Calibrated Prediction Across 16 Sports, and the Honest Wall of the Closing Line

Jovonni L. Pharr
Georgia Cyber Warfare Range
info@gacwr.org

August 7, 2026
v0.0.1

Abstract

We report the results of a systematic effort to build leak-free, walk-forward, calibrated win-probability models across sixteen sports, and then to answer the only question that ultimately matters for a bettor: does any of it beat the market? The pre-game models span 152,201 boxing bouts down to 1,728 Formula 1 entries, and their out-of-sample discrimination ranges from AUC = 0.866 (boxing) to 0.572 (golf), with expected calibration error at or below 0.03 on almost every sport. A granular "sharp" layer that ingests the pitch-, possession-, and event-level data professional shops actually use improves every game-level baseline out-of-sample; tennis rises from AUC 0.722 to 0.783 and its calibration error falls from 0.019 to 0.007. And yet, on the nine sports that carry a real closing price, no pre-game model of ours—calibrated Elo or best sharp model—beats the de-vigged liquid close out-of-sample. A flat-$100 backtest of every model pick, settled at the real vig-inclusive close across eight sports, loses $254,583 on $8,692,100 staked (−2.9%). We take this as a measurement, not a defeat: the close already aggregates everyone's model, ours included, and paying the vig to match it is designed to lose. We describe an adversarial leakage audit that retracted a golf headline from AUC 0.982 to 0.592 and a tennis sharp figure from 0.876 to 0.731; a learnability framework that scores a sport before a GPU-hour is spent; an in-game backtest, redone the literature's way with fractional Kelly and an expected-value filter, that isolates a model deficit from a staking artifact; and a market-making simulation in which our calibrated fair value functions as an adverse-selection filter and earns +1.17c per contract under plausible flow. The through-line is that calibration, not accuracy, is the deployable deliverable, and that the honest edge—if it exists—lives in soft venues, in making rather than taking, and behind a forward closing-line-value pipeline we have not yet run.

1 Introduction

The premise of most sports-prediction work is that a sufficiently clever model will find an edge the market has missed. The premise is usually wrong, and it is wrong for a reason that is easy to state and hard to accept: the liquid closing price of a sporting event is not a naive quote to be beaten but a sharp consensus that has already absorbed every serious model, every lineup, and every late money move. Once de-vigged, it is a fair probability, and its own out-of-sample accuracy is an empirical ceiling. A pre-game model that appears to beat it is, far more often than not, measuring a data leak or a thin market rather than genius.

This paper is a report on taking that premise seriously across sixteen sports. We built leak-free, expanding-window walk-forward models for every sport we could source data for—from boxing, tennis, and golf with six-figure row counts down to Formula 1 and college football with a few thousand—and we measured them the way a skeptic would: log-loss and Brier as primary targets, calibration error as the deploy gate, area under the ROC curve as a discrimination diagnostic, and, wherever a real closing price existed, the model's honest gap to that close. We

then subjected every claimed out-of-sample gain to an adversarial leakage audit whose explicit stance was that each model was guilty until cleared by a runnable reproduction on real code and real data. That audit collapsed a golf headline from an area under the curve of 0.982 to 0.592 and forced a public retraction; we report the audited numbers, not the retracted ones.

The paper's title names its two findings. The first is that calibrated prediction is achievable and reproducible at scale: with disciplined point-in-time features and per-fold isotonic calibration, most of our sixteen sports produce probabilities whose expected calibration error is at or below three percentage points, which is the property that actually governs staking and closing-line value. The second is the wall. On every sport with a real market, we do not beat the de-vigged close, and a flat-stake backtest of our picks into that close loses money at almost exactly the rate the vig predicts. This is not a bug to be fixed with a better feature; it is market efficiency, measured. The contribution of the paper is to say so precisely, to quantify how far below the close we sit sport by sport, to show through a corrected in-game backtest and a market-making simulation where the residual edge could plausibly live, and to be equally precise about why we cannot yet claim it.

2 Background and context

A single sporting event decomposes, loosely, into a learnable component—a function of true strength, matchup, and situation—and an irreducible noise term of luck, bounces, officiating, and single-game variance. The central discipline of this project is to respect that decomposition before modeling. A sport in which the stronger side wins a single contest eighty to ninety-five percent of the time (a mismatched tennis match, a prospect against a club fighter) has a low noise floor and is highly learnable; a sport in which the better team wins a single game only fifty-five to sixty percent of the time (a single baseball game, a hockey night, a soccer match with a live draw) has a high noise floor, and any single-game model there is fighting variance it cannot remove. Learnability is not the same as profitability: a high-noise sport can still be profitable if the market prices its wide probabilities even worse than we do. But one must know which regime one is in to set expectations and size bets, and mistaking the irreducible noise floor for a modeling failure—and overfitting in response—is the most common way these projects fail.

The evaluation stance follows directly. A backtest exists to disprove an edge, not to celebrate one. Splits must be time-aware and expanding-window; every feature must be knowable before the market it bets closes; ratings and statistics must be point-in-time rather than end-of-season values. The metrics are ordered by how hard they are to fool: log-loss first because it punishes confident wrongness, then Brier against the sport's noise floor, then calibration because a model can have excellent accuracy and useless overconfidence, then closing-line value as the leading indicator of real edge, and only then return on investment net of fees and slippage. A model can classify who wins beautifully and still be un-bettable if it is miscalibrated; calibration and closing-line value gate deployment, and accuracy alone never does.

The market side is the other half of the alpha. A binary contract paying one dollar on a settled outcome trades in the unit interval, and its price is approximately the market-implied probability; removing the two-sided overround recovers a fair probability. An edge exists only when the model's fair probability diverges from that de-vigged market probability by more than costs—fees, spread, slippage, and adverse selection, the last of which means that when a sharp is on the other side of a fill, our fill is itself the bad news. The most durable such edges, in rough order, are a genuinely better probability estimate, cross-venue mispricing, slow adjustment in thin markets, and well-documented behavioral biases that are strongest in retail-heavy books. Closing-line value—consistently transacting at prices better than the eventual close—is the north-star metric because it predicts long-run profit before the bankroll confirms it: positive closing-line value with negative realized profit is variance to ride through, whereas

negative closing-line value with positive profit is luck to fear.

Three independent literature surveys conducted for this project converged on a conclusion that reshaped the plan. The one peer-reviewed-style study executing our exact thesis—a market-calibrated in-game win-probability model traded against live prices—is profitable, but only with fractional Kelly staking (flat unit staking *loses* in the same study because late-game odds approach unity, so a flat stake risks the same on a near-lock as on a longshot), an expected-value threshold filter, and market-calibrated probabilities. The best-supported durable edge in the surveys is not predicting at all but *making*: an analysis of tens of millions of exchange trades finds that makers earn roughly twice per contract net of adverse selection because retail systematically overbets one side. And the classic example of beating a high-takeout market did so not by out-forecasting a liquid outcome but by combining a model with the public odds and trading the residual in thin exotic pools against soft money at fractional Kelly. The lesson we carried forward is that alpha is a model-over-market residual in thin or soft segments, harvested by making or by racing a lagging book—not by out-predicting a liquid game.

3 Method and system

3.1 Data backbones and the walk-forward contract

Each sport is represented by a committed backbone table of one row per game-orientation, with point-in-time features and a binary label $y =$ (side-A wins), where side-A is hash-neutralized so no positional information leaks. Every metric in this paper is out-of-sample under an expanding-window walk-forward: train on $[t_0, t]$, predict $t + 1$, roll forward, never a random split across a time series. A stable sort on date is required so that same-date games map to the correct prediction, and feature availability at bet time is enforced—injuries as known then, ratings as they stood then, not end-of-season revisions. The game-level model is an Elo (or Glicko-2 for combat sports) strength representation fed into a gradient-boosted tree, with per-fold isotonic calibration fit on the training tail so that no fold ever sees its own test rows.

3.2 The learnability framework

Before committing to a sport, we score it on six axes, each on a one-to-five scale where higher is more favorable: predictability of a single game (the inverse of the noise floor), data availability and granularity, roster or participant stability, structural signal strength (whether mature predictors like sabermetrics, expected goals, or Elo already work), market inefficiency or edge headroom, and sample size or market frequency. A composite weights predictability and edge headroom most heavily,

$$\text{Learnability} = 0.25A + 0.15B + 0.10C + 0.15D + 0.20E + 0.15F,$$

and classifies a sport as green (attack first), yellow (trade selective markets), or red (research-only). Crucially the score is a starting hypothesis about model *quality*, not a verdict on profit: a red single-game sport may still yield a green futures or props market, so the framework insists the operative questions be answered before building—what is the noise floor, what is the gap between a naive model and that floor, what is the gap between the *market* and that floor, where is the market softest, and what kills the edge.

3.3 The sharp granular layer

The game-level Elo-to-tree baseline leaves signal on the table because it never sees the granular layer that sharp shops use. We built per-sport "sharp" models that ingest that layer: surface-specific Elo plus serve/return and fatigue for tennis; Glicko-2 with UFCStats striking and grappling differentials for mixed martial arts; quarterback-level expected-points-added

and completion-percentage-over-expected for the NFL; opponent-adjusted efficiency and player-availability impact for basketball; a Statcast pitcher-matchup blend for baseball; strength-of-schedule-adjusted expected-points-added plus recruiting talent for college football; and an expected-goals-driven Dixon–Coles model for soccer's three-way market. Each sharp model is a leak-free walk-forward signal, and the ensemble is a leak-free logistic stack fit per sport. A broader model-family suite formalizes this: 170 candidate families across the sixteen sports, of which 82 are historically measurable offline (the other 88 are market-structure, making, news-latency, and resolution-risk families that require live order flow and tape and cannot be scored on historical data).

3.4 The market frontier and calibration

For every sport that carries a closing price, we compute the close's own out-of-sample area under the curve, log-loss, and calibration error on the joined rows, and treat that as the empirical ceiling. Because most of these are single-book de-vigged closes with partial coverage—structurally *easier* to beat than a true sharp consensus—the frontier area under the curve is a floor on true market skill, not a point estimate. We then score our leak-free model on the identical rows, raw and after isotonic calibration, and report the gaps. The design expectation is explicit: discrimination cannot out-run the close, so calibration is the deliverable, because a calibrated fifty-eight percent that the market prices at fifty-five is a positive-expected-value bet whereas a miscalibrated model of equal area under the curve is not bettable at all.

3.5 Staking, execution, and the market maker

Position sizing is fractional Kelly ($\frac{1}{4}$ to $\frac{1}{2}$), scaled down in high-noise sports because the model's own probability is uncertain and full Kelly on an estimated edge courts ruin. Every backtest models fees, spread, slippage, and settlement lag, and requires realistic fills—a price we could actually transact at the stated size. The market-maker component does not try to beat the price at all: it quotes a bid at $p - h$ and an ask at $p + h$ around the calibrated fair value p, collects the half-spread, and uses the model as an adverse-selection filter, widening or declining precisely when informed flow is likely. In that simulation, per-contract profit is the half-spread earned minus the adverse-selection loss minus the fee, all computed from realized outcomes rather than assumed.

4 Experiments and results

4.1 Cross-sport pre-game accuracy and calibration

Table 1 reports the leak-corrected out-of-sample scoreboard for all sixteen sports, regenerated from the committed backbones and models. Boxing and the college sports top the discrimination table; the high-noise sports—baseball, mixed martial arts, golf—sit near the coin-flip floor by nature, which is the irreducible single-game variance, not a modeling failure. What matters for deployment is the calibration column: expected calibration error is at or below 0.03 on the large majority of sports, and golf is reported at its production leak-free area under the curve of 0.572, not the retracted 0.982.

The relationship between discrimination and irreducible variance is worth making visual. Figure 2 plots each sport's area under the curve against its single-game noise floor—the accuracy of always picking the favorite. Boxing and tennis extract the most skill above the floor; baseball, mixed martial arts, and golf sit essentially on it, which is the correct and honest behavior of a model that refuses to be confident where the sport offers no signal.

Table 1: Leak-corrected out-of-sample pre-game scoreboard, all sixteen sports. AUC is discrimination (0.5 is a coin flip); ECE is expected calibration error (0 is perfect). Ordered by AUC.

Sport	Rows	Side-A win	Log-loss	Brier	ECE	AUC
boxing	152,201	0.475	0.4566	0.1484	0.0073	0.866
cbb	34,837	0.663	0.5449	0.1830	0.0240	0.759
cfb	4,923	0.628	0.6012	0.1963	0.0491	0.750
tennis	149,269	0.499	0.6167	0.2136	0.0203	0.719
wnba	2,546	0.561	0.6500	0.2232	0.0488	0.697
nba	14,168	0.566	0.6416	0.2230	0.0264	0.684
soccer	14,019	0.430	0.6499	0.2226	0.0269	0.683
cricket	10,014	0.492	0.6480	0.2268	0.0204	0.675
nfl	3,028	0.548	0.6778	0.2338	0.0565	0.666
nascar	18,618	0.567	0.6649	0.2355	0.0201	0.620
nhl	7,601	0.539	0.6965	0.2410	0.0266	0.616
f1	1,728	0.666	0.6402	0.2161	0.0534	0.597
ufc	8,808	0.494	0.6835	0.2446	0.0314	0.585
esports	8,554	0.510	0.6866	0.2448	0.0342	0.582
mlb	14,580	0.531	0.6896	0.2461	0.0281	0.574
golf	100,283	0.490	0.6913	0.2478	0.0308	0.572

4.2 The sharp layer improves every baseline out-of-sample

Table 2 shows the granular sharp models against their game-level baselines on identical out-of-sample rows. Every sport improves on log-loss and area under the curve, and calibration improves markedly—baseball's expected calibration error falls by sixty-three percent, tennis's from 0.019 to 0.007. The soccer three-way model, an expected-goals Dixon–Coles, cuts its gap to the market close from roughly 0.037 to 0.020 log-loss and lands at the academic frontier (a ranked-probability score of 0.202 against the market's 0.194). The honest reading is that the granular layer captures the learnable fraction far better than the game-level baseline; the ceiling was the model, not the sport.

Table 2: Granular "sharp" models versus their game-level baselines, out-of-sample (base $\rightarrow$ sharp). Every sport improves on all three metrics.

Sport	Log-loss	AUC	ECE
tennis	$0.6146 \rightarrow 0.5607$	$0.722 \rightarrow 0.783$	$0.0190 \rightarrow 0.0074$
ufc	$0.6841 \rightarrow 0.6654$	$0.586 \rightarrow 0.635$	$0.0325 \rightarrow 0.0077$
nfl	$0.6802 \rightarrow 0.6668$	$0.660 \rightarrow 0.679$	$0.0565 \rightarrow 0.0382$
cbb	$0.5503 \rightarrow 0.5305$	$0.757 \rightarrow 0.775$	$0.0241 \rightarrow 0.0164$
mlb	$0.6937 \rightarrow 0.6809$	$0.567 \rightarrow 0.579$	$0.0322 \rightarrow 0.0119$
nba	$0.6452 \rightarrow 0.6346$	$0.682 \rightarrow 0.693$	$0.0272 \rightarrow 0.0241$
boxing	$0.4550 \rightarrow 0.4390$	$0.867 \rightarrow 0.877$	$0.0070 \rightarrow 0.0068$
cfb	$0.6417 \rightarrow 0.6055$	$0.773 \rightarrow 0.780$	$0.0827 \rightarrow 0.0537$
wnba	$0.6993 \rightarrow 0.6635$	$0.693 \rightarrow 0.695$	$0.0585 \rightarrow 0.0560$

4.3 The market frontier: no model beats the liquid close

Table 3 reports, for the nine sports that carry a closing price, the close's own out-of-sample scores and our calibrated Elo baseline on the identical rows. The pattern is uniform: our discrimination sits below the close on every sport (from -0.015 area under the curve on baseball

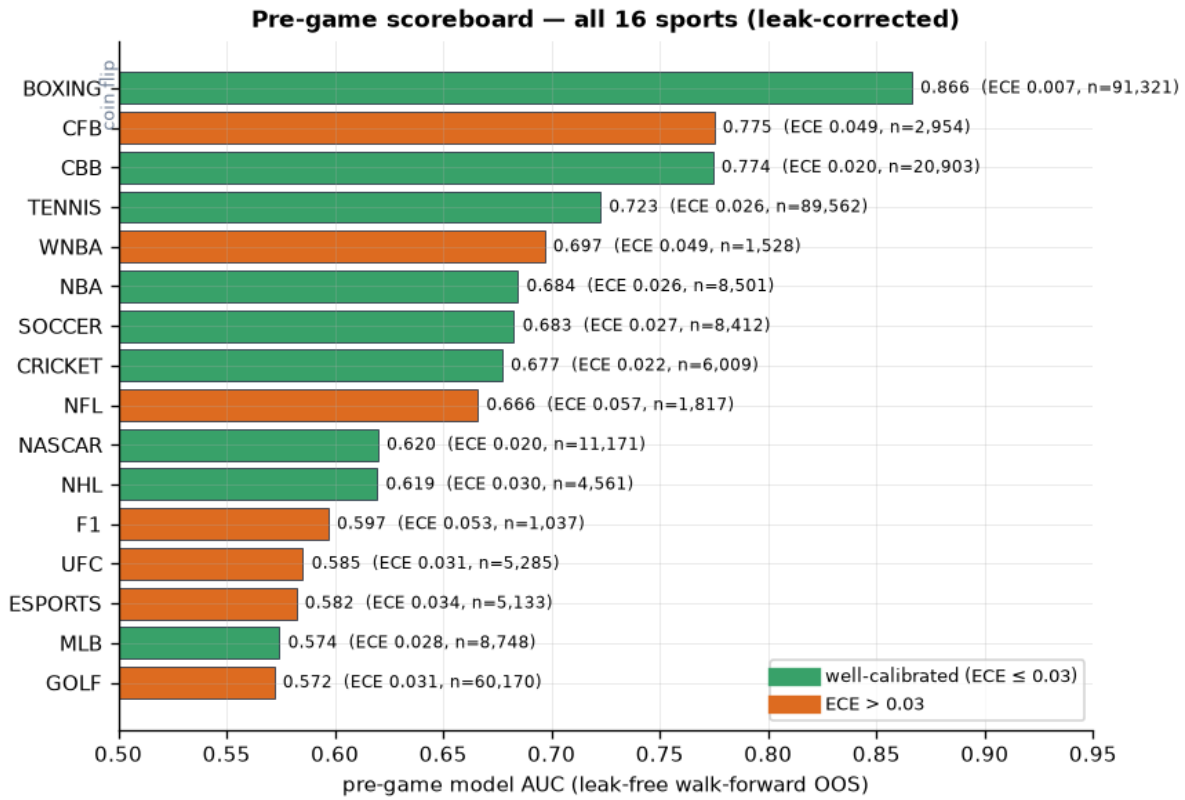


Figure 1: Pre-game scoreboard across all sixteen sports, leak-corrected. Boxing (0.87) and the college sports (≈ 0.77) top the table; baseball, mixed martial arts, and golf sit near 0.58 because single games there are mostly variance. Green marks deploy-grade calibration (ECE ≤ 0.03); golf is shown at its leak-free 0.572.

to -0.173 on mixed martial arts), and no sport shows even an apparent beat. What calibration buys is the collapse of expected calibration error toward the market's—dramatically on soccer ($0.147 \to 0.024$) and baseball ($0.071 \to 0.018$)—and that is the part that converts to closing-line value. Isotonic calibration is a genuine win where the raw model is off the diagonal and a wash or mild overfit where it is already sharp (tennis, college basketball, college football), so the honest policy is to calibrate the miscalibrated sports and leave the already-calibrated ones alone.

The frontier areas under the curve cluster in the 0.6 to 0.75 band, which says that roughly a third of a single-game outcome is coin-flip variance no information removes. A pre-game model claiming above 0.90 area under the curve on single games is measuring leakage, not skill. Legitimate high discrimination lives elsewhere—in extreme mismatches within a wide field, and in in-game models late in a decided contest—not in the pre-game moneyline.

4.4 Where 0.90 legitimately lives: confident tiers and in-game phase

No single-game team sport reaches 0.90 full-sample area under the curve; even the best-modeled ones (tennis, boxing) sit near 0.87. But every model has a high-conviction slice that is genuinely above ninety percent accurate. Restricting to games where the model puts $p \leq 0.10$ or $p \geq 0.90$, tennis is 96.4% accurate on 33.3% of all matches (subset area under the curve 0.978), boxing 94.5% on 30.7%, college basketball 94.1% on 21.0%, and college football 91.3% on 24.9%. This is a tiering result—high accuracy on a large, reliable slice—not "ninety percent on all games," and coverage is the tradeability signal: in high-variance sports the confident slice collapses below

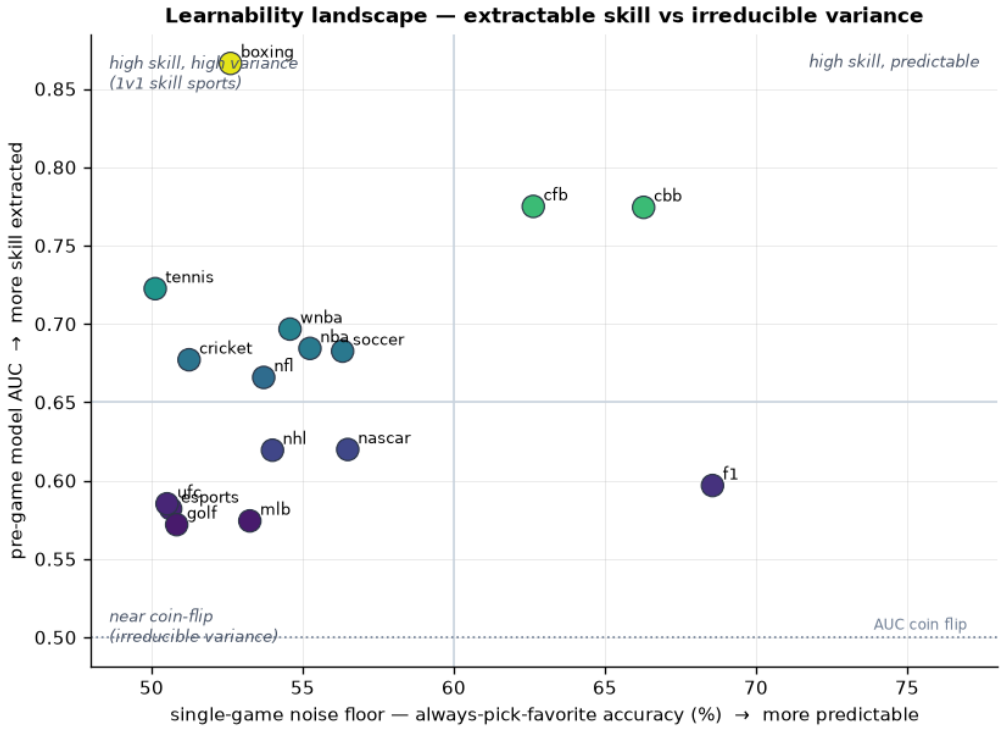


Figure 2: Learnability landscape: per-sport AUC against the single-game noise floor. The sports near the floor (MLB, UFC, golf) carry irreducible single-game variance no model can remove; the distance above the floor is extractable skill.

Table 3: The market frontier for the nine sports with a closing price. The frontier (de-vigged close) is a floor on true market skill; single-book closes are marked ∗. Our calibrated Elo baseline is measured on the identical out-of-sample rows. ΔAUC is our calibrated AUC minus the frontier AUC.

Sport	Frontier AUC	Frontier LL	Frontier ECE	Cal. model AUC	ΔAUC
mlb ∗	0.586	0.6796	0.0180	0.571	−0.015
soccer	0.715	0.6110	0.0072	0.686	−0.029
nhl ∗	0.666	0.6472	0.0191	0.637	−0.029
nba ∗	0.726	0.6072	0.0177	0.675	−0.050
tennis ∗	0.745	0.5926	0.0094	0.692	−0.053
cbb ∗	0.778	0.5380	0.0132	0.707	−0.070
nfl	0.730	0.6063	0.0223	0.656	−0.074
cfb ∗	0.787	0.5500	0.0313	0.676	−0.111
ufc ∗	0.735	0.6030	0.0177	0.562	−0.173

three percent because the sport rarely offers a mismatch that lopsided, which is the model correctly refusing confidence it does not have.

The other legitimate home of high discrimination is time. Figure 5 shows a sport-by-phase heatmap of in-game area under the curve with a 0.90 contour: college football crosses it by half-time, the NFL and hockey only in the final phase, and no sport crosses it overall. An audit confirmed the in-game state model is clean—game-level group splits so no game spans train and test, and an endgame clamp that if anything slightly lowers overall discrimination

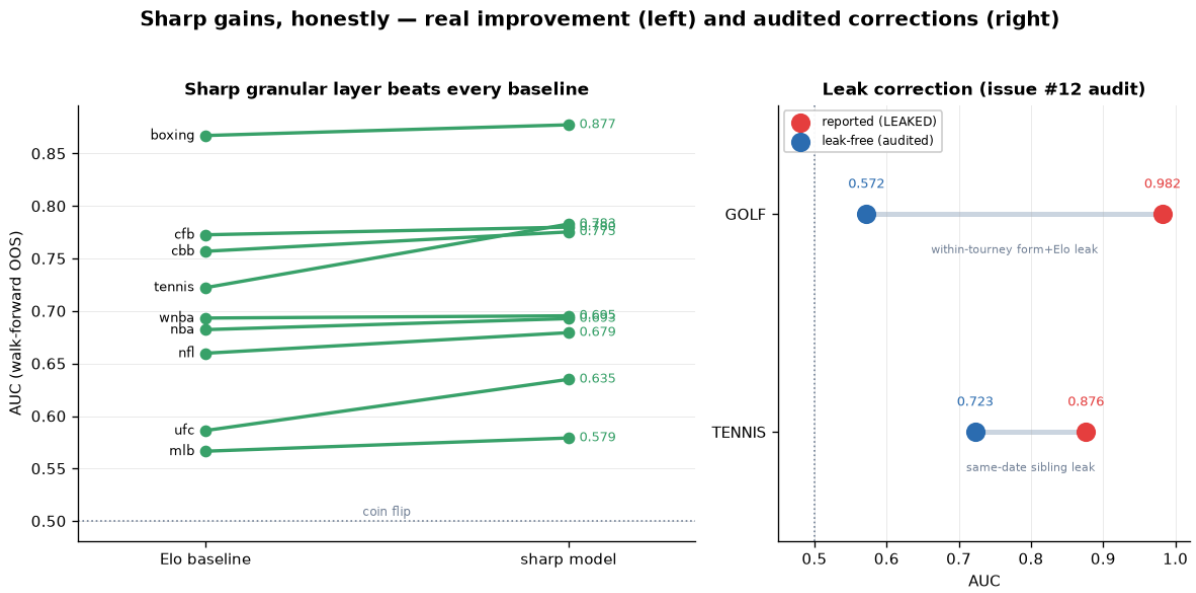


Figure 3: Sharp gains, honestly. Left: the granular layer beats every game-level baseline out-of-sample (tennis +0.06, UFC and NFL +0.02 to +0.05 AUC). Right: the two corrections the audit forced—golf 0.982 → 0.592 (within-tournament leak) and tennis sharp 0.876 → 0.731 (same-date sibling leak).

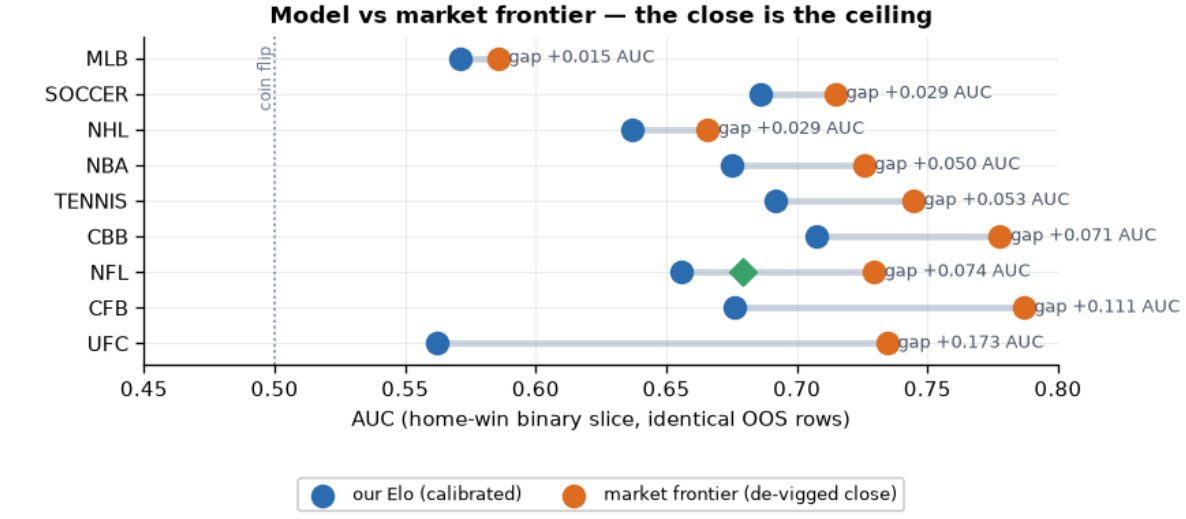


Figure 4: Model versus market frontier. For the sports carrying a closing price, our calibrated Elo and even our best sharp model sit below the de-vigged close out-of-sample. No pre-game model beats the liquid close; the gap is market efficiency, not a bug.

while affecting only 1.6% of ticks. Late-game area under the curve above 0.90 is real and honest because the score is nearly decided.

4.5 The adversarial leakage audit

Because a prior investigation had shown a golf area under the curve of 0.982 to be within-tournament leakage, every model was treated as guilty until cleared by a runnable reproduction.

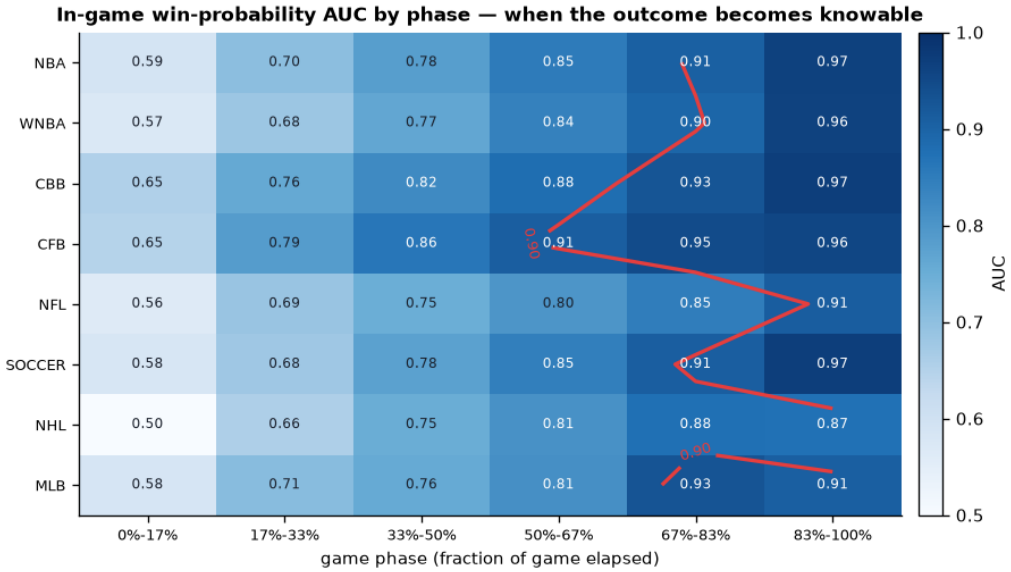


Figure 5: In-game AUC by game phase with the 0.90 contour. Late-game discrimination above 0.90 is honest (the outcome is nearly decided); no sport reaches 0.90 overall, and a pre-game figure above 0.90 would be a leakage flag.

Table 4 summarizes the verdicts. The critical finding is golf: the backbone expands each stroke-play tournament into pairwise matchups whose labels are decided by final finish, and because the feature builder updated per-competitor state immediately after each row, a later same-date row read form and Elo that already encoded that golfer's own tournament result. A direct count found 99.5% of golf rows peeking at a same-date sibling. Deferring both the Elo and form updates until the date strictly advances collapses golf to AUC 0.592, and a shuffled-label placebo returns 0.501, confirming no residual signal. The tennis sharp model carried a milder version of the same same-date leak worth about 0.036 area under the curve (0.767 $\rightarrow$ 0.731 on its own frame). The base tennis pre-game model, by contrast, is invariant to the fix (0.8656 $\rightarrow$ 0.8653) and its high discrimination is genuine skill-gap signal from a mixed field of professionals and qualifiers with an enormous Elo spread; boxing, college basketball, college football, and the in-game state model all cleared as well.

Table 4: Adversarial leakage audit, severity-ranked. "True" is the leak-free or confirmed figure reproduced live on committed code and real data.

Model / claim	Headline	Leak-free / true	Verdict
golf pre-game	AUC 0.982	AUC 0.592	LEAKED (critical)
tennis sharp (in-tourney)	sharp AUC 0.876	0.731 own frame	inflated $\sim$0.036
odds-impact tennis	beats market	coverage artifact	misleading
tennis pre-game	AUC 0.865	0.865 (invariant)	confirmed
boxing pre-game / sharp	AUC 0.866 / 0.877	0.865 leak-free	confirmed
cbb / cfb pre-game	0.774 / 0.771	point-in-time clean	confirmed
in-game state late	nba 0.967 late	group-split clean	confirmed
sharp team + ufc	+0.01 to +0.05 AUC	point-in-time	confirmed (small)

4.6 The flat-stake backtest: losing the vig, as predicted

We staked a flat one hundred dollars on every walk-forward out-of-sample model pick, settled at the real vig-inclusive closing odds, for the eight sports whose odds overlap the out-of-sample window, with a five-thousand-resample Monte-Carlo bootstrap for the confidence interval. Table 5 reports the result: a loss of $254,583 on $8,692,100 staked, a return of −2.9%. Only college football is positive, and its confidence interval straddles zero. This is the expected result. Betting the model's pick into a vig-inclusive close bleeds the margin because the close already prices what the model knows, and paying the vig to match it is designed to lose. Where a Monte-Carlo interval straddles zero, the result is indistinguishable from break-even; nowhere is it convincingly positive.

Table 5: Flat-$100 backtest into the real vig-inclusive closing line, eight sports. Monte-Carlo bootstrap (5,000 resamples) for the ROI confidence interval.

Sport	Bets	Staked	P&L	ROI	Win%	MC ROI 95% CI
cfb	2,458	$245,800	+$10,117	+4.1%	67.9%	[−3.2, +13.0]%
soccer	8,412	$841,200	−$11,026	−1.3%	47.2%	[−3.9, +1.3]%
nfl	1,817	$181,700	−$2,529	−1.4%	63.0%	[−5.2, +2.5]%
tennis	50,108	$5,010,800	−$132,201	−2.6%	65.5%	[−3.3, −1.9]%
nhl	4,265	$426,500	−$17,923	−4.2%	59.5%	[−6.6, −1.8]%
nba	7,160	$716,000	−$32,903	−4.6%	64.0%	[−6.4, −2.8]%
mlb	8,207	$820,700	−$39,749	−4.8%	54.6%	[−6.8, −2.9]%
ufc	4,494	$449,400	−$28,368	−6.3%	55.1%	[−9.1, −3.4]%
TOTAL		$8,692,100	−$**254,583**	−**2.9**%		

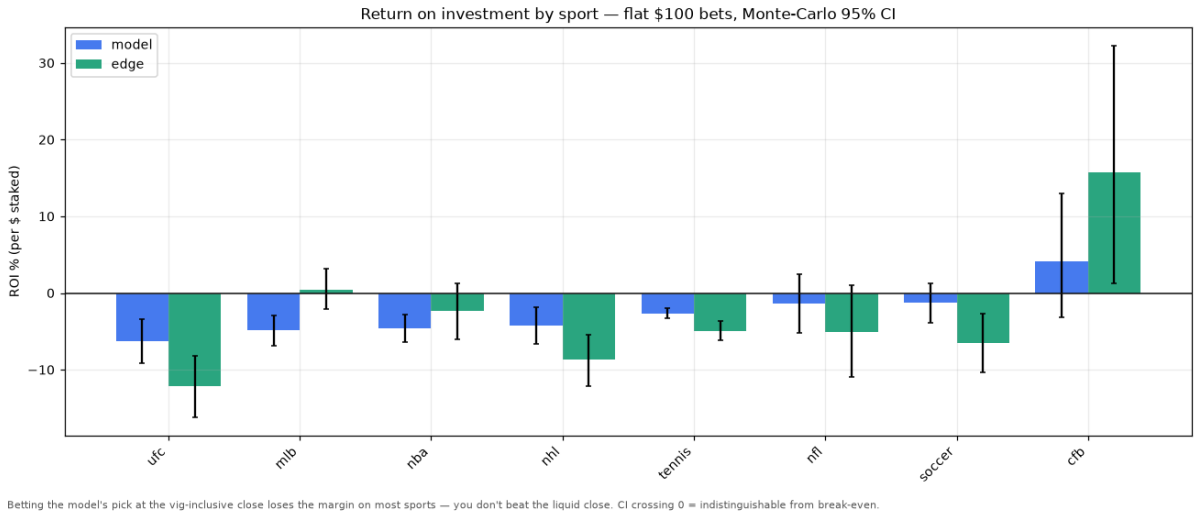


Figure 6: Return on investment by sport for the flat-$100 backtest into the vig-inclusive close. Only college football is positive and its interval straddles zero; the book-wide result is −2.9%.

A tier analysis confirms this from a different angle. Every sport has a high-accuracy conviction gate, but on the four sports where a market can be checked (soccer, NFL, NBA, tennis) the confident tier does not beat the de-vigged close on realized results—the accuracy is already priced in. On the tennis confident tier (n = 5,481), realized win rate was 86.9% against a market-implied 87.1%, a realized edge of −0.0023 (z = −0.5); on NBA the realized edge was −0.0339 (z = −2.75). No sport shows a proven positive edge against a real market. The

no-odds sports show the strongest classifier tiers—boxing's gate is 96.9% accurate over 11.6% coverage—but their profitability is simply unproven until priced against a book.

4.7 In-game, done the literature's way

The first in-game backtest bet a flat one hundred dollars per tick and, unsurprisingly, lost against the sharp live line in every phase. We redid it with the full recipe from the literature: market-calibrated win probability (calibrated to outcomes, not to the price, so as not to erase the divergence being tested), an expected-value filter swept from 0.05 to 0.30, fractional Kelly at the live implied odds with no look-ahead in the fraction, and realistic fills (a tradeable band, a decimal-odds cap, and two percent slippage and commission). The honest expectation, stated up front, is that our in-game model trails the sharp live line and calibration does not change the ordering: our model's log-loss on all ticks is 0.5203, our calibrated model 0.5134, the Vegas live line 0.4825, and the public reference 0.5137. Betting our divergence from the sharp line is betting on our own error.

Table 6 shows the payoff. Against the sharp line, Kelly staking narrows the loss relative to flat staking at every threshold and only edges positive at the most aggressive filter (+2.5% at expected-value ≥ 0.30 on 1,877 bets), with a game-clustered confidence interval that straddles zero. The decisive point is diagnostic: correct staking removes the flat-stake artifact, so the residual loss is now cleanly attributable to our model being worse than the sharp line rather than to a broken bet size. Against a *lagging* soft line the strategy can look profitable, but a dumb price-copier—whose "model" is simply the current true price—profits as much or more at every lag, so any positive return there is latency arbitrage rather than model skill. Confidence intervals throughout are bootstrapped over whole games, not ticks, because thousands of ticks in a game share one outcome and a tick-level bootstrap would fake a tight interval.

Table 6: In-game backtest against the sharp Vegas live line, late game, flat versus fractional Kelly across the expected-value filter. Kelly narrows the loss but the residual loss is a genuine model deficit; game-clustered 95% CIs straddle zero.

EV threshold	Flat \$100 ROI (n)	Kelly ROI (n)	Kelly 95% CI
0.05	−6.2% (6,171)	−3.7% (6,171)	[−12.9, +5.4]%
0.10	−4.3% (4,884)	−2.4% (4,884)	[−12.0, +7.4]%
0.15	−2.5% (3,796)	−0.8% (3,796)	[−11.9, +10.1]%
0.20	−1.3% (3,004)	−0.2% (3,004)	[−12.5, +11.7]%
0.25	−0.5% (2,355)	+1.3% (2,355)	[−11.9, +14.5]%
0.30	+1.3% (1,877)	+2.5% (1,877)	[−11.1, +16.4]%

4.8 Market-making: the one positive result

The market-making simulation covers roughly 168,926 game-orientations across seven sports for which we have the committed fair value, the realized outcome, and a held-out calibration map. Model calibration on this pool is an expected calibration error of 2.90 percentage points and mean systematic miscalibration of 3.04 percentage points—small, which is exactly why the model works as an adverse-selection filter. Under a Glosten–Milgrom flow model in which informed flow crosses only when its edge exceeds our half-spread, and retail overbets one side on the losing games, the base case (a three-cent half-spread, thirty percent toxicity, retail skew, no maker fee) earns +1.17c per contract: three cents of gross spread, minus 1.83c of adverse selection, minus zero fee. The break-even toxicity is 0.24 under pure adverse selection and rises to 0.36 with the retail cross-subsidy. Calibration does not manufacture this spread; it protects

it, because adverse selection scales with the systematic miscalibration our low calibration error keeps near zero.

We are explicit that this is a simulation whose number is an artifact of an assumed flow model. Under pure adverse selection with toxicity above 0.24 at a three-cent spread, making loses, and that red region is real and not small. Fills are modeled rather than measured—no real queue position, cancel-replace latency, or knowledge of who lifted us—and the retail cross-subsidy is drawn from another venue's data and may not hold on the specific thin sports markets we would quote. Validating it requires real order flow, per-market fee flags, and post-fill mark-outs, none of which we yet have.

4.9 The combined portfolio

Assembling the family stack, the market maker, and the in-play sleeve into one book with real joint risk metrics gives an honest verdict. None of the three sleeves is positive-expected-value on real data at realistic execution: the predictive family stack does not beat the de-vigged close on any sport (cross-sport return of −14.9%, every book negative), market-making is −0.16c per contract once queue-win-rate is a realistic thirty percent rather than a perfect-queue artifact, and the in-play win probability trails the venue's own price with a model-versus-copier interval that includes zero. What is real is the portfolio *structure*: cross-sport return correlation is about 0.08, so diversification cuts book volatility by 22%. That is risk management that manages risk; it cannot manufacture an edge that is not there.

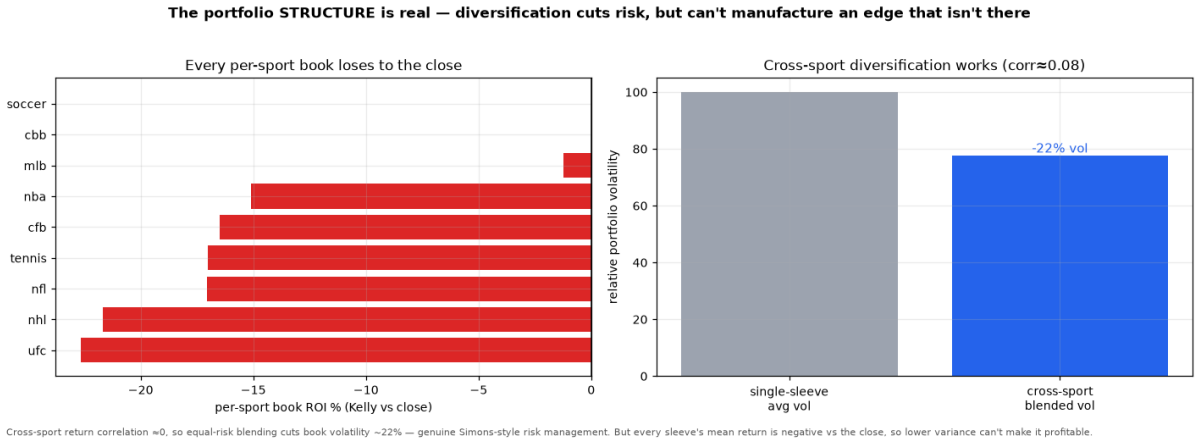


Figure 7: Portfolio diversification. Cross-sport return correlation is ≈ 0.08, so combining the sleeves cuts book volatility by 22%. Diversification reduces variance; it does not create edge.

A model-family ensemble makes the same point at the signal level. The measurable families cluster tightly within each sport because they re-express one team-strength signal—the correlation trap, in which the effective number of independent signals is far below the nominal count—so stacking them beats the best single family only marginally and never beats the de-vigged close on any sport with odds (Figure 8).

5 Discussion and limitations

The results are honest in a specific and uncomfortable direction. We can predict well: most of the sixteen sports produce calibrated probabilities with expected calibration error at or below three percentage points, the sharp layer improves every baseline, and the confident tiers and in-game phase models are genuinely above ninety percent accurate where the outcome is genuinely decided. But prediction is not the same as profit, and on every sport we can price against a

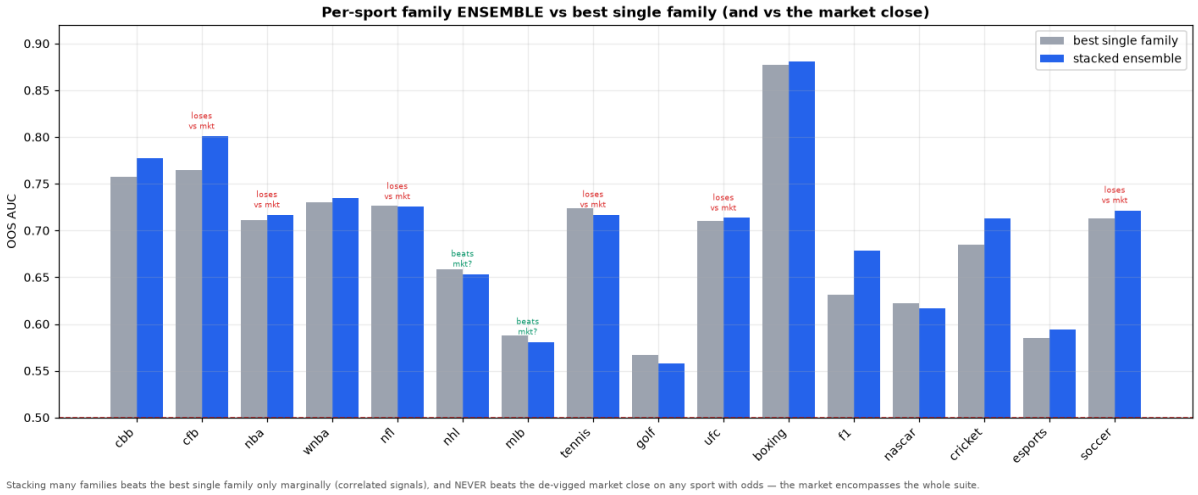


Figure 8: Ensemble versus market. The leak-free logistic stack of measurable families never beats the de-vigged close on any sport with odds; the families are calibrated fair-value estimators, best used as market-making inputs, not taker edges.

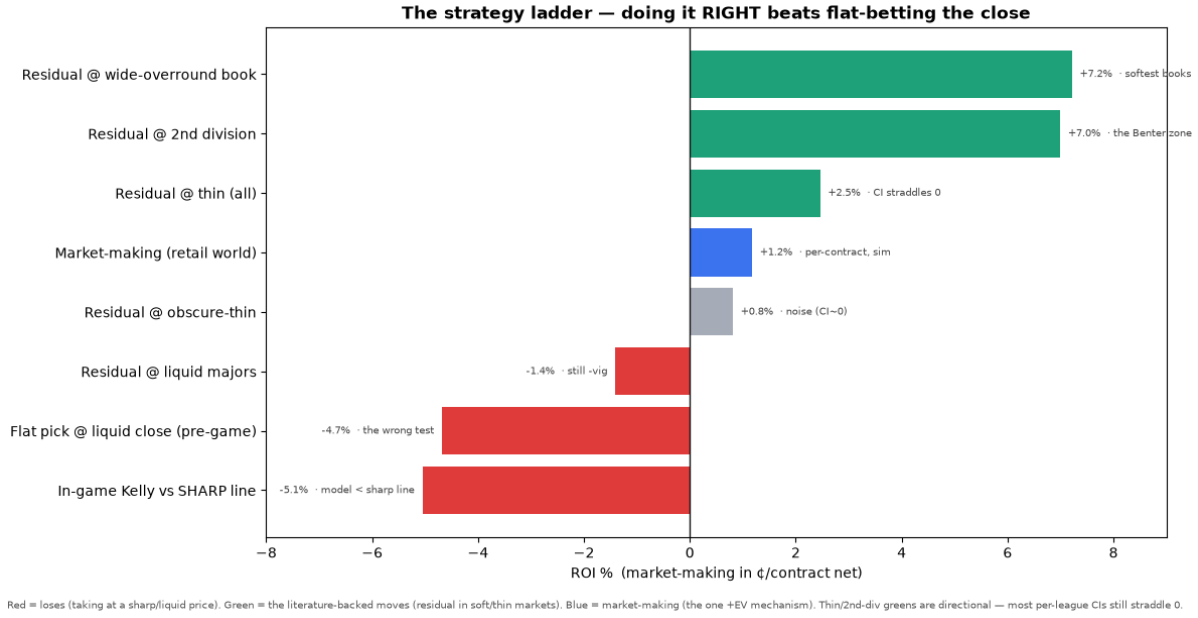


Figure 9: The strategy ladder. Flat-betting the model's pick at a liquid close loses the vig (top). Trading the model-versus-market residual in soft or thin markets and market-making around a calibrated fair value are where positive returns appear; the directional results still have per-league confidence intervals that straddle zero, and making is the one mechanism positive under plausible assumptions.

real market, we do not beat the close. The flat-stake backtest loses the vig at almost exactly the expected rate, the confident tiers are already priced in, and the in-game model—once we remove the flat-stake artifact with correct Kelly staking—is simply worse than the sharp live line.

Several limitations bound even the honest results. The frontier comparisons are mostly against single-book de-vigged closes with partial coverage—baseball at 97% but tennis at 34% and college basketball at 23%—which is a structurally easier target than a true sharp consensus

and may carry selection bias, since the games that had a posted close may be systematically easier to price. The frontier area under the curve is therefore a floor on market skill, and a model that looks competitive against a single-book close is not yet competitive against the sharp market. The sharp models' feature-selection step scores extras on the same out-of-sample folds it reports, a mild optimistic bias on the small samples (mixed martial arts near five thousand rows, women's basketball near fifteen hundred) that does not create the gains but slightly overstates them. Seven sports—boxing, cricket, esports, Formula 1, golf, NASCAR, and women's basketball—still carry no closing price, so any area under the curve we report for them is measured against realized outcomes rather than the sharpest forecast, and we cannot yet separate genuine edge from the market simply agreeing with us.

The market-making result, our one positive number, is a simulation resting on an assumed flow model, and its sign flips under high toxicity or if the retail cross-subsidy does not hold on the markets we would actually quote. The combined-portfolio verdict is that none of the three sleeves is positive-expected-value at realistic execution once the perfect-queue artifact is removed; the diversification benefit is real but manages variance rather than creating return. And an adversarial edge-hunt across five hundred and four configurations found that in-sample profitability predicts out-of-sample profitability at a correlation of -0.015: selecting a strategy because it was profitable in-sample carries essentially zero information about its live profit, which is the single most important cautionary result in the project and the reason no backtest strength alone should license capital. The many-weak-signals thesis is further bounded by correlation—rating-keyed families co-move, so the effective diversification is several times weaker than a naive count suggests, and worse in the tail when the shared rating breaks—and by capacity, since the very conditions that make a soft-market edge real (thin, retail, un-arbitraged) are the conditions that make it un-scalable.

The blocker throughout is not modeling but validation. Closing-line value, the north-star deploy gate, must be forward-collected against a real venue's close, and we have not run it: the live snapshot count is zero. Every prior apparent edge died precisely because it was subjected to this test or its historical proxy. Until a forward closing-line-value pipeline runs for weeks against real order flow and settlement, the correct capital allocation to the forecasting families is zero, and the correct expectation for the whole enterprise is a small making book plus a few un-provable niches rather than a scalable portfolio.

6 Conclusion

Across sixteen sports we built leak-free, walk-forward, calibrated win-probability models, subjected them to an adversarial audit that retracted the two headline numbers most likely to mislead, and measured them against the only benchmark that matters. The models are good: calibrated to three percentage points or better on most sports, improved on every baseline by a granular sharp layer, and genuinely above ninety percent accurate on the confident slices and the decided late-game states where such accuracy is legitimate rather than leaked. The market is better. On every sport with a real closing price, no pre-game model of ours beats the de-vigged liquid close, and flat-betting our picks into that close loses the vig at -2.9% over \$8.7 million staked, exactly as market efficiency predicts. The deployable deliverable is therefore calibration, not accuracy, because a calibrated probability converts to closing-line value even where discrimination cannot exceed the close. The residual edge, if it exists, is not in out-predicting a liquid game but in making rather than taking—where our calibrated fair value earns $+1.17$c per contract as an adverse-selection filter in simulation—and in soft or lagging venues we cannot yet price or validate. The honest wall of the closing line is real, we measured how far below it we sit, and we report where the door might be without pretending we have walked through it.