

Volatility in Prediction Markets: A Structural Approach*

Weiye Xi[†] Ciamac C. Moallemi[‡] Mallesh Pai[§] Shouqiao Wang[¶]

Abstract

Forward-looking volatility forecasts are central inputs to derivatives pricing, market making, risk management, and volatility-linked trading strategies, with ARCH and GARCH models serving as the canonical workhorses. Such models are natural in standard asset markets, where prices are positive-valued stochastic processes and volatility is typically inferred from return dynamics. Prediction markets have a different structure: prices are bounded probabilities, payoffs are binary, and contracts resolve at known deadlines.

We develop and estimate a volatility model tailored to binary prediction markets. The model combines two economic mechanisms: a Wright-Fisher deadline-resolution component, capturing how remaining binary uncertainty is forced to resolve over time, and a Glosten-Milgrom order-flow component, capturing volatility from informed trading as reflected in spreads and volume. Using a large panel of Kalshi contracts, we show that these structural variables carry substantial forecasting power. Plain ARCH/GARCH benchmarks are dominated by structural specifications; combining the structural model with residual GARCH dynamics gives the best overall forecasts. The model also provides an interpretable measurement framework: volatility is highest near fifty-fifty prices, rises near resolution, and varies across categories with the timing and discreteness of information arrival. Economics contracts are closer to smooth deadline-resolution dynamics, while sports contracts exhibit more event-concentrated, jump-like behavior. Across major categories, category-specific fitting does not systematically improve out-of-sample performance, suggesting that the structural specification transfers beyond the pooled headline result.

JEL classification: G14, G17, C58, D84.

1. Introduction

Prediction-market prices are increasingly used as real-time probability forecasts for economic, political, weather, sports, and entertainment events. Exchange-traded platforms such as Kalshi and Polymarket have grown substantially in the past several years, with recent financial press coverage documenting rapid increases in trading volume [The Wall Street Journal, 2026]. Prediction-market prices are now routinely referenced as market-implied probabilities. However, a price is only a level forecast: it says what the market currently believes, not how likely that belief is to change. A contract trading at a given price may represent a stable consensus of genuine uncertainty, or it may sit immediately before an information event that is likely to move the price sharply. Distinguishing

*CCM is supported by the Briger Family Digital Finance Lab at Columbia University. CCM is a research advisor for Paradigm and Uniswap Labs.

[†]Columbia University. wx2305@columbia.edu.

[‡]Columbia University, Paradigm, and Uniswap Labs. ciamac@gsb.columbia.edu.

[§]Tempo Labs. [mallesh.pai@gmail.com](mailto:malles.pai@gmail.com).

[¶]Columbia University. shwang27@gsb.columbia.edu.

these two cases requires a forward-looking model of conditional volatility, not a backward-looking realized-volatility summary.

In traditional financial markets, such volatility forecasts are central inputs to derivatives pricing, market making, and portfolio risk measurement. Prediction markets create analogous needs, but the object being forecast is a probability rather than an asset return. Traders and risk managers need prediction intervals for future probability moves, market makers need a measure of adverse-selection risk when quoting, and downstream users of prediction-market probabilities need to know whether a quoted probability is likely to remain stable over their decision horizon.

The prediction-market setting, however, is different from the standard asset-market setting. ARCH and GARCH models are natural workhorses for volatility forecasting in ordinary financial assets, where prices are positive-valued stochastic processes and volatility is typically inferred from return dynamics [Engle, 1982, Bollerslev, 1986]. A prediction-market contract instead has a price that is a bounded probability, a payoff that is binary, and a resolution date that is known in advance. These features impose structure that the former models do not encode. Order-book variables add further information, since spreads and trading activity reflect adverse selection and informed trading.

In this paper, we develop a structural volatility model tailored to this setting. Suppressing the hourly index, the baseline specification writes conditional variance as the sum of two components:

$$h_t^2 = \underbrace{\frac{p_t(1-p_t)}{\tau_t}}_{\text{deadline-resolution channel}} + K \cdot \underbrace{\nu(V_t) \cdot \frac{s_t^2}{4}}_{\text{adverse-selection channel}}$$

The first term is a Wright-Fisher deadline-resolution (DR) component: it captures how remaining binary uncertainty, $p_t(1-p_t)$, is released over the time left to resolution, τ_t . Here p_t is the prediction-market price. The second term is a Glosten-Milgrom adverse-selection (AS) component: it links bid-ask spreads (s_t) to trade-level adverse-selection variance and scales this variance by a monotone transformation of trading activity ($\nu(V_t)$). The parameter K governs the strength of the adverse-selection channel. We refer to this model as DR-AS.

Empirically, we construct an hourly panel of Kalshi binary contracts from August 2021 through April 2026. For each contract-hour, we measure the market-implied probability, bid-ask spread, trading volume, and time to resolution before the next-hour price is realized. Our headline exercise forecasts the conditional scale of active hourly probability updates: the size of the next-hour probability move conditional on an update occurring. Each month, we fit the models on all prior active-update observations and test them on that month’s active-update contract-hours, yielding roughly 880,000 out-of-sample forecasts. Because conditional variance is not directly observed, we evaluate each volatility forecast through the prediction interval it implies for the next-hour price. We score these intervals using the Winkler interval score [Winkler, 1972], which penalizes intervals that are too wide as well as intervals that miss the realized price.

We compare DR-AS with a broad set of structural, dynamic, and hybrid specifications. Plain ARCH/GARCH benchmarks lag the structural specifications: the pure deadline-resolution component already improves substantially on those generic benchmarks, DR-AS with concave volume scaling is the strongest closed-form specification, and GARCH+DR-AS is the best overall model. In headline numbers, the one-parameter closed-form structural specification outperforms a plain GARCH(1,1) by 34% on the volume-weighted Winkler interval score. Appendix D shows that this advantage is not driven by the active-update definition: when zero-update hours are added back to evaluation, and when the filter is removed from both estimation and evaluation, prediction-market structural variables remain well ahead of generic ARCH/GARCH dynamics.

Another benefit of our model is that it decomposes conditional variance into interpretable components. Volatility is highest near fifty-fifty prices, where binary outcome uncertainty $p(1-p)$ peaks, and intensifies as resolution approaches, consistent with the $1/\tau$ deadline-resolution scaling; both patterns are visible in model-free univariate binning (Section 3.4). Spreads and volume are weaker drivers in isolation, but jointly, through the order-flow term, they capture microstructure volatility outside the deadline channel (Section 4).

Finally, we show that the globally fitted GARCH+DR-AS specification is the best or near-best volatility predictor in most major Kalshi categories—including Sports, Politics, Economics, Elections, Crypto, and Entertainment—spanning the bulk of the panel. Although categories differ in absolute volatility level and modeling difficulty, refitting parameters within each category does not systematically improve out-of-sample performance.

The rest of the paper is organized as follows. Section 2 derives the structural model. Section 3 describes the Kalshi panel, the specifications, and the interval-score evaluation. Section 4 reports the main forecast comparison, Section 5 examines category heterogeneity and portability, and Section 6 concludes.

1.1. Related Literature

Our work connects three literatures.

Volatility forecasting. The conditional-volatility literature begins with ARCH and GARCH [Engle, 1982, Bollerslev, 1986] and extends through realized-volatility approaches such as Andersen et al. [2003]. These models are designed for financial return series and capture persistence in conditional second moments. Our setting differs because the traded object is a bounded probability that resolves to a binary payoff at a known deadline.

Market microstructure and information flow. Our order-flow component draws on the adverse-selection view of bid-ask spreads in Glosten and Milgrom [1985]. Related work studies the information content of trades [Easley and O’Hara, 1992, Hasbrouck, 1991]. The use of volume as an information-arrival proxy is related to the mixture-of-distributions tradition [Tauchen and Pitts, 1983, Andersen and Bollerslev, 1998]. In our setting, these ideas enter a binary-contract volatility model through the joint dependence of volatility on spreads and trading activity.

Prediction markets and binary prices. A large prediction-market literature studies when market prices can be interpreted as probabilities or forecasts [Wolfers and Zitzewitz, 2004, Manski, 2006]. Hanson-style market scoring rules provide one mechanism-design foundation for information markets [Hanson, 2003]. Closest to our paper, Archak and Ipeitotis [2010] develop a structural volatility model for binary prediction markets based on latent belief dynamics. We use their implied probit-Brownian variance shape as a benchmark, but differ by using a Wright-Fisher/Bernoulli boundary shape, adding a Glosten-Milgrom order-flow channel, allowing residual ARCH/GARCH dynamics, and evaluating the specifications on a large Kalshi panel. More recently, Dalen [2025] proposes a Black-Scholes-style kernel for prediction-market derivatives based on belief volatility and jump-diffusion dynamics.

2. A Structural Model for Prediction-Market Volatility

This section derives a structural volatility model for a binary prediction-market claim. The market price is modeled as a posterior probability, and its one-step conditional variance is decomposed into two information channels: a Wright-Fisher deadline-resolution channel and a Glosten-Milgrom order-flow channel.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with filtration $(\mathcal{F}_t)_{0 \leq t \leq T}$. The YES claim pays $\Theta \in \{0, 1\}$ at resolution time T , where $\Theta = 1$ means that the YES event occurs. Interpreting $\mathbb{P}$ as the risk-neutral distribution, the prediction-market price is the posterior probability of the YES state:

$$p_t = \mathbb{P}(\Theta = 1 \mid \mathcal{F}_t) = \mathbb{E}[\Theta \mid \mathcal{F}_t].$$

By the tower property, $(p_t)_{0 \leq t \leq T}$ is a bounded $\mathcal{F}_t$ -martingale, and terminal settlement reveals the true state, so $p_T = \Theta \in \{0, 1\}$.

We decompose market information into two observation channels: $\mathcal{F}_t = \mathcal{F}_t^P \vee \mathcal{F}_t^Q$. The P -channel is the Wright-Fisher deadline-resolution channel, motivated below from weak public signals. The Q -channel is the Glosten-Milgrom order-flow channel, derived below from informed trading and bid-ask spreads. Both channels are signals about the same terminal state Θ . The next two subsections develop their per-step conditional variances separately; the final subsection combines them under a conditional-orthogonality restriction.

2.1. Wright-Fisher Deadline-Resolution Channel

Let $\tilde{p}_u = \mathbb{P}(\Theta = 1 \mid \mathcal{F}_u^P)$ denote the market's posterior probability of YES after observing the public-information channel up to information time u . In Appendix A we present a structural learning model in which the market observes a stream of weak binary public signals about the eventual settlement state Θ . Standard generator-convergence arguments yield the diffusion limit

$$d\tilde{p}_u = \sqrt{\tilde{p}_u(1 - \tilde{p}_u)} d\tilde{B}_u^P, \quad (1)$$

which is the neutral Wright-Fisher diffusion in information time u . Here, $(\tilde{B}_u^P)$ is a Brownian motion.

Two features make (1) a natural object for a prediction-market posterior. First, $\tilde{p}_u$ is a martingale on $[0, 1]$, consistent with the requirement that the price be the conditional expectation of a bounded payoff. Second, the diffusion coefficient $\sigma(p) = \sqrt{p(1-p)}$ vanishes at 0 and 1, so the boundary points are absorbing; the only possible limit points are 0 and 1, matching binary settlement. Moreover, it is possible to establish via standard arguments that $\tilde{p}_u \rightarrow \tilde{p}_\infty \in \{0, 1\}$, i.e., the process is absorbed eventually almost surely.

Deadline clock. The SDE (1) absorbs in $\{0, 1\}$ eventually, but not necessarily by any particular calendar time. To force terminal resolution at the deadline T , introduce the deterministic information clock

$$A(t) = \int_0^t \frac{dr}{T-r} = -\log\left(\frac{T-t}{T}\right), \quad A'(t) = \frac{1}{T-t},$$

which accelerates as resolution approaches and satisfies $A(t) \rightarrow \infty$ as $t \uparrow T$. Define the calendar-time public posterior $p_t^P = \tilde{p}_{A(t)}$. After the time change, p_t^P satisfies

$$dp_t^P = \sqrt{\frac{p_t^P(1-p_t^P)}{T-t}} dW_t^P, \quad (2)$$

where (W_t^P) is a calendar-time Brownian motion. Appendix A.3 establishes the calendar-time variance-budget identity

$$\text{Var}[p_\tau^P \mid \mathcal{F}_t^P] = \mathbb{E}\left[\int_t^\tau \frac{p_s^P(1-p_s^P)}{T-s} ds \mid \mathcal{F}_t^P\right], \quad t \leq \tau \leq T. \quad (3)$$

The integrand $p_s^P(1-p_s^P)$ is remaining binary uncertainty and $1/(T-s)$ is the deadline information-release rate. As $\tau \uparrow T$, the right-hand side converges to $p_t^P(1-p_t^P)$, recovering $\text{Var}[p_T^P | \mathcal{F}_t^P] = \text{Var}[\Theta | \mathcal{F}_t^P] = p_t^P(1-p_t^P)$; this formalizes the statement that the deadline clock spends exactly the remaining binary uncertainty by time T .

Related models of prediction-market beliefs. Wright-Fisher and related belief-martingale diffusions on $[0, 1]$ have been used previously to model prediction-market prices and binary forecasts; see, among others, Archak and Ipeitotis [2010], Restocchi et al. [2018], Dalen [2025].

Discretization. The rest of the paper works in discrete time on an hourly observation grid $t_i < t_{i+1}$ with step $\Delta_i = t_{i+1} - t_i$ and time to resolution $\tau_i = T - t_i$. Treating the price as locally constant on $[t_i, t_{i+1})$, an Euler-Maruyama step of (2) gives the one-step conditional variance

$$\text{Var}(\Delta p_i^P | \mathcal{F}_{t_i}^P) \approx \frac{p_{t_i}(1-p_{t_i})}{\tau_i} \Delta_i, \quad (4)$$

where $\Delta p_i^P = p_{t_{i+1}}^P - p_{t_i}^P$. This is the deadline-resolution variance component used throughout the rest of the paper. In the empirical implementation, this public-channel expression is evaluated at the observed full-market price p_{t_i} ; this treats the current information from both channels as summarized by the common market posterior while keeping the two variance contributions separate.

2.2. Glosten-Milgrom Order-Flow Channel

The second channel captures private information revealed through order flow. Fix time t and write

$$p = \mathbb{P}(\Theta = 1 | \mathcal{F}_t^Q)$$

for the order-flow-channel posterior. An information-sensitive order-flow event arrives. The order is either a buy order for the YES claim, denoted by B , or a sell order, denoted by S . With probability α , the trader is informed and observes Θ ; with probability $1 - \alpha$, the trader is uninformed. The informed trader buys YES when $\Theta = 1$ and sells YES when $\Theta = 0$. The uninformed trader buys or sells with probability $1/2$ each. Therefore

$$\begin{aligned} \mathbb{P}(B | \Theta = 1) &= \frac{1 + \alpha}{2}, & \mathbb{P}(B | \Theta = 0) &= \frac{1 - \alpha}{2}, \\ \mathbb{P}(S | \Theta = 1) &= \frac{1 - \alpha}{2}, & \mathbb{P}(S | \Theta = 0) &= \frac{1 + \alpha}{2}. \end{aligned}$$

The likelihood ratio of a buy order is $(1 + \alpha)/(1 - \alpha) > 1$, so a buy order is positive evidence for the YES state. The unconditional buy probability is

$$\mathbb{P}(B | \mathcal{F}_t^Q) = p \frac{1 + \alpha}{2} + (1 - p) \frac{1 - \alpha}{2} = \frac{1 + \alpha(2p - 1)}{2}.$$

Bayes' rule gives the posterior after a buy order:

$$p^+ = \mathbb{P}(\Theta = 1 | B, \mathcal{F}_t^Q) = \frac{p(1 + \alpha)}{1 + \alpha(2p - 1)},$$

and after a sell order,

$$p^- = \mathbb{P}(\Theta = 1 | S, \mathcal{F}_t^Q) = \frac{p(1 - \alpha)}{1 - \alpha(2p - 1)}.$$

A competitive risk-neutral market maker breaks even conditional on order direction: quoting ask price $a = \mathbb{E}[\Theta \mid B, \mathcal{F}_t^Q] = p^+$ and bid price $b = \mathbb{E}[\Theta \mid S, \mathcal{F}_t^Q] = p^-$. The spread is

$$s = a - b = p^+ - p^- = \frac{4\alpha p(1-p)}{1 - \alpha^2(2p-1)^2}.$$

The two possible posterior jumps are

$$p^+ - p = \frac{2\alpha p(1-p)}{1 + \alpha(2p-1)}, \quad p^- - p = -\frac{2\alpha p(1-p)}{1 - \alpha(2p-1)}.$$

Because $p_{\text{after}} = \mathbb{P}(\Theta = 1 \mid \mathcal{F}_t^Q, B/S)$ is a posterior, it is a conditional martingale, and its one-event conditional variance is

$$\begin{aligned} v^Q(p, \alpha) &= \text{Var}(p_{\text{after}} - p \mid \mathcal{F}_t^Q) \\ &= \mathbb{P}(B \mid \mathcal{F}_t^Q)(p^+ - p)^2 + \mathbb{P}(S \mid \mathcal{F}_t^Q)(p^- - p)^2 \\ &= \frac{4\alpha^2 p^2(1-p)^2}{1 - \alpha^2(2p-1)^2} \\ &= \frac{s^2}{4} \left[1 - \alpha^2(2p-1)^2 \right]. \end{aligned} \tag{5}$$

Equation (5) is the exact binary Glosten-Milgrom one-event posterior variance. A parsimonious reduced form uses the first-order adverse-selection scale

$$v^Q(p, \alpha) \approx \frac{s^2}{4},$$

and treats the omitted correction $1 - \alpha^2(2p-1)^2$ as part of the reduced-form empirical scale. Because this correction depends on p and α , the fitted scale should be interpreted as an average price-impact multiplier rather than as a structural estimate of the informed-trader share.

To aggregate per-event variance to a per-step variance over the same hourly grid $[t_i, t_{i+1})$ used in (4), let N_t^Q be the counting process of information-sensitive order-flow events, and assume

$$\mathbb{E}[dN_t^Q \mid \mathcal{F}_t^Q] = K\nu(V_t) dt,$$

where V_t is observed volume, $\nu(V_t)$ maps measured volume into an event-arrival proxy, and $K \geq 0$ is a scale parameter capturing the effective intensity of information-sensitive order flow after controlling for V_t . Treating the per-step state variables as locally constant on $[t_i, t_{i+1})$, the expected number of events is $K\nu(V_{t_i})\Delta_i$, and aggregating the reduced-form per-event variance gives the per-step conditional variance

$$\text{Var}(\Delta p_i^Q \mid \mathcal{F}_{t_i}^Q) \approx K\nu(V_{t_i}) \frac{s_{t_i}^2}{4} \Delta_i. \tag{6}$$

This reduced-form expression treats $s_{t_i}^2/4$, the squared half-spread, as the adverse-selection price-impact scale and lets K absorb the GM correction factor, trade aggregation, and the conversion from volume to information-event intensity.

2.3. Combining the Two Channels in Discrete Time

Write the one-step price innovation on $[t_i, t_{i+1})$ as the sum of the two channel innovations:

$$\varepsilon_i = p_{t_{i+1}} - p_{t_i} = \Delta p_i^P + \Delta p_i^Q,$$

where Δp_i^P and Δp_i^Q are the conditional martingale increments generated by the Wright-Fisher deadline-resolution channel and the Glosten-Milgrom order-flow channel, respectively.

Conditional orthogonality. We impose the identifying restriction that the two channel innovations are conditionally uncorrelated:

$$\text{Cov}(\Delta p_i^P, \Delta p_i^Q \mid \mathcal{F}_{t_i}) = 0.$$

In practice the two channels may be correlated: a public news event can simultaneously move the consensus and trigger informed order flow, and informed trading itself may be timed around expected public information. We adopt the orthogonality restriction for tractability and identification; an unrestricted specification would require pinning down a cross-channel covariance term that the data do not separately reveal.

Variance decomposition. Under conditional orthogonality, the per-step variances of the two channels in (4) and (6) add:

$$h_i^2(K) = \text{Var}(\varepsilon_i \mid \mathcal{F}_{t_i}) = \left[\frac{p_{t_i}(1 - p_{t_i})}{\tau_i} + K \nu(V_{t_i}) \frac{s_{t_i}^2}{4} \right] \Delta_i. \quad (7)$$

This closed-form predictor is the structural object taken to the data in the rest of the paper. It combines a single deadline-resolution term, a single spread-scaled order-flow term, and one non-negative scale parameter K . The choice of activity proxy $\nu(\cdot)$, the estimation of K , and the design of the forecast evaluation are discussed in Section 3; the out-of-sample results are in Section 4.

Under the structural conditional-variance specification, the price innovation can be written in standardized form as

$$\varepsilon_i = h_i(K) z_i, \quad \mathbb{E}[z_i \mid \mathcal{F}_{t_i}] = 0, \quad \text{Var}(z_i \mid \mathcal{F}_{t_i}) = 1, \quad (8)$$

so that $h_i(K)$ is the conditional scale of the price innovation and z_i is a unit-variance standardized shock. This decomposition does not impose a Gaussian distribution, or any other parametric distribution, on z_i . Throughout, we refer to the conditional standard deviation $h_i = \sqrt{h_i^2(K)}$, the square root of the structural conditional variance, as the *volatility* of the one-step price innovation.

3. Data and Forecast Evaluation

This section describes the empirical object and the forecast evaluation. We first construct an hourly Kalshi panel and define the one-hour volatility target. We then summarize the model families in the forecast comparison, describe the expanding-window estimation and interval-score evaluation, and close with model-free stylized facts that corroborate the structural variables of Section 2.

3.1. Hourly forecasting panel

The data are contract-hour observations from Kalshi between August 2021 and April 2026. Each observation is one listed yes/no claim during one calendar hour. Markets with several mutually exclusive outcomes appear on Kalshi as separate binary claims, one per candidate outcome, so the unit of observation is the claim-level contract rather than the broader event. For each contract-hour we observe price information, trading volume, and the best bid and ask at the close of the hour.

Forecast-origin state. Let t_i denote the close of hour i . The main price variable is the end-of-hour mid-quote

$$p_{t_i} = \frac{1}{2}(b_{t_i} + a_{t_i}),$$

where b_{t_i} and a_{t_i} are the best YES bid and ask at the close of hour t_i , normalized to dollars in $[0, 1]$. The quoted spread is $s_{t_i} = a_{t_i} - b_{t_i}$, V_{t_i} is volume traded during the hour ending at t_i , and $\tau_i = T - t_i$ is time to resolution. These are all forecast-origin variables, observed before the next-hour horizon $[t_i, t_{i+1})$. We use the mid-quote rather than the last trade because it reflects the closing order book and is less exposed to bid–ask bounce or stale prints; Appendix E repeats the main comparisons on last-trade prices and finds the same ranking.

Forecast target. At the close of hour i , after observing the order book at t_i , we forecast the signed one-hour innovation

$$\varepsilon_i = p_{t_{i+1}} - p_{t_i}, \quad t_{i+1} = t_i + 1. \quad (9)$$

The conditional scale of this innovation is the empirical object modeled by the structural variance $h_i^2(K)$ in (7). Because the data are hourly, the step length in (7) is $\Delta_i = 1$ throughout.

Analysis filtering. The headline forecasts target the conditional scale of active price updates: the size of the next-hour probability move conditional on an update occurring. We therefore exclude hours whose next-hour innovation is numerically zero ($|\varepsilon_i| \leq 10^{-10}$), and retain contracts with at least 48 hourly observations. This restriction defines the forecasting target. It is not a claim that inactive hours are uninformative. An unconditional hourly model would have to forecast two margins: whether the price updates at all and, conditional on an update, the size of the move. The present paper studies the second margin, which is the margin most directly connected to conditional volatility once information reaches prices.

Appendix D reports two unfiltered-panel checks. The first keeps the active-update estimates fixed and adds zero-update hours back to the test panel. The second removes the zero-update filter from both estimation and evaluation. These exercises show that the central structural hierarchy is not an artifact of the active-update target. After the headline filters and the monthly split below, the pooled out-of-sample panel contains 880,719 contract-hours for closed-form specifications and 878,187 for specifications with lagged ARCH/GARCH recursions.

3.2. Model classes

We compare three broad classes of models, aiming to isolate what prediction-market structure adds. The first class contains closed-form structural predictors, including deadline resolution alone, the binary-market benchmark of Archak and Ipeiritos [2010], and DR-AS. The second contains plain ARCH/GARCH models, which serve as generic time-series benchmarks with no prediction-market state variables. The third adds residual ARCH/GARCH dynamics on top of a structural baseline, asking whether volatility clustering remains after conditioning on price, time to resolution, spread, and volume. Table 1 groups the main specification families by the economic question they address.

For the DR-AS family we compare four activity proxies, $\nu(V) \in \{1, \log(1 + V), \sqrt{V}, V\}$. The constant proxy isolates the spread channel, the concave proxies allow the marginal effect of activity to diminish with volume, and the linear proxy corresponds to volume being proportional to the arrival rate of information-sensitive trades.

For the activity-scaled deadline resolution family, we use $a(V_i) = \nu(V_i)/\bar{V}$ ($\bar{V} > 0$) with $\nu(V) \in \{V, \sqrt{V}, \log(1 + V)\}$. This family asks whether trading activity primarily changes the speed of the deadline clock rather than adding an adverse-selection component. A constant-activity version would amount only to a fitted rescaling of the parameter-free deadline benchmark, so it is not treated as a separate activity-scaled deadline specification.

Table 1: Specification families in the forecast comparison. Models are grouped by their economic role.

Family	Representative specification	Purpose
Deadline resolution	$p(1 - p)/\tau$	Binary uncertainty and fixed-deadline structure only
Archak–Ipeirotis benchmark Archak and Ipeirotis [2010]	$\varphi^2(\Phi^{-1}(p))/\tau$	Latent-belief binary-market volatility benchmark
DR-AS closed form	$p(1 - p)/\tau + K\nu(V)s^2/4$	Adds spread-scaled order-flow variance
Activity-scaled deadline resolution	$a(V)p(1 - p)/\tau$	Tests whether activity rescales the deadline clock
Plain ARCH/GARCH	ARCH(1), GARCH(1,1)	Generic volatility persistence without prediction-market state variables
Additive structural dynamics	ARCH/GARCH + structural baseline	Residual clustering after conditioning on binary-market structure

3.3. Estimation and interval-score evaluation

Each specification produces, for every retained contract-hour i , a volatility forecast h_i , equivalently a conditional variance forecast h_i^2 . We estimate all specifications in a monthly expanding-window design: each calendar month from September 2021 through April 2026 serves once as the test window, with all earlier observations used for training. This yields 56 out-of-sample test months. A contract may appear in both a training window and a later test month, but no observation from a test month is used to fit that month’s parameters.

Parameters are estimated separately for each test month using forecast-origin volume weights. The same weights are used in evaluation, so the reported scores emphasize contract-hours in which more risk is traded. Closed-form DR-AS fits only the nonnegative order-flow scale K ; dynamic specifications add ARCH/GARCH residual dynamics to selected structural baselines. Appendix B provides the quasi-likelihood objective and the implementation details for the dynamic recursions.

The conditional variance is latent, so we evaluate a volatility forecast through the prediction interval it implies for the next-hour price. From the forecast-origin scale h_i , we form a symmetric nominal 95% interval around p_{t_i} , using the normal-reference multiplier $z_{0.975} = \Phi^{-1}(0.975)$ and clipping to the admissible range $[0, 1]$:

$$L_i = \max\{0, p_{t_i} - z_{0.975}h_i\}, \quad U_i = \min\{1, p_{t_i} + z_{0.975}h_i\}.$$

The normal reference is only a common map from scale forecasts to intervals; it is not an assumption that standardized innovations are Gaussian.

Writing $p_i^+ = p_{t_{i+1}} = p_{t_i} + \varepsilon_i$ for the realized next-hour price, we score $[L_i, U_i]$ with the Winkler interval score [Winkler, 1972, Gneiting and Raftery, 2007],

$$\text{IS}_i = (U_i - L_i) + \frac{2}{\eta}(L_i - p_i^+) \mathbb{I}\{p_i^+ < L_i\} + \frac{2}{\eta}(p_i^+ - U_i) \mathbb{I}\{p_i^+ > U_i\}, \quad \eta = 0.05. \quad (10)$$

Lower scores are better. The width term rewards sharpness, while the penalty terms charge for intervals that miss the realized price move. The score therefore penalizes both intervals that are too wide and intervals that are too narrow to cover.

The headline metric is the volume-weighted interval score,

$$\text{VW-IS} = \frac{\sum_i V_{t_i} \text{IS}_i}{\sum_i V_{t_i}}, \quad (11)$$

reported alongside volume-weighted coverage,

$$\text{VW-Cov} = \frac{\sum_i V_{t_i} \mathbb{I}\{L_i \leq p_i^+ \leq U_i\}}{\sum_i V_{t_i}}. \quad (12)$$

Coverage is a calibration diagnostic rather than the ranking metric, since a model can obtain high coverage by making intervals too wide. A useful model posts a low interval score while keeping coverage near the nominal 95% level.

3.4. Stylized facts

Before fitting the multivariate specifications, we ask whether the mechanisms in (7) are visible directly in the data. We sort observations into bins by each candidate driver and compute the volume-weighted root-mean-square next-hour innovation (VW-RMS) within each bin. Appendix C gives the exact binning statistic and additional figures.

Table 2: Model-free driver-strength ratios. Each ratio is the maximum bin VW-RMS over the minimum. The table omits $1/\sqrt{\tau}$, a monotone re-expression of time to resolution rather than a separate primitive driver; the full eight-driver table is in Appendix C.

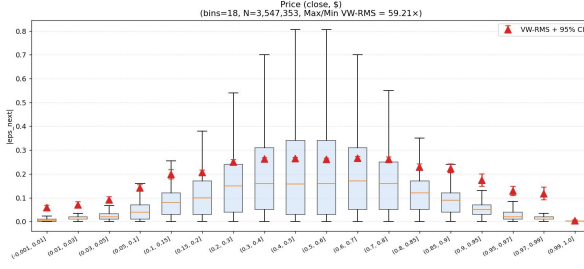
Rank	Driver	Ratio	Interpretation
1	Price p	59.2×	Boundary shape in $p(1-p)$
2	Time to resolution τ	7.4×	Deadline effect
3	Category	5.2×	Proxy for duration / market type
4	Volume V	3.3×	Activity / information-flow proxy
5	Time % to resolution	2.9×	Normalized time-to-resolution proxy
6	Open interest	2.4×	Participation proxy
7	Spread s	1.7×	Adverse-selection / liquidity proxy

Table 2 ranks the candidate drivers by the ratio of the largest bin VW-RMS to the smallest. The strongest patterns are the two primitive inputs of the deadline-resolution variance (4). Price leads by a wide margin: realized volatility peaks near 50 cents and collapses toward the binary boundaries, matching the inverted-U of $p(1-p)$. Time to resolution is the next strongest primitive driver, with volatility rising as resolution approaches. Category also has a large univariate contrast, but as we argue below, it is a composite proxy for market type and duration mix rather than a primitive state variable.

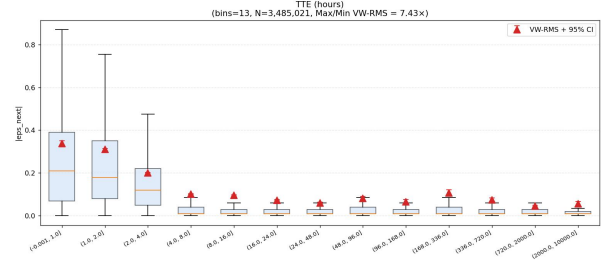
The microstructure variables are weaker in isolation: volume has a moderate univariate contrast and spread is weaker still. This does not contradict the order-flow channel, because DR-AS uses volume and spread jointly through $\nu(V)s^2/4$. Section 4 separates the deadline and order-flow channels in the forecast comparison.

4. Main Results

This section compares the out-of-sample interval forecasts from the model families in Section 3.2. Table 3 reports VW-IS, VW-Cov, and volume-weighted interval width for the main specifications on the active-update mid-quote panel. The point estimates form three economic tiers: plain ARCH/GARCH models perform worst, closed-form structural models perform substantially better, and structural-dynamic models perform best. The lowest VW-IS is attained by GARCH+DR-AS ($\sqrt{V}$). Because the leading structural-dynamic specifications are close in level, we use paired



(a) Price p . Max/Min = 59.2 $\times$.



(b) Time to resolution in hours. Max/Min = 7.4 $\times$.

Figure 1: Model-free binning diagnostics for the two primitive inputs of the deadline-resolution term. Red markers are volume-weighted RMS next-hour innovations within each bin; error bars are 95% within-bin bootstrap confidence intervals.

bootstrap comparisons below to assess whether the top of the ranking is statistically separated rather than a point-estimate artifact.

Table 3: Out-of-sample forecast comparison on the active-update mid-quote panel, ranked by VW-IS. Results are pooled across 56 expanding-window test months. VW-Width is the volume-weighted average interval width. Pairwise bootstrap comparisons for the ranking claims are reported in Table 4.

Specification	VW-IS $\downarrow$	VW-Cov	VW-Width
<i>Structural + residual dynamics</i>			
GARCH+DR-AS ($\sqrt{V}$)	0.4620	0.9592	0.3438
ARCH+DR-AS ($\sqrt{V}$)	0.4757	0.9582	0.3508
GARCH+deadline resolution	0.4757	0.9568	0.3529
GARCH+Archak-Ipeiritis	0.4767	0.9568	0.3528
ARCH+deadline resolution	0.4905	0.9470	0.3448
GARCH+activity-scaled deadline ($\sqrt{V}$)	0.4925	0.9547	0.3573
ARCH+Archak-Ipeiritis	0.4926	0.9477	0.3469
<i>Closed-form structural predictors</i>			
DR-AS ($\sqrt{V}$)	0.5085	0.9439	0.3805
DR-AS ($\log(1 + V)$)	0.5187	0.9395	0.3871
DR-AS (constant activity)	0.5246	0.9369	0.3920
DR-AS (linear V)	0.5263	0.9429	0.3977
Deadline resolution only	0.5829	0.8673	0.3397
Archak-Ipeiritis benchmark	0.6511	0.7862	0.2760
<i>Plain dynamic benchmarks</i>			
GARCH(1,1) plain	0.7675	0.8989	0.3053
ARCH(1) plain	0.7788	0.8969	0.3118

Pairwise ranking evidence. The VW-IS entries in Table 3 are pooled point estimates. To assess the ranking claims, we compute paired bootstrap confidence intervals for score differences. For two specifications A and B , define

$$\Delta_{B-A} = \text{VW-IS}(B) - \text{VW-IS}(A),$$

computed on the common evaluation support. Positive values favor specification A . The bootstrap resamples contract-level clusters with replacement and recomputes both scores on the same resampled contracts in each draw. This paired design uses the fact that competing specifications

are evaluated on the same contract-hours, so much of the sampling variation is common across specifications.

We compute the full within-family pairwise matrix for the leading structural-dynamic specifications and for the closed-form structural specifications. Table 4 reports the comparisons that determine the headline ranking tiers and the structural interpretation of the closed-form model. The purpose is not to attach economic meaning to every lower-tier ordering, but to test whether the main ranking claims are separated from sampling variation.

Table 4: Pairwise bootstrap comparisons that determine the main ranking claims on the active-update mid-quote panel. The reported difference is $\Delta_{B-A} = \text{VW-IS}(B) - \text{VW-IS}(A)$, so positive values favor specification A . Confidence intervals are paired contract-cluster bootstrap intervals.

Specification A	Specification B	Δ_{B-A}	95% CI
<i>Panel A. Structural-dynamic ranking</i>			
GARCH+DR-AS ($\sqrt{V}$)	ARCH+DR-AS ($\sqrt{V}$)	0.0122	[0.0055, 0.0187]
GARCH+DR-AS ($\sqrt{V}$)	GARCH+deadline resolution	0.0137	[0.0075, 0.0204]
GARCH+DR-AS ($\sqrt{V}$)	GARCH+Archak-Ipeirotis	0.0147	[0.0074, 0.0231]
GARCH+DR-AS ($\sqrt{V}$)	GARCH+activity-scaled line ($\sqrt{V}$)	dead- 0.0273	[0.0145, 0.0384]
GARCH+DR-AS ($\sqrt{V}$)	ARCH+deadline resolution	0.0285	[0.0191, 0.0380]
<i>Panel B. Closed-form structural core</i>			
DR-AS ($\sqrt{V}$)	DR-AS ($\log(1 + V)$)	0.0101	[0.0032, 0.0196]
DR-AS ($\sqrt{V}$)	DR-AS (constant activity)	0.0161	[0.0077, 0.0277]
DR-AS ($\sqrt{V}$)	DR-AS (linear V)	0.0177	[0.0089, 0.0256]
DR-AS ($\sqrt{V}$)	Deadline resolution only	0.0744	[0.0603, 0.0899]
Deadline resolution only	Archak-Ipeirotis benchmark	0.0682	[0.0318, 0.1160]
DR-AS ($\sqrt{V}$)	Archak-Ipeirotis benchmark	0.1426	[0.0993, 0.1952]

Notes: Differences are computed on the common evaluation support for each pair. They need not equal the difference between rounded VW-IS entries in Table 3. The full within-family pairwise matrices are computed for the leading structural-dynamic specifications and for the closed-form structural specifications; the table reports the comparisons that identify the headline ranking tiers and structural ingredients.

Panel A shows that GARCH+DR-AS ($\sqrt{V}$) is not merely the lowest point estimate. It is statistically separated from the closest structural-dynamic alternatives, including the same DR-AS baseline with only ARCH dynamics, GARCH added to the deadline-resolution baseline, GARCH added to the Archak-Ipeirotis benchmark, and activity-scaled deadline variants. The full dynamic pairwise matrix also cautions against overinterpreting the runner-up ordering: ARCH+DR-AS, GARCH+deadline resolution, and GARCH+Archak-Ipeirotis form a statistically close runner-up tier.

Panel B shows the same structural pattern in the closed-form horse race. First, within DR-AS, the $\sqrt{V}$ activity proxy is statistically separated from $\log(1 + V)$, constant activity, and linear V . Second, adding the spread-flow term to deadline resolution is not a point-estimate artifact: DR-AS ($\sqrt{V}$) significantly improves on deadline resolution alone. Third, the deadline-resolution shape improves on the Archak-Ipeirotis benchmark, and the full DR-AS specification is separated even more strongly from Archak-Ipeirotis. We therefore treat the DR-AS ($\sqrt{V}$) component as the stable closed-form structural core.

Inactive-hour robustness. The headline comparison conditions on active price updates. Ap-

pendix D restores zero-update hours in two complementary ways. First, we keep the active-update estimates fixed and evaluate the fitted forecasts on the full test panel. Second, we remove the zero-update filter from both estimation and evaluation. In both exercises, the ranking that matters for the paper is preserved: forecasts using structural variables remain in the leading group, DR-AS($\sqrt{V}$) remains the best closed-form specification, and plain ARCH/GARCH benchmarks remain far from the structural specifications. The inactive-hour checks therefore support the interpretation of Table 3: the forecasting gains come from the structural state variables rather than from the active-update sample definition.

4.1. Where the forecast gains come from

With the main ranking and its paired uncertainty established, we read Table 3 through three comparisons: generic dynamics versus structural state variables, the two ingredients inside the structural model, and residual dynamics after conditioning on structure.

Structural variables versus generic dynamics. The central comparison is between plain GARCH(1,1) and the closed-form DR-AS model. Plain GARCH(1,1) attains VW-IS = 0.7675, while DR-AS($\sqrt{V}$) attains 0.5085, a 34% improvement. This is not a close ranking decision: the plain dynamic benchmarks sit far outside the structural tiers in Table 3. This gain comes from the forecast-origin state variables (p, τ, s, V) , not from a richer time-series recursion. The bulk of the forecasting improvement therefore comes from contemporaneous prediction-market state variables rather than from return persistence alone.

The two structural ingredients. Within the closed-form structural family, the progression from the Archak–Ipeiritis benchmark to deadline resolution to DR-AS separates two design choices along this ordering. The Archak–Ipeiritis benchmark uses the probit-Brownian shape $\varphi^2(\Phi^{-1}(p))$; deadline resolution replaces it with the Wright–Fisher/Bernoulli shape $p(1-p)$; and DR-AS adds the order-flow term $K\nu(V)s^2/4$.

Replacing the Archak–Ipeiritis price shape with $p(1-p)$ improves VW-IS from 0.6511 to 0.5829, a 10% reduction. Both shapes vanish at the binary boundaries, but $p(1-p)$ is the variance of the binary payoff itself. Empirically, that simple binary-outcome shape fits the heterogeneous Kalshi panel better than the probit-Brownian shape.

Adding the order-flow term further improves VW-IS from 0.5829 to 0.5085, a 13% reduction. This term enters through the squared half-spread and trading activity: the spread measures the adverse-selection price-impact scale, while volume proxies for the frequency of information-sensitive order flow. Thus the deadline channel captures smooth calendar-time resolution of binary uncertainty, and the order-flow channel captures additional variance from information entering through trading. The paired closed-form comparisons reinforce this interpretation: DR-AS($\sqrt{V}$) is statistically separated from the deadline-only and Archak–Ipeiritis benchmarks, as well as from the activity-scaled-deadline family.

State-space patterns. Figure 2 shows the forecast-score analogue of the model-free patterns in Section 3.4. The gap between plain GARCH and the structural curves shows where binary-market structure helps; the gap between deadline resolution and DR-AS shows where the order-flow term adds information.

The price panel shows the clearest pattern. The structural advantage over plain GARCH is largest near $p = 1/2$, where binary uncertainty is highest, and shrinks toward the boundaries, where little uncertainty remains to be resolved. This is consistent with the $p(1-p)$ shape of the deadline-resolution variance.

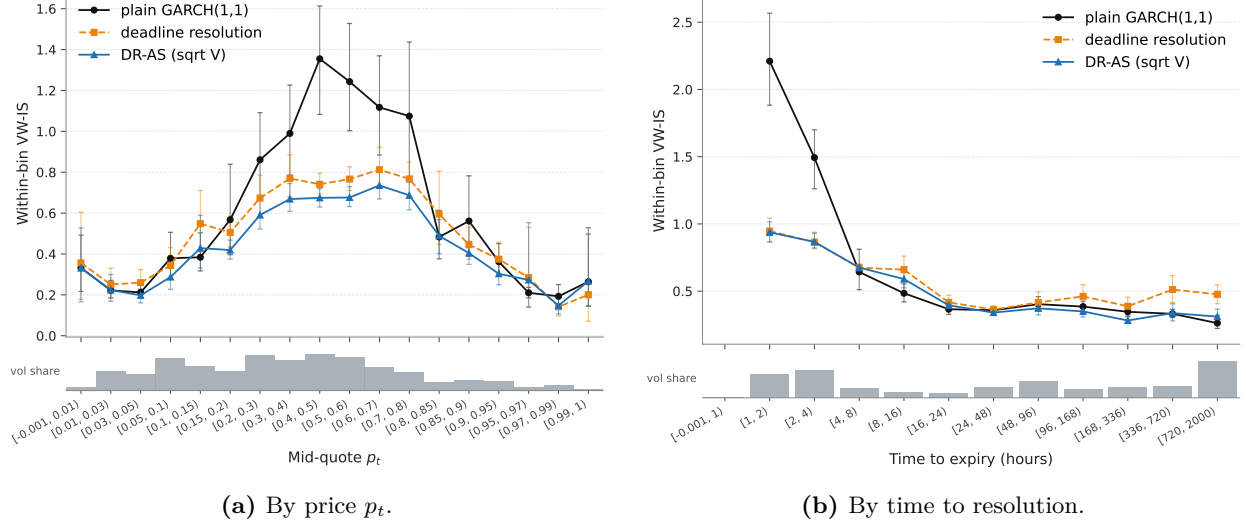


Figure 2: Within-bin out-of-sample VW-IS (lower is better) for plain GARCH, deadline resolution alone, and DR-AS, against price (left) and time to resolution (right). Error bars are 95% within-bin bootstrap confidence intervals.

The time-to-resolution panel separates the two structural components. Plain GARCH performs poorly at short horizons, consistent with its lack of a direct time-to-resolution state variable. Deadline resolution absorbs much of this short-horizon pattern through its $1/\tau$ factor. DR-AS then improves most visibly on deadline-only at longer horizons, where the deadline force is weaker and the additive order-flow term accounts for a larger share of forecast variance. Unlike the deadline term, the order-flow term does not mechanically scale with $1/\tau$; spreads and activity can matter throughout the contract's life.

Residual dynamics. Adding GARCH(1,1) dynamics to the DR-AS baseline lowers VW-IS from 0.5085 to 0.4620, a 9% reduction in the point estimate. Residual time-series predictability remains, but it is secondary. The pairwise comparisons above show that, within the structural-dynamic family, the resulting GARCH+DR-AS ($\sqrt{V}$) specification is statistically separated from the closest alternatives. Relative to plain GARCH, GARCH+DR-AS improves VW-IS by 40%. The closed-form structural baseline accounts for most of that gain, with the residual GARCH layer providing the remaining improvement. GARCH is therefore complementary to the structural state variables, not a substitute for them.

4.2. Activity scaling within DR-AS

The remaining modeling choice within the DR-AS form is how measured activity enters the order-flow channel. Table 5 compares the four activity proxies used in the closed-form DR-AS specifications.

The comparison favors concave activity scaling. The paired closed-form comparisons show that $\sqrt{V}$ is statistically separated from the other activity proxies within the main DR-AS family. We therefore interpret $\sqrt{V}$ as the stable activity proxy in the closed-form structural model. The exact ordering among the remaining proxies is less important: the lower-ranked activity variants contain statistically close comparisons, so the economic conclusion is the superiority of the concave $\sqrt{V}$ scaling rather than a sharp ranking of every alternative proxy.

Table 5: Activity scaling within the closed-form DR-AS family.

$\nu(V)$ proxy	VW-IS	VW-Cov	VW-Width
$\sqrt{V}$	0.5085	0.9439	0.3805
$\log(1 + V)$	0.5187	0.9395	0.3871
constant	0.5246	0.9369	0.3920
V	0.5263	0.9429	0.3977

5. Category Heterogeneity and Model Portability

The pooled comparison in Section 4 treats the Kalshi panel as one forecasting target. This section asks whether that result travels across categories, or instead reflects the large, liquid categories that dominate the volume-weighted criterion. The answer has two parts. The best specifications in major categories remain within the same deadline-resolution/DR-AS family, but categories differ sharply in difficulty and in the concentration of large price moves.

5.1. Category composition

The per-category exercises below use the 11 categories with at least 10,000 active-update observations. The remaining six named categories are shown in Table 6 for sample-composition accounting but are too small for stable category-specific fits.

Table 6: Sample composition by Kalshi category, ranked by active-update observations. The dashed line separates the eleven major categories used in the per-category analysis from the six smaller categories not used. Volume totals are in millions of contracts. The final column reports the best-performing specification in each category under the global fit defined in Section 5.3. A further 1,624 contract-hours carry no category label and are omitted from this table.

Category	Obs	Share of obs	Volume (M)	Share of vol	Best global spec
Sports	217,557	24.8%	2,172	51.6%	GARCH+deadline resolution
Politics	184,472	21.0%	986	23.4%	GARCH+DR-AS ($\sqrt{V}$)
Entertainment	123,059	14.0%	246	5.8%	GARCH+DR-AS ($\sqrt{V}$)
Economics	67,775	7.7%	234	5.6%	GARCH+DR-AS ($\sqrt{V}$)
Elections	66,968	7.6%	262	6.2%	GARCH+DR-AS ($\sqrt{V}$)
Crypto	64,978	7.4%	120	2.9%	GARCH+DR-AS ($\sqrt{V}$)
Mentions	51,167	5.8%	38	0.9%	DR-AS ($\sqrt{V}$)
Science and Technology	30,168	3.4%	41	1.0%	GARCH+DR-AS ($\sqrt{V}$)
Companies	22,420	2.6%	37	0.9%	GARCH+DR-AS ($\sqrt{V}$)
Climate and Weather	15,827	1.8%	13	0.3%	GARCH+deadline resolution
Commodities	14,948	1.7%	16	0.4%	GARCH+DR-AS ($\sqrt{V}$)
Financials	8,242	0.9%	7	0.2%	GARCH+deadline resolution
Exotics	6,578	0.7%	35	0.8%	GARCH+DR-AS ($\sqrt{V}$)
Health	2,241	0.3%	1	0.0%	GARCH+deadline resolution
Social	1,212	0.1%	1	0.0%	GARCH+DR-AS ($\sqrt{V}$)
World	1,102	0.1%	1	0.0%	DR-AS ($\sqrt{V}$)
Transportation	381	0.0%	0	0.0%	GARCH+deadline resolution

The category distribution is highly concentrated. Sports and Politics account for about 46% of active-update observations and 75% of volume in the named-category panel. The 11 major categories account for about 98% of observations and 99% of volume, while the six smaller categories

together have only 19,756 active-update observations and little weight in the volume-weighted criterion.

5.2. Model portability across major categories

The final column of Table 6 shows that the category winners do not fragment into unrelated models. Among the 11 major categories, GARCH+DR-AS($\sqrt{V}$) wins eight, GARCH+deadline resolution wins two, and closed-form DR-AS($\sqrt{V}$) wins one. No major category is best served by a plain ARCH/GARCH specification.

The winning specifications differ in how much structure they use, but all retain the deadline-resolution backbone $p(1-p)/\tau$. They differ only in whether that backbone is augmented by the order-flow term $K\nu(V)s^2/4$, residual GARCH dynamics, or both. The order-flow term appears in the winning specification in nine of the eleven major categories, and GARCH dynamics appear in ten. Category heterogeneity therefore changes the preferred member of the family, not the broad volatility architecture.

5.3. Parameter portability

Model portability asks which specification works across categories. Parameter portability asks whether a given specification needs separate category-level estimates. For each specification, the *global fit* estimates parameters on the pooled training panel and applies them to every category in the test month; the *per-category fit* re-estimates the same specification within each category. The expanding-window splits and metrics are otherwise identical. Because this exercise is restricted to the eleven major categories, score levels are not directly comparable to the pooled panel in Table 3. We measure the effect of category-specific refitting by

$$\Delta\text{VW-IS} = \frac{\text{VW-IS}^{\text{catfit}} - \text{VW-IS}^{\text{global}}}{\text{VW-IS}^{\text{global}}}, \quad \Delta\text{VW-Cov} = \text{VW-Cov}^{\text{catfit}} - \text{VW-Cov}^{\text{global}},$$

so $\Delta\text{VW-IS}$ is a relative change and $\Delta\text{VW-Cov}$ is a change in percentage points. Because VW-IS is a loss, positive $\Delta\text{VW-IS}$ means that per-category refitting worsens the interval score.

Table 7: Global fit versus per-category fit over the eleven major categories. $\Delta\text{VW-IS}$ is the relative change in VW-IS (percent) and $\Delta\text{VW-Cov}$ is the change in VW-Cov (percentage points); positive values mean per-category fit increases the corresponding metric relative to global fit.

Specification	VW-IS global	VW-Cov global	VW-IS per-cat	VW-Cov per-cat	$\Delta\text{VW-IS}$	$\Delta\text{VW-Cov}$
Deadline resolution only	0.5670	0.8678	0.5680	0.8676	+0.19%	-0.02
DR-AS (constant)	0.5176	0.9367	0.5263	0.9273	+1.68%	-0.94
DR-AS ($\log(1+V)$)	0.5114	0.9394	0.5231	0.9273	+2.30%	-1.21
DR-AS ($\sqrt{V}$)	0.5003	0.9440	0.5173	0.9364	+3.38%	-0.76
GARCH+deadline resolution	0.4700	0.9576	0.4781	0.9487	+1.73%	-0.89
GARCH+DR-AS ($\sqrt{V}$)	0.4684	0.9464	0.4730	0.9464	+0.97%	+0.01

Table 7 is a refitting exercise, not a new pooled horse race: it asks whether holding a specification fixed and estimating it separately by category improves out-of-sample performance. It does not. The global fit has lower VW-IS for every specification in the table. The pattern is also present in the largest categories, Sports and Politics, so it is not only a small-category phenomenon. At

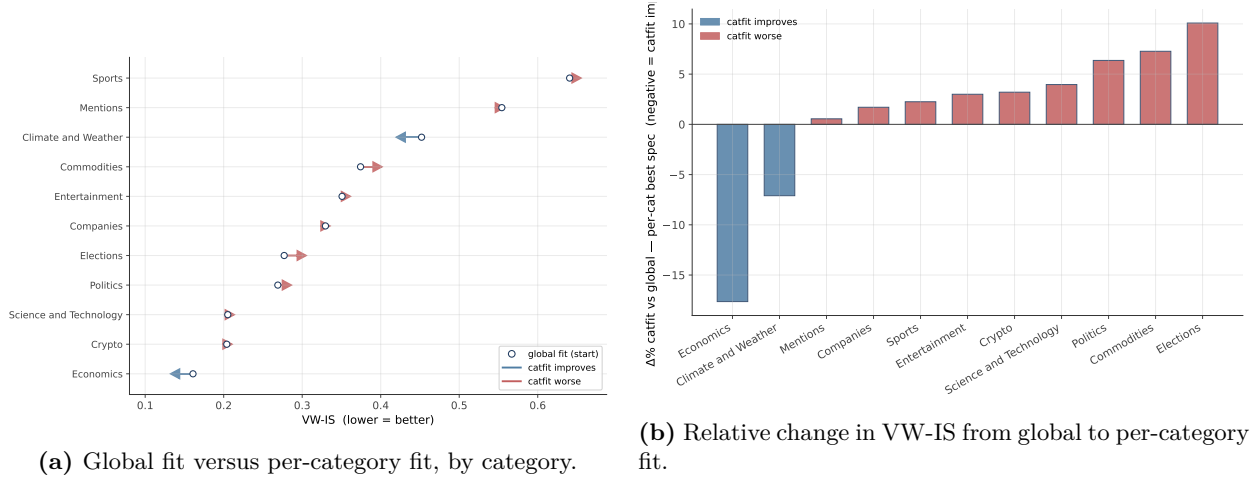


Figure 3: Per-category refitting of each category’s best global specification, across the 11 major categories. For each category, the specification with the lowest global-fit VW-IS is selected and re-estimated within the category. Panel (a): the open circle is the category-level VW-IS under the global fit; the arrow points to the per-category-fit value. Panel (b): the relative change $\Delta\text{VW-IS}$; negative values (blue) indicate that per-category refitting improves the score, positive values (red) indicate that it worsens the score. Per-category refitting improves the score in only two of the eleven major categories.

the category sample sizes available here, pooling therefore appears to be the more robust way to estimate the structural family.

Figure 3 repeats the comparison at the category level, now letting each category use its own best global specification. Even with this extra freedom, refitting improves the score in only two of the eleven major categories—Economics and Climate and Weather—and worsens it in the other nine, with the largest deterioration in Elections and Mentions. The two largest categories by volume, Sports and Politics, both fail to improve under refitting, so the volume-weighted pooled comparison in Table 7 reflects this deterioration. Category-specific calibration does not reliably improve on the global fit.

5.4. Smooth resolution and event-concentrated risk

Portability does not mean equal difficulty. The best global-fit VW-IS ranges from 0.16 in Economics to 0.64 in Sports, a fourfold difference. Figure 4 suggests that this spread is tied to the category’s information environment, not to a need for unrelated model architectures.

Economics is the smooth-resolution endpoint. In Panel A of Figure 4, the four representative structural specifications have nearly identical VW-IS in Economics: deadline resolution alone is already close to the winning specification, and adding the order-flow term or GARCH dynamics improves the score only marginally. Panel B shows the corresponding large-move pattern. Economics is the only major category whose volume-weighted share of large-move hours, 4.9%, is below its observation share, 8.0%. Trading volume is therefore not concentrated in the large-move tail. This fits the category’s economic structure: Economics contracts often resolve around scheduled macroeconomic releases—CPI, employment, GDP, and similar public data prints—where the resolution date is fixed and much of the price movement is the gradual incorporation of public information before the release.

Sports is the event-concentrated endpoint. In Panel A of Figure 4, the four specifications are again close to each other, but at a much higher level. The order-flow term and GARCH

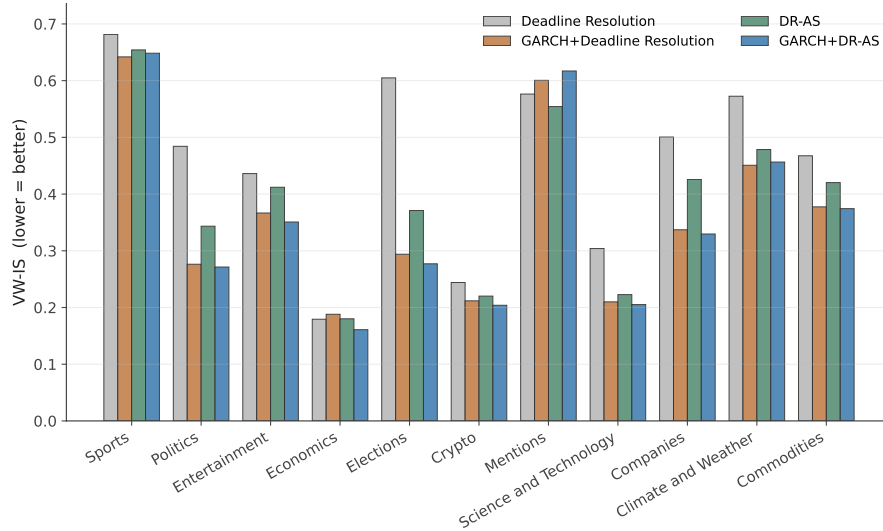
dynamics add only modest improvement over deadline resolution relative to the overall difficulty of the category. Panel B shows why the volume-weighted score is so hard to reduce. Only 7.3% of Sports hours fall in the large-move tail, but those hours carry 48.7% of Sports volume. Thus nearly half of the volume-weighted Sports evaluation is concentrated in jump-like hours. Sports contracts resolve around discrete game states and events—goals, injuries, period breaks, substitutions, and related changes—whose next-hour price impact is hard to anticipate from smooth forecast-origin state variables.

The other major categories lie between these endpoints. Category heterogeneity therefore does not overturn the structural-family result, but it shows where the smooth conditional-variance approximation is weakest. The model is closest to its intended setting in smooth-resolution categories and most incomplete in event-concentrated markets, where an event clock or jump component would be a natural extension.

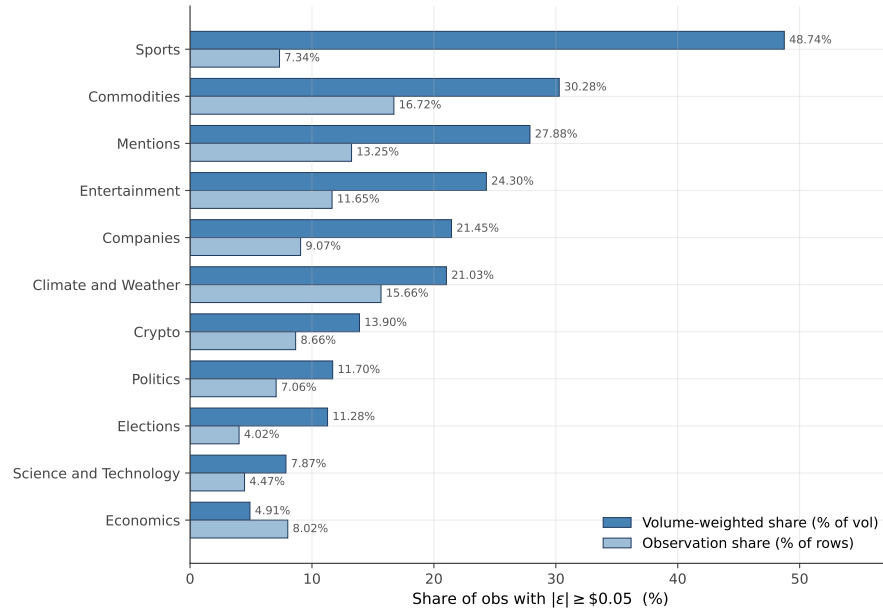
6. Conclusion

This paper builds the case that prediction-market volatility is shaped by features that ordinary asset-price volatility models do not encode: binary payoffs, fixed resolution dates, bounded probability prices, and information-sensitive order flow. In the Kalshi panel, forecasts built from these state variables substantially outperform plain ARCH/GARCH benchmarks, while residual GARCH dynamics provide a further but smaller improvement. The main empirical lesson is therefore not that time-series persistence is irrelevant, but that it works best after conditioning on the structure of the contract.

Several extensions follow naturally. First, the model treats deadline-resolution and order-flow innovations as conditionally orthogonal; future work could allow an explicit covariance or interaction term, capturing the fact that public information events and informed trading often arrive together. Second, our main exercise studies volatility conditional on active price updates. The unfiltered-panel checks show that the structural advantage remains when inactive hours are added back, but a full unconditional model would naturally combine an update/no-update hazard with the conditional scale of price moves. Third, event-heavy markets such as sports suggest the value of adding a jump or event-clock component, especially around scheduled announcements, game states, or other discrete information releases. These extensions would move the model from a parsimonious structural forecasting rule toward a richer description of how beliefs evolve in real time.



(a) Within-category VW-IS by specification.



(b) Observation and volume-weighted shares of large-move hours.

Figure 4: Category regimes across the 11 major categories. Panel A reports within-category VW-IS (lower is better) for four representative specifications: deadline resolution alone, GARCH+deadline resolution, closed-form DR-AS, and GARCH+DR-AS. Panel B reports the share of hourly observations with $|\varepsilon_i| \geq \$0.05$; light bars are row shares and dark bars are volume-weighted shares.

References

- T. G. Andersen and T. Bollerslev. Deutsche mark-dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *Journal of Finance*, 53(1):219–265, 1998.
- T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003.
- N. Archak and P. G. Ipeirotis. Modeling volatility in prediction markets. In *ACM Conference on Electronic Commerce*, pages 197–206, 2010.
- T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.
- Shaw Dalen. Toward black-scholes for prediction markets: A unified kernel and market-maker’s handbook, 2025.
- D. Easley and M. O’Hara. Time and the process of security price adjustment. *Journal of Finance*, 47(2):577–605, 1992.
- R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.
- Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1986.
- Stewart N. Ethier and M. F. Norman. Error estimate for the diffusion approximation of the wright-fisher model. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5096–5098, 1977. doi: 10.1073/pnas.74.11.5096.
- L. R. Glosten and P. R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, 1985.
- Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- R. Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119, 2003.
- J. Hasbrouck. Measuring the information content of stock trades. *Journal of Finance*, 46(1):179–207, 1991.
- S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes*. Academic Press, 1981.
- C. F. Manski. Interpreting the predictions of prediction markets. *Economics Letters*, 91(3):425–429, 2006.
- V. Restocchi, F. McGroarty, E. Gerding, and J. E. V. Johnson. The stylized facts of prediction markets: Analysis of price changes. *Physica A*, 515:159–170, 2018.
- G. E. Tauchen and M. Pitts. The price variability-volume relationship on speculative markets. *Econometrica*, 51(2):485–505, 1983.

- The Wall Street Journal. Why almost everyone loses-except a few sharks-on polymarket and kalshi. *The Wall Street Journal*, May 2026.
- R. L. Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.
- J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

A. Wright-Fisher Structural Learning Model

This appendix gives the structural learning model behind the Wright-Fisher deadline-resolution channel of Section 2.1. Section A.1 sets up a primitive Bayesian learning chain in which the market observes a stream of weak binary public signals about the eventual settlement state. Section A.2 shows that this chain converges weakly to the neutral Wright-Fisher diffusion as the signal strength vanishes. Section A.3 establishes the calendar-time variance-budget identity used to justify the deadline clock.

A.1. Primitive Public-Signal Experiment and Bayesian Learning

We micro-found the Wright-Fisher channel in information time u . The object of learning is the single terminal settlement state Θ , not a sequence of new binary outcomes. For $t < T$, the market does not observe Θ directly. Instead, it observes pre-settlement public signals whose distributions depend on that same terminal state. Thus

$$\tilde{p}_u = \mathbb{P}(\Theta = 1 \mid \mathcal{F}_u^P)$$

is the market's posterior probability that the final settlement will be YES after observing public information up to information time u . When both channels are present, the same local calculation is applied at the current full posterior $p_t = \mathbb{P}(\Theta = 1 \mid \mathcal{F}_t)$.

Over a small information-time interval $[u, u + \Delta u]$, the market observes a binary public micro-signal

$$X_{u+\Delta u} \in \{0, 1\}.$$

Here $X = 1$ means that the newly arrived public information leans toward a YES settlement, and $X = 0$ means that it leans toward a NO settlement. The variable $X_{u+\Delta u}$ is not the settlement outcome itself; it is an intermediate observation about the eventual settlement state. Public signal innovations are assumed conditionally independent across disjoint small intervals given the current information and the true state. Set

$$\delta = \sqrt{\Delta u}.$$

The primitive experiment is a mixture of a weak truth-aligned component and a consensus-baseline component:

$$X_{u+\Delta u} = \begin{cases} \Theta, & \text{with probability } \delta, \\ Z, & \text{with probability } 1 - \delta, \end{cases} \quad Z \mid \mathcal{F}_u^P \sim \text{Bernoulli}(\tilde{p}_u), \quad Z \perp \Theta \mid \mathcal{F}_u^P.$$

The first branch is a reduced-form representation of a public signal component aligned with the eventual settlement state, not literal early settlement. The market observes only $X_{u+\Delta u}$, not the branch from which it came. Equivalently,

$$\mathbb{P}(X_{u+\Delta u} = 1 \mid \Theta, \mathcal{F}_u^P) = (1 - \delta)\tilde{p}_u + \delta\Theta.$$

Ordinary public information is interpreted relative to the current market consensus $\tilde{p}_u$, but each micro-signal has a small tilt toward the true state. If $\Theta = 1$, then

$$\mathbb{P}(X = 1 \mid \Theta = 1, \mathcal{F}_u^P) = \tilde{p}_u + \delta(1 - \tilde{p}_u),$$

and if $\Theta = 0$, then

$$\mathbb{P}(X = 1 \mid \Theta = 0, \mathcal{F}_u^P) = \tilde{p}_u(1 - \delta).$$

Hence

$$\mathbb{P}(X = 1 \mid \Theta = 1, \mathcal{F}_u^P) - \mathbb{P}(X = 1 \mid \Theta = 0, \mathcal{F}_u^P) = \delta > 0.$$

The public signal is therefore genuinely truth-bearing: a YES-leaning signal is more likely when the true state is YES, and a NO-leaning signal is more likely when the true state is NO. The adaptive baseline is not a claim that the truth itself is drawn from Bernoulli($\tilde{p}_u$); it is a statement about how public news is read relative to the market's current expectation.

Predictive distribution, Bayes update, and local moments. Before observing X , the market averages over the two truth states. Using $\mathbb{P}(\Theta = 1 \mid \mathcal{F}_u^P) = \tilde{p}_u$,

$$\begin{aligned} \mathbb{P}(X = 1 \mid \mathcal{F}_u^P) &= \tilde{p}_u[\tilde{p}_u + \delta(1 - \tilde{p}_u)] + (1 - \tilde{p}_u)[\tilde{p}_u(1 - \delta)] \\ &= \tilde{p}_u. \end{aligned}$$

Thus

$$X_{u+\Delta u} \mid \mathcal{F}_u^P \sim \text{Bernoulli}(\tilde{p}_u). \quad (13)$$

This is a predictive statement about the next public signal; the terminal state Θ is fixed, and $\tilde{p}_u$ is the market's posterior probability given current information.

Bayes' rule gives the posterior after a YES-leaning public signal:

$$\begin{aligned} \tilde{p}_{u+\Delta u}^+ &= \mathbb{P}(\Theta = 1 \mid X = 1, \mathcal{F}_u^P) \\ &= \frac{\tilde{p}_u[\tilde{p}_u + \delta(1 - \tilde{p}_u)]}{\mathbb{P}(X = 1 \mid \mathcal{F}_u^P)} \\ &= \tilde{p}_u + \delta(1 - \tilde{p}_u). \end{aligned}$$

After a NO-leaning public signal,

$$\begin{aligned} \tilde{p}_{u+\Delta u}^- &= \mathbb{P}(\Theta = 1 \mid X = 0, \mathcal{F}_u^P) \\ &= \frac{\tilde{p}_u(1 - \tilde{p}_u)(1 - \delta)}{1 - \tilde{p}_u} \\ &= \tilde{p}_u(1 - \delta). \end{aligned}$$

Both cases can be written as the forecast-error update

$$\tilde{p}_{u+\Delta u} - \tilde{p}_u = \delta(X_{u+\Delta u} - \tilde{p}_u). \quad (14)$$

The learning rate multiplying the forecast error is δ ; the realized jump size still depends on $\tilde{p}_u$ through the surprise $X - \tilde{p}_u$.

Using (13) and (14),

$$\begin{aligned} \mathbb{E}[\tilde{p}_{u+\Delta u} - \tilde{p}_u \mid \mathcal{F}_u^P] &= \delta\{\mathbb{E}[X_{u+\Delta u} \mid \mathcal{F}_u^P] - \tilde{p}_u\} = 0, \\ \text{Var}(\tilde{p}_{u+\Delta u} - \tilde{p}_u \mid \mathcal{F}_u^P) &= \delta^2 \text{Var}(X_{u+\Delta u} \mid \mathcal{F}_u^P) \\ &= \tilde{p}_u(1 - \tilde{p}_u)\Delta u. \end{aligned}$$

The public posterior is therefore a zero-drift martingale in information time, and its local variance is proportional to remaining binary uncertainty. The variance $\tilde{p}_u(1 - \tilde{p}_u)$ is the predictive variance of a truth-bearing public signal, not the result of drawing the terminal state from the current posterior.

A.2. Diffusion Limit

We now sketch the diffusion-limit argument by which the chain of Section A.1 converges weakly to the neutral Wright-Fisher diffusion. The arguments are standard.

Set $h = \Delta u$ and $\delta = \sqrt{h}$, and consider the piecewise-constant interpolation of the chain in Section A.1. Conditional on $\tilde{p}_{kh} = p$, the chain has zero conditional drift (so $\tilde{p}$ is a martingale), conditional second moment $\mathbb{E}[(\tilde{p}_{(k+1)h} - p)^2 \mid \tilde{p}_{kh} = p] = h p(1-p)$, and bounded one-step jumps $|\tilde{p}_{(k+1)h} - \tilde{p}_{kh}| \leq \sqrt{h}$. The jump bound implies the Lindeberg condition, ruling out macroscopic jumps in the limit. A Taylor expansion of any $f \in C^2([0, 1])$ then gives uniform convergence of the discrete generator to $Af(p) = \frac{1}{2} p(1-p)f''(p)$, the generator of the neutral Wright-Fisher diffusion of population genetics [Karlin and Taylor, 1981, Ethier and Norman, 1977]. The diffusion coefficient $\sigma(p) = \sqrt{p(1-p)}$ vanishes at the boundary, making 0 and 1 absorbing, and is $\frac{1}{2}$ -Hölder on $[0, 1]$; the Yamada-Watanabe condition $\int_{0+} du/u = \infty$ then yields pathwise uniqueness, so the martingale problem for A is well posed. Combined with compactness of the state space $[0, 1]$, the martingale-problem convergence theorem [Ethier and Kurtz, 1986, Ch. 4, Cor. 8.7] delivers

$$\tilde{p}^h \Rightarrow \tilde{p} \quad \text{in } D([0, \infty), [0, 1]),$$

where the limit $\tilde{p}$ is the neutral Wright-Fisher diffusion

$$d\tilde{p}_u = \sqrt{\tilde{p}_u(1-\tilde{p}_u)} d\tilde{B}_u^P, \quad \tilde{p}_u \in [0, 1],$$

with 0 and 1 absorbing.

A.3. Calendar-Time Variance Budget

We now establish the variance-budget identity (3) used in Section 2.1 to justify the deadline clock. Recall the calendar-time Wright-Fisher SDE

$$dp_u^P = \sqrt{\frac{p_u^P(1-p_u^P)}{T-u}} dW_u^P.$$

Fix $t < T$ and let $\tau \in [t, T)$. Applying Itô's formula to $(p_u^P)^2$ on $[t, \tau]$ and substituting the SDE,

$$(p_\tau^P)^2 - (p_t^P)^2 = 2 \int_t^\tau p_u^P \sqrt{\frac{p_u^P(1-p_u^P)}{T-u}} dW_u^P + \int_t^\tau \frac{p_u^P(1-p_u^P)}{T-u} du. \quad (15)$$

The stochastic integrand

$$H_u = 2p_u^P \sqrt{\frac{p_u^P(1-p_u^P)}{T-u}}, \quad u \in [t, \tau],$$

is continuous and adapted. Since $p_u^P \in [0, 1]$ and $T-u \geq T-\tau > 0$ on $[t, \tau]$,

$$\mathbb{E} \left[\int_t^\tau H_u^2 du \right] \leq \int_t^\tau \frac{4}{T-\tau} du = \frac{4(\tau-t)}{T-\tau} < \infty.$$

Thus $\int_t^\cdot H_u dW_u^P$ is an $\mathcal{L}^2$ martingale and has zero conditional mean given $\mathcal{F}_t^P$. Taking $\mathbb{E}[\cdot \mid \mathcal{F}_t^P]$ on both sides of (15),

$$\mathbb{E}[(p_\tau^P)^2 \mid \mathcal{F}_t^P] - (p_t^P)^2 = \mathbb{E} \left[\int_t^\tau \frac{p_u^P(1-p_u^P)}{T-u} du \mid \mathcal{F}_t^P \right].$$

By the martingale property $\mathbb{E}[p_\tau^P \mid \mathcal{F}_t^P] = p_t^P$, the left-hand side equals $\text{Var}[p_\tau^P \mid \mathcal{F}_t^P]$, yielding the general variance-budget identity

$$\text{Var}[p_\tau^P \mid \mathcal{F}_t^P] = \mathbb{E} \left[\int_t^\tau \frac{p_u^P(1-p_u^P)}{T-u} du \mid \mathcal{F}_t^P \right] \quad (16)$$

for every $\tau \in [t, T)$, which is (3).

It remains to pass $\tau \uparrow T$. The process p_τ^P is bounded in $[0, 1]$ and absorbs at the boundary, so $p_\tau^P \rightarrow p_T^P = \Theta \in \{0, 1\}$. Bounded convergence gives $\mathbb{E}[(p_\tau^P)^2 \mid \mathcal{F}_t^P] \rightarrow \mathbb{E}[\Theta^2 \mid \mathcal{F}_t^P] = \mathbb{E}[\Theta \mid \mathcal{F}_t^P] = p_t^P$, while conditional monotone convergence applies to the right-hand side of (16). We obtain the terminal case

$$\text{Var}[p_T^P \mid \mathcal{F}_t^P] = \mathbb{E} \left[\int_t^T \frac{p_u^P(1-p_u^P)}{T-u} du \mid \mathcal{F}_t^P \right] = p_t^P(1-p_t^P),$$

which formalizes the statement that the deadline clock spends exactly the remaining binary uncertainty by time T .

B. Estimation and Dynamic Recursion Details

The main paper defines the forecasting target and interval-score evaluation. This appendix records the additional implementation details needed to interpret the horse race: how fitted scales are estimated, how ARCH/GARCH dynamics are added to structural baselines, and how the dynamic state is propagated through contracts.

Quasi-likelihood and fitted scales. For each test month m , fitted quantities are estimated using only observations in the expanding training window preceding that month. For a specification with parameter vector θ , the training criterion is the volume-weighted Gaussian quasi-likelihood

$$\hat{\theta}_m \in \arg \min_{\theta} \sum_{i \in \mathcal{T}_m^{\text{train}}} V_i \left\{ \log h_i^2(\theta) + \frac{\varepsilon_i^2}{h_i^2(\theta)} \right\}. \quad (17)$$

$\varepsilon_i = p_{i+1} - p_i$ is the one-hour price innovation. The Gaussian criterion is used only to fit conditional second moments; the interval-score evaluation does not require Gaussian standardized innovations. The same forecast-origin volume weights are used in estimation and in the reported VW-IS.

The deadline resolution and Archak–Ipeiritis benchmark specifications are parameter-free: no fitted multiplicative scale is applied to either benchmark. Closed-form DR-AS estimates only the nonnegative order-flow scale K . Activity-scaled deadline specifications estimate the positive activity normalization $\bar{V}$. These scalar quantities are re-estimated separately in each expanding-window training sample.

Additive structural-dynamic recursions. The plain dynamic benchmarks use lagged squared innovations but no prediction-market state variables:

$$h_i^2 = \omega + \alpha \varepsilon_{i-1}^2 \quad \text{and} \quad h_i^2 = \omega + \alpha \varepsilon_{i-1}^2 + \beta h_{i-1}^2.$$

The structural-dynamic specifications add a structural variance predictor b_i to this recursion:

$$h_i^2 = \omega + \alpha \varepsilon_{i-1}^2 + \beta h_{i-1}^2 + c b_i, \quad c \geq 0, \quad (18)$$

with the ARCH version obtained by setting $\beta = 0$. The lagged shock is the raw lagged price innovation, not a residualized innovation after subtracting the structural baseline. Thus (18) is an additive variance recursion.

The structural baseline b_i is the closed-form structural predictor named in the dynamic specification in Table 1.

For GARCH+DR-AS specifically, the dynamic fit does not first estimate the closed-form DR-AS model and then append a separate GARCH correction. Instead, the structural term enters the additive variance recursion as a structural state variable, with its weight c estimated jointly with the ARCH/GARCH parameters.

Contract-level recursion. The ARCH/GARCH recursion is run separately within each contract. Lagged innovations are taken only from the same contract, and the recursion resets at contract boundaries. The first dynamic observation in a contract is omitted because no valid lagged innovation is available, which is why the dynamic evaluation panel is slightly smaller than the closed-form panel.

For contracts that continue from the training window into a test month, the variance state is carried forward sequentially using only information available before each forecast origin. Parameters are fixed at the beginning of the test month, but the variance state updates as earlier within-contract innovations become observable. Future test observations are never used to estimate that month’s parameters.

Boundary conventions. All model families use common numerical conventions near binary boundaries and near expiration. These conventions are applied uniformly across specifications, so relative performance is not driven by singularities in a particular formula.

C. Additional Model-Free Binning Diagnostics

This appendix reports the full set of model-free binning diagnostics summarized in Section 3.4. These diagnostics are descriptive. They do not estimate DR-AS, do not control for other state variables, and are not used to rank forecasting models. Their purpose is to show how realized price variation moves univariately with the main observable states. The diagnostics use the same active-update innovation target as the main forecasting exercise. They are computed on the full filtered analysis panel, pooling across months, rather than only on the pooled out-of-sample test windows. For each driver X , observations are sorted into pre-specified bins, and the plotted statistic is the volume-weighted root-mean-square next-hour innovation,

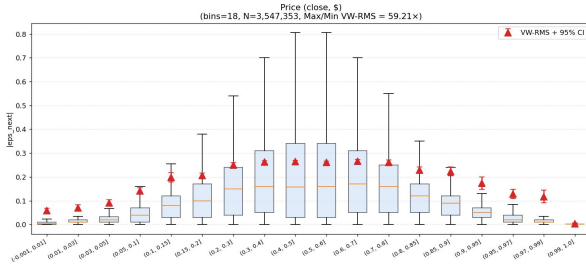
$$\text{VW-RMS}_b = \sqrt{\frac{\sum_{i \in b} V_{t_i} \varepsilon_i^2}{\sum_{i \in b} V_{t_i}}}. \quad (19)$$

The driver-strength ratio is the largest bin VW-RMS divided by the smallest bin VW-RMS. Confidence bands in the figures are computed by within-bin bootstrap resampling.

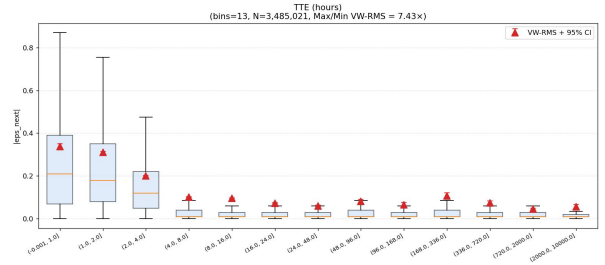
The first two rows, price and time to resolution, are the primitive inputs of the deadline-resolution component. The third row, $1/\sqrt{\tau}$, is included only as a re-expression of the deadline scale and should not be interpreted as a separate primitive driver. Category is also diagnostic rather than structural: it summarizes differences in market type, contract duration, and trading environment.

Table 8: Complete model-free driver-strength ratios. Each ratio is the maximum bin VW-RMS divided by the minimum bin VW-RMS.

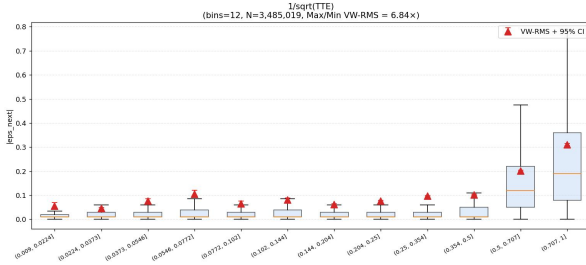
Rank	Driver	Ratio	Interpretation
1	Price p	$59.2\times$	Boundary shape in $p(1-p)$
2	Time to resolution τ	$7.4\times$	Deadline effect
3	$1/\sqrt{\tau}$	$6.8\times$	Re-expression of deadline scale
4	Category	$5.2\times$	Proxy for duration / market type
5	Volume V	$3.3\times$	Activity / information-flow proxy
6	Time % to resolution	$2.9\times$	Normalized time-to-resolution proxy
7	Open interest	$2.4\times$	Participation proxy
8	Spread s	$1.7\times$	Adverse-selection / liquidity proxy



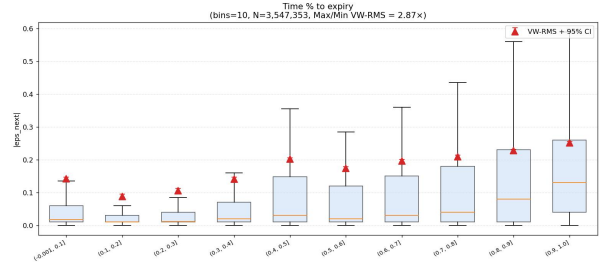
(a) Price p .



(b) Time to resolution in hours.



(c) $1/\sqrt{\tau}$.



(d) Time percent to resolution.

Figure 5: Deadline and time-to-resolution diagnostics. Price and time to resolution are the primitive inputs of the deadline-resolution component. The inverse-square-root and time-percent panels are included as alternative time-to-resolution scalings.

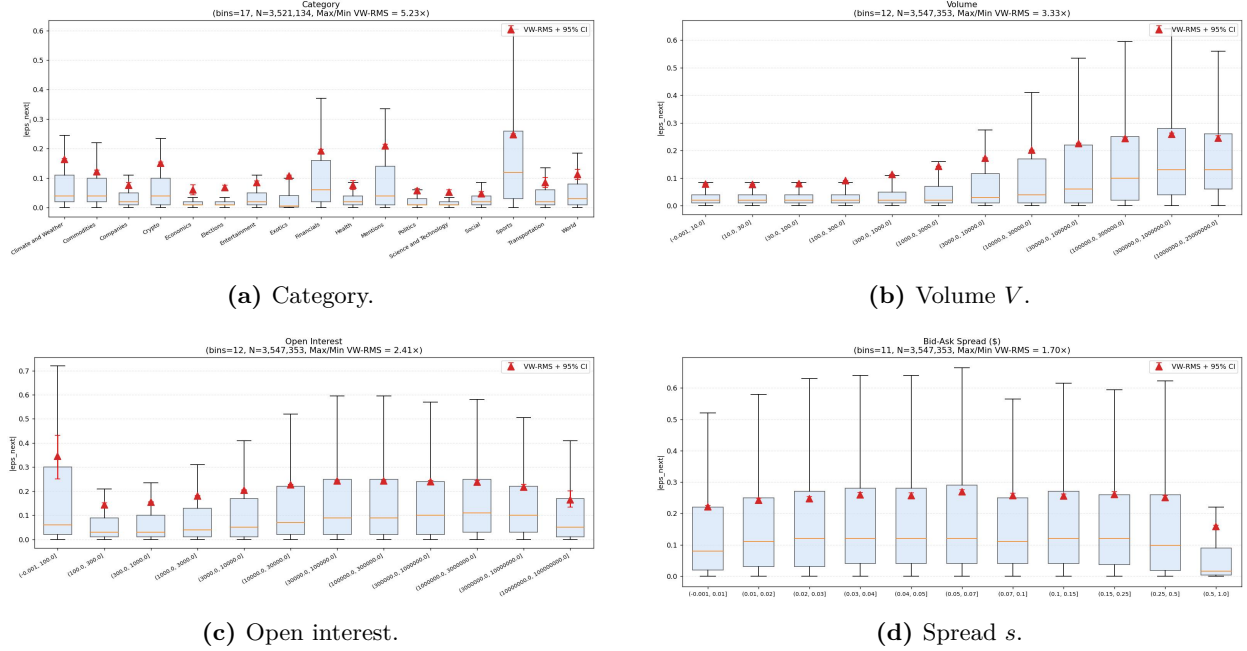


Figure 6: Market-type, activity, participation, and spread diagnostics. Category and open interest are descriptive controls rather than primitive inputs to DR-AS. Volume and spread are the two observable inputs to the order-flow term, where they enter jointly through $\nu(V)s^2/4$.

D. Zero-Update Observations and Unfiltered-Panel Robustness

The hourly panel contains many observations in which the quoted probability does not move over the next hour. These observations are economically meaningful: some are quiet hours with live quotes but little trading, while others are traded hours in which the price does not change over the one-hour horizon. They are not measurement errors. They represent the no-update margin of the market.

The headline exercise focuses on the conditional scale of active price updates: the size of the next price move conditional on an update occurring. An unconditional model of hourly price changes would combine this update-size margin with a separate incidence margin for whether the price moves at all. This appendix asks whether the main structural ranking is sensitive to adding inactive hours back to the forecasting exercise.

We report two checks. The first changes only the evaluation sample: parameters are estimated on the same active-update training observations used in the headline exercise, but the fitted forecasts are evaluated on the full test panel, including zero-update hours. The second removes the zero-update filter from both estimation and evaluation, so that the model is fit and scored on the full hourly panel. Pairwise statements below use the same paired contract-cluster bootstrap design as in the main horse race. For two specifications A and B , we report

$$\Delta_{B-A} = \text{VW-IS}(B) - \text{VW-IS}(A),$$

computed on the common evaluation support. Positive values favor specification A .

D.1. Active-update estimates evaluated on the full test panel

We first hold the headline forecasting rules fixed and change only the test sample. Parameters are estimated exactly as in the active-update exercise. The fitted intervals are then scored on the full test panel, adding observations with $|\varepsilon_i| \leq 10^{-10}$. This design keeps the forecasting rule fixed and isolates the effect of adding inactive hours at evaluation.

The full test panel is much larger than the active-update panel. Because prediction intervals are centered at the forecast-origin price, zero-update observations are often covered and contribute mainly through interval width rather than miss penalties. Consequently, the level of VW-IS and VW-Cov should not be compared mechanically with the active-update baseline. The relevant question is whether the ordering of forecasting rules changes once inactive hours are included at evaluation.

Table 9: Unfiltered-test robustness holding active-update estimates fixed. The active-update column reports the headline VW-IS. The full-test columns evaluate the same fitted forecasting rules on the full test panel, including zero-update hours. Lower VW-IS is better.

Specification	Active-update VW-IS	Full-test VW-IS	Full-test VW-Cov
<i>Structural + residual dynamics</i>			
GARCH+DR-AS ($\sqrt{V}$)	0.4620	0.5705	0.9648
GARCH+deadline resolution	0.4757	0.5707	0.9650
GARCH+Archak-Ipeirotis	0.4767	0.5794	0.9664
GARCH+activity-scaled deadline ($\sqrt{V}$)	0.4925	0.5864	0.9632
ARCH+DR-AS ($\sqrt{V}$)	0.4757	0.5871	0.9643
<i>Plain dynamic benchmarks</i>			
GARCH(1,1) plain	0.7675	0.8460	0.9173
ARCH(1) plain	0.7788	0.8479	0.9193
<i>Closed-form structural predictors</i>			
DR-AS ($\sqrt{V}$)	0.5085	0.6029	0.9777
DR-AS ($\log(1 + V)$)	0.5187	0.6090	0.9777
DR-AS (linear V)	0.5263	0.6112	0.9764
DR-AS (constant activity)	0.5246	0.6129	0.9774
Deadline resolution only	0.5829	0.6201	0.9503
Archak-Ipeirotis benchmark	0.6511	0.6778	0.9054

The score level increases when zero-update observations are added back, but the main hierarchy is preserved. The leading forecasts remain structural-dynamic specifications, closed-form DR-AS remains the best closed-form predictor, and plain ARCH/GARCH remains far below the structural specifications.

Pairwise check. The paired comparisons in Table 10 confirm the interpretation of the full-test table. The two lowest-scoring dynamic specifications, GARCH+DR-AS and GARCH+deadline resolution, are not statistically distinguishable on the full test panel. At the same time, the structural-dynamic tier remains well separated from plain GARCH, and the closed-form DR-AS advantage over deadline-only and alternative activity scaling remains statistically stable.

D.2. Fully unfiltered estimation and evaluation

The second exercise removes the zero-update filter from both estimation and evaluation. This exercise targets the full hourly panel rather than the conditional scale of active updates. Table 11

Table 10: Paired comparisons for the active-fit, full-test exercise. The reported difference is $\Delta_{B-A} = \text{VW-IS}(B) - \text{VW-IS}(A)$, so positive values favor specification A .

Specification A	Specification B	Δ_{B-A} [95% CI]
GARCH+deadline resolution	GARCH+DR-AS ($\sqrt{V}$)	0.0003 [−0.0020, 0.0031]
GARCH+DR-AS ($\sqrt{V}$)	GARCH(1,1) plain	0.2750 [0.2556, 0.2949]
DR-AS ($\sqrt{V}$)	Deadline resolution only	0.0172 [0.0142, 0.0206]
DR-AS ($\sqrt{V}$)	DR-AS ($\log(1 + V)$)	0.0061 [0.0047, 0.0078]

therefore includes the active-update VW-IS as a reference column, not as a same-target comparison. The full-panel columns report the result of fitting and evaluating on the unfiltered panel.

Table 11: Fully unfiltered robustness. The active-update VW-IS column reports the headline active-update result as a reference. The remaining columns remove the zero-update filter from both estimation and evaluation. Lower VW-IS is better.

Specification	Active-update VW-IS	Full-panel VW-IS	Full-panel VW-Cov
<i>Structural + residual dynamics</i>			
GARCH+Archak–Ipeiritis	0.4767	0.5878	0.9522
GARCH+activity-scaled deadline ($\sqrt{V}$)	0.4925	0.5902	0.9512
GARCH+deadline resolution	0.4757	0.5979	0.9450
ARCH+activity-scaled deadline ($\sqrt{V}$)	0.5044	0.6000	0.9592
ARCH+DR-AS ($\sqrt{V}$)	0.4757	0.6022	0.9538
GARCH+DR-AS ($\sqrt{V}$)	0.4620	0.6027	0.9433
<i>Plain dynamic benchmarks</i>			
ARCH(1) plain	0.7788	0.9683	0.8883
GARCH(1,1) plain	0.7675	0.9730	0.8825
<i>Closed-form structural predictors</i>			
DR-AS ($\sqrt{V}$)	0.5085	0.5991	0.9789
DR-AS ($\log(1 + V)$)	0.5187	0.6039	0.9756
DR-AS (constant activity)	0.5246	0.6063	0.9750
DR-AS (linear V)	0.5263	0.6097	0.9808
Deadline resolution only	0.5829	0.6201	0.9503
Archak–Ipeiritis benchmark	0.6511	0.6778	0.9054

The full-panel target gives much more weight to hours in which the price does not update. Even under this target, the leading forecasts are specifications that use prediction-market state variables, plain ARCH/GARCH remains far from the leading group, and DR-AS ($\sqrt{V}$) remains the best closed-form predictor.

Pairwise check. Table 12 reports the paired comparisons most relevant for the fully unfiltered exercise. The top structural-dynamic specifications form a close leading group, while the closed-form DR-AS comparisons remain sharply separated from deadline-only and weaker activity scalings.

Together, the two inactive-hour checks support the interpretation of the main horse race. Adding zero-update hours changes the level of the scores and, in the fully unfiltered exercise, puts additional weight on the no-update margin. But the comparison that matters for the paper is stable: forecasts using structural terms remain the relevant leading class, and closed-form DR-AS remains the strongest closed-form structural predictor.

Table 12: Paired comparisons for the fully unfiltered exercise. The reported difference is $\Delta_{B-A} = \text{VW-IS}(B) - \text{VW-IS}(A)$, so positive values favor specification A .

Specification A	Specification B	Δ_{B-A} [95% CI]
GARCH+Archak-Ipeiotis	GARCH+activity-scaled deadline ($\sqrt{V}$)	0.0024 [−0.0068, 0.0108]
GARCH+Archak-Ipeiotis	GARCH+deadline resolution	0.0101 [0.0073, 0.0129]
GARCH+Archak-Ipeiotis	GARCH+DR-AS ($\sqrt{V}$)	0.0150 [0.0109, 0.0193]
DR-AS ($\sqrt{V}$)	Deadline resolution only	0.0211 [0.0172, 0.0260]
DR-AS ($\sqrt{V}$)	DR-AS ($\log(1 + V)$)	0.0049 [0.0028, 0.0080]

E. Robustness: Last-Trade Prices

The main empirical analysis uses end-of-hour mid-quotes as the price input. This appendix asks whether the main conclusions are sensitive to that choice. We repeat the horse-race comparisons using last-trade close prices instead of mid-quotes, while keeping the same train-test splits, filters, estimation procedure, and interval-score evaluation.

Table 13 reports representative specifications under both price definitions. The qualitative ranking is stable. DR-AS is the strongest closed-form structural predictor under both mid-quote and last-trade pricing. Adding GARCH dynamics to DR-AS gives the best overall model under both price definitions. Plain GARCH remains substantially worse than the structural and structural-dynamic specifications.

Table 13: Robustness to price definition. The main text uses mid-quote prices; this table compares representative mid-quote and last-trade results. Lower VW-IS is better.

Specification	Mid-quote		Last-trade	
	VW-IS	VW-Cov	VW-IS	VW-Cov
Closed-form DR-AS ($\sqrt{V}$)	0.5085	0.9439	0.5006	0.9520
Deadline resolution only	0.5829	0.8673	0.6195	0.8238
Archak-Ipeiotis benchmark	0.6511	0.7862	0.6962	0.7378
Plain GARCH(1,1)	0.7675	0.8989	0.7638	0.8994
GARCH+deadline resolution	0.4757	0.9568	0.4773	0.9549
GARCH+Archak-Ipeiotis	0.4767	0.9568	0.4775	0.9566
GARCH+DR-AS ($\sqrt{V}$)	0.4620	0.9592	0.4681	0.9554

The last-trade pipeline therefore confirms the main conclusion: prediction-market state variables explain a large share of volatility, and the best dynamic model is obtained by adding residual GARCH dynamics on top of the structural DR-AS baseline. The small differences between mid-quote and last-trade scores are consistent with the interpretation that mid-quotes reduce bid-ask bounce and staleness, but the structural ranking is not an artifact of the price definition.