NHL Forecasting and Market Pricing: Chronological Evaluation of Coherent Score Models and Opportunity-Based Player Distributions

Fourth & Value Research Group

fourthandvalue.com

September 26, 2026 · Version 1.1 · Model nhl–v2.1

Technical report; not peer reviewed.

ABSTRACT

We evaluate a replacement NHL forecasting and pricing system using 5,248 regular-season games and 188,883 skater appearances from 2022–23 through 2025–26. Two expanding development folds select an ℓ_2-regularized Poisson team model and opportunity-based negative-binomial player distributions. A common settlement-score distribution generates moneyline, puck-line, and total probabilities; binomial allocation of latent total points preserves the identity goals + assists = points. Final-season joint-score negative log likelihood improves from 3.68481 to 3.68073, but its paired game-bootstrap interval includes zero, and moneyline log loss worsens slightly. Player count log scores improve across all four markets, with the largest absolute reduction for shots (−0.012139 nats per appearance). A separately acquired, fixed monthly sample of 130 games supplies timestamped sportsbook comparisons. Every paired model-versus-market log-loss interval includes zero; a frozen price-selection policy produces only three final-season shadow selections and loses 1.5238 units. Historical publication timestamps are unavailable, player evaluation conditions on participation, and a disclosed feature-correctness repair required rerunning the final period. The evidence supports a more auditable forecasting and price-comparison system, not a demonstrated betting advantage. Market blending and validated recommendations remain disabled.

Keywords: NHL; probabilistic forecasting; negative binomial; chronological validation; market consensus; settlement; point-in-time data; calibration.

1. Introduction: What the Findings Mean

A useful hockey model has to answer two different questions. First, how likely is an outcome, such as a player taking at least three shots? Second, does the price offered by a sportsbook pay enough for that chance? Better answers to the first question do not automatically produce profitable answers to the second. Sportsbooks already incorporate substantial information, and their prices include a margin.

We rebuilt Fourth & Value's NHL system to make those questions measurable. The new system uses earlier games to forecast later games, compares several relatively simple models, and keeps the underlying hockey forecast separate from sportsbook opinion. It also accounts for a push, when an integer betting line exactly matches the result and the stake is returned. Analysts can see the forecast, the available price, and the assumptions that could make the calculation unreliable.

The strongest finding is modest: separating a player's expected ice time from production per minute improved predictions of shots, goals, assists, and points compared with our simple historical-rate baseline. That does not tell us that the model beats a sportsbook. Our much smaller sample of historical sportsbook prices could not establish such an advantage. The three selections produced by a fixed experimental policy in the final season lost money; that sample is too small to support a reliable estimate of future profits or losses.

For readers, the practical result is a research tool, not an automatic betting card. It can help identify prices worth investigating, explain a model's disagreement with a market, and record what an analyst knew at the time. It can also correctly return no recommendation. The technical sections below specify exactly what was implemented, what improved, and what remains unproven.

2. Data, Information Sets, and Inherited Defects

2.1 Observation units and source coverage

The numerical study uses public NHL per-game team and skater summaries, archived as 44 source-response pages with content hashes. Each of four regular seasons contributes 1,312 games. There are 188,883 skater appearances in total and 47,230 in the final season; player rows require positive recorded time on ice. Playoffs and preseason are excluded. Stable NHL game and player identifiers drive joins and grading. Monthly skater partitions prevent the source endpoint's report-size cap from silently truncating a full-season request. The implementation and retained data manifests, rather than names or filenames alone, define the sample [1], [2].

The separate market dataset contains 21 historical snapshots: the 15th of October through April, at 10:30 America/New_York, in each season from 2023–24 through 2025–26. The dates were fixed before acquisition and cover 130 regular-season games. Existing authorized access supplied these snapshots; the study did not purchase an upgrade. The provider exposes historical snapshot and bookmaker update times, which are retained alongside local ingestion times [3]. There is no historical player-price sample, full daily price panel, or closing-price panel in this study.

2.2 Reconstructed and observed information sets

Let t_g denote a decision cutoff for game g, and let a_j denote an observation's availability time. A feature is measurable with respect to the modeled information set only when its inputs satisfy:

$$\mathcal{I}(t_g) = \{x_j : a_j \le t_g\}. \tag{1}$$

For historical box scores, original publication and revision times are absent. We therefore assign a reconstructed availability time of 12:00 UTC on the day following the recorded game date. Features for a date use the 10:30 Eastern cutoff and cannot consume that date's outcomes. Event date, assigned availability, actual ingestion, and publication time are distinct fields; unavailable publication times remain null. This enforces a declared lag, but it cannot recover the exact statistical values originally visible to an analyst. Consequently, the experiment is a *reconstructed predictive evaluation*, not a certified replay of historical data vintages.

The whole-season statistics also lack original puck-drop timestamps. Their date-level morning evaluation cannot certify whether an unusually early international game had already begun. Historical price comparisons and live inference separately use observed start times and exclude games already started. Within a held-out season, completed outcomes may enter rolling histories after their availability cutoff, while fitted regression, dispersion, and overtime parameters remain frozen. This is sequential forecasting with updating covariates, not season-final feature construction.

2.3 Audit findings and unavailable context

The inherited totals feature pipeline used full-sample opponent/home-away averages, and its training path selected same-game shots, points, power-play goals, and goalie save percentage. Those variables could reveal the target game. The inherited player path predominantly used historical rates, Normal/Poisson assumptions, calibration toward consensus rather than realized outcomes, and a default 0.5 probability for an unknown player. The contemporary public board displayed historical references; it did not itself demonstrate an ML betting edge. None of those artifacts or performance claims provides evidence for this study. The five-entry NHL betting ledger also has inconsistent labels and is not used as a validation sample [4].

No timestamped historical starter announcements, line combinations, power-play units, injuries, scratches, or analyst narratives are reconstructed from final-game records. Expected goals, explicit travel distance, teammate effects, and coaching changes are absent. Aggregate team save rate is evaluated only as a historical context proxy; it is not a starting-goalie talent estimate. Missing source coverage limits the model specification and is disclosed rather than replaced by hindsight.

3. Team Scoring and Coherent Settlement

3.1 Recency weighting and shrinkage

For a team's historical quantity x, the feature builder retains at most 164 earlier games and uses exponentially decaying weights with half-life h. For target date d, prior mean m_0, and prior strength K, the smoothed estimate is:

$$w_j = 2^{-(d-d_j)/h}, \quad \tilde{x}(d) = (\Sigma_j w_j x_j + K\, m_0) / (\Sigma_j w_j + K). \tag{2}$$

Core attack and conceding-rate features use $h = 90$ calendar days, $K = 12$ equivalent games, and a fixed prior of three regulation goals. Calendar-day weighting explicitly increases shrinkage after long gaps, including the offseason. These are fixed modeling choices, not priors estimated from the final season. The construction partially pools rates but is not a fully Bayesian hierarchical model with a posterior over team abilities.

3.2 Candidate rate estimators

Five team candidates share the same information cutoffs and settlement transform. The rate baseline uses the lagged attack estimate alone. A multiplicative opponent baseline combines attack, the opponent's conceding rate, a training-sample league normalization, and home adjustment. The selected core model predicts regulation goals using a log link and three standardized inputs: home indicator, lagged attack, and opponent conceding rate. Standardization is fitted on training rows only. With z_i denoting those standardized features, the fitted objective is:

$$\lambda_i = \exp(\beta_0 + z_i^\top \beta),$$
$$\hat{\beta} = \arg\min\, [N^{-1}\Sigma_i(\lambda_i - y_i \log \lambda_i) + (0.3/2)\|\beta\|^2]. \tag{3}$$

The intercept is unpenalized; outcome-only constants are omitted. Scikit-learn's `PoissonRegressor` implements this regularized log-link regression, with a 300-iteration limit; predicted team rates are bounded to [0.2, 8] [5]. The context challenger adds team shots for, opponent shots against, team and opponent save-rate proxies, own power-play percentage, opponent penalty-kill percentage, rest days, and back-to-back indicators. Save proxies include team defense and empty-net effects, while PP/PK inputs are smoothed game percentages rather than opportunity-weighted event estimates.

The machine-learning challenger uses histogram gradient boosting with Poisson loss, 100 maximum iterations, seven leaf nodes, minimum leaf size 80, learning rate 0.05, ℓ_2 penalty 10, and random seed 41. These fixed settings restrict flexibility; no final-season hyperparameter search was conducted [6]. Opponent adjust-

ment here enters through lagged opponent defense and training regression. It is not an iteratively solved schedule-strength rating or a complete hierarchical team-effect model.

3.3 A common score distribution

Conditional on estimated rates, regulation scores are independent Poisson random variables. Let $Q(h,a)$ be their product mass. Every regulation tie is assigned one deciding game-settlement goal. With n_E historical extra-time games and w_H home winners, the training-only home extra-time tendency is $\pi = (w_H + 10)/(n_E + 20)$. The transformation is:

$$
\begin{aligned}
Q(h,a) &= \mathrm{Pois}(h;\lambda_H)\,\mathrm{Pois}(a;\lambda_A),\\
P(n,n) &= 0,\\
P(n+1,n) &= Q(n+1,n) + \pi Q(n,n),\\
P(n,n+1) &= Q(n,n+1) + (1-\pi)Q(n,n).
\end{aligned}
\tag{4}
$$

Other off-diagonal cells retain their mass. The transform preserves total probability and removes final ties. It treats overtime and shootouts identically for mapped full-game moneyline, spread, and total settlement; shootout awards do not enter player scoring counts. Source normalization separately reconciles statistical goals with official settlement scores, including shootout awards and the overtime empty-net standings exception. The model does not forecast the identity of a shootout participant or winner mechanism.

For a home spread ℓ, the win event is $H - A + \ell > 0$; equality is a push. For an Over total L, it is $H + A > L$. Moneyline win probability integrates $H > A$. All probabilities therefore refer to the same score matrix, and the expected settlement total is $\lambda_H + \lambda_A + \Pr(\text{regulation tie})$, up to negligible numerical truncation. They cannot contradict one another merely because separate market classifiers were fitted.

Coherence does not establish realism. Independent regulation counts omit shared pace and game-state dependence. Empty-net scoring is present in regulation targets but has no explicit tactical process. A single regressed extra-time tendency ignores matchup differences. The deployed count routine uses support $0,\ldots,47$, normalizes its mass, and rejects materially insufficient support; the transformed team matrix permits one additional deciding goal.

4. Player Opportunity and Count Distributions

4.1 Separating minutes from production

The player-rate baseline smooths past counts directly using (2), with half-life 90 days and strength 12. For defensemen, fixed prior means are 1.6 shots, 0.12 goals, and 0.30 assists, with 18 minutes of ice time. For forwards or an unknown position, they are 1.9, 0.25, 0.34, and 15 minutes. Points always equal the sum of

the goal and assist means. Position is the most recently observed historical position; a first appearance uses the fixed unknown-position prior.

The opportunity specification instead projects minutes $\hat{T}$ using a 30-day half-life and five prior appearances, and per-minute production $\hat{r}_k$ using a 120-day half-life and 12 prior appearances. Each rate observation is the count divided by that appearance's positive TOI. For statistic k:

$$\mu_k = \hat{T}\,\hat{r}_k, \quad \mu_P = \mu_G + \mu_A. \tag{5}$$

This averages appearance-level rates by recency; it does not pool all historical minutes into one exposure denominator. A brief appearance can therefore contribute a noisy rate. The split can respond to recent usage changes, but it does not infer a specific line assignment, power-play unit, teammate effect, or future injury restriction. Projected TOI is a point estimate, not a separately integrated random variable.

4.2 Overdispersion and scoring dependence

The selected count family is negative binomial with mean μ and dispersion α, so $\mathrm{Var}(Y) = \mu + \alpha\mu^2$. Its gamma–Poisson representation uses shape $r = 1/\alpha$ and success parameter $p = 1/(1+\alpha\mu)$ in SciPy's count parameterization; the $\alpha \to 0$ limit is Poisson [7]. Separate pooled training estimates are used for shots and total points:

$$\hat{\alpha} = \mathrm{clip}_{[0,0.75]}\{\Sigma_i[(y_i-\mu_i)^2-y_i] \,/\, \Sigma_i\mu_i^2\}. \tag{6}$$

This residual-moment estimator can absorb mean misspecification as well as genuine heterogeneity. It is not a player-specific posterior dispersion estimate. To preserve scoring dependence, let $Z \sim \mathrm{Gamma}(1/\alpha, \text{scale } \alpha)$, with mean one. Conditional on Z, draw goals and assists as independent Poisson counts with means $Z\mu_G$ and $Z\mu_A$. Equivalently, draw total points and then allocate:

$$P \sim \mathrm{NB}(\mu_G+\mu_A, \alpha),$$
$$G \mid P=n \sim \mathrm{Binomial}(n, \mu_G/(\mu_G+\mu_A)), \quad A=P-G. \tag{7}$$

Thus $\mathrm{Cov}(G,A) = \alpha\mu_G\mu_A$ before numerical truncation, and goals plus assists equal points for every draw. The implementation computes the marginal count distributions from this allocation. Shots have their own dispersion and are not assigned a validated joint dependence with scoring, teammates, or game totals. Multiplying these marginal probabilities to price a parlay is unsupported.

4.3 Challengers and participation

Four player candidates are compared: historical-rate Poisson, opportunity Poisson, opportunity negative binomial, and an opportunity hurdle distribution. The hurdle candidate fits a training zero-frequency correction for points and a shifted-Poisson positive component, then allocates points to goals and assists. Shots remain opportunity Poisson in that challenger. Shots select their family separately; goals, assists, and points select one common family by their equally weighted mean count log score, preserving coherence.

Every player forecast is conditional on participation under the mapped market's action rules. The evaluation includes actual positive-TOI appearances; it does not claim to have forecast the roster or probability of dressing. Live uncertainty about scratches remains explicit. An ambiguous identity produces no independent probability, and an unresolved missing appearance is not silently graded as a zero. A sourced DNP can support a void under the applicable rule.

5. Markets, Pushes, and Offered-Price Evaluation

5.1 Three distinct probability objects

The independent hockey forecast uses pregame hockey history only. The market benchmark uses paired sportsbook prices only. A market-aware forecast would combine the two on a common probability basis after past-data fitting and later validation. In this release the market weight is zero, so the final published forecast equals the independent one. A missing independent estimate remains null; consensus is never relabeled as a hockey model.

For a valid two-sided book with decimal prices d_1, d_2, define raw inverse prices $q_s = 1/d_s$. Multiplicative de-vig yields:

$$\tilde{p}_s = q_s / (q_1+q_2). \tag{8}$$

The adapter also records the difference from additive de-vig, $q_s - (q_1+q_2-1)/2$, as a method-sensitivity diagnostic. Neither transformation proves that a bookmaker's probability is correct. For push-capable contracts, the inferred probability is conditional on a non-push; it does not identify push mass.

Pairing requires the exact game, player where applicable, market, line, start time, and settlement profile, with both sides observed within five minutes. Spread keys normalize opposite sides to one home-oriented handicap. Contradictory same-time duplicate quotes fail closed. The offered book is excluded from its reference consensus; remaining paired-book probabilities within fifteen minutes of the offer are combined by their median. Market Watch requires at least three other books for its price-comparison signal. Unknown settlement profiles are isolated by book and do not support mapped cross-book EV.

The live Market Watch filter is intentionally a research screen: it requires a positive consensus-based price advantage, not a minimum 2% independent-model EV or agreement with the hockey forecast. Its best-price filter compares payouts only within the identical outcome, line, and settlement contract. A total of 6.0 and a total of 6.5 are different contracts; their half-goal difference must be evaluated through the score distribution and push probability. In the analyst-workflow extension, Market Watch membership and ordering depend on paired prices alone; a model push estimate is not required to display an integer-line price difference. Model-based screening occurs on the separate NHL Top Picks research board. Neither board enables the disabled, stricter validated-recommendation selector.

5.2 Settlement-aware value and minimum prices

Let w, u, and l be unconditional win, push, and loss probabilities given action, with $w+u+l=1$. A one-unit stake at decimal odds d has expected net return:

$$\mathrm{EV}(d) = w(d-1)-l = wd-(1-u),$$
$$d_{\mathrm{fair}} = (1-u)/w, \quad p_{\mathrm{conditional}} = w/(1-u). \tag{9}$$

Consequently, market-model disagreement is measured between $p_{\mathrm{conditional}}$ and the market's conditional probability, while offered-price EV uses all three settlement masses. Consensus alone cannot deliver an unconditional integer-line EV without a push assumption. For illustration, w=0.47, u=0.08, l=0.45 gives fair decimal odds 1.9574 and EV 0.02 at odds 2.00. The 47% win probability is not equivalent to a 47% non-push chance.

Minimum acceptable price uses a fixed target ε=0.02 expected units per unit staked and a finite set of stress scenarios S. Team rates are perturbed by ±10% in all four home/away combinations; player means are perturbed together by ±10%. Each scenario supplies its own push probability:

$$d_{\min} = \max_{s\in S} (1-u_s+\varepsilon)/w_s,$$
$$R(d) = \min_{s\in S}\{w_s \log[1+f(d-1)] + l_s \log(1-f)\}, \quad f=0.0025. \tag{10}$$

The nominal forecast is included. These are transparent stress assumptions, not calibrated confidence bounds, estimated parameter uncertainty, or optimized betting thresholds. The worst-scenario log-growth quantity orders research candidates and penalizes risk at a fixed fraction; it does not authorize wagering. Price shopping, independent hockey disagreement, and their combination are separately labeled. A favorable price on an ordinary outcome can exist without independent predictive superiority.

5.3 Market blending remains a future experiment

A prospective blend utility considers $p_\gamma=(1-\gamma)p_{\text{hockey}}+\gamma p_{\text{market}}$ on the non-push basis, with γ on a 0.05 grid. It requires at least 500 distinct training games, past outcome availability, and leave-offered-book-out references; a separate later period must then validate any fitted artifact. The present development price sample contains only 86 games, so no blend is fitted or evaluated as a production candidate. For integer lines, adopting any such blend would additionally require a declared way to retain or estimate push mass. There is no reported market-aware performance gain.

6. Evaluation Design and Statistical Estimands

6.1 Chronological selection and a disclosed rerun

The first season supplies initial training. Two later seasons form expanding development folds. All sides, players, and related records from a game remain in the same season fold. Candidate selection minimizes mean validation joint-score NLL for teams, count NLL for shots, and mean scoring-market count NLL for the common scoring family. Regularization and feature constants are fixed; no probability remapping is fitted. Dispersion and overtime tendency are estimated only within each training window.

Table 1. Expanding training windows and season-level evaluation windows.

Evaluation	*Parameter fitting*	*Evaluation dates*	*Games*
2023–24	2022-10-07 to 2023-04-14	2023-10-10 to 2024-04-18	1,312
2024–25	2022-10-07 to 2024-04-18	2024-10-04 to 2025-04-17	1,312
2025–26 final	2022-10-07 to 2025-04-17	2025-10-07 to 2026-04-16	1,312

After the algorithm choices were locked, 2025–26 was used for final evaluation. A subsequent correctness audit found that the initial cold-start player feature builder had used the target-game position. The repair replaced it with last-observed position or the fixed unknown-position prior. The same protocol was rerun; model choices and thresholds were unchanged. This disclosure matters: the final season was opened before the repair, so the reported computation is not a pristine untouched test. No new independent final season has yet been observed. Live deployment subsequently refits the selected specification through April 16, 2026; those all-season artifacts do not generate the historical test metrics.

6.2 Proper scores, calibration, and uncertainty

Predictive densities are evaluated using negative log likelihood; binary events use log loss and Brier score. These proper scoring rules reward probabilistic accuracy rather than a chosen win-rate threshold [8]. Quantity diagnostics include MAE and RMSE. For a binary outcome y and forecast p:

$$\mathrm{LL} = -N^{-1}\Sigma[y \log p + (1-y)\log(1-p)],$$
$$\mathrm{BS} = N^{-1}\Sigma(p-y)^2.$$

(11)

We report reliability-bin means and ten-bin expected calibration error, ECE = $\Sigma_b(n_b/N)|\bar{p}_b - \bar{y}_b|$. ECE is bin-dependent, noisy in sparse cells, and not itself a proper scoring rule or a confidence interval. Player distributions also report observed versus predicted zeros, central 5th–95th percentile coverage, and ranked probability score RPS = $\Sigma_k[F(k) - 1\{y \le k\}]^2$ over the retained count support. Different base rates make Brier scores across different markets unsuitable as a direct league table of model quality.

Paired model differences use the same realized observations. Five hundred percentile bootstrap replicates resample entire games, preserving within-game player and market dependence. For replicate game multiplicities c_g, the player-weighted difference is:

$$\Delta^* = [\Sigma_g\, c_g \Sigma_{i \in g}(L_{new,i} - L_{base,i})] \,/\, [\Sigma_g\, c_g n_g].$$

(12)

This addresses intra-game correlation, not all serial dependence across games played by the same teams and players. The intervals are conditional on the observed seasons and selected specification; they do not repeat feature selection, account for unknown data revisions, or provide simultaneous family-wise coverage. The limited 500-replicate resolution is another reason to avoid overinterpreting small endpoint differences.

7. Forecast Results and Ablations

7.1 Development comparisons

Table 2. Development joint-score negative log likelihood, in nats per game. Lower is better.

Candidate	*2023–24*	*2024–25*	*Mean*
rate	3.74192	3.74018	3.74105
opponent	3.72330	3.72427	3.72378
poisson_core	3.72314	3.71863	3.72089
poisson_context	3.71437	3.73204	3.72320
boosting	3.72239	3.74500	3.73370

The attack/opponent/home candidate improves on the attack-only baseline in both development seasons. Regularized core regression has the best average joint log score. Adding the context bundle improves 2023–24 but worsens 2024–25; boosting similarly fails to produce a stable improvement. The selected specifica-

tion therefore excludes the added bundle. This is a grouped ablation: it cannot establish that each individual rest, shot, save-rate, or special-teams feature lacks value. Neither starting-goalie identity nor expected goals was tested.

Table 3. Development count NLL. Scoring averages goals, assists and points with equal weights.

Candidate	Shots 23–24	Shots 24–25	Scoring 23–24	Scoring 24–25
rate_poisson	1.60642	1.57263	0.65227	0.64257
opportunity_poisson	1.60081	1.56580	0.64941	0.63961
opportunity_nb	1.59703	1.56157	0.64934	0.63962
opportunity_hurdle	1.60081	1.56580	0.65277	0.64372

Opportunity separation improves all four player count scores in both development seasons. Negative-binomial dispersion further improves shots in both folds. The scoring-family advantage over opportunity Poisson is extremely small and not uniform across individual market/fold cells; its selection follows the predefined common-family objective. The hurdle construction performs worse on the scoring-family objective. No candidate is chosen for historical returns.

7.2 Final-season quantities and event probabilities

Table 4. Final-season team results, 1,312 games.

Model	Joint NLL	Total MAE	Total RMSE	ML log loss	ML Brier
rate	3.68481	1.8547	2.3151	0.68599	0.24646
poisson_core	3.68073	1.8433	2.3060	0.68715	0.24704

Table 5. Selected-model binary diagnostics. These fixed lines are not historical quoted offers.

Outcome	Log loss	Brier	ECE
moneyline	0.68715	0.24704	0.01844
total_5.5	0.68482	0.24587	0.03684
total_6.5	0.69356	0.25015	0.03976
puck_−1.5	0.60004	0.20553	0.01485

The final joint-score improvement is −0.004079 nats per game, with a game-bootstrap 95% interval [−0.013763, 0.005311]. Total MAE and RMSE decrease slightly. Moneyline log loss and Brier score worsen relative to the rate baseline. We retain the preselected common score model rather than choose a moneyline-specific winner after inspecting the final season. This supports consistency of the experimental procedure, not a claim that the selected model dominates its baseline in every market.

Table 6. Final-season player results, 47,230 appearances per market. Binary thresholds: shots >2.5; all scoring markets >0.5.

Market	Rate NLL	Selected NLL	MAE	RMSE	Brier	ECE
shots	1.56569	1.55355	1.0502	1.3368	0.14827	0.03704
goals	0.45709	0.45492	0.2802	0.4160	0.12097	0.01338
assists	0.65098	0.64798	0.4172	0.5408	0.17179	0.02291
points	0.84880	0.84410	0.5478	0.6823	0.20614	0.03480

Table 7. Paired selected-minus-baseline NLL. Negative favors the selected model; 500 whole-game bootstrap replicates.

Target	Difference	95% percentile interval
joint_score	−0.004079	[−0.013763, 0.005311]
shots	−0.012139	[−0.013178, −0.011103]
goals	−0.002167	[−0.002409, −0.001902]
assists	−0.002996	[−0.003367, −0.002641]
points	−0.004701	[−0.005139, −0.004227]

Player count NLL improves in each market. The largest absolute reduction is for shots, while goals, assists, and points have smaller reductions. These are improvements over a particular simple statistical baseline, not over sportsbook prices. The dataset evaluates all positive-TOI skaters, including many players without a posted prop. Its population therefore differs from the prop board analysts will actually use.

Table 8. Selected player-distribution diagnostics. Zero rates and central discrete interval coverage are percentages.

Market	Observed zero	Predicted zero	90% coverage	RPS
shots	27.21	23.35	97.55	0.69811
goals	84.82	83.72	98.52	0.13970
assists	75.90	74.63	97.99	0.21647
points	65.26	63.03	98.03	0.30358

Calibration remains incomplete. The selected model predicts fewer zero-shot appearances than observed (23.35% versus 27.21%), and understates zeros in all four player markets. Nominal 90% central discrete intervals cover approximately 97–99% of outcomes. Integer quantiles and large zero masses contribute to this conservatism; high coverage alone is not proof of a useful or sharply calibrated distribution. A count-score improvement is compatible with remaining threshold-specific errors.

8. Historical Market and Shadow-Policy Results

8.1 A fixed, limited timestamped comparison

The price sample was acquired after the core predictive study without changing the selected algorithms or policy thresholds. Each season uses parameters fitted to prior seasons. The development-season comparisons are retrospective diagnostics of the specification ultimately selected using both development seasons; they are not an independent validation of that selection. Only 2025–26 is the final-season price diagnostic, subject to the previously disclosed correctness rerun.

For the market benchmark, one home/Over observation is retained per game, market, and exact line, using the freshest mapped offer and its paired leave-book-out reference. Quotes must precede the decision and the game must not have started. Push outcomes are excluded from this conditional binary comparison. Multiple totals lines may therefore contribute several observations for one game; uncertainty remains clustered by game. The sample contains 44, 42, and 44 distinct games by season.

Table 9. Timestamped price diagnostic: conditional non-push log loss. Intervals are hockey minus market; every interval contains zero.

Season / market	*Obs. / games*	*Hockey*	*Market*	*Difference 95% interval*
2023–24 / h2h	44 / 44	0.67437	0.69635	[−0.06105, 0.01479]
2023–24 / spreads	44 / 44	0.71863	0.70646	[−0.03972, 0.06392]
2023–24 / totals	41 / 40	0.69001	0.68718	[−0.02135, 0.02364]
2024–25 / h2h	42 / 42	0.65232	0.65666	[−0.04683, 0.03367]
2024–25 / spreads	42 / 42	0.67764	0.66999	[−0.04990, 0.07026]
2024–25 / totals	45 / 40	0.70312	0.68497	[−0.00266, 0.04453]
2025–26 / h2h	44 / 44	0.70420	0.70547	[−0.03466, 0.03151]
2025–26 / spreads	44 / 44	0.58389	0.57256	[−0.02814, 0.05739]
2025–26 / totals	53 / 43	0.68079	0.68952	[−0.02656, 0.01272]

Every paired interval contains zero. In the final-season sample, the independent model has slightly lower moneyline and totals log loss and higher spread log loss than the market. These mixed, imprecise estimates do not establish market superiority. Reliability bins and Brier scores are supplied in the accompanying evidence JSON, but sparse bin counts make any apparent calibration ranking unstable.

8.2 A shadow policy fixed before price acquisition

The shadow selector accepts at most four one-unit offers per day and one offer per game. It requires a mapped standard settlement profile, quote age at most 30 minutes, nominal independent EV at least 2%, positive worst-scenario log growth, and offered odds meeting the stress-based minimum price. It uses no fabricated historical analyst approvals. The selected offers are graded as wins, losses, or pushes with flat staking and no compounding.

Table 10. Fixed quoted-price shadow policy. One flat unit per selection; turnover equals count. These are not realized bets.

Season	Count	Net units	ROI	Drawdown	ROI 95% interval	Worse price ROI
2023–24	6	3.2407	54.01%	1.0000	[−41.98%, 146.75%]	50.68%
2024–25	8	0.9192	11.49%	3.0000	[−54.17%, 73.51%]	8.36%
2025–26	3	−1.5238	−50.79%	1.5238	[−100.00%, 47.62%]	−52.46%

Table 11. Decimal odds distribution among shadow selections; five quantiles [minimum, 25%, median, 75%, maximum].

Season	Min	25%	Median	75%	Max
2023–24	1.7407	1.9444	2.0250	2.2375	3.2000
2024–25	1.6061	1.7081	1.7945	1.8960	2.1500
2025–26	1.4762	1.6381	1.8000	1.9000	2.0000

The final season contains three selections, three units of turnover, −1.5238 net units, and −50.79% ROI. The bootstrap ROI interval [−100.00%, 47.62%] is far too wide to support a stable profitability estimate and rests on only three game clusters. Reducing each offered decimal price by 0.05 changes final net return to −1.5738 units (−52.46%). Positive development-season returns are not a discovered profitable strategy; thresholds were not revised after seeing these outcomes.

These are *quoted-price shadow simulations*. Historical house-rule versions, account eligibility, limits, rejected bets, and execution are unverified. One selection per game reduces obvious within-game exposure but does not remove dependence across dates or teams. The sample cannot certify executable historical ROI, early-versus-later timing value, or closing-line value. Closing quotes and 16:30 historical snapshots were not collected. When closing data become available, the grading utility requires the identical line and settlement contract, with a paired observed quote within 30 minutes before start; a line change is not silently converted into price-only CLV.

9. Analyst Intervention and Operation

Analysts receive a quote's market, side, line, book, price, and timestamp together with independent and market probabilities on explicitly labeled bases, push mass, fair odds, estimated EV, minimum acceptable price, sensitivity assumptions, model version, freshness, and review status. Driver notes describe actual model inputs such as projected TOI or lagged team rates. A missing goalie or lineup assumption is a reason for review, not an invented numerical penalty.

The review layer is append-only. Each record names the exact forecast and offer, analyst, source URL, source publication time, review time, reason, and overlap with existing features or market information. An explicit override also requires its numerical method and a double-counting explanation. The original probability re-

mains preserved alongside the adjustment. A newly generated forecast has a new identity and does not inherit an old review silently. There is no historical intervention-effect estimate in this report; these records begin a prospective shadow cohort.

A credible prospective comparison would evaluate original and adjusted probabilities on the same eligible records, track abstentions and changes in selection, and grade prices and results separately. Because analysts choose which cases to review, a naive reviewed-versus-unreviewed comparison is confounded by selection. Randomized assignment or a prespecified matched protocol would be needed for a stronger causal claim. An LLM may summarize sourced evidence, but it cannot manufacture news, confidence percentages, or numerical edge from prose.

The subsequent morning-workflow extension operationalizes this sequence without changing the fitted models or the empirical results reported here. A fixed quantitative screen saves at most four same-day candidates, one per game, subject to the stated EV, sensitivity-price, identity, settlement and freshness guards. GPT-6 Astra then critiques bounded, timestamped public-source excerpts, identifying cited evidence, a countercase, possible double counting and missing checks. Exact source-excerpt matching and schema validation constrain its output; they do not certify that its interpretation is correct. Human select, watch or pass decisions remain separate and require an explicit record. No qualitative probability adjustment is learned or applied by this workflow. The original screened cohort and timely human decisions form a prospective shadow evaluation; no historical improvement from this extension is claimed.

Training is separate from daily inference. Versioned artifacts carry feature-schema, dependency, training-window, and checksum metadata. Daily snapshots archive prices, model inputs, sources, and original forecasts. Morning and optional later refreshes target 10:30 and 16:30 Eastern. Market quotes older than 24 hours, model-input checks older than 36 hours, and model artifacts older than 400 days are withheld under the respective guards; live model forecasts are restricted to games within 48 hours. These are operational availability limits, not guarantees of executable prices. The shadow study's 30-minute quote rule is deliberately tighter.

The existing NHL routes and seven supported market types remain compatible. On failure, the interface makes unavailable or stale data explicit; valid fresh market quotes can remain visible while independent forecasts are withheld. All recommendation lists remain empty unless separate validated status and analyst requirements are met. The model does not interact with sportsbook accounts or place wagers.

10. Threats to Validity and Remaining Research

Historical availability. The declared next-day lag prevents direct same-game feature leakage, but cannot prove the absence of revisions in retrospectively downloaded statistics. Date-level games lack original start-time histories, and current schedules cannot reconstruct rescheduling announcements. The corrected rookie-position feature and rerun are disclosed rather than treated as fresh independent evidence.

Distributional misspecification. The team model omits shared regulation dependence, an explicit empty-net process, matchup-specific overtime strength, and starter identity. Player rates use noisy appearance-level exposure ratios, pooled dispersion, fixed rookie priors, and unconfirmed participation. The common scoring family preserves an accounting identity without capturing all hockey mechanisms. Dependence across player shots, scoring, and game outcomes is unvalidated.

Coverage and selection. The final player population differs from the offered-prop population. Monthly date sampling provides only 130 price-covered games; it may miss changes in schedule, market liquidity, or pricing behavior across other days. Standard settlement mappings do not reconstruct every historical rule version or jurisdiction. There are no player-price, full daily, closing-price, or contemporaneous qualitative archives sufficient for the requested comparisons.

Statistical precision. Whole-game bootstrapping preserves within-game relationships, but does not model serial correlation or uncertainty from specification selection. Calibration estimates depend on binning and event prevalence. Very small shadow samples cannot validate profitability, and a fixed ±10% stress test is not a frequentist or Bayesian interval. No post-hoc subset of profitable books, lines, or players is promoted as a strategy.

Next evidence required. Priority should go to prospectively timestamped eligible offers, preserved source values, clear participation/settlement records, and frozen selection policies. A larger authorized historical sample could support separate blend training and later validation; any new feature needs historical availability evidence or prospective shadow status. Starting-goalie, injury, deployment, and shot-quality data should be judged by incremental held-out performance, not by their narrative plausibility. No recommendation threshold is promoted solely because the system is operational.

11. Conclusion for Readers and Analysts

The new NHL system is a better foundation for analysis because its inputs, assumptions, prices, and results can be checked. It produces a consistent picture of game outcomes, treats refunded bets correctly, and shows when an estimate is missing or out of date. Those improvements make the workflow more trustworthy; they do not make every prediction accurate.

Our clearest modeling improvement came from separating how much a player is likely to play from how quickly that player produces shots and scoring. This modestly improved forecasts against our simple historical baseline. Extra team context and a more flexible machine-learning model did not produce consistent improvements in development testing. The final team results were mixed, so complexity was not rewarded for its own sake.

We have not shown that these forecasts beat sportsbook prices. The historical price sample is small, its uncertainty is substantial, and the final experimental selections lost 1.52 units across just three bets in simulation. Neither those losses nor the earlier positive results are enough to predict future profitability. There is also no evidence yet that analyst overrides improve the system.

For now, use the board to compare like-for-like prices, understand why a hockey forecast differs from the market, and identify assumptions that need checking. Treat a displayed positive expected value as a calculation under those assumptions, not proof that a bet is good. The appropriate recommendation can be no bet. A stronger claim will require fresh, recorded decisions and a much larger body of independent evidence.

Appendix A. Reproducibility and Artifact Lineage

This report describes model `nhl-v2.1`, feature schema `nhl-pit-1`, and the implementation merged at commit `e1f91609524f74a6b99cf86cdebb2328cd5e08d1`. Its numerical tables are generated from `evaluation.json`, `paired-differences.json`, and `historical-market-evaluation.json`; `selection-lock.json` records the retained specification. The downloadable evidence JSON includes their SHA-256 hashes, full diagnostics, and the source commit. It contains no API credentials.

The archived environment pins Python 3.11, NumPy 1.26.4, pandas 1.5.3, SciPy 1.17.1, scikit-learn 1.9.0, joblib 1.5.2, requests 2.34.2, and python-dotenv 1.2.3. Public library documentation linked below may reflect a newer release; the repository's pinned environment and source code govern reproduction. From a clean checkout, the following commands restore retained source responses and rerun the calculations without acquiring new odds:

```
python3.11 -m venv .venv
.venv/bin/pip install -r requirements-nhl.txt
.venv/bin/python scripts/nhl/v2/restore.py
OPENBLAS_NUM_THREADS=1 OMP_NUM_THREADS=1 \
  .venv/bin/python scripts/nhl/v2/evaluate.py
.venv/bin/python scripts/nhl/v2/report.py
OPENBLAS_NUM_THREADS=1 OMP_NUM_THREADS=1 \
  .venv/bin/python scripts/nhl/v2/market_evaluate.py
python3 scripts/research/build_nhl_paper.py
```

Raw training responses are preserved in `artifacts/nhl/training-sources.tar.gz`; historical quotes reside under `artifacts/nhl/historical-odds/`. Compressed player and team prediction files retain outcomes and scores. Restoring the archive checks content hashes; a newly downloaded revision is a different data vintage. Reproduction needs no GPU and is expected to fit the existing infrastructure; machine speed and dependency installation determine elapsed time. The operational runbook documents failure handling, retention, and rollback [9].

References and Research Materials

[1] Fourth & Value Research Group (2026). NHL v2.1 implementation and evaluation artifacts. Versioned primary research materials.

[2] National Hockey League. NHL Statistics. Team and skater per-game summary endpoints; retrieved September 26, 2026. Exact responses and source manifests archived with this study.

[3] The Odds API. V4 documentation: historical odds. Provider timestamps and historical snapshot semantics; accessed September 26, 2026.

[4] Fourth & Value Research Group (2026). Audit and decision log, source inventory, and model card.

[5] Scikit-learn developers. PoissonRegressor API documentation. Accessed September 26, 2026; fitted implementation pinned to scikit-learn 1.9.0.

[6] Scikit-learn developers. HistGradientBoostingRegressor API documentation. Accessed September 26, 2026.

[7] SciPy developers. Negative-binomial distribution reference. Accessed September 26, 2026; computation pinned to SciPy 1.17.1.

[8] Gneiting, T., and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), 359–378. DOI: 10.1198/016214506000001437.

[9] Fourth & Value Research Group (2026). Build, operate, review, and roll back NHL v2 and timestamped historical market diagnostic.

> Research status: experimental forecasts in all seven markets. No validated betting advantage, analyst-intervention benefit, or market-blend improvement is established. This report is an internal technical publication, not external peer review.