A Market-Anchored Multifactor Framework for Moneylines and Player Props

Hierarchical event generation, calibrated market residuals, robust pricing, and constrained portfolio growth

Research design and mathematical strategy paper | Prepared October 18, 2026

Status: proposed methodology, not an empirical performance claim. No historical model has been fitted, no live prices have been collected, and no out-of-sample profitability has been established in this paper. The four aggregate benchmark rows supplied as input are the only performance observations available. This is not a peer-reviewed paper.

Abstract

We formulate a sports-wagering research architecture that separates three tasks: forecasting the joint distribution of game and player outcomes; detecting mispricing relative to executable odds; and allocating risk under estimation error. The design combines dynamic team strength, player opportunity, conditional efficiency, opponent interactions, latent game environments, and timestamped market information. Its principal claim is methodological: a model should be promoted only when it improves paired out-of-sample scoring and produces positive net economic results under realistic execution. Mathematical complexity is not evidence of an edge.

The supplied benchmark is audited using within-market Brier skill. NHL and NFL show the largest proportional improvements over their respective baselines; the benchmark does not establish that soccer is the most informative or profitable market. We propose hierarchical count and possession models, a regularized market-residual ensemble, held-out probability calibration, robust expected-value filters, and joint-scenario portfolio optimization. Every performance target and staking constant below is a research hypothesis or illustrative control, not a promised outcome.

Keywords: probabilistic forecasting; market residuals; hierarchical Bayes; counting processes; calibration; Kelly growth; distributional robustness; temporal validation.

1. Audit of the supplied benchmark

Market	Player-games	Model BS	Baseline BS	Brier skill
Soccer anytime goal	50,885	0.0724	0.0745	2.82%
NHL anytime goal	47,081	0.1200	0.1258	4.61%
NBA points O/U	25,973	0.2432	0.2489	2.29%
NFL anytime TD	6,140	0.1297	0.1356	4.35%

The rows above are transcribed from the supplied benchmark data, not independently verified. Their labels do not identify the season, player eligibility universe, baseline construction, information cutoff, settlement rules, uncertainty intervals, or whether the predictions were genuinely out of sample. Source: provided benchmark data.

```
BS = (1/n) sum_i (p_i - y_i)^2
BSS = 1 - BS_model / BS_reference
```

These are binary-event scores on the single-event scale. Multiclass Brier scores use a different convention, so soccer three-way moneyline scores must not be compared directly with these binary prop values. Proper scores evaluate forecast distributions, but comparisons are directly interpretable only on the same forecast situations. Gneiting and Raftery (2007), Strictly Proper Scoring Rules, Prediction, and Estimation.

An 8% event forecast at the correct constant base rate has expected Brier score 0.08(0.92) = 0.0736, without any player discrimination. A balanced binary event forecast at 50% has expected score 0.25. Therefore a soccer score of 0.0724 does not, by itself, imply a stronger model than NBA points at 0.2432. Within the supplied rows, proportional gains are NHL 4.61%, NFL 4.35%, soccer 2.82%, and NBA 2.29%; they still cannot rank economic profitability because the baselines differ. Calculations from the benchmark data.

```
BS = reliability - resolution + uncertainty
uncertainty = y_bar * (1 - y_bar)
```

The decomposition separates calibration error, discrimination across probability strata, and inherent event uncertainty. For empirical binned decompositions, binning introduces approximation when forecasts differ within bins; report the binning scheme. A lower total score is not synonymous with better calibration alone. Proper-score decomposition literature.

A fair claim of beating the benchmark requires its row-level outcomes, forecasts, and evaluation cohort. A new model scoring below a benchmark number on another season or player universe is not a head-to-head victory. The benchmarks contain no moneyline benchmark.

2. Objectives, estimands, and the central distinction

Let F_t be the information actually available at decision time t; Y_i the contract outcome; O_i the executable decimal payout including returned stake; and c_i a conservative deterministic cost allowance per unit stake. Decimal odds are O = 1 + A/100 for positive American odds A, and O = 1 + 100/|A| for negative A.

```
p_i = P(Y_i = 1 | F_t)
R_i = O_i * Y_i - 1 - c_i
E[R_i | F_t] = O_i * p_i - 1 - c_i
p_break_even = (1 + c_i) / O_i
```

This payout formula assumes win/loss settlement with no push or void. In a three-state contract, EV = p_win*(O-1) - p_loss - c if the cost applies to every submitted stake, including pushes. Actual fees can have state-dependent rules; implement the real payoff vector rather than blindly applying a cost constant. The embedded bookmaker margin is already reflected in O and must not be subtracted again.

Forecast quality and betting value are distinct. A highly accurate 80% forecast loses expectation at decimal odds 1.20 because 1.20*0.80 - 1 = -4%. A 40% forecast is favorable at 2.70 because 2.70*0.40 - 1 = 8%. Improving average Brier score does not ensure improvement among the specific wagers selected at the quoted prices. These are hypothetical arithmetic examples, not market findings.

Pre-register two co-primary objectives: paired predictive improvement against a strong contemporaneous reference, and positive net portfolio return after execution costs. Logarithmic growth, downside risk, and capacity are secondary economic objectives; win rate alone is not an objective.

3. Data design: prevent information and settlement leakage

Use a bitemporal database: event time records when a fact occurred; availability time records when the system could have known it. Join on availability time. A corrected injury label, postgame minutes figure, or revised lineup must not

overwrite the historical forecast snapshot.

Data layer	Minimum fields and restrictions
Odds and execution	Book, timestamp, market ID, side, threshold, decimal price, available limit, quote age, accepted price, rejected orders, and settlement policy. Save every eligible contract, including no-bets.
Team and schedule	Opponent-adjusted efficiency; rest; travel; venue; surface; altitude; weather; expected starting lineup; lineup announcement timestamps.
Player opportunity	Minutes, shifts, routes, carries, targets, shot opportunities, line or unit assignment, and role uncertainty. Do not use realized exposure as a pregame feature.
Player efficiency	Shot quality, finishing, shot mix, conversion, foul or penalty rates, and opponent interaction terms; shrink small samples toward role-level priors.
Targets and rules	Final statistics and original bet settlement; overtime, shootout, starting-player requirements, DNP, own goals, passing-TD exclusions, pushes, postponements, and voids.

Capture the full eligible player universe rather than only players who eventually participated. Correct lineup/void handling is part of the contract, not a post hoc deletion rule. Missing historical odds are not missing at random: rare or weakly covered players may be systematically absent. Limit conclusions to the universe with genuinely archived, executable markets.

Market inputs should come from timestamped price feeds, not promotional pick articles. Sportsbook prices are necessary execution data even when independent event statistics drive the model. There is no defensible historical betting return without historical thresholds, both-side prices where applicable, and realistic acceptance constraints.

4. Build a market reference without treating it as truth

```
u_bk = 1 / O_bk
q_bk^(norm) = u_bk / sum_j u_bj
Power alternative: q_bk = u_bk^a_b, with sum_k u_bk^a_b = 1
```

Normalize each complete, mutually exclusive market inside each book before combining books. The power parameter a is solved separately for the market. Compare normalization and power adjustment prospectively; normalization is transparent but does not model favorite-longshot distortions. Clarke, Kovalchik, and Ingram (2017), Adjusting Bookmaker's Odds to Allow for Overround.

For NBA over/under, the two sides must refer to the same player and threshold. For soccer regulation moneylines, include home, draw, and away. An anytime-goal board listing different players is not a partition: several players can score, so never normalize those probabilities to sum to one. Use a same-player yes/no pair if available, a separately validated margin model, or explicitly mark the reference as unavailable.

```
logit(q_i) = sum_b w_b * logit(q_bi), sum_b w_b = 1
q_ik = exp(sum_b w_b log(q_bik)) /
       sum_j exp(sum_b w_b log(q_bij))  [multiclass]
```

Estimate weights using earlier data and quote quality, not future outcomes. Avoid counting copied books as independent opinions. Construct a leave-execution-book-out reference where possible; otherwise label the overlap. Staleness, disagreement, and quote age are features and uncertainty signals, not automatically exploitable errors. Execute at the actual offered O, not at a de-vigged price.

5. Shared multifactor state-space model

```
theta_team,t = rho * theta_team,t-1 + eta_team,t
eta_team,t ~ Normal(0, Q)
log(lambda_team,g) = alpha + attack_team,t
  - defense_opp,t + home_effect + beta^T X_g
X_g = {lineup, rest, travel, weather, pace, matchup}
```

Center attack and defense states within each league to identify the model. Pool home effects, player effects, and interaction coefficients with hierarchical priors. Estimate dynamic-state filtering only from the past, with smoothing reserved for training diagnostics rather than live feature construction.

```
beta_j ~ Normal(0, tau_j^2)
player_effect_i ~ Normal(role_mean_i, role_sd_i^2)
Z_g ~ scenario law for pace, game state, lineup, weather
```

The latent environment Z_g induces dependence between moneylines and props. It should include uncertain availability and game script, not merely a single projected total. Posterior predictive simulation must draw parameter states, lineup/exposure uncertainty, and event randomness in separate layers. Separate epistemic uncertainty about p from ordinary event variance p(1-p).

Use a small, regularized interaction set initially: rest by rotation depth; player usage by absent teammates; shot location by opponent defense; route role by coverage; and weather by passing or kicking exposure. Add nonlinear learners as challengers only after a transparent baseline survives time-ordered validation. A factor is not valuable simply because it is intuitively plausible.

6. Moneylines from settlement-consistent game distributions

Soccer: use a dynamic goal model with a low-score dependence correction. Dixon and Coles (1997) developed a Poisson-regression framework with dynamic team performance and tested a historical betting application. That historical result does not establish present profitability.

```
P(G_H=x,G_A=y) = tau(x,y;lambda_H,lambda_A,rho)
                      * Pois(x;lambda_H) * Pois(y;lambda_A)
p_H = sum_(x>y) P(x,y); p_D = sum_(x=y) P(x,y)
p_A = 1 - p_H - p_D
```

Use the defined Dixon-Coles correction, constrained to yield nonnegative probabilities, or another explicitly normalized joint score law. Regulation soccer moneylines require the draw state. Draw-no-bet and qualification markets have different payoff maps and cannot reuse the same binary target.

```
NHL p_home,full = P(H>A after regulation)
                 + P(H=A after regulation)*p_home,OT/SO
NBA/NFL p_home = E[1{home wins under contract rules}|F_t]
```

For NHL, model overtime and shootout win mechanisms separately, and exclude shootout scoring from a player prop when its rules require that. For NBA, simulate possessions and scoring rather than forcing soccer-like counts. For NFL, simulate drives, scoring type, and overtime/tie rules; regular-season tie treatment must match the particular contract. Derive both game and player quantities from a coherent shared event path.

A robust challenger is a market-anchored team classifier using the same decision-time information. The structural model is valuable only if it adds calibrated information beyond that reference. Report results separately by sport and market, not as a pooled win-rate headline.

7. Anytime goals: separate opportunity and conversion

```
N_i | M_i,Z_g ~ NegBin(mean=M_i*r_i(Z_g), size=k_i)
log r_i = player_shot_effect + role + opponent + beta^T X
C_ij ~ Bernoulli(s_ij)
logit s_ij = shot_quality_ij + finishing_i + goalie_opp
G_i = sum_(j=1 to N_i) C_ij
```

M_i denotes uncertain exposure, such as minutes or time on ice. C_ij is the conversion of a generated shot. In pregame forecasting, future shot locations and quality are random draws from a conditional shot-quality distribution; using their realized values would leak the outcome process. Include power-play role, line partners, starting probability, substitutions, and empty-net context as sport-specific covariates.

```
If G_i|Lambda_i ~ Poisson(Lambda_i):
p_i(anytime) = 1 - E[exp(-Lambda_i) | F_t]
If Lambda_i ~ Gamma(k, scale=mu/k):
p_i(anytime) = 1 - (k/(k+mu))^k
```

Do not replace the first expression with 1-exp(-E[Lambda_i]) unless intensity is effectively fixed. Since exp(-x) is convex, Jensen's inequality gives 1-E[exp(-Lambda)] <= 1-exp(-E[Lambda]); an uncertain-intensity plug-in can overstate scorer probabilities at the same mean intensity. This is a mathematical implication of the specified mixture, not evidence that every real scorer market is overpriced.

Independent player count models can yield incoherent team totals. For production joint simulation, generate team scoring opportunities first and allocate eligible scorer identities using hierarchical categorical shares with a residual/own-goal category where needed. Alternatively use a joint marked event process. Validate marginal anytime probabilities after allocation, not only before it.

Benchmark-specific hypotheses: test whether low-xG players have positive signed residuals y-p after conditioning on minutes and role; compare finishing-heavy and exposure-heavy variants. For NHL goalie effects, compare no-goalie, pooled-goalie, and interaction variants. Keep a goalie feature only if it improves held-out predictions or portfolio outcomes under the pre-registered criterion. The benchmark watch-outs are hypotheses supplied as input, not independently confirmed findings.

8. NFL touchdowns: hierarchical scoring opportunity allocation

```
D_g | Z_g ~ team drive distribution
K_d | Z_g ~ {no score, FG, rush TD, pass TD, other}
s_i,d,type = exp(v_i,d,type) / sum_j exp(v_j,d,type)
p_i(TD) = 1 - E[ product_d (1-a_i,d) | F_t ]
```

a_i,d is the conditional chance that player i scores a qualifying TD on drive d. The product form assumes conditional independence of drive-level scoring indicators given the simulated state; otherwise estimate the probability directly from complete simulated paths. Allocate rushing and receiving touchdowns separately. Include non-offensive touchdowns only for markets that count them.

Features should represent goal-line carries, route participation, end-zone targets, short-yardage substitutions, team touchdown opportunities, injury uncertainty, and script-dependent playing time. A quarterback passing touchdown is not a quarterback anytime touchdown unless the contract explicitly says otherwise. The eligible scorer identities must sum consistently within each generated team scoring event.

The lead-running-back overpricing claim should be tested against price and event residuals, not applied as a blanket fade. A back can be overpredicted by a model but still be underpriced by an unusually generous book. The relevant decision is posterior scoring probability versus executable break-even probability. Source of the proposed watch-out: supplied benchmark.

9. NBA points: simulate the full scoring distribution

```
M_i ~ minutes distribution conditional on availability/script
N_i ~ possession/opportunity law conditional on M_i and pace
(A_2,A_3,A_FT) ~ coherent opportunity and shot-type process
X_2|A_2,pi_2 ~ Binomial(A_2,pi_2)
X_3|A_3,pi_3 ~ Binomial(A_3,pi_3)
X_FT|A_FT,pi_FT ~ Binomial(A_FT,pi_FT)
Points_i = 2*X_2 + 3*X_3 + X_FT
p_over(L) = P(Points_i > L | F_t)
```

Use jointly generated team possessions, on-court lineups, player shot shares, and foul/free-throw sequences. Free throws are not simply an independent third shot type competing one-for-one with field-goal attempts. Random shooting probabilities induce beta-binomial-like dispersion if a beta mixing law is appropriate. Include foul trouble, blowouts, overtime, teammate absences, and substitution uncertainty in the exposure model.

A projected mean of 25 points is not enough to price over 24.5. Two distributions can share that mean and have different threshold probabilities. Test tail coverage, count-distribution log score, and Brier score at archived actual lines. One fitted points distribution should produce monotone probabilities across alternate lines: if L1 < L2, then p_over(L1) >= p_over(L2).

For the benchmark's extreme-probability concern, first inspect held-out reliability by probability band, line, player role, and exposure uncertainty. Do not impose arbitrary global shrinkage solely to make a chart look calibrated. Each calibration map must be learned without touching the final evaluation block. Source of the concern: benchmark data.

10. Market-residual learning and probability calibration

```
s_i = structural forecast probability
logit(p_raw,i) = logit(q_i)
  + gamma_m * [logit(s_i)-logit(q_i)] + h_m(X_i)
h_m = regularized residual learner; m = market family
```

This parameterization nests the market reference when gamma and h are zero. The model must demonstrate incremental information rather than receive an arbitrary conviction weight. Train h jointly with strong regularization; use chronological out-of-fold structural predictions to avoid training on fitted probabilities. Correlated market and fundamental information is not new evidence twice.

```
p_cal = sigmoid(a*log(p_raw) - b*log(1-p_raw) + c)
a >= 0, b >= 0; identity map: a=b=1, c=0
```

This monotone beta-calibration family can handle deviations that a simple logistic rescaling cannot. Clip only for numerical stability using a fixed epsilon. Fit on a distinct chronological calibration window; compare identity, logistic/temperature, and beta maps using earlier inner-validation folds, not the final test. Kull, Silva Filho, and Flach (2017), Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers.

Use market-specific calibrators with partial pooling when sample sizes are small. For a three-way moneyline, use a simplex-preserving multiclass calibration map rather than calibrating each outcome independently without

renormalization. Report performance in both the full forecast universe and the selected bet subset.

11. Robust selection: only wager when the price survives uncertainty

```
p_lower,i = conservative lower bound for latent p_i
e_lower,i = O_i*p_lower,i - 1 - c_i
Eligible_i = 1{e_lower,i > delta_m} * quality_gates_i
```

A posterior percentile is a model-conditional credible bound, not a guaranteed frequentist coverage statement and not a guarantee against misspecification. Validate its behavior on held-out data; add stress scenarios for lineup, exposure, and calibration drift. Do not use a binary outcome quantile as a probability confidence bound. Delta_m is a pre-registered reserve for unmodeled execution and estimation risk; tune it only inside training/validation.

Quality gates should reject stale or unmatched quotes, unresolved player eligibility, missing decisive inputs, unacceptable lineup uncertainty, inadequate executable limits, and contracts outside the model's support. An apparent edge generated by a data mismatch is not a bet. Abstention is an explicit action.

```
O_min = (1 + c_i + delta_m) / p_lower,i
```

This is a usable price floor for binary settlement. Recompute after every odds change. A selection can be attractive at +350 and unattractive at +275 without any change in player quality. Separate quote selection from acceptance: a rejected order does not enter the realized return series, but its rejection must remain in the execution audit.

12. Worked pricing examples: hypothetical, not forecasts

Contract	Odds	p / p lower	Cost	EV / lower EV	Quarter robust Kelly
Moneyline	2.1000	0.500 / 0.485	0.5%	4.50% / 1.35%	0.3067%
Anytime scorer	4.5000	0.240 / 0.225	0.5%	7.50% / 0.75%	0.0534%
NBA over	1.9091	0.550 / 0.530	0.3%	4.70% / 0.88%	0.2426%

Each row assumes no push, a deterministic per-stake cost, and no correlation with other holdings. These are invented inputs solely to demonstrate the equations. The scorer's 7.5% central estimated return falls to 0.75% under its lower probability bound; large displayed upside is not the same as a robust edge. Any positive reserve delta above 0.75% would reject that example.

```
For p_lower and binary return {O-1-c, -1-c}:
f_K = max(0, (O*p_lower - 1 - c) /
              ((O-1-c)*(1+c)))
f_illustration = 0.25 * f_K, before portfolio constraints
```

The formula follows by differentiating p_lower*log(1+f*(O-1-c)) + (1-p_lower)*log(1-f*(1+c)). It requires O-1-c > 0 and feasible nonnegative wealth in each outcome. With c=0 it reduces to max(0,(O*p_lower-1)/(O-1)). Quarter-Kelly plus a lower bound is intentionally conservative; it is a design choice, not a theorem giving an optimal universal multiplier.

13. Portfolio construction: correlated wagers are one risk system

```
R_s = vector of net unit-stake returns in joint scenario s
wealth_s(f) = 1 + R_s^T f
maximize_f  min_(w in U) sum_s w_s*log(wealth_s(f))
subject to f >= 0; wealth_s(f) >= epsilon
           sum_i f_i <= exposure_cap
           A*f <= game/player/team/market caps
```

Use joint paths from the same game simulator. A moneyline, team total, scorer prop, and opposing-player under can all load on one game-script factor. Do not estimate portfolio risk by treating those wagers as independent. The ambiguity set U can be a calibrated neighborhood of scenario probabilities, such as KL(w||w0) <= r; its radius must be stress-tested on earlier blocks, not chosen to maximize final-test profits.

Sun and Boyd, Distributional Robust Kelly Gambling (manuscript dated December 2018), formulate worst-case log-growth optimization over uncertain outcome distributions, with tractable finite-outcome cases. This motivates the risk framework but does not demonstrate a sports-market edge.

```
Small-exposure approximation:
E[log(1+R^T f)] ~= mu^T f - 0.5*f^T E[RR^T]*f
E[RR^T] = Cov(R) + mu*mu^T
```

The second-order matrix is the second moment, not merely covariance; ignoring the mean outer product is an additional approximation. Use the exact scenario log objective when feasible. Shrink noisy correlation estimates and include unobserved shared-driver stress scenarios. Single-bet Kelly calculations are screening tools, not final allocations.

```
L_s = -R_s^T f
CVaR_alpha(L) = min_z [z + (1/(1-alpha))*
                       sum_s w0_s*max(L_s-z,0)]
```

A one-period CVaR limit controls modeled tail loss, not full-path maximum drawdown. Evaluate bankroll paths separately with adverse sequences and regime shifts. Illustrative pilot limits might cap any one bet at 0.25% of a ring-fenced research bankroll, a game at 0.75%, and total unsettled worst-case stake at 2%; these are conservative design placeholders, not personalized staking recommendations. The bankroll can still suffer substantial losses.

14. Parlays: price the joint event, not the excitement

```
p_joint = P(Y_1=1,...,Y_K=1 | F_t)
EV_parlay = O_parlay*p_joint - 1 - c_parlay
p_joint = product_j p_j only under independence
```

Use shared-path simulation for same-game combinations and for common league-wide drivers. Estimate the joint event directly or apply a separately validated dependence model; an arbitrary correlation multiplier is not a probability model. A same-game price can already charge heavily for positive dependence, so correlation itself is not value.

Even independent positive-EV legs can produce high-loss-frequency tickets and poor growth at an oversized stake. For scale, ten consecutive losses have probability 0.76^10 = 6.43% when each independent bet has a 24% win chance. This is a fixed sequence calculation, not the probability of at least one such streak over an entire season. Retain singles as the primary research portfolio; allow a parlay only when its actual joint payout survives the same robust EV, capacity, and concentration tests.

15. Experimental design: what would establish improvement?

Use nested, rolling-origin evaluation. Outer test blocks are chronological and untouched; inner blocks select model structure, regularization, calibration, thresholds, and risk controls. Keep all quotes, alternate lines, and props for a game inside the same outer split. Historical data from before the test game can legitimately appear in training even when it involves the same teams or players.

Purge training observations whose outcome or information availability overlaps the test decision time. Apply a justified embargo for delayed reporting or label horizons, not an arbitrary long embargo merely because rolling features exist. Fit preprocessing, imputers, feature selection, team states, and calibrators inside each fold.

```
d_i = (p_new,i-y_i)^2 - (p_ref,i-y_i)^2
H0_predictive: E[d_i] >= 0
H1_predictive: E[d_i] < 0
```

Use paired resampling by game and sufficiently long calendar blocks to preserve within-game and temporal dependence. Estimate intervals for the mean score differential; report sensitivity to block length and the uncertainty associated with model selection. If many markets or variants are tested, pre-register multiplicity control or a hierarchy of confirmatory tests. Never select the best sport after seeing its final-test return and call that an unbiased result.

Benchmark against: a past-only base-rate model; a regularized fundamental model; timestamp-matched de-vigged consensus; structural models; market-residual models; and calibrated full models. Closing consensus is an informative hindsight comparator or diagnostic, not an input to earlier bets. Reproduce the supplied engine on identical events if its row-level predictions become available.

```
ROI_net = sum_i stake_i*R_i / sum_i stake_i
CLV_log,i = log(O_accepted,i*q_close,i)  [binary, no push]
```

The CLV expression values the accepted payout at a specified de-vigged closing probability. Define closing time, source composition, rules, and margin removal consistently; it is not realized profit or proof the closing market is correct. Report stake-weighted and unweighted results, fixed-stake results before optimized staking, order fill rates, quoted versus accepted prices, maximum drawdown, tail loss, and capacity.

Run a fixed latency scenario and adverse price/slippage tests; enforce recorded limits. No cherry-picked best-price execution unless the price was available and executable at that timestamp. Include pushes and voids under actual contract rules. State whether returns are pre-tax or after jurisdiction-specific tax effects; the worked examples here are pre-tax.

Record every attempted model and strategy configuration. Bailey, Borwein, Lopez de Prado, and Zhu, The Probability of Backtest Overfitting, explain the danger of selecting a best-performing strategy from many historical trials. Nested time tests and a final forward paper portfolio remain necessary; a correction statistic does not replace them.

16. Statistical power and meaningful targets

```
n_eff ~= ((z_(1-alpha)+z_(power))*sigma_R / mu_R)^2
```

For a one-sided 5% test, 80% power, independent approximately even-money unit-stake returns with standard deviation near 1, and a true mean return of 2%, the normal approximation gives roughly 15,457 wagers. This is an illustrative power calculation, not a universal requirement; longshot variance, correlated exposures, adaptive selection, and multiple testing can require materially more evidence. Observations counted as player-games are not the same as independent bets or game clusters.

For Brier improvement, required effective sample size depends on the variance of paired d_i and its dependence structure. The benchmarks do not supply that variance, so their sample counts cannot establish significance. A prospective target could be a 5% relative reduction from each model score, but no such reduction has been achieved here:

Market	Benchmark BS	Illustrative target: 0.95 x BS
Soccer anytime goal	0.0724	0.068780
NHL anytime goal	0.1200	0.114000
NBA points O/U	0.2432	0.231040
NFL anytime TD	0.1297	0.123215

These targets have meaning only for matched events or carefully specified common evaluation distributions. Primary promotion gates should instead require a statistically supported negative paired score difference and economically meaningful positive net return under pre-registered execution assumptions. If results are inconclusive, the correct result is inconclusive, not a higher-confidence staking plan.

17. Ablations, diagnostics, and implementation sequence

Experiment	Question and rejection condition
Exposure uncertainty	Does integrating minutes/role improve scoring and tail reliability versus plug-in exposure? Reject extra complexity if it fails held-out tests.
Goalie / finishing effects	Do pooled effects help beyond shot quality and team defense? Remove redundant effects rather than keeping them for narrative appeal.
Market anchor	Does any fundamental residual survive beyond the same-time consensus? A market-only model is a serious benchmark, not an embarrassment.
Extreme NBA probabilities	Does held-out calibration improve log score and selected-bet behavior without hiding opportunity errors? Inspect probability tails with uncertainty intervals.
Joint simulation	Does dependence modeling change portfolio drawdown or parlay prices under independent validation? Test shared game factors and alternate specifications.
Execution stress	Does estimated profit survive quote delays, realistic limits, rejections, worse fills, and conservative costs? If not, reject the trading strategy.

Implementation sequence: (1) archive rules, quotes, and availability-time features; (2) reproduce simple baselines and validate settlements; (3) build sport-specific generative models; (4) fit market residuals and held-out calibration; (5) produce immutable walk-forward forecasts; (6) evaluate price selection with fixed stakes; (7) add conservative portfolio optimization; (8) operate a forward paper portfolio; (9) consider a small ring-fenced pilot only after the pre-registered evidence and operational controls pass.

```
for each historical decision time t:
    data = records with availability_time <= t
    predict joint outcomes using models trained before t
    calibrate using only earlier held-out forecasts
```

```
join exact contract, rules, threshold, and accepted price
stress probabilities and compute conservative EV
choose no-bet or constrained joint portfolio
freeze forecast, selection, rejected orders, and limits
settle after the event; never rewrite frozen records
```

Unit tests: probabilities sum to one on each mutually exclusive market; alternate-line tails are monotone; team and player event counts reconcile; contract exclusions settle correctly; there are no impossible lineup exposures; simulated dependence is stable under larger samples; folds contain no same-game quote leakage; no realized minutes or future closing prices enter pregame features. Validate simulation convergence before interpreting small probability differences.

18. Conclusion and evidentiary boundary

The plausible source of sustainable value is a small amount of incremental information about opportunity, conditional efficiency, and market execution that survives stringent validation. Dynamic multifactor mathematics can organize that information and reduce specific modeling errors. It cannot create information, eliminate bookmaker margin, guarantee acceptance at stale prices, or convert an overfit backtest into a durable edge.

This paper provides a falsifiable design to challenge the supplied benchmark, not a demonstrated replacement for it. A genuine claim of beating that engine requires matched row-level forecasts; a genuine claim of beating oddsmakers requires actual timestamped prices, settlement-consistent returns, realistic execution, and independent forward evidence. Until those exist, the correct operating decision is research and paper validation, not assumed profitability.

Research risk boundary: use only legal markets and a discretionary, loss-limited research budget; never borrow to chase a modeled edge. Stop or suspend a model when data integrity, calibration, or execution assumptions fail. No staking rule removes the possibility of loss.