HEC PARIS

Master in International Finance

RESEARCH PAPER

Academic Year 2025 - 2026

Where Crowds Get It Wrong: Probability-Band Mispricing on Polymarket

Florian LINK[1] and Tom HOMMOLA[2]

Under the supervision of: Prof. Daniel SCHMIDT

President of the Jury: Prof. Jean-Pierre FRANCOIS

Keywords: prediction markets; Polymarket; market efficiency; favourite-longshot bias; inverse longshot bias; betting markets

☒ *PUBLIC REPORT* ☐ *CONFIDENTIAL REPORT*

[1] 50 Dudley Court, London WC2H 9RF, United Kingdom; +49 162 4102244; Florian.link@hec.edu
[2] 50 Dudley Court, London WC2H 9RF, United Kingdom; +4915120840561; Tom.hommola@hec.edu

Where Crowds Get It Wrong: Probability-Band Mispricing on Polymarket

Florian Link and Tom Hommola

Abstract

This thesis tests the favourite-longshot bias on Polymarket using 4,363 cleanly resolved binary sports contracts across the NBA, NHL, NCAA basketball, and four soccer leagues from August 2025 to April 2026. We apply a hold-to-maturity, per-integer touch-once calibration test comparing each contract's traded price against its realised resolution frequency, using Wilson intervals and by-market block-bootstrap inference. We find no classical longshot bias. Instead, prices show a band-conditional inverse longshot bias: mid-probability contracts are underpriced, with a mean deviation of +0.020 peaking at +0.042 near 0.45, while favourites price near fair. Soccer drives the effect, which clearly weakens after February 2026.

Table of Contents

Where Crowds Get It Wrong: Probability-Band Mispricing on Polymarket

1. Introduction

1.1 Motivation

The favourite-longshot bias is one of the most durable empirical regularities in behavioural finance and one of the most directly attributable to a formal theory. Cumulative prospect theory (Tversky & Kahneman 1992) predicts that people do not treat probabilities at face value. Instead, they distort them: small probabilities are treated as larger than they are, and large probabilities as smaller. Prelec (1998) provides the axiomatic foundation and the standard one-parameter functional form, and Barberis & Huang (2008) extend the mechanism into asset pricing by showing that securities with lottery-like payoff distributions are over-priced in equilibrium. The prediction is sharp: in any market where traders can bet on unlikely outcomes, longshot prices should exceed their true win rate, and heavy favourites should be underpriced.

Snowberg & Wolfers (2010) provide the canonical confirming evidence for this prediction on US horse-racing parimutuel markets and explicitly favour probability weighting over a risk-loving preference as the underlying mechanism. The result has been replicated across a wide range of sportsbook and parimutuel venues and is one of the few stylised facts that prospect theory generates and that the data routinely deliver. The conditions under which the bias appears most cleanly are quite specific: a single shared pool rather than a continuous orderbook, no repricing during the event, predominantly retail participants, and one bet per person per event. Whether the same behavioural prediction holds when those conditions are relaxed is an open empirical question.

Polymarket is a natural laboratory for re-testing this prediction at scale. The venue offers real money, a continuous double-sided orderbook, unambiguous binary resolution, and tens of thousands of resolved markets across sports, politics, and macroeconomic outcomes. Reichenbach & Walther (2025) exploit this setting and report that, across 124 million trades, there is no aggregate favourite-longshot bias on Polymarket. Their result is a clean rejection of the classical prediction at the venue level. However, it raises an additional question, which the thesis addresses: Does the overall picture conceal a more specific mispricing that only emerges when you examine particular price ranges or market types?

1.2 Research question and contribution preview

We test the classical longshot bias on Polymarket sports prediction markets, measuring calibration at the level of individual percentage-point price intervals. he central question is whether the Kahneman-Tversky

prediction – that low-probability tokens are systematically over-priced – holds once we examine results separately by price range, sport, and the market conditions before and after February 2026. The classical bias serves as the formal null hypothesis. We treat the null as fully specified by Tversky & Kahneman (1992), Prelec (1998), and Barberis & Huang (2008), and we evaluate it against two alternatives: a bias running in the opposite direction across the price spectrum, and the near-zero aggregate result reported by Reichenbach & Walther (2025).

Across 4,363 sports markets, the data exhibit a band-conditional inversion of the classical Kahneman-Tversky prediction. Low-probability and mid-range tokens resolve to one at a rate persistently above their implied probability, indicating systematic under-pricing across essentially the entire price range below 0.90. The deviation is largest in the mid-range, peaking at +0.042 around a coin-flip, and fades toward the heavy-favourite end where it is statistically indistinguishable from zero. The classical favourite-longshot bias – under which longshots are over-priced and favourites under-priced – is not present in any price region. The aggregate result is close to the null reported by Reichenbach and Walther (2025), but this masks a structural pattern rather than confirming uniform efficiency. The heavy-favourite end shows almost no mispricing and accounts for a large share of observations, so it dilutes the genuine under-pricing at the longshot end when averaged across the full sample. We refer to this one-sided pattern as the band-conditional, inverse longshot bias, and treat its documentation, its replication across sport categories, and its decomposition across the pre- and post-February-2026 microstructure regimes as the central empirical contribution of the thesis.

1.3 Scope and constraint

The empirical scope is restricted to sports prediction markets on Polymarket. The sample covers NBA, NHL, NCAA Basketball, and soccer (EPL, LaLiga, Bundesliga, Serie A), observed at tick level from August 2025 through April 2026. The window is approximately nine months and includes a structural break in February 2026 that we document independently in the project's regime-change diagnostics. The break is visible across multiple market statistics and across all sports. We report every calibration result separately for the two periods to ensure our headline findings are not driven by conditions specific to one of them. Non-sports markets - political, macroeconomic, and crypto-resolution markets - are out of scope.

The thesis is descriptive: it documents, statistically tests, and decomposes the calibration anomaly. The natural follow-on question - whether the documented mispricing can be translated into a tradeable strategy under Polymarket’s buy-only constraint and microstructure costs - is reserved for the Discussion (§7) and Future Work (§8.4). Where the discussion proposes such a strategy, it is presented as a research design that the present thesis does not execute, rather than as a tested contribution.

1.4 Methodological signature

The methodological backbone is a per-band, per-sport calibration test. Each market is treated as a single observation, weighted by whether its price ever touched the band of interest during the market lifetime. Uncertainty is quantified with Wilson intervals at the cell level and with a by-market block bootstrap that resamples whole markets rather than individual ticks, preserving the correlation between price observations within the same market. We also verify that the inversion is not a data-labelling artefact by checking that the first and second outcome tokens give symmetric results, and we replicate every headline finding on the pre- and post-February-2026 subsamples to confirm it is not specific to one market regime.

1.5 Contribution claim

Our contribution sits in the space between two reference points: Reichenbach and Walther's (2025) aggregate null on Polymarket and Snowberg and Wolfers's (2010) classical confirmation of the longshot bias on US horse-racing parimutuels. We have identified Polymarket sports betting markets as a natural venue where these two strands of literature collide.

We make one precise contribution. We extend Reichenbach and Walther (2025) by breaking their aggregate result down by price range, sport, and time period. In our data, the near-zero aggregate result conceals meaningful structure: tokens are under-priced across the longshot and mid-range portions of the price spectrum, with an average edge of around two percentage points and a peak in the mid-range, while heavy favourites are priced close to fair. A subset of per-integer buckets reject a zero edge under Bonferroni correction on the pooled sample. The resulting picture is qualitatively different from the smoothed aggregate and runs counter to the classical Kahneman-Tversky prediction. We also document a structural shift in the calibration pattern around February 2026; we report this as a descriptive finding and are cautious about its interpretation, as the mechanism behind the regime break remains speculative.

1.6 Roadmap

Section 2 surveys the relevant literature on probability weighting, the favourite-longshot bias, prediction-market efficiency, and the behavioural mechanisms that we will draw on in the discussion. Section 3 formalises the hypotheses and states the criteria we use to decide between them. Section 4 describes the data, the sport-level coverage, the structural break in February 2026, and the cleaning and band-construction conventions. Section 5 details the calibration methodology, the three calibration variants, and the bootstrap and multiple-testing machinery. Section 6 presents the calibration results per band, per sport, and per regime. Section 7 discusses the candidate mechanisms - with particular attention to live-event-driven fading

- and reconciles our findings with the classical theory and with Reichenbach & Walther (2025). Section 8 concludes and identifies the natural extensions of this work.

2. Literature Review

This thesis sits at the intersection of four research strands: the prospect-theoretic foundations of probability weighting (§2.1), the favourite-longshot bias documented in betting markets (§2.2), the broader literature on prediction-market efficiency (§2.3), and the recent micro-empirical work on Polymarket (§2.4). For each, we set out what it predicts, what it has found, and where it leaves a gap that our calibration evidence can fill.

2.1 Probability weighting and prospect theory

The theoretical anchor for any test of the longshot bias is prospect theory, initiated by Kahneman and Tversky (1979). That paper argues that people evaluating risky outcomes do not behave as classical expected utility theory predicts. Instead, they apply a distorted weighting to probabilities: small probabilities are over-weighted and moderate-to-high probabilities are under-weighted relative to their true values. This feature was designed to explain well-documented experimental anomalies – the Allais paradox, the certainty effect, the reflection effect – that expected utility could not accommodate. The 1979 paper identifies the core regularity but is limited to simple two-outcome bets and does not specify a weighting function that works over arbitrary probability distributions.

Tversky and Kahneman (1992) resolved this with cumulative prospect theory. The key change is technical but consequential: rather than applying distorted weights directly to each outcome's probability, weights are applied to the cumulative probability – the probability of doing at least as well as a given outcome. This seemingly small adjustment removes a flaw in the original model – whereby irrational choices could appear rational under the old framework The estimated weighting function takes an inverted-S shape: it rises steeply for probabilities near zero and flattens out for probabilities near one, meaning agents behave as if low-probability events are more likely than they are, and high-probability events less likely. This shape generates the prediction we test in this thesis: in a market of prospect-theoretic agents, low-probability outcomes should trade above their true probabilities and high-probability outcomes below them – the classical longshot bias.

Prelec (1998) provides the axiomatic foundation that the empirical literature has largely standardised on. Working from a small set of behavioural axioms – compound invariance (the idea that if two gambles feel

equally attractive, they should still feel equally attractive when both are made conditional on the same outside event; a consistency requirement across contexts) and subproportionality (the idea that scaling a probability down by a fixed fraction distorts perceived likelihood more when the starting probability is already small than when it is large) among them – he derives a one-parameter weighting function $w(p) = \exp(-(-\ln p)^{\alpha})$ that satisfies the key qualitative features (inverted-S, fixed point near 1/e, asymmetry) of the empirically observed weighting function. The Prelec form is now ubiquitous in applied work for two reasons: it is parsimonious, with a single curvature parameter α, and it has clean limiting behaviour at $p = 0$ and $p = 1$. For our purposes, the Prelec function provides a quantitative benchmark: with α near 0.65, the modal estimate in the experimental literature, a 10% event is weighted as if its probability were closer to 20%. Read mechanically, this translates into a prediction that prices on 10% Polymarket events should trade roughly twice their true probability – a prediction that can be directly tested.

Taken together, these contributions deliver a sharp, joint prediction. Probability weighting is well-established as a feature of individual decision-making under risk; it has a tractable functional form in Prelec (1998); and Barberis and Huang (2008) show that its implications extend from psychology experiments to financial asset pricing, where securities with lottery-like payoffs are over-priced in equilibrium. The joint prediction for a market in binary contingent claims – that low-probability tokens are systematically over-priced relative to realised win rates – is unambiguous. This is the null hypothesis we test. We are not the first to take this prediction to a betting venue; the relevant precedent is the favourite-longshot bias literature, to which we turn next.

2.2 The favourite-longshot bias in betting markets

The favourite-longshot bias has been documented in betting markets for the better part of a century. Thaler and Ziemba (1988) provide what remains a useful early synthesis. Surveying the parimutuel literature they catalogue the bias as a robust feature of racetrack betting across geographies and time periods. The paper discusses two competing explanations – risk-loving preferences (the Weitzman interpretation) versus probability misperception (the Kahneman-Tversky reading) – and finds the evidence more consistent with the latter, particularly given that the bias persists across lotteries and exotic bet structures where a pure risk-preference story is harder to sustain. Its broader contribution is an empirical taxonomy that distinguishes bookmaker markets, parimutuel pools, lotteries, and exotic-bet structures, establishing that the bias is not a peculiarity of any single venue.

The canonical empirical paper for our purposes is Snowberg and Wolfers (2010). Using a large sample of US horse-racing parimutuel races, they document a pronounced classical longshot bias: longshots are systematically over-bet and favourites systematically under-bet at the implied-probability frontiers of the

market. The methodological contribution of the paper, and the reason it is the natural anchor for our work, is that the authors design their empirical strategy explicitly to distinguish two competing explanations: a "neoclassical" risk-loving story in which bettors derive positive utility from variance, and a "behavioural" story in which bettors systematically misperceive probabilities along the Kahneman-Tversky lines. Their preferred reading favours the misperception explanation, and they argue that the magnitude and shape of the bias are quantitatively closer to what prospect-theoretic weighting predicts than to what plausible parametrisations of risk-loving preferences would imply.

Cain, Law, and Peel (2000) provide complementary evidence from UK fixed-odds bookmaker markets. The bias they document is qualitatively the same direction as on US parimutuel: longshots are over-priced, favourites are under-priced. The fact that the bias survives a change of microstructure - from a parimutuel pool, in which the price clears once at race time, to a fixed-odds bookmaker, in which a profit-maximising market-maker posts prices - is one of the empirical robustness checks that the broader betting-market literature has accumulated.

The theoretical framework explaining when the favourite-longshot bias should and should not appear under different market microstructures is developed in Ottaviani and Sørensen (2008, 2010). Their work demonstrates that the bias can arise endogenously under quite different model assumptions: in their parimutuel model, the bias emerges as an equilibrium feature of betting against a non-informative pool. Critically - and this is the point that we draw on in §3.3 - the theoretical conditions under which the bias appears most cleanly are those of a single clearing pool with no in-event repricing and limited informed-trader entry. Where these conditions are relaxed – as on a continuous orderbook with active in-game trading and competition from sharp participants – the framework predicts the bias to be much weaker or absent.

The picture that emerges from this strand is unambiguous on the descriptive side: the favourite-longshot bias is well-replicated on parimutuel venues and on traditional bookmaker odds boards, the qualitative direction is identical across both, and the bias has been interpreted - most prominently by Snowberg and Wolfers - as a real-money confirmation of the prospect-theoretic prediction laid out in §2.1. The conditions under which the bias appears, however, are not incidental. Single-pool clearing, the absence of in-event repricing, and retail-heavy participation are part of the explanation, not a side note. The interest in Polymarket sports as a test venue is precisely that none of these conditions hold there. Polymarket sports operate on a continuous double-sided orderbook with in-event repricing on every score change, with a mixed participant base in which sharp informed traders are visibly active, and with capital that flows freely across markets and across sports. Whether the classical bias survives transplantation into this microstructure is the empirical question this thesis addresses.

2.3 Prediction-market efficiency

The prediction-market literature has, since the mid-2000s, taken something of an opposite tack from the favourite-longshot literature. Where the latter is, in essence, a catalogue of well-documented inefficiencies, the prediction-market literature is largely concerned with explaining why these markets work as well as they do - why their prices are, in many settings, surprisingly good probabilistic forecasts of the underlying events.

Manski (2006) takes a more theoretically pointed view. He shows that interpreting a prediction-market price as a probability is not automatic – it requires assumptions about how participants assess risk and how their beliefs are distributed across the population. Under specific conditions, the price equals the true probability; outside them, the price reflects a weighted average of participant beliefs that need not equal the average belief. This matters for our work because it formalises the gap between observed price and true probability that the longshot literature treats as empirical: even in the absence of biases, the price-as-probability interpretation is conditional, not automatic.

Berg, Forsythe, Nelson, and Rietz (2008) is the long-run empirical companion to the above. The Iowa Electronic Markets, run by the University of Iowa from the late 1980s onwards, are the most-studied prediction-market venue in the literature. Across more than a decade of election cycles, Berg et al. document that IEM prices consistently track realised election outcomes well, generally beating polls in terms of forecasting accuracy and exhibiting good calibration. The IEM is in many ways a clean laboratory: small dollar limits, a tightly defined event space, and an academic operator. Its findings establish a benchmark for what well-calibrated prediction markets can deliver. The contrast with retail-heavy parimutuel betting markets is informative, and is part of the reason the prediction-market literature has tended to draw more optimistic conclusions about market efficiency than the racetrack-betting literature has done.

Tetlock (2008) introduces the dimension of liquidity. Working with TradeSports and similar venues, he documents that calibration improves systematically with market liquidity: markets in which dollar volume is high, in which a critical mass of participants is active, and in which the order book is thick exhibit smaller predictable biases than thinly traded markets. This is intuitive and consistent with information-aggregation theory.

Page and Clemen (2013) sharpen the temporal dimension. They examine how the calibration of prediction-market prices evolves as time to resolution approaches, finding that prices become better calibrated as the resolution date nears and as information accumulates. The implication for our setting is that any test of calibration ought to be conditional on the time-to-resolution slice in which the observations sit - a point we revisit when we describe our intra-market tick-level calibration construction.

Taken together, this strand establishes that prediction markets are capable, but conditionally so. Aggregate price accuracy is well-documented; what is less well-documented is whether that accuracy holds uniformly across differences in probabilities, liquidities, market types, and time horizons – or whether it is the product of offsetting deviations that cancel in aggregate. The contributions that follow speak directly to that question.

2.4 Reichenbach and Walther (2025) - the most direct precedent

Reichenbach and Walther (2025) is the single most direct empirical precedent for our work. Their paper analyses on the order of 124 million trades on Polymarket. Their headline finding, and the one we engage with most directly, is that they find no aggregate longshot bias on Polymarket prices. In their main calibration analysis, the plot of implied probabilities against realised resolution rates does not exhibit the systematic deviation pattern that the classical betting-markets literature would predict. This is a striking result. On its face, it implies that Polymarket - or at least the aggregate cross-section of Polymarket trades - does not replicate the well-documented favourite-longshot bias of parimutuel and bookmaker venues.

The R&W result is, however, an aggregate result. Their analysis is performed on the pooled cross-section of trades; while they do decompose by timing decile within the market lifecycle, they do not decompose by event category (sports versus politics versus crypto versus macro), or by implied-probability band conditional on event type or regime. They also frame their result as a calibration finding rather than as a tradeability finding: their methodological apparatus is appropriate for testing whether the price tracks the truth, but it is not designed to quantify intra-market mispricing or to evaluate whether deviations are large enough to support a tradeable strategy net of microstructure costs.

3. Theoretical Framework and Hypothesis Definition

3.1 Probability weighting and the longshot prediction

As established in §2.1, the classical longshot prediction descends from cumulative prospect theory and its axiomatic formalisation in Prelec (1998). We do not re-derive the theory here; we state its properties in the form required for the empirical tests that follow.

The Prelec weighting function $w(p) = \exp(-(-\ln p)^{\alpha})$ satisfies $w(0) = 0$ and $w(1) = 1$ at its endpoints and departs systematically from the identity in between. For small probabilities the function is concave, so that $w(p) > p$; for large probabilities it is convex, so that $w(p) < p$. The two regions meet at a single interior fixed point, typically located between 0.30 and 0.40 in empirical estimates. At $\alpha \approx 0.65$ – the modal estimate

in the experimental literature – a true probability of 0.10 maps to a decision weight of approximately 0.20, and a true probability of 0.90 maps to approximately 0.80.

These properties translate directly into a price prediction on binary prediction markets. A YES token at price p on a market with true resolution probability q is fairly priced if and only if p = q. If participants price tokens using their distorted perception of probability rather than the true probability, the price will be too high when q is small and too low when q is large. The resulting mispricing follows the inverse-S shape of the weighting function: prices are too high for longshots, too low for heavy favourites, with a crossover in between. Barberis and Huang (2008) confirm that this logic extends to asset-pricing equilibrium: securities with lottery-like payoff distributions are over-priced even when they trade alongside conventional assets, which applies directly to low-priced YES tokens.

The classical prediction, restated in testable form: across low implied-probability bands, market prices should systematically exceed true resolution rates; across high implied-probability bands, they should fall short. The mispricing should be non-monotone in implied probability, change sign once, and produce the signature inverse-S pattern in calibration space. This is the null hypothesis we take to the data.

3.2 Why Polymarket sports may differ

The theoretical conditions under which cumulative prospect theory generates a classical longshot bias – retail dominance, single-pool clearing, no in-event repricing, and limited arbitrage capacity – are absent from Polymarket sports. The Ottaviani-Sørensen framework implies that relaxing these conditions does not merely weaken the classical bias but that it can reverse it.

In a continuous orderbook with in-game repricing, casual participants overreact to events during the match. Sharp participants observe both the event and the reaction, and profit by fading overshoots toward the conditional fair price – a mechanism documented empirically on Betfair's in-play soccer markets by Croxson and Reade (2014), whose orderbook structure closely resembles Polymarket's. If fading is sufficiently fast and well-capitalised, the time-averaged traded price of a heavy-favourite token can sit above its true resolution probability, and the time-averaged price of a longshot token below. This inverts the classical pattern.

We do not claim that probability weighting has disappeared from participant preferences. We claim the narrower point that the equilibrium prices in this venue need not reproduce the classical prediction, because the assumptions under which that prediction follows from those preferences are violated. The direction and magnitude of any deviation is therefore an empirical question.

3.3 Formal hypotheses

We define the empirical object of interest as follows. For an implied-probability band $B \subseteq [0, 1]$, let $\pi(B)$ denote the realised win rate conditional on a market's implied probability falling within B at some point during its trading life.

Let $m_B = (\inf B + \sup B)/2$ denote the midpoint of B. The signed deviation between realised resolution rate and band midpoint is $d(B) = \pi(B) - m_B$. We use the band midpoint rather than the trade-level implied probability as the comparator because our calibration object is defined at the band level; this is a modelling choice that trades granularity for robustness to within-band price variation, and we discuss its sensitivity in §5. We state three competing hypotheses against this object.

H0 - Classical longshot bias. For low bands, $d(B) < 0$; specifically, $d(B) < 0$ for $B \subseteq [0.10, 0.30]$. For high bands, $d(B) > 0$; specifically, $d(B) > 0$ for $B \subseteq [0.70, 0.90]$. The signed deviation is non-monotone in implied probability, negative at the low end, positive at the high end. This is consistent with the prediction of Tversky and Kahneman (1992), Prelec (1998), and Barberis and Huang (2008), reproduced on horse-racing data by Snowberg and Wolfers (2010).

H1 - Inverse longshot bias. The pre-registered alternative. For low bands $d(B) > 0$; for high bands $d(B) < 0$. The signed deviation is non-monotone in the opposite direction to H0. Low-probability tokens are under-priced, heavy favourites are over-priced, and the aggregate signed deviation may be close to zero because the two halves cancel.

H2 - Efficient market null. $d(B) \approx 0$ for all B, with bootstrap confidence intervals covering zero in every band. This reproduces the aggregate null of Reichenbach and Walther (2025) at the band-conditional level and is the conclusion the calibration literature on the Iowa Electronic Markets (Wolfers and Zitzewitz 2004; Berg, Forsythe, Nelson and Rietz 2008) most often supports.

The decision rule we apply is the same for all three. Bootstrap confidence intervals for d(B) must exclude zero, with Bonferroni correction across the $4 \times N$ cells defined by four implied-probability bands and N sports. The pattern must replicate in at least four of seven sports. The direction must be preserved, qualitatively, across both regime subsamples defined by the structural break in February 2026.

The hypotheses are stated in purely descriptive terms. They concern the alignment between price and realised resolution rate; they do not assert anything about whether the deviation is exploitable as a trading strategy. The translation from a calibration anomaly to a tradeable expression - and the additional conditions (execution costs, fill rates, depth, walk-forward validation) that such a translation requires - is reserved for the Discussion (§7) and Future Work (§8.4).

4. Data

4.1 The Polymarket venue

We test the longshot-bias hypothesis on Polymarket, a decentralised prediction-market platform. Each event question is resolved as a binary market with two complementary outcome tokens, conventionally labelled YES and NO. For the markets we are looking at the tokens represent one of the teams, meaning the market question is not “Will team X win?” but “Will team X or will team Y win?” (though soccer markets use the former and thus have 3 separate markets for one game. The two tokens are minted as a matched pair against one USDC of collateral, and their orderbook prices sum to approximately one, so a price of p on the YES token is directly interpretable as a market-implied probability that the underlying event resolves in favour of YES. Trading takes place through a continuous central limit-order book, with a spread of typically one or two cents in the liquid sports markets which we are researching.

Resolution is deterministic and on-chain. For every market in our sample, settlement is administered through UMA's Optimistic Oracle: the market creator posts a resolution proposal, a challenge period elapses, and absent a dispute the market settles to one USDC for the winning token and zero for the losing token. The terminal price observation we use in §4.4 to assign outcome labels is therefore the direct product of this resolution mechanism.

Polymarket is well-suited to a calibration test for four reasons. First, the venue uses real money: positions are USDC stakes that cannot be reset, preserving the incentives that a behavioural-finance test of probability weighting requires. Second, liquidity in headline sports categories is genuine, with order-book depth and tick frequencies that allow microstructure analysis. Third, price discovery is continuous across the full pre-event and in-play window, so we observe an entire probability path rather than only a market-open snapshot. Fourth, the participant base is heterogeneous, blending retail bettors with funded systematic traders which is precisely the population that the Tversky-Kahneman framework assumes when it predicts that systematic over-weighting of small probabilities can survive in equilibrium.

4.2 Sample construction

Our empirical universe consists of last-traded-price (LTP) paths from 5,102 sports betting markets - moneyline contracts in NBA, NHL, NCAA Basketball, and the four major European football leagues (EPL, LaLiga, Bundesliga, SerieA), together with the soccer draw legs that share a fixture with those moneylines. Both panels are scraped directly from the historical Polymarket blockchain. We record the timestamp and the price of both sides of the binary outcome whenever a trade occurs on either token. Assuming that deviations from both tokens summing to 1 are minimal, inconsistent, and short-lived, we work throughout with just the first-outcome token price; any market where the mean sum of the two token prices is ≤ 0.98 is dropped from the sample.

The data is split into two panels separated by the February 2026 microstructure regime change documented in §4.3. The legacy panel covers August 2025 through early February 2026 and contributes 2,850 markets to the universe; the recent panel covers late January 2026 through April 2026 and contributes 2,252 markets. Markets with fewer than two valid ticks and markets whose final YES price falls strictly between 0.01 and 0.99 (typically cancelled or postponed matches) are excluded; the resulting working sample is 4,363 cleanly-resolved markets, 85.5% of the universe. Per-category counts are tabulated in §4.6.

4.3 The February 2026 regime break

While analysing the data, we found that Polymarket sports markets underwent a structural change in microstructure around the start of February 2026. The shift shows up cleanly and consistently across all seven sports of the thesis universe: per-market tick counts and direction-reversal counts both rise sharply, while the realised range of prices visited is essentially unchanged. Table 4.1 (see appendix) reports the per-sport medians of two summary statistics - lifetime tick count and direction-reversal count on a one-cent grid - in the legacy and recent panels.

To test the shift formally we apply Mann-Whitney U and Kolmogorov-Smirnov two-sample tests to seven path-feature distributions on the pooled thesis universe (2,233 legacy resolved markets, 2,130 recent resolved markets). Five of the seven features - tick count, duration, tick density, reversal count, and per-tick volatility - reject distributional invariance at the Bonferroni-corrected threshold across the 49 sport-feature cells ($\alpha\_bonf \approx 0.001$) by both tests. Only the two outcome-shape features, the lifetime price range and the absolute open-to-close move, are stable across the regime. The pattern is therefore one of denser, choppier intra-market activity with unchanged realised outcomes - exactly what one expects from a venue-level participation change rather than a scraping pipeline change, which would produce uniform per-sport multipliers instead of the category-dependent multipliers in Table 4.1.

We treat the break as a real structural feature of the venue. In §6 we report calibration results both pooled and split by regime so that any apparent inverse-longshot effect can be assessed against the possibility that the two halves of the sample reflect mechanically different market environments.

4.4 Resolution rule and cleaning

We assign a binary outcome to each market on the basis of its final tick: if the final YES price $p\_T \geq 0.99$ the market is labelled outcome = 1; if $p\_T \leq 0.01$ it is labelled outcome = 0; otherwise the market is excluded as ambiguous. The rule is deliberately strict - UMA-resolved tokens converge to exactly one or zero on settlement, so prices materially inside the unit interval at the final observed tick signal either incomplete scraping, an open dispute, or a market that has not yet resolved at the time of capture.

Of the 5,102 markets in the working sample, 4,363 resolve cleanly (85.5%) and 739 (14.5%) are excluded as ambiguous. Resolution rates are highest in NBA and NCAA Basketball (above 95%) and lowest in the soccer panels (around 75%-85%), reflecting the fact that the soccer draw legs are more often partially captured in our scrapes than the home and away legs they share a market with. For the soccer three-leg files we retain a derived match_id so that, in robustness checks, observations from the same underlying fixture can be clustered. The headline calibration treats the three legs as independent observations; we discuss this choice as a known caveat in §4.7.

4.5 Per-market summary representation

Every analysis in the thesis is computed from the raw tick streams, not from a downstream aggregate. In a one-time preprocessing pass the loader reads each market's CSV file tick by tick and emits a per-market summary row that records exactly the quantities the downstream tests consume: the binary resolution outcome (from the final tick under the rule in §4.4), the canonical category and market-type labels, the regime label, the open, close, minimum, and maximum YES price, the total tick count and duration, the count of direction reversals on a one-cent grid, and the standard deviation of one-tick returns.

For the per-cent calibration test in §5 the summary additionally stores, for each of 100 per-cent buckets of width 0.01 spanning [0, 1], a single boolean indicating whether the YES price ever entered that bucket. This boolean is the exact input the touch-once test requires: the test asks, for each bucket B, what fraction of markets that visited B resolved $1, and the boolean records the visit directly. The summary is therefore lossless for our headline test: no information used by the calibration analysis is discarded between the tick stream and the summary row.

The advantage of the summary representation is purely computational. Reading 5,102 CSV files tick by tick is the slow step; the test itself, once the summary is built, runs in seconds. Summaries are computed once per panel and cached as Parquet under notebooks/outputs/; downstream notebooks load the parquet rather than re-scanning the underlying files. Where a follow-up question would require information the summary does not preserve - for example, the exact dwell time of the price inside a bucket, or the side of each trade - the analysis re-reads the raw tick streams. None of the headline analyses in this thesis require that.

4.6 Categories and market types

Composition of the working sample by category and panel is summarised in Table 4.2 (see appendix).

The category mix differs across panels in two ways that matter for the calibration analysis. First, NBA is the smallest panel in the sample on the legacy side, contributing 239 markets, compared with several hundred for every other sport; the recent panel partially compensates with 362 NBA markets observed at much higher per-market tick density (§4.3). Second, NCAA Basketball is the only category where the recent panel matches or exceeds the legacy panel in count, a consequence of the recent window straddling the NCAA tournament.

Soccer requires explicit treatment because association football is a three-outcome event (home / draw / away). On Polymarket each outcome is exposed as its own binary YES/NO token, of which one settles at $1. The draw leg in particular has a markedly different calibration profile from the home and away legs, since a draw is rarely the modal pre-match implied outcome but resolves true at a positive base rate. We carry a market_type = draw_leg label through the summary so that the draw legs can be inspected separately in robustness analysis.

4.7 Limitations of the data

Several features of the data constrain what can be tested. First, NBA contributes a smaller legacy panel (239 markets) than the other six sports in the sample. Pooled NBA results combine the legacy and recent NBA panels to reach n = 601, comparable to Soccer-SerieA at n = 486; per-panel NBA statements therefore carry larger sampling error than the same statements on the NCAA Basketball and soccer panels and are reported alongside their Wilson intervals.

Secondly, our sample is sports-only. The thesis claim is therefore explicitly conditional on the sports category, and the headline finding cannot be generalised mechanically to non-sports markets. We return to this point in §8.

Thirdly, the data are price-only. We observe inside best bid and best ask through the First Outcome Price and Second Outcome Price fields, but we do not observe the side of each trade and cannot attribute volume to one side of the book. This precludes any test that relies on trade-direction, including the disposition effect and order-flow imbalance literatures; our calibration test does not require trade-direction data, but related mechanisms are out of scope as a consequence.

Lastly, we lack reliable volume and depth metadata at the tick level. We collected aggregate market volume but were not able to collect the time series of order-book depth for historical markets, so volume-conditional calibration is left as a robustness extension.

These last two limitations prevent us from testing whether our findings are tradable in practice. A mispricing identified in the historical data may not be exploitable in live markets, due to, for example, slippage caused by a thin order book.

5. Methodology

5.1 Overview and design choices

The empirical design is built around a single research question and a single descriptive pillar: does the classical favourite-longshot bias predicted by cumulative prospect theory survive on Polymarket sports prediction markets, and if it does not, what direction the deviation takes. The method is a hold-to-maturity calibration test that compares the implied probability of a token at each point in its life with the realised resolution outcome. The descriptive pillar averages per-market realised outcomes within each implied-probability bucket and reports the signed deviation from the bucket midpoint.

The translation of the calibration anomaly into a tradeable strategy - i.e. whether the band-conditional mispricing identified here is exploitable under realistic execution constraints - is treated as a downstream question and is reserved for the Discussion (§7) and Future Work (§8.4). The present methodology section therefore does not specify entry rules, exit rules, position sizing, or backtest procedures; the calibration test and the trading test are conceptually distinct objects, and conflating them risks importing in-sample selection and overfitting concerns into what is at its core a descriptive statistical question about realised frequencies.

We deliberately keep the calibration design simple. Each market contributes one observation per per-cent bucket of price it visited during its life. We report Wilson 95% confidence intervals and a by-market bootstrap as a robustness check, with a Bonferroni multiplicity correction across buckets. The same procedure is then run separately on the pre- and post-February-2026 panels to test whether any observed deviation is regime-conditional.

5.2 Probability bands and calibration definition

We partition the unit interval into 100 buckets of width 0.01. For each bucket B = [b/100, (b+1)/100) we record the set $\mathcal{M}(B)$ of markets whose YES price entered B at any tick during the market's lifetime, the bucket's implied probability m_B = b/100 (its lower edge), and the realised resolution rate

$$\pi(B) = (1 / |\mathcal{M}(B)|) \Sigma_{i \in \mathcal{M}(B)} y_i,$$

where $y_i \in \{0, 1\}$ is the binary resolution outcome of market i under the rule in §4.4. The signed deviation at bucket B is $d(B) = \pi(B) - m_B$.

We use the bucket's lower edge as the implied probability rather than its midpoint because Polymarket prices live predominantly on the 1¢ grid: roughly 89% of the ticks in our sample sit at exactly b/100 for some integer b, with the remainder being sub-cent prices clustered at the extremes of the unit interval. The modal touched price inside a bucket [b/100, (b+1)/100) is therefore b/100 itself, and the lower edge is the minimum-bias representative of the within-bucket price distribution. Using the midpoint b/100 + 0.005 would impart a systematic +0.005 bias on the implied side and a corresponding −0.005 bias on every d(B).

For each per-cent bucket B we compute π(B) and d(B) as defined in. Each market contributes at most one observation per bucket regardless of how long the YES price spent there, which controls within-market autocorrelation at the cost of one source of information (time spent at a given price). The interpretation is operational: of all markets that ever traded at price p within ± 0.005, what fraction resolved $1?

Under the null of classical longshot bias H0, we expect $d(B) < 0$ at low B and $d(B) > 0$ at high B. Under the inverse hypothesis H1, we expect $d(B) > 0$ at low B and $d(B) < 0$ at high B. Under H2 (efficient calibration extending Reichenbach and Walther, 2025), we expect $d(B) \approx 0$ uniformly in B after multiple-testing correction. The thesis is structured as a test of H1 against the joint alternative of H0 and H2.

5.3 Statistical inference

We use two confidence-interval procedures and a multiplicity correction. For each bucket B we report the Wilson score 95% confidence interval for π(B), treating per-market outcomes as independent Bernoulli trials. For k successes in n trials at $z = 1.96$,

$$CI_{95}(\pi) = ((k/n + z^2/2n) \pm z \sqrt{(k(n - k)/n + z^2/4)}) / (1 + z^2/n).$$

Since per-market observations are exchangeable within each bucket and each market contributes at most once, the binomial assumption is exact and the Wilson interval is well-specified. As a robustness check we additionally report a by-market bootstrap 95% confidence interval (10,000 resamples) on the touch-once buckets, which preserves any within-market correlation structure via the resampling unit.

For multiplicity correction we restrict attention to buckets with $n \geq 30$ markets touched, which in the pooled filtered universe is $K = 100$ buckets. The Bonferroni-corrected per-bucket threshold is $\alpha_{bonf} = 0.05 / K = 0.0005$, corresponding to a per-cell two-sided z critical of about 3.48. We also report Benjamini-Hochberg false discovery rate at $q = 0.05$ as a less conservative comparator that is more powerful when several buckets

deviate in the same direction (as expected under H1). Consistent with the step-down prescriptions of Romano and Wolf (2005), we treat the strict Bonferroni cutoff as the binding threshold for headline cell claims.

The decision criterion for the central H1 claim is as follows. H1 is supported if (i) at least one low bucket exhibits $d(B) > 0$ and at least one high bucket exhibits $d(B) < 0$, both with Bonferroni-corrected $p < 0.05$ in the pooled sample; (ii) the same directional pattern holds in at least four of the seven sport partitions; and (iii) the pattern is preserved qualitatively in both the pre- and post-February-2026 subsamples. Failure of (i) collapses the claim to H2; failure of (ii) narrows the claim to the sports in which it holds; failure of (iii) re-casts the result as regime-conditional.

5.4 Sanity checks and robustness

Three diagnostic checks accompany the calibration tables.

YES/NO symmetry. By construction, the realised win-rate of the YES token at price p equals one minus the realised win-rate of the complementary NO token at price $1 - p$, because $y^{NO} = 1 - y^{YES}$ and the two tokens always cohabit a market. Any observed departure from this symmetry indicates a defect in outcome detection or in the price-to-bucket assignment. The check passes by construction in our framework but is reported as a sanity figure to confirm the implementation is correct.

Initial versus final calibration. The opening-tick calibration and the closing-tick calibration are reported side by side. By construction the closing-tick calibration should be near-perfect: at resolution, the YES price is approximately equal to the binary outcome, so $\pi\_close(B) \approx m\_B$ at the two extreme buckets and undefined elsewhere. Any material divergence at the extremes is a quality-control flag indicating ambiguous resolutions or labelling errors.

Maximum-price-reached. For each market we record the lifetime maximum of the YES price and report the realised win-rate per maximum-price bucket. This diagnoses whether markets that ran up to a high price but failed to resolve $1 contribute disproportionately to the favourite-band over-pricing, distinguishing a "fooled-by-spikes" mechanism from steady-state mispricing.

Empirical results are reported in §6. The behavioural mechanism reconciling the findings with cumulative prospect theory, and a sketch of the buy-only operational expression of the inversion, are discussed in §7.

6. Results

This section reports the empirical findings of the hold-to-maturity calibration test described in §5. We begin with the headline per-integer touch-once calibration on the pooled working sample (§6.1), then examine the shape of the deviation across the full probability universe at per-cent resolution (§6.2) and decompose it sport by sport (§6.3). We next condition on the February 2026 microstructure break and find it to be efficiency-improving rather than bias-strengthening (§6.4). The diagnostic and sanity checks specified in §5.4 are carried through each of these results, and §6.5 draws the evidence together against the formal hypotheses H0, H1, and H2.

6.1 Headline finding - inversion of the classical longshot prediction

Pillar 1 asks one descriptive question: across the 4,363 cleanly-resolved markets of the working sample, when a market's first outcome token (FOT) traded at an implied price p, how often did that token go on to resolve $1? To test this, we choose unit intervals which are partitioned into 100 per-cent buckets of width 0.01. A market contributes one observation to bucket B if its FOT price entered B at any point in its life, and $\pi(B)$ is the fraction of those markets that resolved $1. Figure 6.1 (see appendix) plots $\pi(B)$ against the bucket's implied probability, its lower edge, per the convention of §5.2, for all 100 buckets.

The picture is unambiguous, and it is the central empirical result of the thesis. The realised-rate curve lies above the 45° identity line across essentially the entire price range below 0.90: tokens trading at low and middling implied prices resolve $1 more often than their price implies. The curve meets the diagonal only in the heavy-favourite region above roughly 0.92. This is the opposite of what the classical favourite longshot bias of Tversky and Kahneman predicts, under which longshots are over-bet and should resolve $1 less often than their price implies i.e. realised points below the diagonal at low p.

Table 6.1 condenses the 100 buckets into ten price deciles. A decile cell is the population-weighted mean of the ten per-cent signed deviations $d(B) = \pi(B) - \text{implied}(B)$ that the decile spans. Each decile figure is therefore an average of integer-level mispricings, computed bucket by bucket and then aggregated, not a mispricing measured on a coarse price.

The magnitudes are modest and orderly. Every decile from [0.00, 0.10) through [0.80, 0.90) carries a positive mean edge, rising from +0.017 at the lowest decile to a maximum of +0.033 at [0.40, 0.50) and easing back to +0.010 by [0.80, 0.90); only the heavy-favourite decile [0.90, 1.00) is marginally negative, at −0.002. Ten of the 100 per-cent buckets reject a zero edge after Bonferroni correction, every one of them at a low or middling price. The population-weighted mean edge across the whole spectrum is +0.020.

6.2 The shape of the inversion across the probability universe

The decile table smooths over structure that the per-integer grid was built to expose. Figure 6.2 (see appendix) plots the signed deviation d(B) directly against implied price for all 100 buckets, with the per-bucket Wilson interval and a data-driven segmentation of the curve into four phases.

The four phases are economically distinct. In the long-shot ramp, implied prices from 0.00 to 0.26, the deviation is positive but still climbing, averaging +0.022: under-pricing is already present at the very bottom of the price ladder but has not yet reached full strength. The mid-range plateau, 0.26 to 0.48, is where the inversion is largest: the deviation averages +0.032 and reaches its per-integer maximum of +0.042 at an implied price of 0.45. The favourite decline, 0.48 to 0.94, is a long, gradual fade in which the deviation trends monotonically downward from roughly +0.032 through zero, averaging +0.014 over its considerable width. The heavy-favourite tail, 0.94 to 0.99, is the only phase in which the deviation is negative, averaging −0.004, a slight over-pricing of near-certain favourites that, as the next paragraph shows, is not statistically separable from fair.

Three features of the curve carry interpretive weight. First, the inversion is not a longshot-corner phenomenon. The classical favourite-longshot literature locates its bias in the extreme tails of the price distribution; here the largest mispricing sits squarely in the middle of the range, around a coin-flip, and the under-pricing extends continuously from the lowest bucket all the way to roughly 0.90. Second, the crossing into negative territory is gradual rather than sharp: the curve drifts down through the entire favourite decline and only reaches zero near 0.94, so any partition of the price axis into discrete "longshot" and "favourite" bands is an analytical convenience, not a feature of the data. Third, the statistical weight is asymmetric. All ten Bonferroni-significant buckets lie at or below an implied price of 0.46 - four in [0.01, 0.05), three in [0.14, 0.19), and three between [0.28, 0.46) while no heavy-favourite bucket reaches significance. To the resolution our sample affords, the favourite side of the market is efficiently priced; the inverse-longshot signal is carried entirely by the longshot and mid-range buckets.

This shape constrains interpretation. A market that systematically under-prices every contract below a coin-flip while pricing favourites correctly is not committing the symmetric probability-weighting error that cumulative prospect theory predicts, in which the over- and under-weighting of tail probabilities mirror each other. It looks instead like a one-sided reluctance to pay up for unlikely outcomes: bettors discount longshot and middling contracts relative to their realised frequency, but do not correspondingly overpay for favourites. Section 6.3 shows that this one-sided shape is shared by most of the sports in the universe, and §6.4 shows that it is fading over time.

6.3 Cross-sport heterogeneity

The pooled curve aggregates seven sports with materially different calibration profiles. Figure 6.3 (see appendix) plots the per-integer calibration curve separately for each sport; Table 6.2 (see appendix) summarises each sport as the population-weighted mean per-integer edge in four price regions - longshot [0.00, 0.40), coin-flip [0.40, 0.60), slight-favourite [0.60, 0.85), and heavy-favourite [0.85, 1.00). As in §6.1, every cell is an average of per-integer deviations, not a re-pooled coarse estimate.

The four European football leagues carry the cleanest and strongest inverse-longshot signal in the data. Every one of the sixteen soccer cells in Table 6.2 is positive: soccer YES tokens resolve $1 more often than their price implies at every price level, the heavy-favourite region included, where the pooled curve is already flat. The effect is largest in the longshot and coin-flip regions, where the typical soccer league shows an edge near +0.055, and it remains positive but smaller, roughly +0.01 to +0.02, among heavy favourites. LaLiga is the most strongly mispriced league, peaking at +0.076 in the coin-flip region; SerieA carries the largest longshot edge, +0.064. That the signal survives into the favourite region for soccer but not for the pooled sample already shows that the favourite-side efficiency of the pooled curve is a North-American-sports property, not a universal one.

A structural feature of soccer and its Polymarket markets offers a plausible explanation for why its deviation is both larger and more uniform than any North American sport's. An NBA, NHL, or NCAA Basketball moneyline is a two-outcome event whose two tokens are structurally symmetric; each is simply one team winning, and the pair partitions the entire outcome space. A soccer match, by contrast, has three outcomes: home win, draw, away win. The soccer contracts in our sample are binary markets written on a single one of those outcomes, so the YES token captures just one slice of a three-way event while the NO token absorbs both the draw and the opposing side. We conjecture that bettors are systematically reluctant to pay full value for a contract framed as a specific outright win that must also overcome the standing possibility of a draw: the YES side reads as the narrower, less likely call and is persistently under-bid relative to its realised frequency, while the compound "draw-or-opponent" NO side is correspondingly over-bid. The draw, absent from every North-American sport in the sample, is thus the natural candidate for the one thing soccer shows that the others do not: a positive edge that holds even among heavy favourites for the “Yes” token.

NCAA Basketball and NHL reproduce the inverse pattern through the longshot and the middle of the price range, but not at the top. NCAA Basketball shows the largest non-soccer cell anywhere in the table, +0.040 in the coin-flip region, alongside a +0.019 slight-favourite edge, before turning mildly negative among heavy favourites at −0.013. NHL is positive but weaker through the longshot and coin-flip regions, at +0.025 and +0.003, and essentially flat thereafter. Both sports, in other words, support H1 precisely where

the pooled signal is strong and fade to fair or marginally over-priced exactly where the pooled signal also fades.

NBA is the single exception in the universe. All four NBA regions are negative, between −0.002 and −0.025, with no positive cell anywhere. This suggests NBA moneyline tokens are, if anything, mildly over-priced across the whole spectrum.

Two readings are admissible. NBA may genuinely be the best-calibrated sport on the venue; alternatively, with n = 601 the smallest sport total in the sample, NBA estimate may simply be too noisy to resolve a small positive edge. A second mechanism, complementary to the contract-structure effect above, is grounded in the sports themselves. Soccer and ice hockey are low-scoring games: goals are rare, one to three per side, and each one moves the win probability in a large, sudden jump. Between goals the scoreline is frozen, but the true win probability is not, since a single reversing goal can arrive at any moment. A market that anchors on the current scoreline will tend to treat a lead as more decisive than it is, over-pricing the team ahead and under-pricing the team behind, and the trailing side is precisely the low-priced longshot token whose realised win rate the data place above its implied price.

Basketball is a different regime. An NBA game turns over a basket roughly every half-minute and finishes past two hundred combined points, so win probability evolves as a smooth, densely-sampled process the order book can track almost in real time. Leads are accumulated gradually and priced faithfully, leaving little room for the lead-anchoring error, consistent with NBA being the one sport in the sample with no longshot under-pricing. NCAA Basketball, the apparent counter-example, fits the same gradient once the college game is distinguished from the professional one. A men's college contest runs forty minutes to the NBA's forty-eight and a markedly lower combined score, so each basket carries more weight and the win probability advances in larger, less frequent steps. College basketball therefore sits partway between the NBA and the goal-sports on the scoring-density spectrum. A second, market-side factor reinforces it. College basketball betting markets draw lower volume and thinner participation than NBA markets, so whatever reversibility the scoreline does carry is priced less accurately than in the deep, heavily traded NBA markets.

We offer this as a plausible reading of the cross-sport gradient, not a tested claim: isolating it would require the in-game state data that the per-market summary does not retain. The broader conclusion is unaffected: the inverse-longshot direction holds in six of the seven sports in the longshot region, and in the favourite region is soccer-led rather than universal.

6.4 The February 2026 efficiency shift

The working sample straddles the February 2026 microstructure break documented in §4.3. Because that break coincides with a sharp rise in trading activity, the calibration result must be checked for regime-dependence: is the inverse-longshot pattern a stable property of the venue, or an artefact of one half of the sample? We recompute the per-integer calibration separately on the legacy panel (pre-break) and the recent panel (post-break). Figure 6.4 (see appendix) overlays the two edge curves and plots their per-bucket difference.

The two curves separate cleanly. The legacy edge curve sits well above the recent edge curve across almost the whole price range: the underpricing of longshot and mid-range tokens was larger before the break than after it. The right-hand panel makes the direction explicit - $\Delta d(B)$ is negative in the large majority of buckets, meaning the recent panel is the better calibrated of the two. Table 6.3 (see appendix) quantifies the shift per sport as the population-weighted mean Δd in the longshot region [0.00, 0.40) and the heavy-favourite region [0.85, 1.00).

The dominant pattern is convergence toward efficiency. Longshot under-pricing weakened in six of the seven sports, every sport except NBA, whose +0.006 longshot shift is statistically indistinguishable from zero at the regime sample size, and the pooled longshot Δd is -0.042, an attenuation of roughly two-thirds of the pooled longshot edge documented in §6.1.
The heavy-favourite region tells a milder version of the same story. NBA, NHL, and NCAA Basketball all show positive Δd, meaning the small favourite over-pricing of the legacy panel has moved toward fair, while the four soccer leagues retain, and slightly deepen, a modest favourite over-pricing of around -0.04. At the per-bucket level, only four of the 100 buckets show a Bonferroni-significant two-proportion shift across the break, and all four sit in the extreme longshot corner [0.00, 0.04), exactly where the legacy under-pricing was largest.

This explanation is offered as a correlation, the most plausible reading of a coincidence in timing, not as a causal claim the data can prove. We observe neither trader identities nor order provenance, so we cannot directly attribute the activity increase to bots, still less to AI-built bots; the growth of retail AI tooling is a venue-external trend whose link to Polymarket participation is inferential. What the data do establish is the conjunction: a structural increase in activity and an improvement in calibration occurring together, in the same quarter, across all seven sports.

6.5 Implications for the formal hypotheses

The thesis rests on three mutually exclusive hypotheses about the calibration curve, stated formally in §3.4. H0, the classical favourite-longshot bias, predicts $d(B) < 0$ for longshots and $d(B) > 0$ for favourites. H1,

the inverse-longshot hypothesis, predicts the opposite - $d(B) > 0$ for longshots and $d(B) \leq 0$ for favourites. H2, efficient calibration, predicts $d(B) \approx 0$ uniformly in B. Figure 6.5 (see appendix) places each hypothesis side by side with the observed per-integer edge curve.

H0 is rejected decisively. The classical prediction requires the deviation curve to lie in the negative half-plane for longshots; the observed curve is positive at every bucket below an implied price of 0.92, and the four lowest-price Bonferroni-significant buckets exclude a zero or negative edge by several standard errors. Not one longshot bucket behaves as H0 requires. The classical favourite-longshot bias does not survive on Polymarket sports moneylines in any price region.

H1 is supported in direction. The observed curve occupies the H1 region, positive deviation for longshots, near-zero or marginally negative for favourites, across its entire length, and the directional pattern holds in six of the seven sports of §6.3. Two qualifications attach. The magnitude is modest: a population-weighted mean edge of +0.020 and a per-integer peak of +0.042, an order of magnitude below the inverse-longshot effects reported for the unfiltered, multi-category drafts. And the favourite-side half of the H1 prediction, strictly negative deviations among heavy favourites, is met only weakly, by a tail that is small and not statistically separable from zero. H1 is therefore best stated precisely: robust longshot and mid-range under-pricing, with favourite over-pricing that is marginal at most.

H2 is rejected at the per-bucket level but is nearly recovered in aggregate, and the tension between those two statements is itself a result. Ten individual buckets reject a zero edge under Bonferroni correction, so the venue is not efficiently calibrated cell by cell. Yet the population-weighted mean edge is only +0.020, and a trader attempting to harvest the deviation meets it one bucket at a time, against execution costs and per-bucket sampling error. The market is inefficient in a statistically demonstrable but economically thin sense.

Taken together with the regime evidence of §6.4, the overall verdict is sharper than any single hypothesis label. The pooled sample rejects H0 outright and supports H1 in direction; the regime conditioning then shows H1 itself attenuating over time, the pooled longshot edge losing roughly two-thirds of its magnitude across the February 2026 break. The most accurate one-sentence characterisation is therefore not that "Polymarket sports markets exhibit an inverse-longshot bias", but rather that Polymarket sports moneyline markets displayed a modest, soccer-driven inverse-longshot bias that was progressively arbitraged toward efficiency in the period preceding the venue's early-2026 structural change, resulting in a finding that is descriptively consistent with H1, dynamically convergent with H2, and at no point reconcilable with H0.

7. Discussion

The findings presented in Section 6 yield a directional verdict that warrants careful interpretation. The classical longshot bias hypothesis (H0); H1 receives directional support across the aggregate sample and, with varying magnitude, across the majority of sport categories examined; and H2 is recovered only at the aggregate level, where it appears to emerge largely as an artefact of opposite-signed deviations offsetting one another across sub-samples rather than as evidence of genuine market-wide efficiency. The purpose of this section is not to reiterate these results but to examine what they imply. We consider, in turn, what the precise shape of the observed bias reveals about the underlying pricing mechanism, why the asymmetry between the two tokens appears to reflect structural rather than incidental features of the market, and how the cross-sport and cross-regime patterns refine the broader interpretation.

7.1 From Hypothesis to Finding: What the Data Actually Show

This thesis set out to test whether the classical longshot bias replicates on Polymarket sports prediction markets. The classical prediction, grounded in cumulative prospect theory and confirmed empirically on parimutuel and bookmaker venues, is that low-probability tokens should trade above their true resolution probability because retail participants systematically over-weight small probabilities in their decision-making. The data reject this. Across the pooled sample of 4,363 resolved markets, low-probability tokens resolve to one at a rate consistently above their implied price. They are underpriced, not over-priced.

This directional inversion is the headline result, but it is not, on its own, the most informative one. Two features of the shape of the deviation, visible in the per-cent calibration curve in Figure 6.2, carry more interpretive weight than the band-level average alone. The first is that the underpricing does not peak at the extreme low end of the probability range, as a symmetric inversion of the classical bias would predict. Instead it rises from near zero at very low implied probabilities, reaches its highest values in the mid-range roughly between 0.26 and 0.48, and persists with meaningful if diminishing magnitude well above the 50/50 midpoint. The second feature is that this pattern is not symmetric across the two tokens in each market. The first-outcome token is the one that is systematically underpriced. By construction, given that the two token prices sum to approximately one, the second-outcome token is correspondingly overpriced. This is not a market-wide calibration anomaly in which both sides are equally affected. It is an asymmetry between token positions within the same market.

These two features together define the interpretive challenge this discussion addresses. A market-wide, symmetric calibration anomaly might be explained by a uniform shift in participant preferences or by a data

artefact. An asymmetry between the first and second token, concentrated in the mid-range of the probability distribution, requires a structural explanation: something about what the first-outcome token represents that distinguishes it from the second. The remainder of this discussion develops that explanation, beginning with the most direct structural candidate.

7.2 Why the First-Outcome Token Is Underpriced

Across all North American sports markets in the sample, the first-outcome token represents the away team. This is a structural fact about how markets are labelled on the platform, and it is consequential for interpreting the calibration asymmetry. The away team's token is systematically underpriced throughout the lower and mid-range probability bands; the home team's token is correspondingly overpriced. A natural candidate explanation is that market participants overvalue home court or home field advantage when forming their pre-game probability assessments, inflating the price of the home team's token beyond what its true win probability warrants and leaving the away team's token trading below its true resolution rate.

We present this as a hypothesis rather than a confirmed finding. A direct test would require regressing the per-market signed deviation on a home advantage proxy for each sport, which we have not conducted in the current study. We identify the direct test as the most targeted follow-up this paper motivates and return to it in Section 8.4.

7.3 Why Soccer Amplifies the Signal

The cross-sport results in Section 6.3 show that the underpricing signal is strongest and most consistent across the four soccer leagues in the sample, while North American sports show weaker and less uniform patterns. This heterogeneity is not incidental. Two structural features of soccer markets on Polymarket, both absent from the North American sports in the sample, amplify the underpricing of the first-outcome token in the same direction as the home advantage mechanism described in 7.2.

The first structural feature is the compound nature of the NO token in soccer markets. Soccer has three possible match outcomes: home win, draw, and away win. On Polymarket, each outcome is listed as a separate binary market. A participant buying the NO token on a home win market is not placing a clean bet that the away team wins. They are buying a token that resolves to one if either the away team wins or the match ends in a draw. A participant who treats this as equivalent to an away win bet will systematically overpay for it, because they are acquiring draw probability without pricing it. The inflated NO token price directly compresses the YES token price of the home side below its true win probability, generating

underpricing of the first-outcome token by a mechanism entirely separate from home advantage overvaluation.

The second structural feature is the low-scoring nature of soccer. A single goal in a football match produces a substantially larger shift in conditional win probability than a single scoring event in basketball or hockey. A team that is 1-0 down at 70 minutes retains a genuine and non-trivial probability of drawing or winning, but the visible scoreline is stark and retail participants anchor on it. The market price of the trailing team's token tends to fall further than the true conditional probability warrants, because participants treat the deficit as more terminal than it is. This scoreline anchoring dynamic is a within-game amplifier of the underpricing: it is not a pre-game pricing error like home advantage overvaluation, but a live-market overreaction that depresses the first-outcome token during the event itself and creates the space for the calibration deviation we measure.

The NBA provides a useful structural contrast. NBA markets in the sample have the highest tick density of any sport, with median lifetime tick counts substantially above all other categories in both panels (Table 4.1). Continuous scoring means that individual scoring events produce small incremental shifts in win probability rather than large step changes, reducing the scope for scoreline anchoring. There is no compound token structure, since each market resolves as a clean binary win or loss. The NBA is the sport in the sample where the structural conditions for the bias are least favourable, and the signal behaves accordingly.

7.4 The Regime Trajectory: What the Efficiency Shift Tells Us

Section 6.4 documents that the underpricing signal attenuates meaningfully after the February 2026 structural break. The pooled signed deviation in the longshot band shrinks by approximately two thirds between the legacy and recent panels, and the pattern is directionally consistent across sports. This is a dynamic result, and it is interpretively significant independent of what caused the increase in market activity that accompanied the break.

An inefficiency that decays in the direction of zero as participation rises is consistent with one reading above all others: the mispricing was real, it was detectable by sufficiently attentive participants, and capital entered to exploit it. As more informed activity was directed at these markets, the gap between the away team's traded price and its true resolution probability narrowed. This is the standard Grossman-Stiglitz dynamic: the expected return to gathering information and trading on it draws in capital until the edge is competed down to the point where it no longer compensates for the effort. The regime trajectory is therefore not a complication for the main finding. It is supporting evidence for it, as a data artefact or a calibration

error won't move under increased participation, but a real pricing inefficiency will.

What drove the participation increase itself lies outside the scope of this thesis, and no causal claim is made here. It is nonetheless worth noting that the February 2026 break is contemporaneous with a period of rapid adoption of general-purpose AI coding assistants. The mechanism, if operative, would be participatory rather than informational: tools of this kind sharply lower the cost for a non-specialist to build and operate an automated trading bot, so the marginal new participant entering a sports prediction market in early 2026 may have been more likely to be running one than the marginal participant a few months earlier. A larger population of automated, price-sensitive traders is precisely the kind of force that arbitrages a static mispricing toward fair value, surfacing first as more frequent quoting and denser tick paths (see table 7.1 in the appendix), and then as tighter calibration. Whether that is what occurred here is a question the current data cannot answer.

7.5 The First-Outcome Token: An Unresolved Structural Question

Throughout this study, the calibration analysis is conducted on the first-outcome token only. This choice was made on the basis of the mechanical symmetry between the two token prices: because YES and NO prices sum to approximately one by construction, the calibration properties of the first token imply those of the second. The symmetry check in Section 5.4 confirms this relationship holds in the data. What it does not confirm is whether the two tokens are structurally equivalent in terms of what they represent, and the results suggest they are not.

The home advantage hypothesis developed in Section 7.2 provides the primary structural account: the first-outcome token is the away team's token, and systematic overvaluation of home advantage by market participants depresses the away token below its true resolution probability. This explains the direction of the asymmetry and its concentration in the mid-range. A secondary question nonetheless arises from the token-ordering design of the platform itself. If the first-outcome token is the default or most prominently displayed option when a participant opens a market, attention and salience effects might be expected to attract excess buying toward it, inflating rather than depressing its price. The data show the opposite: the first token is under-priced throughout the lower probability range. This is itself informative. The fact that the first token is not over-priced argues against a pure salience or attention-based account and points back toward something structural about team identity, consistent with the home advantage hypothesis.

8. Conclusion

This section draws together the findings of the thesis, assesses what they contribute to the prediction-market calibration literature, and identifies the boundaries of what the current analysis can and cannot establish. Sections 8.1-8.4 cover the headline result, contributions, limitations, and directions for future work in turn.

8.1 Headline

This thesis set out to test whether the classical longshot bias, the prediction, grounded in cumulative prospect theory and confirmed empirically by Snowberg and Wolfers (2010) on parimutuel venues, that low-probability tokens trade above their true resolution rate, survives on Polymarket sports prediction markets. Across 4,363 cleanly-resolved markets spanning NBA, NHL, NCAA Basketball, and four major European football leagues, it does not. Low-probability tokens resolved to one at a rate higher than their implied price, inconsistent with the classical over-pricing prediction and consistent with a market in which casual over-weighting of small probabilities is not the dominant pricing force.

The pattern is, however, more nuanced than a clean inversion of the classical bias would suggest. Calibration across the mid-range of the probability spectrum is broadly well-behaved, and the deviation at the longshot end, while directionally consistent and statistically meaningful in the pooled sample, is better characterised as a failure of the classical prediction than as a strong mirror image of it. A further structural finding concerns the asymmetry between the two outcome tokens within the same market: the first-outcome token is systematically priced differently from its complement, a pattern that adds a within-market dimension to the calibration anomaly and points toward structural, rather than purely behavioural, sources of mispricing. The population-weighted mean edge is 0.020, peaking at 0.032 in the mid-range band, with European football leagues driving the strongest signal (longshot edges of 0.054-0.064) and the pattern attenuating substantially across the February 2026 microstructure regime break.

8.2 Contributions

The thesis makes two substantive contributions to the empirical literature on prediction-market calibration. The first is descriptive. By decomposing Polymarket sports calibration into per-band, per-sport, and per-regime cross-sections, we show that the aggregate null documented by Reichenbach and Walther (2025), no longshot bias across Polymarket, is a consequence of opposite-signed deviations approximately cancelling, not an absence of mispricing. Classical cumulative prospect theory is rejected as the pricing mechanism: low-probability tokens are under-priced, not over-priced, across the pooled sample, with the signal concentrated in the mid-range of the probability distribution (0.26-0.48) and led by the four European football leagues. The finding is directionally consistent across all seven sports categories.

The second contribution concerns the structural source of the anomaly. The calibration asymmetry is not market-wide: it is concentrated in the first-outcome token within each market, with the second-outcome token correspondingly over-priced by arithmetic complement. This token-level asymmetry points toward structural rather than purely behavioural explanations - specifically the home advantage bias hypothesis and the structure of the soccer markets, where the compound nature of the NO token embeds draw probability that casual participants systematically misprice. The February 2026 regime shift reinforces this interpretation: the underpricing signal attenuates by approximately two thirds as participation increases, a pattern consistent with informed capital entering and competing the edge toward zero - the standard Grossman-Stiglitz dynamic.

8.3 Limitations

The most consequential limitation of this thesis is the absence of order book and volume data. The calibration is estimated entirely from price paths - sequences of last-traded prices - with no visibility into the depth of the book at each price level, the volume transacted at each tick, or the composition of order flow in terms of maker and taker activity. This matters because a price may appear persistently mispriced in a calibration sense while being effectively inaccessible: if the volume available at the longshot band is thin, or if the spread between best bid and ask is wide relative to the measured edge, the apparent anomaly may not be reachable in practice. Without trade-side data we cannot distinguish a genuine pricing inefficiency from a structural feature of how prices move through a thinly traded region of the order book. Relatedly, the absence of flow composition data means we cannot test the live-event-driven informed-trader fading mechanism directly, nor assess whether the first-outcome token asymmetry reflects persistent mispricing or a transient feature of how liquidity is supplied around event resolution.

A second limitation is scope. The entire analysis is restricted to sports prediction markets on a single venue, Polymarket, across seven league and sport categories chosen for data availability. Whether the calibration pattern documented here extends to political, macroeconomic, or other event-type markets on Polymarket, or to sports prediction markets on other continuous-orderbook venues, is unknown. The divergence in signal strength across sports, with European football leagues showing longshot edges two to three times larger than North American team sports, already suggests the result is sensitive to the scoring structure and liquidity profile of the underlying market. The current sample does not span enough variation in these dimensions to support a general claim about continuous-orderbook prediction markets as a class.

8.4 Future Work

The most direct extension of this work is to implement and test a trading strategy built around the documented calibration signal. The natural starting point is a buy-only strategy targeting first-outcome tokens in the longshot band at or near market open, holding through resolution, an approach motivated directly by the finding that these tokens resolve at a systematically higher rate than their implied price. The central empirical question is how much of the raw calibration edge survives once Polymarket's trading fees, bid-ask spreads at entry, and execution slippage in a thinly-traded book are accounted for. A credible answer requires either live execution or a tick-level replay with fill assumptions calibrated to actual book depth, and should include walk-forward validation across market conditions rather than a single in-sample backtest.

Resolving the data limitations identified in 8.3 is the second priority. Acquiring order book snapshots and volume-at-price data would allow a direct test of the token-level asymmetry mechanism and give a sharper estimate of the accessible edge in each probability band. It would also allow the informed-trader fading hypothesis to be assessed empirically, by examining whether the calibration deviation is concentrated in the in-play window or already present at market open, rather than left as a proposed candidate mechanism. Additionally, further investigating the home-team bias would allow to better understand the potential causes of this mispricing, as discuss in 7.2.

References

Bailey, D. H., and López de Prado, M. (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. Journal of Portfolio Management, 40(5), 94-107.

Bali, T. G., Cakici, N., and Whitelaw, R. F. (2011). Maxing out: stocks as lotteries and the cross-section of expected returns. Journal of Financial Economics, 99(2), 427-446.

Barberis, N., and Huang, M. (2008). Stocks as lotteries: the implications of probability weighting for security prices. American Economic Review, 98(5), 2066-2100.

Berg, J. E., Forsythe, R., Nelson, F., and Rietz, T. (2008). Results from a dozen years of election futures markets research. Handbook of Experimental Economics Results, 1, 742-751.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. Quarterly Journal of Economics, 127(3), 1243-1285.

Cain, M., Law, D., and Peel, D. (2000). The favourite-longshot bias and market efficiency in UK football betting. Scottish Journal of Political Economy, 47(1), 25-36.

Camerer, C. F. (1989). Does the basketball market believe in the 'hot hand'? American Economic Review, 79(5), 1257-1261.

Croxson, K. and Reade, J.J. (2014). Information and Efficiency: Goal Arrival in Soccer Betting. *The Economic Journal*, 124(575), pp. 62-91.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. Journal of Finance, 53(6), 1839-1885.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. Journal of Political Economy, 98(4), 703-738.

Hong, H., and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. Journal of Finance, 54(6), 2143-2184.

Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. Econometrica, 47(2), 263-291.

Kumar, A. (2009). Who gambles in the stock market? Journal of Finance, 64(4), 1889-1933.

Manski, C. F. (2006). Interpreting the predictions of prediction markets. Economics Letters, 91(3), 425-429.

Odean, T. (1998). Are investors reluctant to realize their losses? Journal of Finance, 53(5), 1775-1798.

Ottaviani, M., and Sørensen, P. N. (2008). The favorite-longshot bias: an overview of the main explanations. In Hausch, D. B., and Ziemba, W. T. (eds.), Handbook of Sports and Lottery Markets, 83-101. Elsevier.

Ottaviani, M., and Sørensen, P. N. (2010). Noise, information, and the favorite-longshot bias in parimutuel predictions. American Economic Journal: Microeconomics, 2(1), 58-85.

Page, L., and Clemen, R. T. (2013). Do prediction markets produce well-calibrated probability forecasts? Economic Journal, 123(568), 491-513.

Prelec, D. (1998). The probability weighting function. Econometrica, 66(3), 497-527.

Reichenbach, F., and Walther, M. (2025). Exploring decentralized prediction markets: accuracy, skill, and bias on Polymarket. SSRN Working Paper No. 5910522.

Romano, J. P., and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. Econometrica, 73(4), 1237-1282.

Shefrin, H., and Statman, M. (1985). The disposition to sell winners too early and ride losers too long. Journal of Finance, 40(3), 777-790.

Shleifer, A., and Vishny, R. W. (1997). The limits of arbitrage. Journal of Finance, 52(1), 35-55.

Snowberg, E., and Wolfers, J. (2010). Explaining the favorite-longshot bias: is it risk-love or misperceptions? Journal of Political Economy, 118(4), 723-746.

Tetlock, P. C. (2008). Liquidity and prediction market efficiency. Working paper, Columbia University.

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

Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297-323.

Wolfers, J., and Zitzewitz, E. (2004). Prediction markets. Journal of Economic Perspectives, 18(2), 107-126.

9. Appendix

This appendix bundles material that supports the main text but does not belong in it. We describe in turn the graphs and tables referred to in the text (A.0) and a glossary (A.1).

A.0: Graphs and Tables

Table 4.1:

Category	Ticks (legacy)	Ticks (recent)	Ratio	Reversals (legacy)	Reversals (recent)	Ratio
NBA	2,016	4,629	2.30×	585	1,508	2.58×
NHL	828	1,884	2.28×	237	624	2.63×
NCAA Basketball	182	389	2.14×	53	117	2.21×
Soccer-EPL	525	1,040	1.98×	130	318	2.45×
Soccer-Bundesliga	141	616	4.37×	36	184	5.11×
Soccer-LaLiga	190	833	4.38×	52	255	4.90×
Soccer-SerieA	145	680	4.69×	38	198	5.20×
Pooled	316	922	2.92×	77	275	3.57×

Table 4.1. Median per-market lifetime tick count and one-cent reversal count, by sport and panel. Every category shows a multiplicative increase in both statistics across the February 2026 boundary, with ratios ranging from 1.98× to 4.69× on tick count and 2.21× to 5.20× on reversal count.

Table 4.2:

Category	Legacy	Recent	Total
NBA	239	362	601
NHL	555	273	828
NCAA Basketball	687	664	1,351
Soccer - EPL	288	287	575
Soccer - LaLiga	409	252	661
Soccer - Bundesliga	374	226	600
Soccer - SerieA	298	188	486
Total	2,850	2,252	5,102

Table 4.2. Working sample by category and panel. Soccer rows include both moneyline contracts and the companion draw legs that share a fixture.

Figure 6.1:

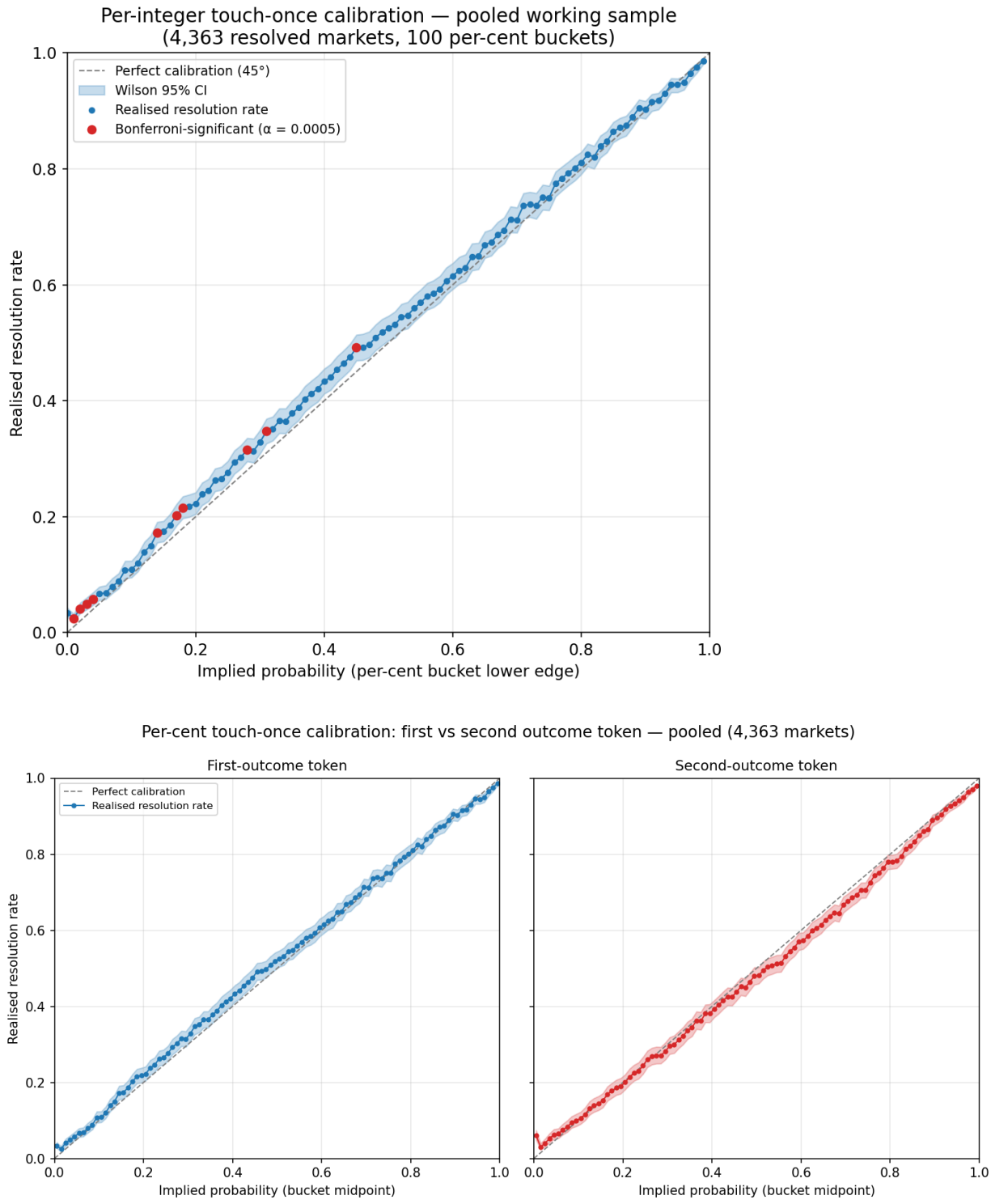

Figure 6.1. Per-integer touch-once calibration on the pooled working sample (4,363 resolved markets). Each point is one per-cent bucket; its height is the realised resolution rate of all markets that touched that price. The shaded band is the per-bucket Wilson 95% confidence interval. Red points mark buckets whose deviation from the 45° identity is significant after Bonferroni correction across the 100 buckets (α = 0.0005).

Table 6.1:

Price decile	Mean implied	Markets touched	Realised rate	Mean per-integer edge
[0.00, 0.10)	0.043	18,377	0.060	+0.017
[0.10, 0.20)	0.145	17,807	0.169	+0.024
[0.20, 0.30)	0.246	19,491	0.274	+0.029
[0.30, 0.40)	0.344	19,308	0.375	+0.031
[0.40, 0.50)	0.445	19,145	0.477	+0.033
[0.50, 0.60)	0.544	18,267	0.563	+0.019
[0.60, 0.70)	0.644	16,571	0.660	+0.015
[0.70, 0.80)	0.744	15,356	0.757	+0.013
[0.80, 0.90)	0.844	14,575	0.854	+0.010
[0.90, 1.00)	0.947	14,710	0.945	−0.002

Table 6.1. Per-integer touch-once calibration by price decile, pooled working sample. Contracts are grouped into ten price deciles by traded price. For each decile the table reports the mean implied (traded) price, the number of distinct market touches contributing, the realised resolution rate, and the mean per-integer signed edge, defined as realised rate minus implied price.

Figure 6.2:

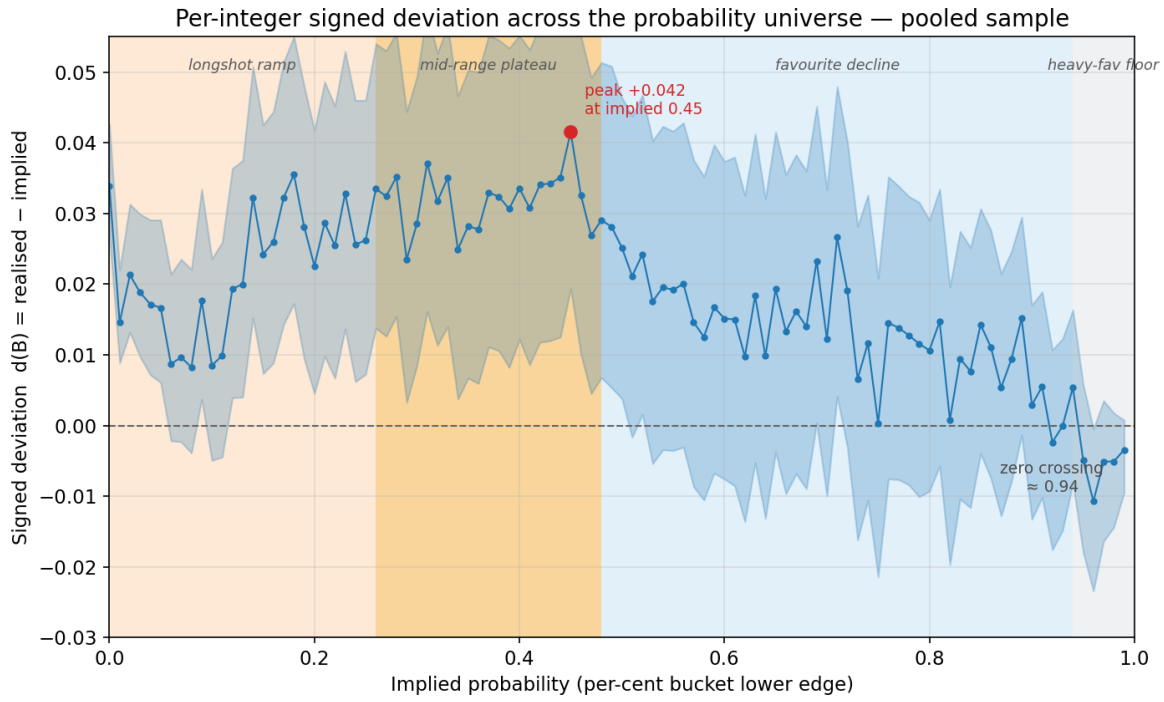

Figure 6.2. Per-integer signed deviation d(B) = π(B) − implied across the pooled working sample, with the per-bucket Wilson 95% interval. The horizontal line is exact calibration; shading marks the four data-driven phases described in the text.

Figure 6.3:

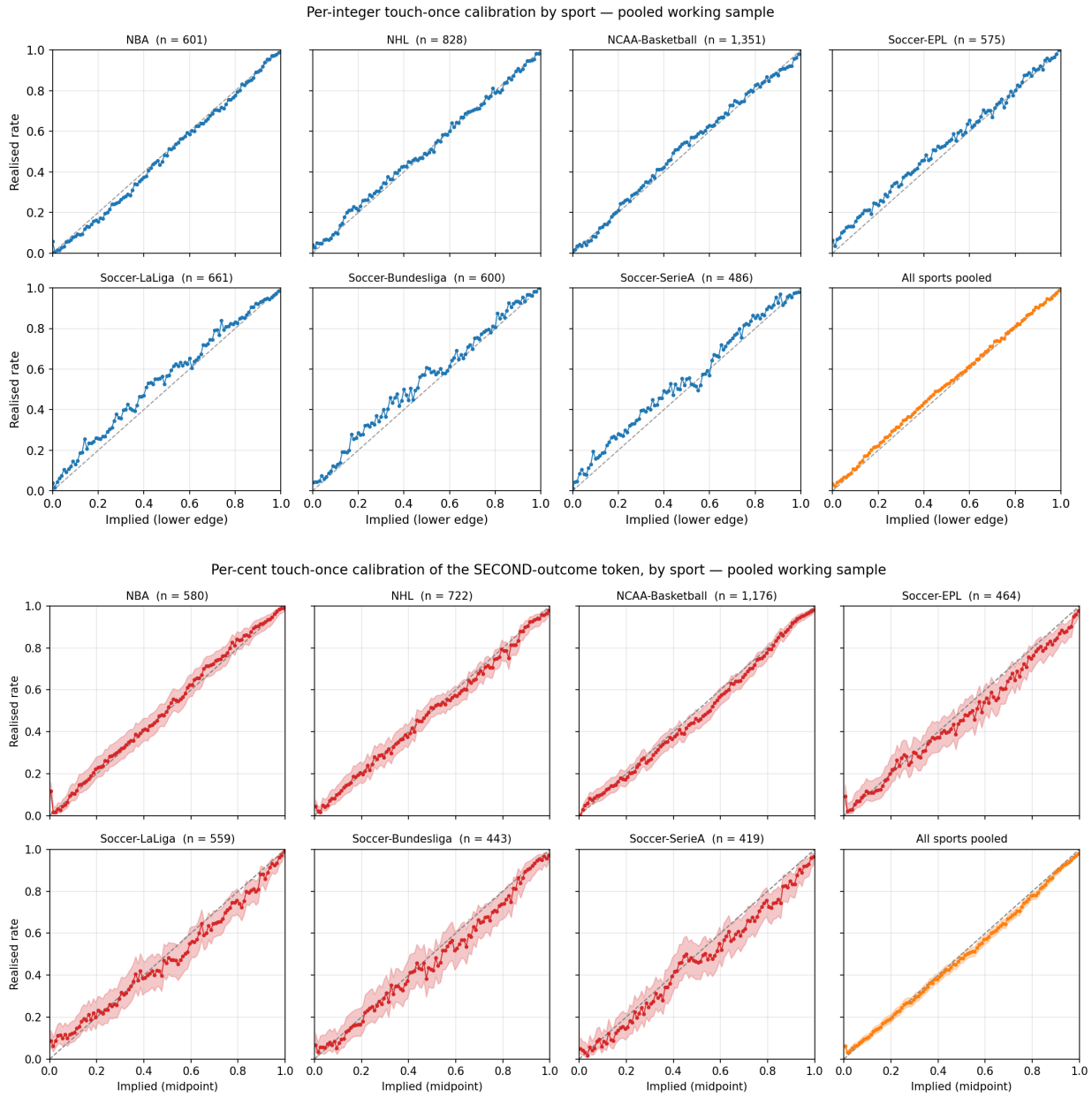

Figure 6.3. Per-integer touch-once calibration by sport, pooled working sample. Each panel plots realised resolution rate against implied price for one sport; the dashed line is exact calibration. The lower-right panel repeats the all-sports pooled curve for reference.

Table 6.2:

Sport	**Longshot [0, 0.40)**	**Coin-flip [0.40, 0.60)**	**Slight fav [0.60, 0.85)**	**Heavy fav [0.85, 1.00)**
NBA	−0.025	−0.010	−0.016	−0.002

NHL	+0.025	+0.003	+0.001	−0.004
NCAA-Basketball	+0.012	+0.040	+0.019	−0.013
Soccer-EPL	+0.054	+0.045	+0.011	+0.011
Soccer-LaLiga	+0.057	+0.076	+0.040	+0.011
Soccer-Bundesliga	+0.056	+0.054	+0.025	+0.025
Soccer-SerieA	+0.064	+0.033	+0.045	+0.020

Table 6.2. Population-weighted mean per-integer edge by sport and price region, pooled (legacy + recent) working sample. Each cell averages the per-cent deviations d(B) within the region, weighted by the number of markets touching each bucket.

Figure 6.4:

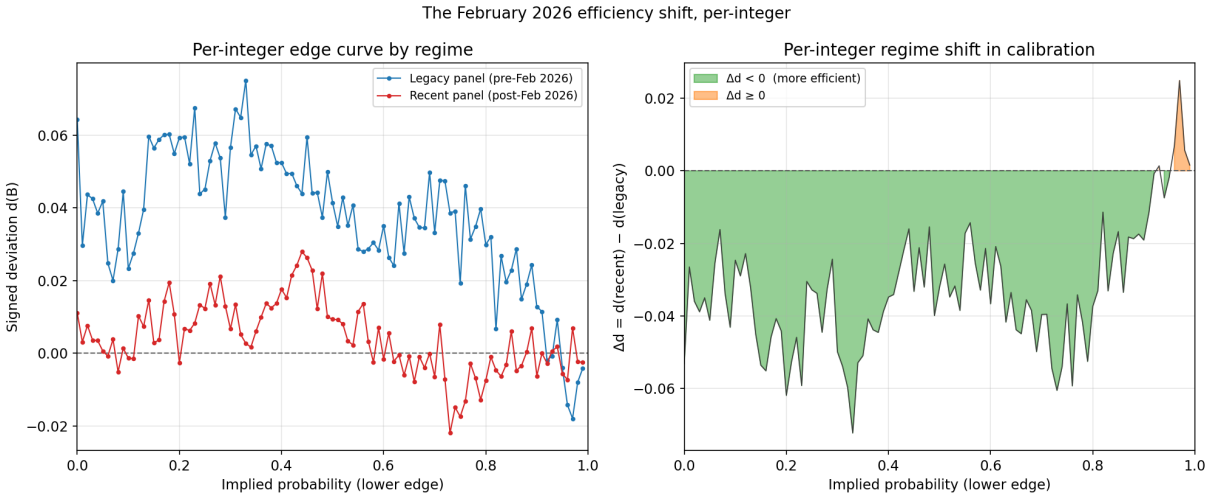

Figure 6.4. Left: per-integer signed deviation d(B) for the legacy and recent panels. Right: the per-bucket regime shift Δd(B), shaded green where Δd < 0 - that is, where the recent panel sits closer to exact calibration.

Figure 6.5:

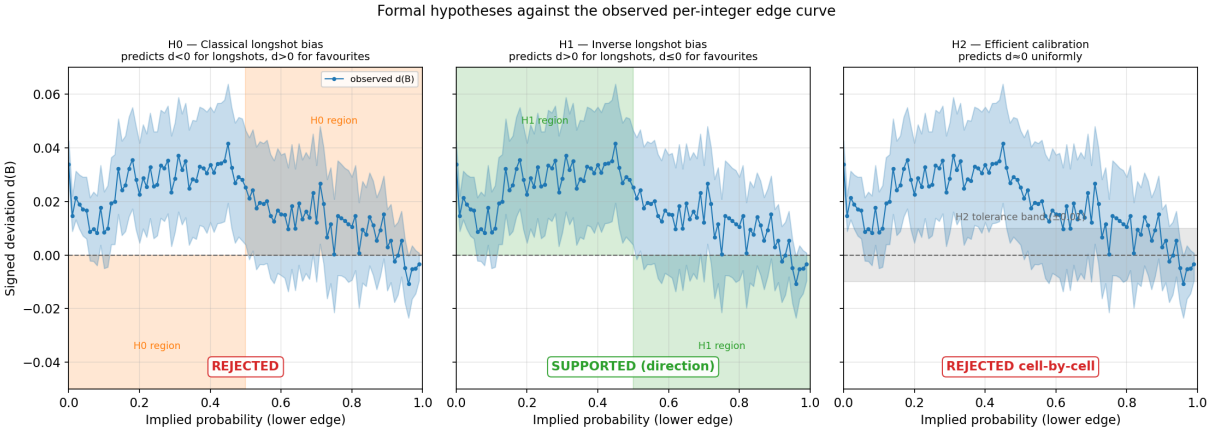

Figure 6.5. The three formal hypotheses against the observed pooled edge curve. Each panel shades the region of price-deviation space that the hypothesis predicts the curve should occupy; the observed curve and its Wilson band are overlaid.

Table 7.1:

Sport	Δd longshot [0, 0.40)	Δd heavy fav [0.85, 1.00)
NBA	+0.006	+0.022
NHL	−0.049	+0.028
NCAA-Basketball	−0.039	+0.023
Soccer-EPL	−0.074	−0.037
Soccer-LaLiga	−0.038	−0.052
Soccer-Bundesliga	−0.023	−0.039
Soccer-SerieA	−0.119	−0.033
Pooled	−0.042	−0.007

Table 7.1. Population-weighted mean regime shift Δd by sport and price region. A negative longshot Δd means longshot under-pricing weakened across the break; a positive heavy-favourite Δd means favourite over-pricing weakened.

Figure 7.1:

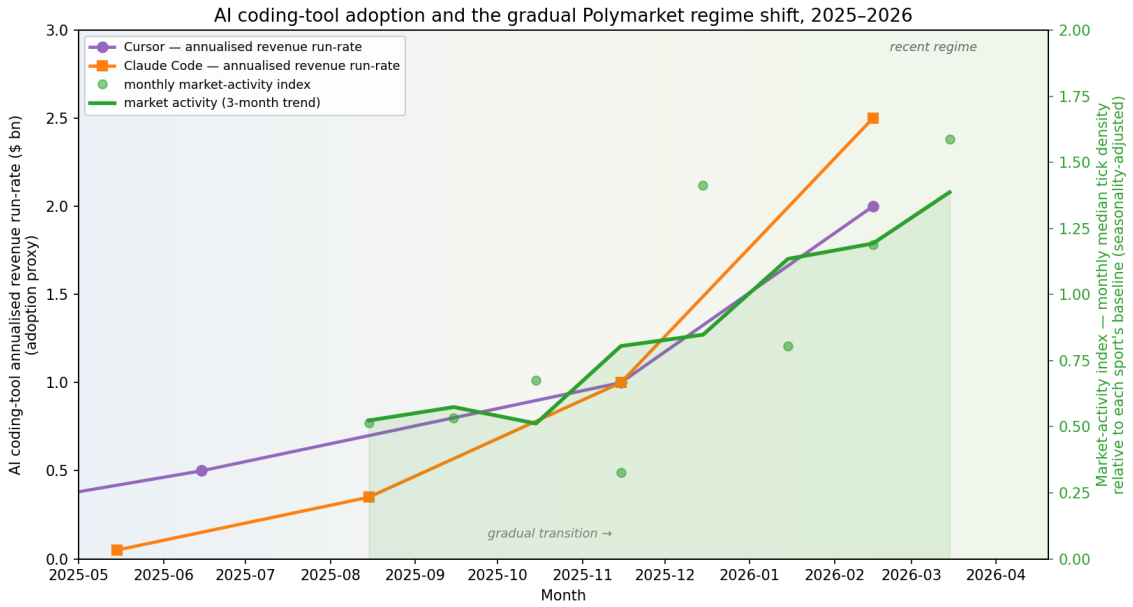

Figure 7.1. AI coding-tool adoption - annualised revenue run-rate, a public proxy for adoption - against the Polymarket microstructure break. Cursor and Claude Code revenue milestones are from press and company reporting (see text); market activity is the per-sport-averaged tick density of the legacy and recent panels. Sources: Reuters and company reporting

A.1: Glossary

Term	Explanation
Ambiguous market	A market whose final price falls strictly between 0.01 and 0.99 (typically a cancelled or postponed match). Excluded from the sample (~14.5% of the universe).
Benjamini-Hochberg / FDR	A less conservative alternative to Bonferroni that controls the expected fraction of false discoveries rather than the probability of any false discovery. More powerful when many buckets deviate in the same direction.
Bonferroni correction	Adjusts significance thresholds when running many simultaneous tests by dividing α by the number of tests. Used here across 100 per-cent buckets.
Bootstrap / by-market block bootstrap	Estimating uncertainty by repeatedly resampling the data with replacement and recomputing the statistic. The by-market variant resamples whole markets rather than individual ticks, preserving within-market autocorrelation.
Buy-only constraint	On Polymarket, traders can only open long positions. Betting against an outcome requires buying the complementary token (NO), not shorting YES.
Calibration	Whether a market's prices match the actual frequency of outcomes. A perfectly calibrated market priced at 0.30 sees events resolve YES exactly 30% of the time.
Coin-flip region / mid-range plateau	The price range around 0.40-0.60, where the inverse-longshot deviation reaches its peak in this thesis (+0.042 at implied 0.45).
Cumulative prospect theory (CPT)	The 1992 extension of prospect theory (Tversky & Kahneman). Applies probability distortion to cumulative rather than raw probabilities, producing a mathematically consistent model and the standard reference in empirical work.
Decision weight	The distorted probability a CPT agent uses internally instead of the objective probability. Exceeds the true probability for small p; falls short for large p.

Direction reversal	A change in price direction of at least one cent (up after down, or vice versa). Used to measure within-market choppiness and to document the February 2026 regime break.
Draw leg	In soccer, the binary YES/NO market for the match ending in a draw, listed separately from the home-win and away-win markets.
Edge / signed deviation d(B)	The gap between what happened and what the price implied: $d(B) = \pi(B) - \text{implied}(B)$. Positive edge means the token was under-priced; negative means over-priced.
Fading	Trading against a recent price move in expectation that it has overshot fair value. Informed traders fade retail overreactions to in-game events, pushing prices back toward fair.
Favourite-longshot bias (classical)	The well-documented finding that bettors over-pay for longshots and under-pay for heavy favourites. The classical direction is longshots over-priced, favourites under-priced. This thesis finds the opposite.
First Outcome Token (FOT)	The YES-side token in a binary market; the contract that pays $1 if the named team wins. For North American sports this is the away team by platform convention.
H0 / H1 / H2	The three competing hypotheses. H0 (classical longshot bias): $d(B) < 0$ for longshots, $d(B) > 0$ for favourites. H1 (inverse longshot bias): the opposite. H2 (efficient calibration): $d(B) \approx 0$ uniformly, reproducing Reichenbach & Walther (2025).
Implied probability	The market's estimate of how likely an event is. On Polymarket, the YES token price is directly interpretable as a probability (a price of 0.30 implies a 30% chance).
Informed trader / sharp	A well-capitalised, price-sensitive participant trading on genuine information or systematic analysis. Contrasts with the retail or noise trader.
Inverse-S weighting	The shape of the CPT probability weighting function: concave for small probabilities (overweighting them) and convex for large ones (underweighting them), crossing the identity near $p \approx 0.30$-0.40.

Kolmogorov-Smirnov / Mann-Whitney U	Non-parametric two-sample tests used to formally confirm that path-feature distributions differ across the February 2026 regime break. KS compares full distribution shapes; MW tests whether one sample tends to produce larger values than the other.
Last-traded price (LTP)	The price at which the most recent trade occurred. The thesis works exclusively from LTP paths scraped from the Polymarket blockchain.
Legacy panel / recent panel	The two subsamples separated by the February 2026 microstructure break. Legacy: August 2025-early February 2026 (2,850 markets). Recent: late January-April 2026 (2,252 markets).
Longshot	A token with a low implied probability of resolving YES – a small chance of a large payoff.
Lottery-like asset	A security with a small probability of a large positive payoff. Low-priced YES tokens are the prediction-market analogue. Barberis & Huang (2008) show such assets are over-priced in CPT equilibrium – the prediction this thesis tests and rejects.
Microstructure	The mechanics of how a market operates: order-matching rules, participant composition, repricing frequency. The thesis argues that Polymarket's continuous orderbook microstructure explains why the classical longshot bias does not survive there.
Moneyline contract	A bet on which team wins outright, with no point spread. The simplest sports-betting format and the primary contract type in this sample.
Multiple-testing / multiplicity correction	Adjusting for the inflation of false-positive risk that arises when running many simultaneous tests. Bonferroni and FDR are the two methods applied here.
NO token	The complement to the YES token in a binary market. Pays $1 if the event does not resolve YES. YES and NO prices sum to approximately $1 at all times.
Parimutuel market	A betting structure where all wagers pool together, the operator takes a fixed cut, and the remainder is divided among winning bettors. Prices are set by the

	crowd and fixed at event close. The canonical venue for classical longshot bias research – structurally different from Polymarket's continuous orderbook.
Polymarket	A decentralised prediction market platform where users trade binary outcome tokens with real USDC collateral on a continuous central limit-order book. Resolution is administered on-chain via UMA's Optimistic Oracle.
Population-weighted mean	An average in which each per-cent bucket is weighted by the number of markets behind it, so high-volume buckets count proportionally more than sparse ones.
Prelec function	The one-parameter probability weighting function $w(p) = \exp(-(-\ln p)^{\alpha})$. With $\alpha \approx 0.65$ (the modal experimental estimate), a 10% event is weighted as roughly 20%. Provides the quantitative benchmark for the longshot bias prediction tested here.
Probability band / bucket	A small slice of the [0, 1] price interval. The thesis uses 100 per-cent buckets of width 0.01.
Probability weighting function w(p)	A function that converts objective probabilities into the decision weights CPT agents use. Produces the inverse-S shape that generates the classical longshot bias prediction.
Prospect theory	Kahneman & Tversky's (1979/1992) model of decision under risk. Agents distort probabilities via a weighting function and weight losses more heavily than equivalent gains. The theoretical foundation for the longshot bias prediction.
Realised win-rate $\pi(B)$	The fraction of markets that visited bucket B and resolved YES. The primary empirical object: $\pi(B)$ compared against the implied probability gives the signed deviation d(B).
Regime conditioning	Re-running every calibration result separately on the legacy and recent panels to verify findings are not an artefact of one microstructure environment.

Resolution / resolution rule	Resolution is the moment a market settles on-chain (YES = $1, NO = $0). The thesis's resolution rule: final YES price ≥ 0.99 → outcome 1; ≤ 0.01 → outcome 0; otherwise excluded as ambiguous.
Slight-favourite / heavy-favourite region	Price ranges in the upper half of the spectrum. Used in Table 6.2: slight favourite [0.60, 0.85), heavy favourite [0.85, 1.00).
Tick / tick density	A tick is a single recorded trade or price update. Tick density is the number of ticks per market per unit of time. Rising tick density across the February 2026 break is the primary quantitative signature of the regime change.
Touch-once calibration	The thesis's main aggregation rule: each market contributes one observation to a bucket if its YES price ever entered that bucket during its life, regardless of how long it spent there. Controls within-market autocorrelation.
Two-proportion z-test	Tests whether two binomial success rates (here: realised win-rates in legacy vs recent panels at the same bucket) are statistically different.
UMA Optimistic Oracle	The on-chain service Polymarket uses to settle markets. A proposer posts the resolution outcome; if unchallenged during a waiting period, it is accepted and the market pays out.
USDC	A US-dollar-pegged stablecoin used as Polymarket's settlement currency. Real-money stakes denominated in USDC preserve the financial incentives a behavioural-finance test requires.
Wilson confidence interval	A confidence interval for a proportion that performs well at small samples and extreme win-rates. Used throughout in preference to the normal approximation.