501$	0.002997
Predicted-top rank correlation	$+0.00011247$	0.002997
Raw utility	$+0.00001732$	0.006993
True-top Recall@K	$+0.00010448$	0.002997
Score-dependent utility	$+0.00010247$	0.005994
Overall Spearman	$+0.00006198$	0.011988
Spearman + equal-allocation utility	$+0.00007605$	0.004995
Spearman + NDCG	$+0.00007894$	0.002997
Spearman + Recall	$+0.00008851$	0.014985
Spearman + Recall + NDCG + equal utility	$+0.00007145$	0.008991

Table U4: All three chronological domains: changes in next-block allocation gain across the full 13-context comparison. Domain-specific raw units are not comparable across columns.

Historical context	Electricity	Traffic	Weather
All-direct	$+0.7798$	$+0.0000568$	$+0.00369$
All-direct-raw	$+0.2048$	$+0.0000529$	$+0.00448$
NDCG@K	-2.9877	$+0.0001024$	-0.01302
Within-predicted-top capture	$+1.1617$	$+0.0001050$	-0.00337
Predicted-top rank correlation	-1.5942	$+0.0001125$	-0.00990
Raw utility	-0.6255	$+0.0000173$	-0.01220
True-top Recall@K	-2.9484	$+0.0001045$	-0.01144
Score-dependent utility	$+1.5004$	$+0.0001025$	-0.01002
Overall Spearman	-6.3094	$+0.0000620$	-0.01216
Spearman + equal-allocation utility	-0.4397	$+0.0000760$	-0.01075
Spearman + NDCG	-0.8338	$+0.0000789$	-0.00312
Spearman + Recall	-1.4894	$+0.0000885$	-0.01344
Spearman + Recall + NDCG + equal utility	-0.3343	$+0.0000714$	-0.00459

Table U5: All three chronological domains: changes in next-block raw total utility for the same 13 contexts. These are point estimates, not adjusted significance claims.

Historical context	Electricity	Traffic	Weather
All-direct	$+1.5819$	$+0.0005508$	$+0.00083$
All-direct-raw	$+0.3531$	$+0.0000939$	$+0.00232$
NDCG@K	-10.8846	$+0.0008472$	$+0.00998$
Within-predicted-top capture	$+26.2445$	$+0.0000314$	$+0.04606$
Predicted-top rank correlation	-6.7075	$+0.0007136$	$+0.01952$
Raw utility	-11.5525	-0.0000790	-0.00274
True-top Recall@K	-12.2857	$+0.0001350$	-0.01359
Score-dependent utility	-2.8161	$+0.0006068$	$+0.00056$
Overall Spearman	-20.1771	$+0.0001022$	-0.01919
Spearman + equal-allocation utility	-1.4292	$+0.0001593$	$+0.00144$
Spearman + NDCG	-1.9633	$+0.0001485$	$+0.00547$
Spearman + Recall	-3.8728	$+0.0001380$	-0.00288
Spearman + Recall + NDCG + equal utility	-0.8557	$+0.0000872$	$+0.00106$

U.4 STATIC-DOMAIN SENSITIVITY

The six static domains use five repeated outer splits and 15 historical contexts each. Table U6 gives effect-sign and model-switch counts. The overlapping splits and contexts are dependent; this inventory is not a binomial test or pooled cross-domain effect.

Table U6: Static-domain sensitivity: effect-sign counts across 15 contexts and five repeated splits; the final column counts contexts that changed the model.

Domain	Allocation $+/-/0$	Raw total $+/-/0$	Changed model
Diabetes	2/13/0	9/6/0	15/15
Drug-AChE	5/10/0	6/9/0	15/15
Energy Efficiency	6/0/9	6/0/9	6/15
ESOL	9/4/2	11/2/2	13/15
Superconductivity	4/2/9	3/3/9	6/15
Wine	0/13/2	1/12/2	13/15

The results are mixed, with adverse contexts in Diabetes, Drug-AChE and Wine, favourable contexts in Energy Efficiency and ESOL, and unchanged selections under many Energy Efficiency and Superconductivity rules. They do not support a uniformly better utility selector.

U.5 TRAFFIC MATCHED-MODEL DIAGNOSTIC

Models are paired when their historical Spearman, Recall@K, NDCG@K and equal-allocation utility differ by no more than one blockwise standard error; each pair is then oriented by historical W_T. The analysis covers 2,882 pairs over 120 future blocks. Inference first aggregates within blocks, rather than treating the model pairs as independent. Table U7 reports higher-historical-W_T minus lower-historical-W_T differences.

At comparable historical direct metrics, larger historical W_T is associated with larger next-block allocation gain and raw utility. This conditional association is not an independent selector experiment.

Table U7: Traffic matched-model diagnostic: next-block mean differences and 95% MBB interval status.

Next-block endpoint	Mean difference	95% MBB interval
Recall	$+0.00457$	Excludes zero
NDCG	$+0.001997$	Includes zero
Equal-allocation utility	$+0.000229$	Excludes zero
Allocation gain	$+0.0000835$	Excludes zero
Raw total utility	$+0.000313$	Excludes zero

V IDENTIFICATION AND TRANSPORT CONDITIONS

V.1 WHAT HISTORICAL FEEDBACK IDENTIFIES

Fix the candidate population and evaluation regime. Let Y be the outcome under that regime, let O indicate that it is observed, and let C contain the pre-outcome policy variables and all evaluated model scores. The candidate scores are frozen before the audit outcomes used for selection, or evaluated under a valid sample-splitting scheme. Consistency means that the recorded outcome equals the target Y when $O = 1$. If the action changes the outcome rather than merely observing it, the target regime and the corresponding causal assumptions must be stated separately.

With complete or delayed-complete feedback, the completed candidate cohort can be evaluated directly. This is the regime of the fully labelled Drug benchmark and the recorded full-universe returns used in the financial evaluation. With selectively observed outcomes, an unweighted audit of $O = 1$ rows generally targets that observed cohort, not the full candidate population.

Proposition 8 (Identification under selective feedback)**.** *Assume consistency, $O \perp Y \mid C$, and $e(C) = P(O = 1 \mid C) > 0$ almost surely. Then for every integrable h,*

$$E\left[\frac{O}{e(C)}h(C,Y)\right] = E[h(C,Y)]. \tag{24}$$

Consequently, the full-candidate joint law, its declared outcome cutoff, and its conditional rank correlations are identified wherever the required rank variances are positive.

Proof. Consistency permits the observed outcome to be replaced by Y in the terms with $O = 1$. Ignorability gives $E[O \mid C, Y] = e(C)$, so

$$E\left[\frac{O}{e(C)}h(C,Y)\right] = E\left[\frac{E[O \mid C,Y]}{e(C)}h(C,Y)\right] = E[h(C,Y)].$$

Applying this identity to indicators identifies the joint law. The outcome quantile, selected marginal ranks and their correlation are functionals of that same law.

Positive observation probability is an identification condition, not by itself a finite-sample precision guarantee. Stable inverse weighting requires adequate overlap and suitable estimation regularity. When a deterministic assay rule gives some candidates zero observation probability, this route cannot identify the unrestricted full-pool W_T. Report an observed-policy/cohort target, or obtain a randomised audit sample with positive inclusion probabilities over the target population. Such sampling does not make the previously unobserved outcomes of a fixed finite cohort known individually.

V.2 The cutoff and ranks must also be weighted

For an estimated positive propensity $\widehat{e}(C_i)$, put $v_i = O_i/\widehat{e}(C_i)$ and $V_n = \sum_i v_i$. Unobserved outcomes may use a fixed sentinel; their zero-weight terms do not depend on that value. When $V_n > 0$, the weighted full-candidate empirical law is

$$\widehat{P}^v(A) = V_n^{-1}\sum_i v_i \mathbf{1}\{(C_i, Y_i) \in A\}.$$

Compute the outcome cutoff $\widehat{c}_q$ from this law. For the exact probability-q convention, set

$$\widehat{\tau}_q(y) = \mathbf{1}\{y > \widehat{c}_q\} + \widehat{\gamma}_q \mathbf{1}\{y = \widehat{c}_q\}, \qquad \sum_i v_i \widehat{\tau}_q(Y_i) = qV_n,$$

where $\widehat{\gamma}_q \in [0,1]$ allocates the required boundary mass. An include-all-boundary-ties target instead uses $\widehat{\gamma}_q = 1$ and its actual selected mass. These are distinct declared conventions; this discussion does not change the empirical rule in Definition 1 or any completed experiment.

Let $b_i = v_i\widehat{\tau}_q(Y_i)$ and $B_n = \sum_i b_i > 0$. For either the score or outcome variable Z, its selected weighted mid-distribution is

$$\widehat{M}_{Z,b}(z) = B_n^{-1}\sum_i b_i\left[\mathbf{1}\{Z_i < z\} + \tfrac{1}{2}\mathbf{1}\{Z_i = z\}\right].$$

The estimator for model j is the weighted Pearson correlation

$$\widehat{W}^v_{T,j} = \mathrm{Corr}_b\big(\widehat{M}_{F_j,b}(F_j), \widehat{M}_{Y,b}(Y)\big). \tag{25}$$

Weights therefore enter membership, both ranks and the correlation; there is no correction obtained by multiplying an already computed complete-case Spearman by an inverse propensity. Nonpositive estimated propensities, zero selected mass or zero rank variance require a prespecified handling rule, not an undefined division.

Consistency of this plug-in calculation additionally requires a weighted uniform law of large numbers, a stable weighted cutoff and boundary rule, and nondegenerate limiting rank variances. Under these conditions the weighted selected law and rank moments converge, hence so does the correlation. The identification identity alone supplies neither unbiasedness of this nonlinear estimator nor a confidence interval under date dependence.

V.3 FROM HISTORICAL ESTIMATES TO FUTURE MODEL QUALITY

Identification of past feedback does not guarantee future stability. Use one common target for all models and periods: the same top fraction, selection/tie convention, observation regime and date-aggregation rule. A population W_T and an expected finite-cohort statistic are not interchangeable targets without a separate estimation bridge.

Proposition 9 (Conditional historical-selection regret)**.** *Let a finite candidate family $\mathcal{M}$ be fixed before the audit outcomes used for selection. For model j, let w_j^H be its historical target, $\widehat{w}_j$ its estimate from matured history, and w_j^F its future target under the same evaluation convention. Suppose, on one joint event for all $j \in \mathcal{M}$,*

$$|\widehat{w}_j - w_j^H| \le \epsilon, \qquad |w_j^F - w_j^H| \le \Delta.$$

Let $\mathcal{A} \subseteq \mathcal{M}$ be a nonempty eligible set determined using available historical information, and use a fixed tie rule to set

$$\widehat{j} \in \arg\max_{j \in \mathcal{A}} \widehat{w}_j, \qquad j^* \in \arg\max_{j \in \mathcal{A}} w_j^F.$$

Then, on that event,

$$w_{j^*}^F - w_{\widehat{j}}^F \le 2(\epsilon + \Delta). \tag{26}$$

Proof. The uniform assumptions imply $|w_j^F - \widehat{w}_j| \le \epsilon + \Delta$ for every candidate. Since $\widehat{w}_{j^*} - \widehat{w}_{\widehat{j}} \le 0$,

$$\begin{aligned} w_{j^*}^F - w_{\widehat{j}}^F &= (w_{j^*}^F - \widehat{w}_{j^*}) + (\widehat{w}_{j^*} - \widehat{w}_{\widehat{j}}) + (\widehat{w}_{\widehat{j}} - w_{\widehat{j}}^F) \\ &\le 2(\epsilon + \Delta). \end{aligned}$$

Uniform coverage of the original family makes this argument valid for the history-selected eligible subset as well.

If the historical error event has probability at least $1 - \alpha$ and the uniform drift event at least $1 - \beta$, the conclusion has probability at least $1 - \alpha - \beta$. Separate unadjusted intervals for individual models do not automatically provide the required simultaneous event. For date-wise data, the estimation bound must account for shared dates and cross-model dependence. Future performance may be a conditional next-block target given current information; marginal stationarity alone does not make the conditional drift vanish.

The bound compares models within the same eligible set. It does not bound the loss from excluding a future-best model by the historical floor, certify future retrieval or risk constraints, or guarantee return, Sharpe or raw allocation utility. The chronological experiments assess those outcomes separately; they do not estimate or certify every premise of (26).