Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Prediction Markets

Nam Anh Le[1,*]

[1]Center for Applied Data Science and Artificial Intelligence, College of Technology, National Economics University, Vietnam

*Corresponding author. 11220245@st.neu.edu.vn

Abstract

Prediction markets are increasingly used as probability forecasting tools, yet their usefulness depends on calibration: whether a contract trading at 70 cents truly implies a 70% probability. Using 292 million trades across 327,000 binary contracts on Kalshi and Polymarket, this paper shows that calibration is a structured, multidimensional phenomenon. On Kalshi, calibration decomposes into four components (a universal horizon effect, domain-specific biases, domain-by-horizon interactions and a trade-size scale effect) that together explain 87.3% of calibration variance. The dominant pattern is persistent underconfidence in political markets: prices are chronically compressed toward 50%, and this bias generalises across both exchanges. However, the trade-size scale effect, whereby large trades are associated with amplified underconfidence in politics on Kalshi ($\Delta = 0.53$, 95% confidence interval $[0.29, 0.75]$), does not replicate on Polymarket ($\Delta = 0.11$, $[-0.15, 0.39]$), suggesting platform-specific microstructure. A Bayesian hierarchical model confirms the frequentist decomposition with 96.3% posterior predictive coverage. Consumers of prediction market prices who treat them as face-value probabilities will systematically misinterpret them, and the direction of misinterpretation depends on what is being predicted, when and by whom.

Keywords Bayesian hierarchical models, calibration, crowd wisdom, favourite–longshot bias, information aggregation, prediction markets

1. Introduction

On the evening of 5 November 2024, millions of people worldwide refreshed prediction market websites to track the US presidential election. Polymarket showed one candidate trading at 62 cents. News outlets reported this as a "62% probability of winning." Social media amplified the number. Policymakers cited it. But what does 62 cents actually mean? If the market is well *calibrated*, contracts that trade at 62 cents should correspond to events that occur 62% of the time. If it is not, the price is a distorted signal, and the millions of people relying on it are being misled.

This paper investigates whether prediction markets are well calibrated, and discovers that the answer is: it depends. It depends on *what* is being predicted, *when* you look, and *who* is trading. These three dimensions (domain, time horizon and trade size) interact in structured, predictable ways that are characterised here using the two largest publicly available prediction market datasets assembled to date.

Prediction markets are financial exchanges where contracts pay $1 if a specified event occurs and $0 otherwise [Forsythe et al., 1992, Rhode and Strumpf, 2004]. A contract trading at price $p \in (0, 1)$ is commonly interpreted as encoding a crowd-sourced probability p that the event will happen. This interpretation rests on substantial theoretical work linking market equilibrium prices to aggregated beliefs [Wolfers and Zitzewitz, 2004, Manski, 2006, Arrow et al., 2008]. The broader claim that decentralised aggregation mechanisms can outperform centralised experts draws on a long intellectual tradition [Galton, 1907, Hayek, 1945, Surowiecki, 2004, Hong and Page, 2004]. Prediction markets have been deployed for electoral forecasting [Forsythe et al., 1992, Berg et al., 2008, Rothschild, 2009], economic

policy, pandemic monitoring [Hanson, 2003], and corporate decision-making [Cowgill and Zitzewitz, 2015], and their use has surged dramatically in recent years.

The quality of these probability forecasts is assessed through *calibration*: the property that events assigned probability p occur with empirical frequency p. A substantial literature examines prediction market calibration, with broadly positive but mixed conclusions. Page and Clemen [2013] found a favourite–longshot bias, an underconfidence pattern where favourites are underpriced and longshots are overpriced (first documented in horse-race betting by Griffith, 1949, and formalised by Ali, 1977; Thaler and Ziemba, 1988; Snowberg and Wolfers, 2010), that worsens with time to expiration. Rothschild [2009] found reasonable calibration for US elections but systematic biases at extreme probabilities. Berg et al. [2008] showed that the Iowa Electronic Markets outperformed polls (see also Servan-Schreiber et al., 2004; Erikson and Wlezien, 2008; Leigh and Wolfers, 2006, for competing assessments). Tetlock and Gardner [2015] demonstrated that structured forecasting tournaments can identify "superforecasters" whose calibration exceeds both markets and conventional experts. More recently, Liu et al. [2024] proposed Bayesian approaches to recalibration, and Satopää et al. [2014] developed logit-based methods for combining probability predictions. Dreber et al. [2015] and Camerer et al. [2016] demonstrated that prediction markets can forecast the replicability of scientific findings.

A critical limitation of this literature is the implicit assumption that calibration is a *domain-agnostic* property of the aggregation mechanism. A market forecasting tomorrow's football score and a market forecasting the next president are treated as facing the same calibration challenges. This paper demonstrates that they do not. Using 292 million trades across 327,000 binary contracts spanning six knowledge domains on two major exchanges (Kalshi and Polymarket), calibration is decomposed into four components that together explain 87.3% of observed calibration variance. Cross-platform validation confirms that the dominant pattern, persistent underconfidence in political markets, generalises beyond any single exchange, while the trade-size scale effect proves platform-specific. The decomposition reveals that crowds are not universally wise or foolish; their forecast quality depends systematically on the epistemic structure of what they are predicting.

1.1. Summary of findings

The central finding is that prediction market calibration decomposes into four interpretable components:

A universal horizon effect (μ). All domains share a tendency toward *underconfidence*, with prices compressed toward 50%, at long time horizons. The mean cell-level calibration slope $\mu(\tau)$, averaged across all domain–size cells at each horizon, rises from 0.99 (within one hour of resolution) to 1.32 (beyond one month). This accounts for 30.2% of calibration variance. At long horizons, markets systematically understate the probability of the favoured outcome: a contract trading at 70 cents one month out corresponds to a true probability closer to 75%. This pattern holds on both Kalshi and Polymarket.

Domain-specific structural biases (α). Politics is the clear outlier: its domain intercept ($+0.15$) is far above all other domains, indicating persistent underconfidence (prices chronically too compressed) at nearly all time horizons. Weather (-0.09) and Entertainment (-0.09) exhibit the opposite pattern: their prices are too extreme, reflecting overconfidence. This component accounts for 14.6% of variance. The political bias is confirmed on Polymarket (mean slope 1.31 vs Kalshi 1.64), establishing it as a structural property of political prediction markets rather than a single-platform artefact.

Domain-by-horizon interactions (β). This is the single largest domain-specific component at 26.0% of variance. Domains follow genuinely different calibration *trajectories*. Political markets are underconfident at nearly all horizons (slopes 0.93–1.83), with prices persistently too compressed. Sports markets are well calibrated at short-to-medium horizons (slopes 0.90–1.10) but become sharply underconfident beyond one month (slope 1.74). Weather markets are *overconfident* at short horizons (slopes 0.69–0.97), with prices too extreme relative to base rates, before transitioning to underconfidence at longer horizons.

A trade-size scale effect (γ). In political markets, large trades (over 100 contracts) exhibit calibration slopes of 1.74, compared to 1.19 for single-contract trades, a gap of 0.53 (95% bootstrap CI $[0.29, 0.75]$). Larger trades are associated with *more compressed* prices. In sports markets, no such gap exists ($\Delta = 0.07$, 95% CI $[-0.07, 0.26]$). This component accounts for 16.5% of variance. Notably, this effect does not replicate on Polymarket ($\Delta = 0.11$, 95% CI $[-0.15, 0.39]$), suggesting it reflects Kalshi's specific market microstructure rather than a universal feature of political prediction markets.

1.2. Related work

This work connects to several research areas. The prediction market efficiency literature examines whether market prices aggregate dispersed information. Hayek [1945] provided the conceptual foundation; the efficient markets hypothesis [Fama, 1970] and the impossibility of informationally efficient prices under costly information acquisition [Grossman and Stiglitz, 1980] frame the theoretical debate; Plott and Sunder [1988] provided early experimental evidence; Wolfers and Zitzewitz [2004] formalised the link between prices and beliefs; Arrow et al. [2008] argued for broader adoption. Manski [2006] cautioned that prices may not equal mean beliefs under heterogeneous preferences, though Wolfers and Zitzewitz [2006] showed divergence is typically small. Chen and Plott [2002] demonstrated experimentally that information aggregation can be remarkably efficient even with few traders. Ottaviani and Sørensen [2008] provided a theoretical foundation for the favourite–longshot bias in parimutuel markets. The decomposition presented here characterises precisely when and where aggregation succeeds or fails, and connects the dominant miscalibration pattern (underconfidence in political markets) directly to Manski's theoretical concern about heterogeneous beliefs compressing prices.

The forecast calibration literature provides the measurement framework for this study. Lichtenstein et al. [1982] established the foundations of probability calibration research, building on earlier work by Brier [1950] on verification of probability forecasts, Murphy and Winkler [1977] on the reliability of probability assessments, Dawid [1982] on the well-calibrated Bayesian, and DeGroot and Fienberg [1983] on the comparison and combination of forecasters. The logistic recalibration approach [Platt, 1999, Gneiting and Raftery, 2007, Winkler, 1996, Bickel, 2007] parameterises calibration through a slope in logit space. A slope exceeding one, conventionally termed underconfidence or the favourite–longshot bias, indicates that forecasts are insufficiently extreme. Satopää et al. [2014] developed a logit model for combining multiple probability predictions. Liu et al. [2024] developed Bayesian boldness–recalibration, and Palley and Satopää [2024] addressed biased priors in crowd forecasts. The present study extends this work by modelling calibration as a structured function of multiple covariates.

The Bayesian hierarchical modelling literature provides the inferential framework. Hierarchical models for structured data are widely used in epidemiology, ecology and small-area estimation [Gelman et al., 2013, Hogg et al., 2024, Gelman and Hill, 2007]. Hamiltonian Monte Carlo [Neal, 2011, Betancourt, 2017] and the No-U-Turn Sampler [Hoffman and Gelman, 2014] have transformed posterior computation, enabling the fitting of complex hierarchical structures with efficient exploration of high-dimensional parameter spaces. Bayesian change-point methods [Assareh et al., 2013] and MCMC convergence diagnostics [Mengersen et al., 1999, Brooks and Gelman, 1998] inform the computational strategy employed here. Tetlock [2005] and Tetlock and Gardner [2015] provide the broader intellectual context for understanding when and why crowd forecasts succeed or fail.

Finally, the scale-dependent miscalibration finding connects to the market microstructure literature. Kyle [1985] predicts that large trades convey private information and should improve prices (see also Glosten and Milgrom, 1985; Easley and O'Hara, 1987, on the relationship between trade size and information content). The finding that large trades are associated with *more* compressed prices in political markets suggests a different mechanism: opposing large bets from confident partisans cancel each other out, pulling prices toward 50% rather than toward truth.

The forecasting tournament literature provides an important benchmark. The Good Judgment Project [Mellers et al., 2014] demonstrated that structured forecasting teams can achieve remarkable accuracy, and Baron et al. [2014] showed that extremising transformations improve calibration, a finding directly relevant to the underconfidence documented here. More recently, DellaVigna and Pope [2018] used prediction markets to benchmark beliefs about experimental outcomes, and Metaculus and other platforms have generated large-scale calibration datasets [Karger et al., 2023]. This literature establishes that calibration quality depends on the structure of the forecasting problem, a principle formalised here through domain-specific decomposition.

1.3. Plan of the paper

Section 2 describes the data from both exchanges. Section 3 presents the empirical calibration landscape, establishes three stylised facts, and validates each on Polymarket. Section 4 diagnoses potential artefacts. Section 5 introduces the decomposition model and addresses confounding concerns. Section 6 develops a Bayesian hierarchical model. Section 7 discusses implications. Section 8 concludes.

Table 1. Summary statistics by domain: Kalshi (cutoff 31 December 2025).

Domain	Markets	Trades	Contracts	Resolved	Med. vol.	Base rate
Sports	55,637	43.2M	12.7B	98.1%	76	41.3%
Crypto	76,181	6.5M	742.9M	99.1%	35	40.7%
Politics	6,609	4.9M	2.2B	94.1%	127	40.2%
Finance	38,058	4.3M	677.2M	99.0%	38	37.7%
Weather	26,911	4.4M	279.1M	99.5%	74	24.0%
Entertainment	7,212	1.5M	174.5M	96.7%	60	38.0%
Total	**210,608**	**64.7M**	**16.8B**	**98.6%**	**47**	**38.1%**

Note: Resolved (%) computed over markets past their close date. Base rate is the percentage of resolved markets where outcome = yes. Median volume is the median number of trades per market.

2. Data

2.1. Data source

Data are drawn from two major prediction market exchanges using the pre-collected dataset of Becker [2026]. The primary analysis platform is Kalshi, regulated by the US Commodity Futures Trading Commission (CFTC), which operates a central limit order book for binary event contracts paying $1.00 if a specified event occurs and $0.00 otherwise. Contracts trade at prices between $0.01 and $0.99. For cross-platform validation, the analysis uses Polymarket, a decentralised exchange operating on the Polygon blockchain. Polymarket uses a hybrid automated market maker (AMM) and central limit order book (CLOB) mechanism. Unlike Kalshi, Polymarket is unregulated, globally accessible, and allows pseudonymous trading via blockchain wallets. These structural differences (regulatory environment, trader demographics, market mechanism) make Polymarket an informative validation platform: patterns that replicate across both exchanges are unlikely to be platform-specific artefacts.

The Kalshi dataset comprises 64.7 million trades across 210,608 binary contracts, representing approximately 16.8 billion contracts traded, with a cutoff date of 31 December 2025. For each trade, the dataset includes the contract identifier, execution price (in cents), number of contracts, the side taken by the trade initiator, and the execution timestamp (millisecond precision). For each market, the dataset includes the contract identifier, event category, resolution status and outcome (yes or no), and close time. Of the 210,608 markets, 98.6% of those past their close date have resolved with a definitive yes/no outcome (Table 1). The Polymarket dataset comprises 227.6 million trades across 116,000 resolved contracts (61.3 billion contracts traded), with the same cutoff date. Polymarket trade timestamps are derived from Polygon block numbers with approximately 3-hour noise, a limitation that affects the two shortest time bins (Section 3).

2.2. Domain classification

Kalshi organises contracts into events via a hierarchical ticker structure. Markets are classified into six knowledge domains (see supplementary material for full classification rules): *Sports* (professional leagues including NFL, NBA, MLB and NHL), *Politics* (elections, electoral college outcomes, government policy), *Crypto* (cryptocurrency price contracts), *Finance* (equity indices, interest rates, economic indicators), *Weather* (temperature records, precipitation, natural events) and *Entertainment* (awards, media, culture). The classification uses a deterministic mapping from event ticker prefixes, ensuring reproducibility. For Polymarket, which has no structured ticker namespace, markets are classified using compiled regular expression patterns applied to market titles. This yields three comparable domains (Sports, Crypto, Politics); Finance is excluded due to thin coverage (2,516 Polymarket markets vs 38,058 on Kalshi), and Weather and Entertainment have negligible Polymarket presence. The 42.5% of Polymarket markets classified as 'Other' reflects its long tail of bespoke markets (celebrity events, technology launches, meme markets). Tables 1 and 2 provide summary statistics.

Several features merit comment. Sports dominates Kalshi by volume (43.2 million trades, 66.7% of the total) but Politics commands disproportionate contract value (2.2 billion contracts from only 6,609 markets, representing 13.3% of total volume from 3.1% of markets). Politics also has the highest median volume per market (127 trades),

Table 2. Summary statistics by domain: Polymarket (three comparable domains; cutoff 31 December 2025).

Domain	Markets	Trades	Contracts	Resolved	Med. vol.	Base rate
Sports	25,340	49.1M	21.1B	90.0%	129	34.0%
Crypto	73,918	125.3M	10.7B	98.8%	674	45.0%
Politics	14,225	45.7M	24.6B	81.1%	452	30.7%
Total	**113,483**	**220.1M**	**56.4B**	—	**417**	**38.2%**

Note: Polymarket comparison is restricted to three domains with sufficient coverage; Finance is excluded due to thin coverage (2,516 markets). Weather and Entertainment have negligible Polymarket presence. The 42.5% of Polymarket markets classified as 'Other' reflects its long tail of bespoke markets. Median volume is the median number of trades per market.

reflecting intense engagement with a smaller number of high-profile events. The distribution of trade sizes is heavily right-skewed across all domains: the median trade involves 40 contracts, but 0.15% of trades (those exceeding 10,000 contracts) account for approximately 15% of total contract volume. Polymarket (Table 2) shows a strikingly different composition: $4.0\times$ more trades overall but $0.8\times$ fewer markets, reflecting $8.0\times$ higher median per-market volume. The most consequential difference is in Politics, where Polymarket has $9.3\times$ more trades (45.7M vs 4.9M) and $11\times$ more contracts (24.6B vs 2.2B), suggesting substantially deeper price discovery.

2.3. Measuring calibration

Calibration is measured using logistic recalibration [Platt, 1999, Niculescu-Mizil and Caruana, 2005]. For a collection of N trades indexed by i, each with market price $p_i \in (0, 1)$ and binary outcome $y_i \in \{0, 1\}$, the following logistic regression is fitted by maximum likelihood:

$$\text{logit}\big(P(y_i = 1)\big) = a + b \cdot \text{logit}(p_i), \tag{1}$$

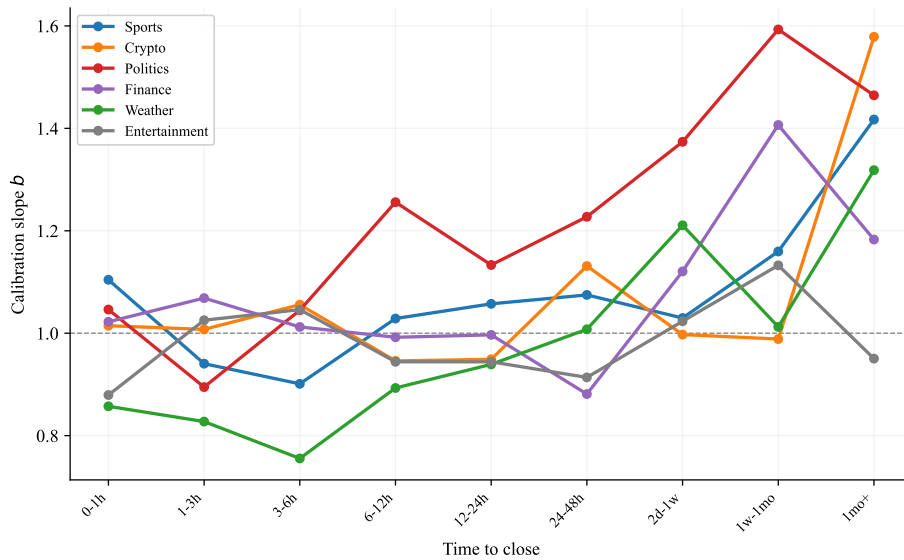where $\text{logit}(x) = \log\{x/(1 - x)\}$. The log-likelihood is

$$\ell(a, b) = \sum_{i=1}^{N} \big[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)\big], \tag{2}$$

with $\pi_i = \sigma(a + b \cdot \text{logit}(p_i))$ and $\sigma(x) = \{1 + \exp(-x)\}^{-1}$. The slope b captures calibration quality. When $b = 1$, the market is perfectly calibrated: prices correspond to true probabilities. When $b > 1$, the market is *underconfident*: prices are insufficiently extreme, compressed toward 50%. A 70-cent contract actually corresponds to a true probability *greater* than 70%; the favourite is underpriced and the longshot is overpriced. This is the classic favourite–longshot bias [Griffith, 1949, Page and Clemen, 2013]. Conversely, $b < 1$ indicates *overconfidence*: prices are too extreme, overstating the probability gap between favourites and longshots. The intercept a captures directional bias (systematic over- or under-prediction of "yes" outcomes); the focus here is on the slope as the primary calibration measure.

The analysis is restricted to trades with prices between 5 and 95 cents, excludes markets with fewer than 10 trades, requires at least 200 trades per analysis cell, and applies mild L_2 regularisation ($C = 10$; Hosmer et al., 2013). These choices are conservative; results are robust to alternative price ranges including $[2, 98]$ and $[10, 90]$ (Appendix A).
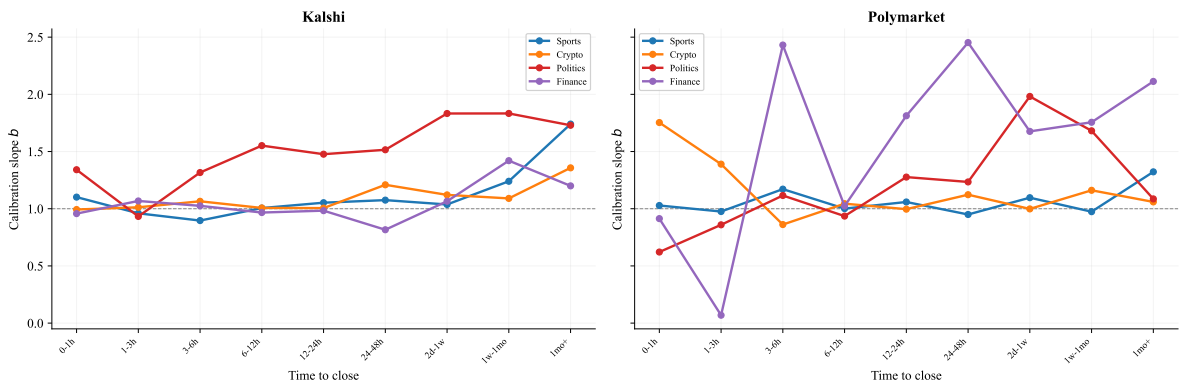
2.4. Analysis dimensions

For each trade, time-to-resolution is computed as $\tau = \text{close\_time} - \text{trade\_time}$ and discretised into nine bins: $[0, 1\text{h})$, $[1\text{h}, 3\text{h})$, $[3\text{h}, 6\text{h})$, $[6\text{h}, 12\text{h})$, $[12\text{h}, 24\text{h})$, $[24\text{h}, 48\text{h})$, $[2\text{d}, 1\text{w})$, $[1\text{w}, 1\text{mo})$ and $[1\text{mo}, \infty)$. Trade sizes are discretised into four bins: Single (1 contract), Small (2–10), Medium (11–100) and Large (>100). The full analysis grid comprises $6 \times 9 \times 4 = 216$ cells, all satisfying the minimum sample-size requirement. A total of 58.7 million Kalshi trades and 220.1 million Polymarket trades enter the calibration analysis after price filtering. Because Polymarket timestamps have approximately 3-hour noise (from block-number bucketing), the two shortest time bins (0–1h and 1–3h) are

Figure 1 Calibration slope b versus time-to-resolution, one line per domain. Slopes above 1 indicate underconfidence (prices compressed toward 50%); slopes below 1 indicate overconfidence (prices too extreme).

Figure 2 Cross-platform calibration slope trajectories: (A) Kalshi vs (B) Polymarket. The dominant pattern—political underconfidence—replicates across exchanges.

unreliable for cross-platform comparison; all Polymarket time-horizon results below use the seven reliable bins (3–6h through 1mo+).

3. The calibration landscape

Figure 1 plots calibration slopes against time-to-resolution for each domain. The detailed slopes are reported in Table 3. Slopes above 1 indicate underconfidence (prices compressed toward 50%); slopes below 1 indicate overconfidence (prices too extreme). Two features are immediately apparent: domains differ sharply, and these differences reflect different *shapes* of calibration trajectory, not mere level shifts.

Table 3. Logistic recalibration slopes by domain and time-to-resolution. Values above 1.0 indicate underconfidence (prices compressed); below 1.0 indicate overconfidence (prices too extreme).

Domain	0–1h	1–3h	3–6h	6–12h	12–24h	24–48h	2d–1w	1w–1mo	1mo+
Politics	1.34	0.93	1.32	1.55	1.48	1.52	1.83	1.83	1.73
Sports	1.10	0.96	0.90	1.01	1.05	1.08	1.04	1.24	1.74
Crypto	0.99	1.01	1.07	1.01	1.01	1.21	1.12	1.09	1.36
Finance	0.96	1.07	1.03	0.97	0.98	0.82	1.07	1.42	1.20
Weather	0.69	0.84	0.74	0.87	0.91	0.97	1.20	1.20	1.37
Entertainment	0.81	1.02	1.00	0.92	0.89	0.84	1.07	1.11	0.96

3.1. Three stylised facts

Stylised Fact 1: All markets are underconfident about the distant future. At long time horizons, prices in every domain become compressed toward 50%, insufficiently extreme relative to the true outcome frequencies. The universal horizon component $\mu(\tau)$, computed as the mean slope across all domain–size cells at each time bin (Section 5), rises from 0.99 (0–1 hour) to 1.32 (beyond one month); note that this cell-level mean differs from the simple average of the six aggregate domain slopes in Table 3, which are contract-weighted pooled estimates. This pattern is the classic favourite–longshot bias [Griffith, 1949, Thaler and Ziemba, 1988, Shin, 1991, Snowberg and Wolfers, 2010] documented at unprecedented scale: at long horizons, favourites are systematically underpriced and longshots are overpriced. It persists after restricting to high-volume markets (Appendix A), is consistent with theoretical predictions about favourite–longshot bias worsening with time to expiration [Page and Clemen, 2013], and is confirmed on Polymarket across all three comparable domains.

   Stylised Fact 2: Domains follow different calibration trajectories (the full posterior β matrix is in Table 13). This is the key insight. Political markets exhibit persistent underconfidence at nearly all horizons (slopes 0.93–1.83), with prices chronically compressed toward 50%. A 70-cent contract in Politics one week before resolution corresponds to a true probability of approximately 83%, not 70%. Sports markets are well calibrated at short-to-medium horizons (slopes 0.90–1.10 from 0 to 48 hours) but become sharply underconfident at long horizons, reaching 1.74 beyond one month. Weather markets exhibit the opposite pattern: *overconfidence* at short horizons (slopes 0.69–0.97 within 48 hours), where prices are too extreme. A journalist asking "Should I trust this prediction market price?" receives a fundamentally different answer depending on the subject matter and timing. These domain-specific trajectories replicate on Polymarket: Politics is underconfident (mean slope 1.31), Sports is near-calibrated (1.08), and Crypto is mildly underconfident (1.05).

   Stylised Fact 3: In political markets, large trades are associated with greater price compression. On Kalshi, large trades (over 100 contracts) in political markets produce calibration slopes of 1.74, compared to 1.19 for single-contract trades (Table 4). The gap of 0.53 (95% bootstrap CI $[0.29, 0.75]$) is both statistically significant and practically consequential. In sports markets, the corresponding gap is 0.07 ($[-0.07, 0.26]$). Because prediction market prices weight traders by position size, these large, poorly calibrated trades are associated with disproportionate influence on the price. However, the effect is platform-specific: on Polymarket, the Politics gap shrinks to 0.11 ($[-0.15, 0.39]$), which is not statistically significant. The scale effect appears to reflect Kalshi's specific microstructure rather than a universal property of political prediction markets.

4. Diagnosing potential artefacts

4.1. Is political underconfidence driven by a subset of markets?

Political markets encompass diverse subcategories. If underconfidence were driven by a single subcategory, the finding would be narrower than claimed. Examining 10 political subcategories with sufficient data (see supplementary material), the pattern is found to be broadly distributed. Electoral College contracts show the strongest underconfidence (slopes 1.53–2.87 across time bins), while Trump Administration contracts span a wider range (0.54–1.64). Other Politics (1.42–2.38), Governor (1.19–4.02) and NYC Mayor (1.12–3.18) all exhibit persistent underconfidence.

Table 4. Calibration slopes by domain and trade size (Kalshi). Δ(L−S) is the mean of time-bin-specific slope differences between the Large and Single bins (the bootstrap estimand). On Polymarket, the Politics Δ is $+0.11$ $[-0.15, +0.39]$ (not significant); Sports $+0.01$ $[-0.20, +0.17]$; Crypto $+0.09$ $[-0.16, +0.36]$.

Domain	Single	Small	Medium	Large	Δ(L−S)
Politics	1.19	1.22	1.37	1.74	$+0.53$
Sports	1.00	1.01	1.01	1.01	$+0.07$
Crypto	1.03	1.03	1.02	1.00	-0.02
Finance	1.10	1.08	1.05	1.05	-0.05
Weather	0.96	0.94	0.91	0.89	-0.07
Entertainment	0.98	1.02	1.00	0.99	$+0.01$

4.2. Composition effects in the 1–3 hour bin

Table 3 shows that Politics achieves its lowest slope (0.93) at the 1–3 hour horizon, the only time bin where Politics appears slightly overconfident rather than underconfident. Disaggregation reveals a composition effect. Trump Administration markets comprise 63% of trades in this bin and have a moderate slope of 1.08. Meanwhile, Electoral College contracts (7.4% of trades, slope 1.81) and Other Politics (5.8%, slope 2.17) remain strongly underconfident but are diluted by the Trump Administration majority. Leave-one-out analysis confirms the mixing: removing Trump Administration (the moderate majority) drops the aggregate from 0.93 to 0.88, revealing that the remaining subcategories are highly heterogeneous: some strongly underconfident (Electoral College 1.81, Other Politics 2.17), others overconfident (Biden Administration -0.14, a negative slope based on only 3.6% of trades in this bin, likely reflecting thin, stale markets during the administration transition). Their opposing biases average to a misleadingly low aggregate. Removing Electoral College (the most underconfident minority) drops the aggregate to 0.69. The low aggregate slope reflects subcategory mixing, not a genuine regime shift.

4.3. Weighting sensitivity and the scale effect

Prediction market prices implicitly weight traders by position size. To disentangle the market price (contract-weighted) from the crowd belief distribution (trade-weighted), the two weighting schemes are compared. In Politics, contract-weighted slopes exceed trade-weighted slopes by an average of 0.33 across all time bins, with the gap reaching 0.54 at the 2-day-to-1-week horizon. Contract-weighted prices are *more* compressed than trade-weighted prices: larger positions are associated with prices further from truth. In Sports, the corresponding gap averages 0.06 and is concentrated at the longest horizon (0.35 at 1 month+, near zero elsewhere). The influence of large trades is amplified by the heavy right skew in trade sizes: the median trade involves 40 contracts, but just 0.15% of trades (those exceeding 10,000 contracts) account for approximately 15% of total contract volume, confirming that a small number of large trades exert disproportionate influence on contract-weighted prices. On Polymarket, the Politics weighting gap collapses to $+0.05$, a seven-fold reduction, and oscillates in sign across time bins, independently confirming that the scale effect is platform-specific.

5. The decomposition model

5.1. Framework

The patterns in Section 3 suggest an additive structure. The calibration slope is modelled as θ observed in cell (d, τ, s), where d indexes domain, τ indexes time-to-resolution bin and s indexes trade-size bin:

$$\theta(d, \tau, s) = \mu(\tau) + \alpha_d + \beta_d(\tau) + \gamma_d(s) + \varepsilon, \tag{3}$$

where $\mu(\tau)$ is a *universal horizon function* capturing the shared tendency toward underconfidence at long horizons; α_d is a *domain intercept* capturing persistent bias (positive for Politics, negative for Weather); $\beta_d(\tau)$ is a *domain-by-horizon interaction* capturing how each domain's trajectory deviates from the universal curve; and $\gamma_d(s)$ is a *scale effect* capturing how trade size modulates calibration within each domain. The model is identified by the following

constraints:

$$\sum_{d=1}^{D} \alpha_d = 0, \tag{4}$$

$$\sum_{d=1}^{D} \beta_d(\tau) = 0 \quad \text{for each } \tau, \tag{5}$$

$$\sum_{s=1}^{S} \gamma_d(s) = 0 \quad \text{for each } d. \tag{6}$$

These ensure that the universal horizon function $\mu(\tau)$ absorbs the cross-domain mean at each horizon, the domain intercepts α_d absorb the cross-horizon mean for each domain, and the scale effects are mean-centred within each domain.

The additive specification is chosen for interpretability and parsimony. In logit space, calibration slopes are unbounded and approximately continuous, making additive decomposition on the logit-transformed calibration measure a natural first-order approximation. A multiplicative specification (modelling log slopes) was also considered; the two yield nearly identical fits because the slopes are concentrated in the range $[0.7, 1.9]$, where additive and log-additive structures are approximately equivalent. A non-domain-specific size–horizon interaction term was tested (Section 5.4) and found to explain only 2.6% additional variance, insufficient to justify the 54 additional parameters it requires. The Bayesian hierarchical model in Section 6 provides a complementary check: its continuous parameterisation of the scale effect ($\delta_d \times \log s$) differs structurally from the categorical specification here, yet the two approaches yield parameter estimates within 0.005 of each other (Table 6), supporting the robustness of the additive decomposition to functional-form assumptions.

5.2. Variance decomposition

The decomposition is estimated using sequential projection, analogous to Type I sums of squares. Let $\bar{\theta}$ denote the grand mean of all 216 observed slopes. The total sum of squares is

$$\mathrm{SS}_{\mathrm{tot}} = \sum_{d,\tau,s} \left[\theta(d,\tau,s) - \bar{\theta} \right]^2, \tag{7}$$

and the residual sum of squares after fitting (3) is

$$\mathrm{SS}_{\mathrm{res}} = \sum_{d,\tau,s} \left[\theta(d,\tau,s) - \hat{\theta}(d,\tau,s) \right]^2, \tag{8}$$

where $\hat{\theta}(d,\tau,s) = \hat{\mu}(\tau) + \hat{\alpha}_d + \hat{\beta}_d(\tau) + \hat{\gamma}_d(s)$. Each component's marginal contribution is computed sequentially: $R^2_\mu = 1 - \mathrm{SS}^{(\mu)}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$, then $R^2_\alpha = R^2_{\mu,\alpha} - R^2_\mu$, and so on. Table 5 reports the variance explained. The model accounts for 87.3% of calibration variance across 216 cells. The model has 72 parameters (9 for μ, 5 for α, 40 for β, 18 for γ) for 216 observations, yielding an adjusted R^2 of 0.810. It is worth noting, however, that the Bayesian hierarchical model in Section 6, which imposes partial pooling through hierarchical priors and thus penalises effective model complexity, achieves 96.3% posterior predictive coverage (Section 6.3), confirming that the four-component structure captures genuine signal rather than noise.

The domain-by-horizon interaction β is the single largest domain-specific component (26.0%). Knowing *which* domain a market belongs to and *when* you observe it is more informative about calibration quality than either piece of information alone. A domain-agnostic model, however sophisticated, misses the largest source of calibration variation.

Because Type I decomposition is order-dependent, verification is conducted with Type II and Type III sums of squares. The interaction terms are identical across all three types: β accounts for 26.0% and γ for 16.1% regardless of decomposition order. Only the main effects shift between types, as expected when time and domain share variance.

Table 5. Variance decomposition of calibration slopes (216 cells, Type I).

Component	Interpretation	Marginal R^2	Cumulative R^2
$\mu(\tau)$	Universal horizon effect	0.302	0.302
α_d	Domain intercept	0.146	0.448
$\beta_d(\tau)$	Domain $\times$ horizon interaction	0.260	0.708
$\gamma_d(s)$	Domain $\times$ size effect	0.165	0.873
ε	Residual	0.127	—

Additionally, weighting cells by inverse estimation variance yields the weighted decomposition

$$\hat{\phi}_{\mathrm{WLS}} = \arg\min_{\phi} \sum_{d,\tau,s} w_{d\tau s}\big[\theta(d,\tau,s) - \hat{\theta}(d,\tau,s;\,\phi)\big]^2, \quad w_{d\tau s} = 1/\mathrm{SE}^2_{d\tau s}, \tag{9}$$

where $\phi = (\mu, \alpha, \beta, \gamma)$ collects all model parameters. This yields a total R^2 of 0.995, with the horizon effect μ dominating (0.74 vs 0.30 unweighted). This confirms that the four-component model explains nearly all variance in precisely estimated cells; the 12.7% residual in the unweighted decomposition is concentrated in noisy, low-volume cells (Appendix A). The domain intercepts replicate on Polymarket: Politics remains the clear positive outlier (mean slope 1.31 at reliable horizons), while Sports (1.08) and Crypto (1.05) cluster near perfect calibration, the same qualitative ranking as on Kalshi (Appendix C reports full cross-platform comparison tables).

5.3. Statistical significance

All components are highly significant (Table 10). Domain intercepts differ ($F(5, 144) = 33.16$, $p < 10^{-16}$). Domain-specific trajectories are necessary ($F(40, 144) = 7.40$, $p < 10^{-16}$). The domain-by-size interaction is confirmed ($F(18, 144) = 10.42$, $p < 10^{-16}$).

The Politics scale effect (the slope difference of 0.53 between large and single-contract trades) is formally defined as

$$\Delta_d = \frac{1}{T} \sum_{\tau=1}^{T} \big[\theta(d, \tau, s_{\mathsf{L}}) - \theta(d, \tau, s_{\mathsf{S}})\big], \tag{10}$$
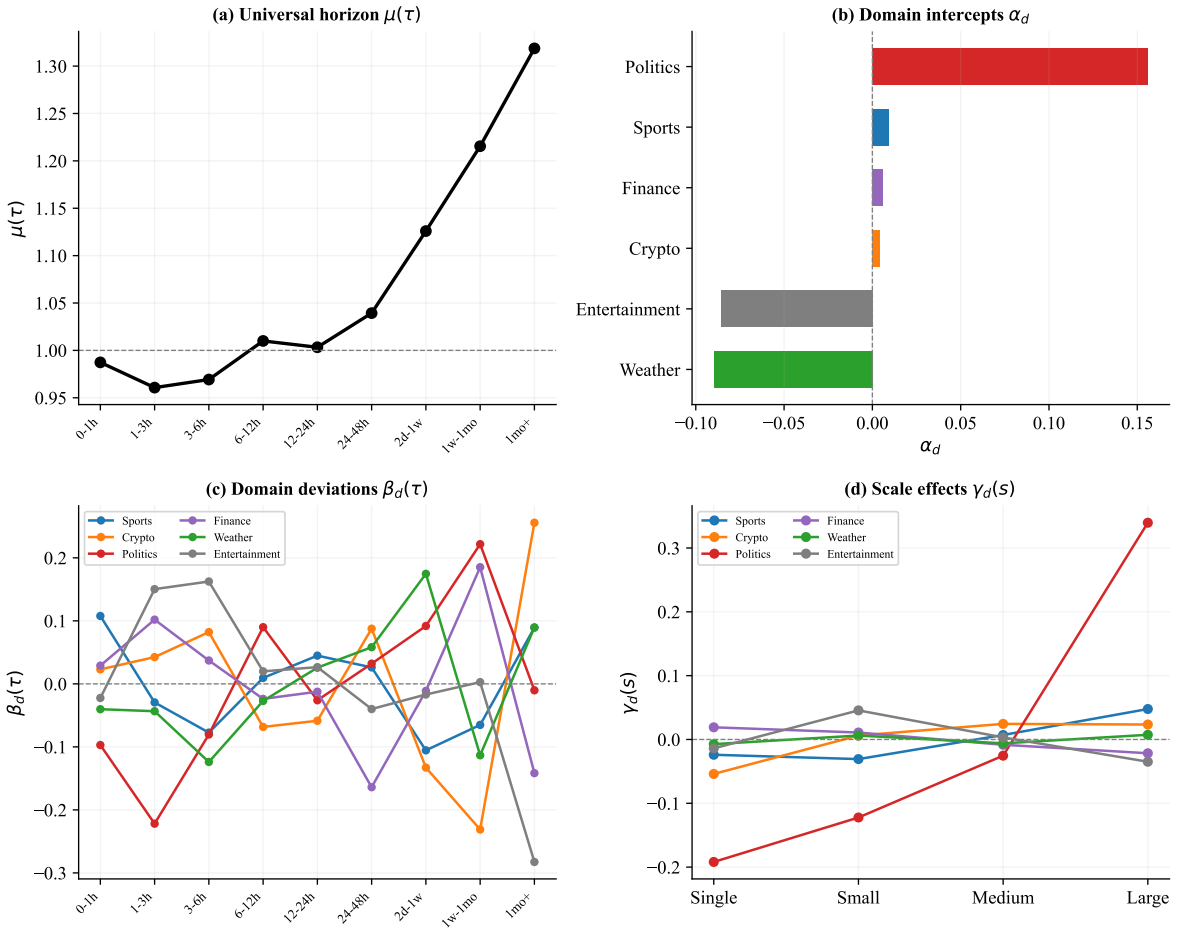
where s_{L} and s_{S} index the Large and Single trade-size bins respectively, and $T = 9$ is the number of time bins. This within-horizon averaging ensures that the estimand is not confounded by compositional differences across horizons. The 95% trade-level bootstrap CI is $[0.29, 0.75]$. Because trades within the same market are serially correlated, a market-clustered bootstrap is also computed (resampling entire markets rather than individual trades). The clustered CI is wider, as expected with fewer effective resampling units, but still excludes zero: $[0.13, 1.29]$. The Sports scale effect remains null under both methods (trade-level $[-0.07, 0.26]$; clustered $[-0.03, 0.05]$).

5.4. Is the scale effect confounded with time horizon?

If large trades concentrate at long horizons, the scale effect could be an artefact of horizon confounding. Three checks rule this out. First, in Politics, large trades have *shorter* median horizons (213 hours) than single-contract trades (862 hours); confounding would therefore bias *against* finding a scale effect, not in favour of it. Second, adding a non-domain-specific size–horizon interaction to the model explains only 2.6% additional variance; the domain-specific scale effect γ remains substantial ($R^2 = 0.132$ vs 0.161 without the interaction). Third, within Politics, the scale effect is positive in all nine time bins, ranging from $+0.12$ (1–3 hours) to $+0.81$ (2 days–1 week). The scale effect is not an artefact of horizon confounding (Tables 9 and 10).

6. Bayesian hierarchical model

The additive decomposition in Section 5 is descriptive: it treats estimated slopes as data without accounting for their estimation uncertainty. A Bayesian hierarchical model is now developed that provides coherent uncertainty quantification and partial pooling.

Figure 3 Four-panel decomposition of calibration slopes. (A) Universal horizon effect $\mu(\tau)$. (B) Domain intercepts α_d. (C) Domain-by-horizon interactions $\beta_d(\tau)$. (D) Domain-by-size effects $\gamma_d(s)$.

6.1. Model specification

For the observed calibration slope θ_{obs} in cell (d, τ, s), the model specifies three levels. At the observation level:

$$\theta_{\mathrm{obs}}(d, \tau, s) \sim \mathcal{N}\big(\mu(\tau) + \alpha_d + \beta_d(\tau) + \delta_d \cdot \tilde{s},\ \sigma^2\big), \tag{11}$$
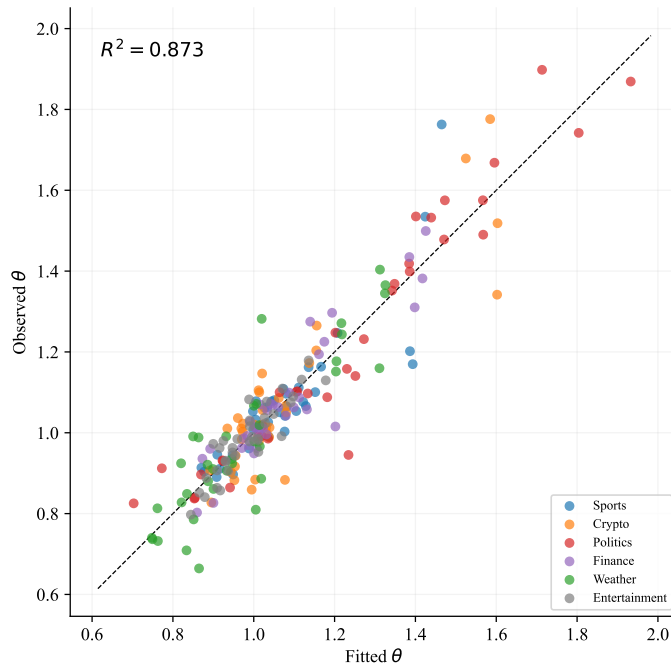
where $\tilde{s} = \log s - \overline{\log s}$ is the centred log trade size for bin s. At the prior level, each component receives a hierarchical structure:

$$\mu(\tau) \sim \mathcal{N}(1.0,\ 0.5^2), \quad \tau = 1, \ldots, 9, \tag{12}$$
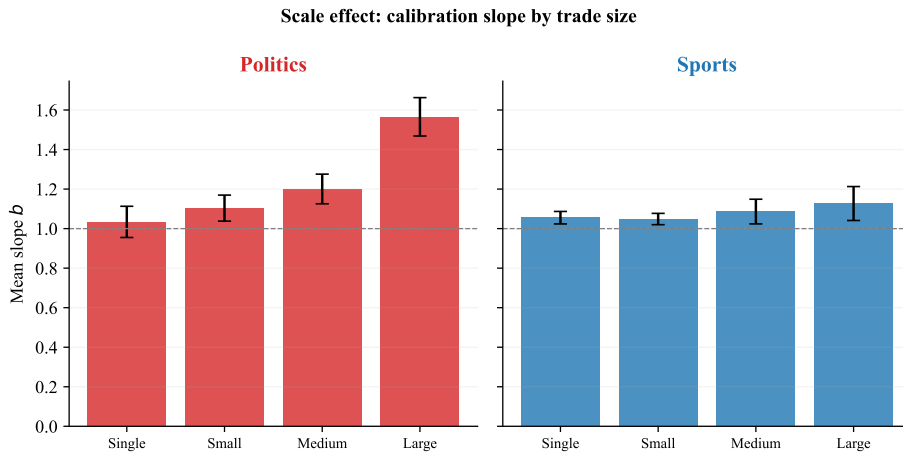
$$\alpha_d \sim \mathcal{N}(0,\ \sigma_\alpha^2), \quad \sum_{d=1}^{6} \alpha_d = 0, \tag{13}$$

$$\beta_d(\tau) \sim \mathcal{N}(0,\ \sigma_\beta^2), \quad \sum_d \beta_d(\tau) = 0 \ \forall \, \tau, \tag{14}$$

$$\delta_d \sim \mathcal{N}(0,\ \sigma_\delta^2). \tag{15}$$

Figure 4 Observed versus fitted calibration slopes $(R^2 = 0.873)$. Each point is one of 216 analysis cells.
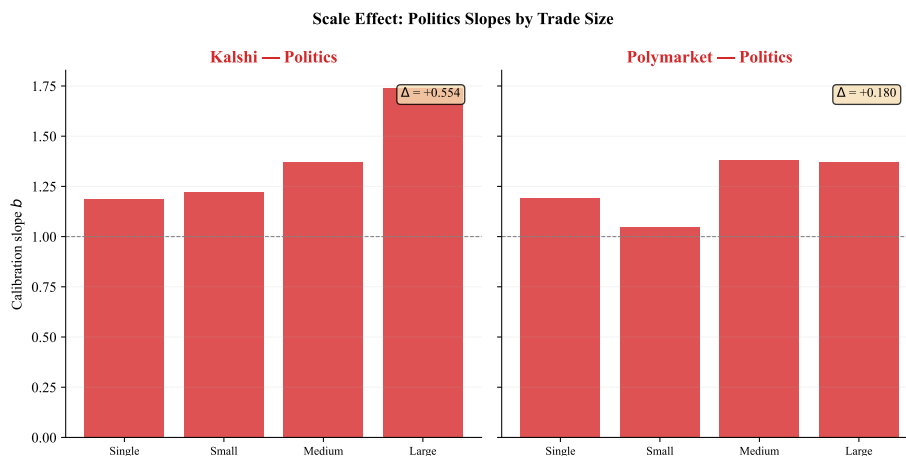

Figure 5 Calibration slopes by trade size: (A) Politics versus (B) Sports. In Politics, larger trades are associated with more compressed prices (higher slopes). In Sports, no such gradient exists.

The prior for $\mu(\tau)$ is centred at perfect calibration $(b = 1)$. At the hyperprior level:

$$\sigma_\alpha, \ \sigma_\beta, \ \sigma_\delta, \ \sigma \ \overset{\text{ind.}}{\sim} \ \text{Half-Cauchy}(0, 1). \tag{16}$$

The non-centred parameterisation $\alpha_d = \sigma_\alpha \cdot \alpha_d^{\text{raw}}$, with $\alpha_d^{\text{raw}} \sim \mathcal{N}(0, 1)$, is used for all hierarchical effects to improve sampling efficiency. Sum-to-zero constraints apply to all domain-level parameters.

Figure 6 Cross-platform scale effect in Politics. (A) Kalshi shows a monotonic increase from Single (1.19) to Large (1.74), $\Delta = +0.53$ $[0.29, 0.75]$. (B) Polymarket shows a non-monotonic pattern, $\Delta = +0.11$ $[-0.15, +0.39]$ (not significant).

Table 6. Bayesian posterior summaries for domain intercepts α_d.

Domain	Post. mean	SD	95% CI	Freq.
Politics	$+0.151$	0.015	$[+0.122, +0.179]$	$+0.156$
Sports	$+0.010$	0.015	$[-0.020, +0.039]$	$+0.009$
Crypto	$+0.005$	0.015	$[-0.024, +0.034]$	$+0.004$
Finance	$+0.006$	0.015	$[-0.023, +0.035]$	$+0.006$
Weather	-0.086	0.015	$[-0.115, -0.057]$	-0.090
Entertainment	-0.085	0.015	$[-0.114, -0.056]$	-0.086

Note: Freq. column reports the corresponding frequentist estimate. Maximum discrepancy is 0.005 (Politics).
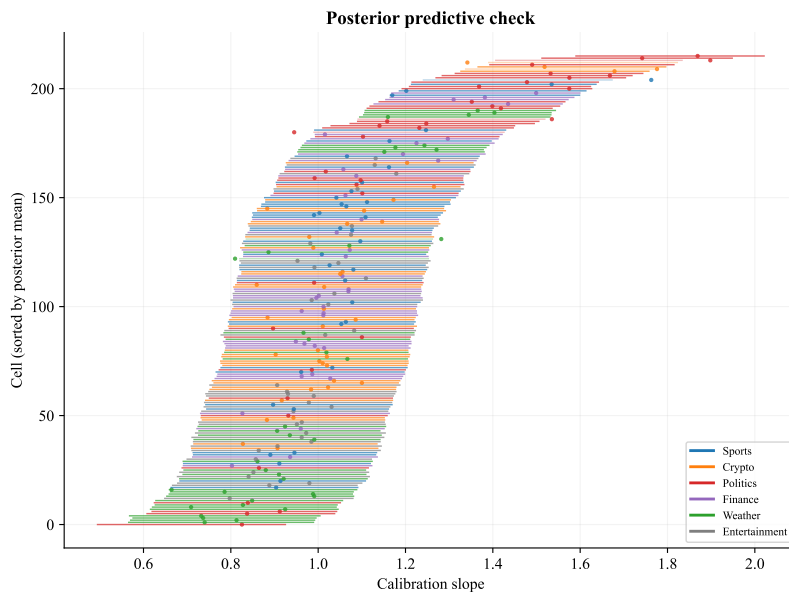
6.2. Results

Posterior inference is conducted via Hamiltonian Monte Carlo [Neal, 2011, Betancourt, 2017] using NumPyro [Phan et al., 2019] (4 chains, 4,000 iterations with 2,000 warmup). The maximum split-$\hat{R}$ [Vehtari et al., 2021, Brooks and Gelman, 1998] is 1.000, the minimum bulk effective sample size is 4,070, and no divergent transitions occurred.

Table 6 reports domain intercepts. Politics is the clear outlier: posterior mean $+0.151$ (95% credible interval $[0.122, 0.179]$), entirely above zero, confirming persistent underconfidence. Weather (-0.086, $[-0.115, -0.057]$) and Entertainment (-0.085, $[-0.114, -0.056]$) are significantly below zero, indicating overconfidence. Sports ($+0.010$), Crypto ($+0.005$) and Finance ($+0.006$) are indistinguishable from zero.

The maximum parameter discrepancy between approaches is 0.005 (Politics), confirming that the hierarchical priors provide appropriate shrinkage without distortion. It is worth noting that the frequentist model treats $\gamma_d(s)$ as categorical (four size bins), while the Bayesian model parameterises the scale effect as $\delta_d \times \log(s)$, a continuous linear specification; the close agreement between the two despite this parameterisation difference supports the robustness of the decomposition. The scale sensitivity δ confirms the scale effect: Politics has $\delta = +0.088$ (95% CI $[0.072, 0.103]$), the only domain with a credible interval entirely above zero, indicating that larger trades are associated with more compressed, less calibrated prices.

6.3. Posterior predictive check

Of 216 cells, 208 (96.3%) have observed slopes within their 95% posterior predictive intervals, close to the nominal 95% coverage (Table 14 reports per-domain breakdowns). The 8 cells outside are not concentrated in any single domain (Figure 7).

Figure 7 Posterior predictive check: observed versus posterior-predicted calibration slopes with 95% intervals. Of 216 cells, 208 (96.3%) fall within their prediction intervals.

7. Discussion

7.1. Why do domains differ?

The finding that domain-by-horizon interactions constitute the largest calibration component (26.0%) demands explanation. Three interpretive hypotheses are offered.

The bilateral cancellation hypothesis. Political markets attract traders with strong, opposing convictions. In a presidential race, partisans on both sides may trade aggressively, each confident their candidate will win. These opposing bets partially cancel: a large buy order at 65 cents meets a large sell order, and the price stays near 50–60% even when the true probability is substantially higher or lower. The result is systematic price compression (underconfidence) that worsens as larger, more conviction-driven traders enter the market. This connects directly to Manski's (2006) theoretical prediction (see also Ottaviani and Sørensen, 2008) that heterogeneous beliefs and risk preferences can cause prices to diverge from mean beliefs, and explains why the scale effect is confined to Politics: the other domains lack the bipolar conviction structure that drives bilateral cancellation. Cross-platform evidence strengthens this interpretation: the domain-level underconfidence ($b > 1$) replicates on Polymarket, confirming that bilateral cancellation is a structural feature of political prediction, but the trade-size amplification does not, consistent with Polymarket's pseudonymous blockchain environment fragmenting whale positions across multiple wallets and its $11\times$ deeper liquidity absorbing large trades more efficiently (Tables 16–17). Note that Politics on Kalshi, despite comprising only 7.6% of trades, accounts for 13.3% of contract volume, consistent with larger position sizes driven by conviction.

The signal over-reaction hypothesis. Weather markets are uniquely overconfident at short horizons, the only domain where prices are too extreme. This likely reflects over-reaction to meteorological signals. When a forecast predicts a storm tomorrow, traders push prices too far: they overshoot what climatological base rates would justify. Weather's low base rate (24.0% in Table 1, reflecting threshold exceedance contracts) means that over-reacting to "yes" signals is particularly costly. At longer horizons, where meteorological signals are weaker and base rates dominate, weather markets converge to the universal underconfidence pattern.

The information convergence hypothesis. Sports markets are well calibrated at short-to-medium horizons because information is continuous, quantifiable and publicly observable: game scores update in real time, player statistics are tracked in detail, injury reports follow regular schedules. This generates smooth convergence to truth with minimal

disagreement, so that prices reach appropriate extremes. At long horizons (1 month+), sports forecasting becomes more speculative, information advantage declines, and the same favourite–longshot underconfidence that affects all domains takes hold (slope 1.74). The sports trajectory illustrates a general principle: calibration quality tracks the richness of the available information environment [cf. Sunstein, 2006].

7.2. Practical implications for prediction market consumers

Prediction markets are increasingly cited by journalists, embedded in news coverage and referenced by policymakers. The findings suggest domain-specific caution. For any (domain, horizon, trade-size) combination, the estimated slope $\hat{\theta}$ transforms a raw market price $p \in (0, 1)$ into a recalibrated probability:

$$p^* = \sigma\big(\hat{\theta} \cdot \mathrm{logit}(p)\big) = \frac{p^{\hat{\theta}}}{p^{\hat{\theta}} + (1-p)^{\hat{\theta}}}, \tag{17}$$

where $\sigma(\cdot)$ is the logistic function and the second equality follows from algebraic simplification. When $\hat{\theta} > 1$ (underconfidence), $p^* > p$ for $p > 0.5$ and $p^* < p$ for $p < 0.5$: the recalibrated probability is more extreme than the raw price. When $\hat{\theta} < 1$ (overconfidence), the recalibrated probability is more moderate. The full 216-cell calibration matrix is provided as supplementary material.

For political markets. A price of 70 cents in a political market one week before resolution does not mean 70%. Applying (17) with the domain- and horizon-specific slope $\hat{\theta} \approx 1.83$ yields

$$p^* = \frac{0.70^{1.83}}{0.70^{1.83} + 0.30^{1.83}} \approx 0.83. \tag{18}$$

Political prediction market prices are systematically compressed toward 50%: favourites are underpriced and longshots are overpriced. Journalists who report these prices at face value systematically understate the confidence that should be attached to the leading outcome.

For sports markets. At horizons under one week, sports market prices are reasonably trustworthy (slopes 0.90–1.10). Beyond one month, the favourite–longshot bias appears strongly (slope 1.74).

For weather markets. At short horizons, weather prices are if anything too extreme, over-reacting to signals. Consumers of short-horizon weather contracts should recognise that the price may overstate the probability of the predicted outcome. At moderate horizons, weather markets are among the best calibrated.

7.3. Implications for market design

Position limits in political markets. The scale effect suggests that large positions are associated with amplified price compression in political markets. Position limits or progressive fee structures would reduce the influence of conviction-driven bilateral trading.

One-person-one-vote aggregation. The finding that trade-weighted calibration is substantially better than contract-weighted calibration in Politics (mean gap 0.33) suggests that equal-weight aggregation may produce better probabilities in politically polarised domains. Prediction polls [Atanasov et al., 2017, Pennock et al., 2001] that weight each forecaster equally may outperform wealth-weighted markets for political questions.

Domain-specific credibility indicators. Exchanges could display calibration track records alongside prices, such as a traffic-light system indicating whether a given domain–horizon combination has historically been well calibrated, compressed, or over-reactive.

7.4. Implications for forecasting research

These results challenge domain-agnostic calibration studies. The decomposition shows that pooling across domains obscures the dominant source of calibration variation. Future studies should stratify by domain and time horizon. The four-component framework extends naturally to election models, expert platforms and machine learning probability outputs.

7.5. Limitations

Several limitations should be noted. First, although core findings are validated on two exchanges, both are US-centric platforms; prediction markets in different regulatory environments (e.g. Betfair in the UK; see Smith et al., 2006) may exhibit different patterns. The Polymarket comparison is limited to three of six domains and affected by timestamp noise at short horizons.

Second, the domain classification is coarse. Within Politics, subcategories have different dynamics (Section 4). A finer-grained decomposition would require a more complex model.

Third, only trades are observed, not trader identities. The bilateral cancellation interpretation of the scale effect is consistent with the data but is a hypothesis. Alternative explanations (institutional hedging, algorithmic artefacts) cannot be ruled out without richer data.

Fourth, the logistic recalibration model is one approach to measuring calibration. Reliability diagrams [Murphy and Winkler, 1977], proper scoring rules [Gneiting and Raftery, 2007, Brier, 1950] and nonparametric calibration curves provide complementary perspectives that future work should integrate.

Fifth, whether these patterns are stable over time is an open question. The decomposition is designed to be re-estimated as new data becomes available.

8. Conclusion

Prediction markets are powerful information aggregation mechanisms, but their power is not uniform. This paper shows, using data from two structurally different exchanges, that calibration varies systematically across knowledge domains, time horizons and trade sizes. Four components account for 87.3% of this variation on Kalshi, and the dominant pattern, political underconfidence, replicates on Polymarket.

All crowds share a tendency toward the favourite–longshot bias at long horizons: prices are compressed toward 50%, understating the probability of favoured outcomes. But some crowds, particularly those trading on political events, exhibit this compression at nearly every horizon, and it is most pronounced among the traders who wager most heavily. This is interpreted as bilateral cancellation: opposing partisan bets pull prices toward 50%. This interpretation is consistent with the broader behavioural finance literature on how probability weighting distorts risk assessment [Kahneman and Tversky, 1979, Tversky and Kahneman, 1992]. Cross-platform evidence supports this: the domain-level bias replicates on Polymarket, but the trade-size amplification does not, consistent with pseudonymous blockchain trading fragmenting whale positions. Other crowds, those trading on weather, display the opposite bias at short horizons, over-reacting to signals and pushing prices too far. Sports crowds, anchored by continuous information and disciplined by professional participants, come closest to the calibration ideal at short-to-medium horizons.

These findings have practical consequences. Millions of people now consult prediction markets. If they take prices at face value, they are systematically misled—and the direction of their error depends on what they are looking at. In political markets, the true probability of the favoured outcome is substantially higher than the price suggests. In short-horizon weather markets, it may be lower.

The wisdom of crowds is real, but it has a structure. Understanding that structure is essential for using it wisely.

Data availability statement

The Kalshi and Polymarket data used in this study were collected using the data collection framework and pre-collected dataset of Becker [2026], available at `https://github.com/Jon-Becker/prediction-market-analysis/`. The raw data are also publicly accessible through the Kalshi API (`https://trading-api.readme.io/`) and the Gamma API/Polygon blockchain indexer. The analysis code, domain classification rules and the full 216-cell calibration matrix are available at `https://github.com/namanhz/prediction-market-calibration` and as supplementary material. All results can be reproduced from the deposited code and the publicly accessible datasets.

Supplementary material

Supplementary material is available online at the journal website. It includes (a) the full 216-cell calibration matrix as a CSV file, (b) domain classification rules (560+ pattern rules) and (c) cross-platform comparison tables.

Acknowledgements

Funding

Conflict of interest

The author declares no conflict of interest. The author has no financial interest in Kalshi, Polymarket or any prediction market exchange.

References

M. M. Ali. Probability and utility estimates for racetrack bettors. *Journal of Political Economy*, 85:803–815, 1977.

K. J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. R. Neumann, M. Ottaviani, T. C. Schelling, R. J. Shiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. R. Varian, J. Wolfers, and E. Zitzewitz. The promise of prediction markets. *Science*, 320:877–878, 2008.

H. Assareh, R. Noorossana, and K. L. Mengersen. Bayesian change point estimation in Poisson-based control charts. *Journal of Industrial Engineering International*, 9:32, 2013.

P. Atanasov, P. Rescober, E. Stone, S. A. Swift, E. Servan-Schreiber, P. Tetlock, L. Ungar, and B. Mellers. Distilling the wisdom of crowds: prediction markets vs. prediction polls. *Management Science*, 63:691–706, 2017.

J. Baron, B. A. Mellers, P. E. Tetlock, E. Stone, and L. H. Ungar. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11:133–145, 2014.

J. Becker. The microstructure of wealth transfer in prediction markets, 2026. URL `https://jbecker.dev/research/prediction-market-microstructure`. Dataset and analysis framework available at `https://github.com/Jon-Becker/prediction-market-analysis/`.

J. E. Berg, F. D. Nelson, and T. A. Rietz. Prediction market accuracy in the long run. *International Journal of Forecasting*, 24: 285–300, 2008.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

J. E. Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4:49–65, 2007.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

C. F. Camerer, A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351:1433–1436, 2016.

K.-Y. Chen and C. R. Plott. Information aggregation mechanisms: concept, design, and implementation for a sales forecasting problem. Working Paper No. 1131, California Institute of Technology, 2002.

B. Cowgill and E. Zitzewitz. Corporate prediction markets: evidence from Google, Ford, and Firm X. *Review of Economic Studies*, 82: 1309–1341, 2015.

A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–610, 1982.

M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32:12–22, 1983.

S. DellaVigna and D. Pope. Predicting experimental results: who knows what? *Journal of Political Economy*, 126:2410–2456, 2018.

A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112:15343–15347, 2015.

D. Easley and M. O'Hara. Price, trade size, and information in securities markets. *Journal of Financial Economics*, 19:69–90, 1987.

R. S. Erikson and C. Wlezien. Are political markets really superior to polls as election predictors? *Public Opinion Quarterly*, 72: 190–215, 2008.

E. F. Fama. Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25:383–417, 1970.

R. Forsythe, F. Nelson, G. R. Neumann, and J. Wright. Anatomy of an experimental political stock market. *American Economic Review*, 82:1142–1161, 1992.

F. Galton. Vox populi. *Nature*, 75:450–451, 1907.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, 2007.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, 3rd edition, 2013.

L. R. Glosten and P. R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14:71–100, 1985.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

R. M. Griffith. Odds adjustments by American horse-race bettors. *American Journal of Psychology*, 62:290–294, 1949.

S. J. Grossman and J. E. Stiglitz. On the impossibility of informationally efficient markets. *American Economic Review*, 70:393–408, 1980.

R. Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5:107–119, 2003.

F. A. Hayek. The use of knowledge in society. *American Economic Review*, 35:519–530, 1945.

M. D. Hoffman and A. Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

J. Hogg, J. Cameron, S. Cramb, P. Baade, and K. Mengersen. A two-stage Bayesian small area estimation approach for proportions. *International Statistical Review*, 92:455–482, 2024.

L. Hong and S. E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101:16385–16389, 2004.

D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. Wiley, Hoboken, 3rd edition, 2013.

D. Kahneman and A. Tversky. Prospect theory: an analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

E. Karger, J. T. Monrad, B. A. Mellers, and P. E. Tetlock. Reciprocal scoring: a method for forecasting unanswerable questions. *Judgment and Decision Making*, 18:e3, 2023.

A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53:1315–1335, 1985.

A. Leigh and J. Wolfers. Competing approaches to forecasting elections: economic models, opinion polling and prediction markets. *Economic Record*, 82:325–340, 2006.

S. Lichtenstein, B. Fischhoff, and L. D. Phillips. Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment Under Uncertainty: Heuristics and Biases*, pages 306–334. Cambridge University Press, Cambridge, 1982.

Y. Liu, T. Gneiting, and A. E. Raftery. Boldness-recalibration for binary event predictions. *The American Statistician*, 78:426–436, 2024.

C. F. Manski. Interpreting the predictions of prediction markets. *Economics Letters*, 91:425–429, 2006.

B. Mellers, L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, P. Atanasov, S. A. Swift, T. Murray, E. Stone, and P. E. Tetlock. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25:1106–1115, 2014.

K. L. Mengersen, C. P. Robert, and C. Guihenneuc-Jouyaux. MCMC convergence diagnostics: a review. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 415–440. Oxford University Press, Oxford, 1999.

A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society, Series C*, 26:41–47, 1977.

R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, Boca Raton, 2011.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, New York, 2005. ACM.

M. Ottaviani and P. N. Sørensen. The favorite–longshot bias: an overview of the main explanations. In D. B. Hausch and W. T. Ziemba, editors, *Handbook of Sports and Lottery Markets*, pages 83–101. North-Holland, Amsterdam, 2008.

L. Page and R. T. Clemen. Do prediction markets produce well-calibrated probability forecasts? *The Economic Journal*, 123:491–513, 2013.

A. B. Palley and V. A. Satopää. Robust recalibration of aggregate probability forecasts using meta-beliefs. *International Journal of Forecasting*, 40:1421–1434, 2024.

D. M. Pennock, S. Lawrence, C. L. Giles, and F. A. Nielsen. The real power of artificial markets. *Science*, 291:987–988, 2001.

D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019.

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularised likelihood methods. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, 1999.

C. R. Plott and S. Sunder. Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica*, 56:1085–1118, 1988.

P. W. Rhode and K. S. Strumpf. Historical presidential betting markets. *Journal of Economic Perspectives*, 18:127–141, 2004.

D. Rothschild. Forecasting elections: comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, 73:895–916, 2009.

V. A. Satopää, J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30:344–356, 2014.

E. Servan-Schreiber, J. Wolfers, D. M. Pennock, and B. Galebach. Prediction markets: does money matter? *Electronic Markets*, 14:243–251, 2004.

Table 7. Variance decomposition under alternative specifications.

Specification	$\mu\ R^2$	$\alpha\ R^2$	$\beta\ R^2$	$\gamma\ R^2$	Total R^2
Baseline $[5, 95]$, $C = 10$	0.302	0.146	0.260	0.165	0.873
Volume ≥ 100	0.302	0.146	0.260	0.165	0.873
Price $[10, 90]$	0.295	0.097	0.333	0.137	0.861
Price $[2, 98]$	0.308	0.157	0.231	0.190	0.885
Price $[1, 99]$	0.279	0.158	0.242	0.183	0.861
$C = 1$	0.302	0.146	0.260	0.165	0.873
$C = 100$	0.302	0.146	0.260	0.165	0.873

Total R^2 is stable across all price ranges (0.861–0.885) and regularisation strengths (identical at $C = 1, 10, 100$). Volume filtering produces identical results. The wider price range $[2, 98]$ slightly increases total R^2 to 0.885; the domain ranking is preserved.

Table 8. Scale effect: trade-level versus market-clustered bootstrap.

Domain	Bootstrap method	Mean Δ	95% CI	Sig.?
Politics	Trade-level	+0.53	$[0.29, 0.75]$	Yes
Politics	Market-clustered	+0.59	$[0.13, 1.29]$	Yes
Sports	Trade-level	+0.07	$[-0.07, 0.26]$	No
Sports	Market-clustered	+0.01	$[-0.03, 0.05]$	No

Table 9. Size $\times$ horizon confounding diagnostic: median hours to close by domain and trade size.

Domain	Single	Small	Medium	Large
Sports	2.3	7.6	1.5	96.3
Crypto	202.2	5.7	0.6	5,936.7
Politics	862	1,505	1,488	213
Finance	1.7	427.1	0.5	4.5
Weather	14.9	15.8	24.2	156.6
Entertainment	144.3	0.5	0.6	55.5

In Politics, large trades have shorter median horizons (213 hours) than single-contract trades (862 hours), meaning any horizon confounding would bias against finding a scale effect.

H. S. Shin. Optimal betting odds against insider traders. *The Economic Journal*, 101:1179–1185, 1991.

M. A. Smith, D. Paton, and L. V. Williams. Market efficiency in person-to-person betting. *Economica*, 73:673–689, 2006.

E. Snowberg and J. Wolfers. Explaining the favorite–longshot bias: is it risk-love or misperceptions? *Journal of Political Economy*, 118:723–746, 2010.

C. R. Sunstein. *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press, Oxford, 2006.

J. Surowiecki. *The Wisdom of Crowds*. Doubleday, New York, 2004.

P. E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, 2005.

P. E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, New York, 2015.

R. H. Thaler and W. T. Ziemba. Anomalies: parimutuel betting markets: racetracks and lotteries. *Journal of Economic Perspectives*, 2:161–174, 1988.

A. Tversky and D. Kahneman. Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992.

A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P. C. Bürkner. Rank-normalisation, folding, and localisation: an improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16:667–718, 2021.

R. L. Winkler. Scoring rules and the evaluation of probabilities. *Test*, 5:1–60, 1996.

J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18:107–126, 2004.

J. Wolfers and E. Zitzewitz. Interpreting prediction market prices as probabilities. NBER Working Paper No. 12200, 2006.

Table 10. F-test derivation and effect sizes.

Source	SS	df	MS	F	Partial η^2
α (domain)	1.435	5	0.287	33.16	0.535
β (domain×time)	2.562	40	0.064	7.40	0.673
γ (domain×size)	1.587	18	0.088	10.42	0.566
Residual	1.247	144	0.009	—	—

All p-values are below 10^{-16}. Partial $\eta^2 = \text{SS}_{\text{component}}/(\text{SS}_{\text{component}} + \text{SS}_{\text{residual}})$.

Table 11. Prior specification.

Parameter	Prior	Dim.	Constraint
$\mu(\tau)$	$\mathcal{N}(1.0, 0.5)$	9	None
σ_α	$\text{HalfCauchy}(1.0)$	1	> 0
α_{raw}	$\mathcal{N}(0, 1)$	5	Sum-to-zero
σ_β	$\text{HalfCauchy}(1.0)$	1	> 0
β_{raw}	$\mathcal{N}(0, 1)$	40	Doubly centred
σ_δ	$\text{HalfCauchy}(1.0)$	1	> 0
δ_{raw}	$\mathcal{N}(0, 1)$	6	None
σ	$\text{HalfCauchy}(1.0)$	1	> 0

The model uses non-centred parameterisations for all hierarchical effects. The domain intercepts satisfy a sum-to-zero constraint. The interaction matrix β is doubly centred.

Table 12. Hyperparameter posterior summaries.

Parameter	Mean	SD	95% CI	Interpretation
σ_α	0.097	0.035	$[0.048, 0.183]$	Moderate spread
σ_β	0.098	0.011	$[0.079, 0.122]$	Well-constrained
σ_δ	0.036	0.013	$[0.017, 0.067]$	Small
σ	0.097	0.006	$[0.087, 0.109]$	Residual ≈ 0.10

Maximum $\hat{R}$ across all parameters is 1.0000; minimum effective sample size is 4,070.

A. Robustness checks

B. Bayesian model specification and diagnostics

C. Cross-platform validation details

This appendix provides full tabular results for the cross-platform comparison between Kalshi and Polymarket. Polymarket trade timestamps are derived from Polygon block numbers via a bucketed lookup table with approximately 6-hour granularity, introducing approximately 3 hours of noise. Bins 0–1 hour and 1–3 hours are therefore unreliable for cross-platform comparison and are excluded from mean slope calculations.

The Polymarket Politics 2 days to 1 week bin (slope 1.982, 4.32 million trades) is a genuine outlier exceeding even the Kalshi value at the same horizon (1.833). This likely reflects a concentration of 2024 US election daily-resolution markets whose peak trading fell 2–7 days before resolution. Several limitations also merit note: the regex-based domain classifier assigns 42.5% of Polymarket markets to "Other"; Weather and Entertainment are absent; the count field represents contracts per on-chain transaction rather than full position size; and no market-clustered bootstrap has been run for Polymarket.

Table 13. Posterior β matrix (domain × time interaction).

Domain	0–1h	1–3h	3–6h	6–12h	12–24h	24–48h	2d–1w	1w–1mo	1mo+
Sports	+.082	−.030	−.066	+.005	+.032	+.020	−.081	−.050	+.088
Crypto	+.013	+.021	+.051	−.058	−.051	+.063	−.106	−.181	+.249
Politics	−.072	−.173	−.066	+.069	−.019	+.028	+.072	+.170	−.008
Finance	+.027	+.077	+.027	−.014	−.006	−.118	−.003	+.146	−.135
Weather	−.031	−.040	−.101	−.022	+.018	+.044	+.132	−.088	+.088
Entertainment	−.019	+.144	+.156	+.021	+.027	−.036	−.015	+.004	−.281

The β matrix captures each domain's slope deviation from the universal horizon mean $\mu(\tau)$ plus the domain intercept α.

Table 14. Posterior predictive coverage by domain.

Domain	Cells	Within 95%	Coverage	Mean PPC p-value
Sports	36	35	97.2%	0.48
Crypto	36	35	97.2%	0.49
Politics	36	33	91.7%	0.47
Finance	36	35	97.2%	0.50
Weather	36	35	97.2%	0.49
Entertainment	36	35	97.2%	0.51

Table 15. Cross-platform calibration slopes (Δ = Polymarket $-$ Kalshi). Bins marked $\dagger$ are unreliable due to timestamp noise.

	Sports			Crypto			Politics		
Bin	Kalshi	Poly	Δ	Kalshi	Poly	Δ	Kalshi	Poly	Δ
0–1h$\dagger$	1.101	1.028	-0.073	0.993	1.753	$+0.760$	1.341	0.621	-0.720
1–3h$\dagger$	0.960	0.976	$+0.016$	1.013	1.390	$+0.377$	0.933	0.859	-0.073
3–6h	0.897	1.171	$+0.275$	1.065	0.862	-0.203	1.317	1.116	-0.201
6–12h	1.006	1.001	-0.005	1.007	1.044	$+0.037$	1.552	0.936	-0.616
12–24h	1.053	1.059	$+0.006$	1.006	0.996	-0.009	1.477	1.277	-0.200
24–48h	1.075	0.950	-0.126	1.209	1.123	-0.087	1.515	1.234	-0.281
2d–1w	1.037	1.097	$+0.059$	1.121	0.998	-0.123	1.833	1.982	$+0.149$
1w–1mo	1.240	0.974	-0.266	1.090	1.161	$+0.071$	1.833	1.681	-0.152
1mo+	1.740	1.322	-0.418	1.357	1.060	-0.298	1.730	1.086	-0.644
Mean (rel.)	1.150	1.082	-0.068	1.114	1.049	-0.065	1.637	1.313	-0.324

Table 16. Cross-platform scale effect by domain and trade size.

Domain	Platform	Single	Small	Medium	Large	Δ(L$-$S)
Sports	Kalshi	1.002	1.010	1.007	1.013	$+0.011$
Sports	Polymarket	1.120	1.036	1.065	1.038	-0.082
Crypto	Kalshi	1.028	1.025	1.023	1.004	-0.024
Crypto	Polymarket	1.006	1.066	1.059	1.063	$+0.057$
Politics	Kalshi	1.188	1.224	1.373	1.741	$+0.554$
Politics	Polymarket	1.193	1.046	1.381	1.373	$+0.180$

Table 17. Cross-platform whale effect bootstrap.

Platform	Domain	Δ(L$-$S)	95% CI	Sig.?
Kalshi (cell)	Politics	$+0.531$	$[+0.288, +0.747]$	Yes
Kalshi (clustered)	Politics	$+0.589$	$[+0.132, +1.289]$	Yes
Polymarket (cell)	Politics	$+0.113$	$[-0.151, +0.395]$	No
Polymarket (cell)	Sports	$+0.006$	$[-0.199, +0.170]$	No
Polymarket (cell)	Crypto	$+0.092$	$[-0.158, +0.358]$	No

D. Additional supplementary appendices

The following materials are provided in the supplementary data: (a) domain classification rules for both Kalshi (560+ ticker-prefix patterns) and Polymarket (compiled regular expression patterns on market titles); (b) the full 216-cell calibration matrix as a CSV file, with observed slopes, standard errors, fitted decomposition components

and residuals; and (c) political subcategory analysis, including slope ranges for 10 subcategories and a Simpson's paradox diagnostic at the 1–3 hour horizon.