The Price is (Almost) Right: A Calibration Analysis of Polymarket Pricing Data

Aidan M. Walker

Advisor: Professor Daniel E. Rigobon

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science in Engineering

Department of Operations Research and Financial Engineering

Princeton University

April 2026

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Aidan M. Walker

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Aidan M. Walker

Abstract

Prediction markets have emerged as a prominent mechanism for aggregating crowd beliefs into probability forecasts for real-world events. The world's largest prediction market, Polymarket, explicitly claims that its trading prices are probabilities, a claim that invites rigorous empirical scrutiny. This thesis presents a dedicated, multi-domain, multi-horizon calibration study of Polymarket, analyzing 188,509 resolved binary markets and 1,875,570 daily price observations spanning November 2022 through December 2025.

Calibration is assessed through reliability diagrams, Brier Score decomposition, and logistic recalibration across both market domain and time-to-resolution horizon. Uncertainty quantification employs an event-level clustered bootstrap approach that corrects for the non-independence of Polymarket's correlated multi-market event structure, a methodological gap identified in comparable recent literature.

We find that Polymarket prices demonstrate genuine forecasting skill overall, achieving a Brier Skill Score of 0.398. However, this magnitude varies substantially by domain, ranging from 0.264 in Sports (worst) to 0.565 in Politics (best). The fitted overall logistic recalibration slope $b = 1.112$ confirms broad underconfidence consistent with the well-documented Favorite-Longshot Bias (FLB): prices are systematically compressed toward 0.50, with longshots overpriced and favorites underpriced. This bias is universally present across all market domains and worsens at longer time horizons. Clustered bootstrapping significantly widens confidence intervals relative to naive inference: while the core miscalibration finding proves robust, two horizon bands lose statistical significance, showing that ignoring clustering would produce incorrect inferences. These findings replicate and extend FLB patterns documented on prior prediction market platforms, and suggest that domain- and horizon-specific recalibration could improve the interpretation of raw Polymarket prices.

Acknowledgments

First of all, I would like to extend my gratitude to my advisor, Professor Daniel E. Rigobon. Thank you for putting up with me for the last 8 months. I truthfully do not know if I could have reached this point without your advice, guidance, and patience throughout this process. I appreciate you.

I have the incredible fortune of having many important people in my life that helped me get to where I am today, and for that I will always be grateful:

To my parents, I owe everything to you. Thank you for always putting me before you. I am proud to be your son every day.

To my siblings, for always making me laugh and keeping me humble. I wouldn't want it any other way.

To my roommates, I will forever look back on these days and nights. This school would be an entirely different place without you.

To my girlfriend, who puts a smile on my face every day and always brings out the best in me.

To all my friends from school and from home, who have been by my side through it all.

To my grandparents, for teaching me the importance of family. And sending me lots of chocolate at school.

And finally to my dog, which needs no further explanation. Woof!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

On Monday, July 28, 2003 it was revealed that the U.S. Defense Advanced Research Projects Agency (DARPA) was planning to go live with a new project: an online futures trading market that could hopefully find use as a new form of intelligence gathering. This Policy Analysis Market, or PAM for short (see Figure 1.1), would allow speculators to bet on the likelihood of certain global events occurring (Hanson 2003). Advertised as "A Market in the Future of the Middle East," so-called "Special Event Securities" might include hypothetical scenarios such as a regime change within some specified timeframe, significant U.S. military activity in the region, or the assassination of a political leader (Looney 2004). By creating an open market for these hypothetical events, the goal was to see if U.S. intelligence officials might be better equipped to predict geopolitical events of interest (Yeh 2006).

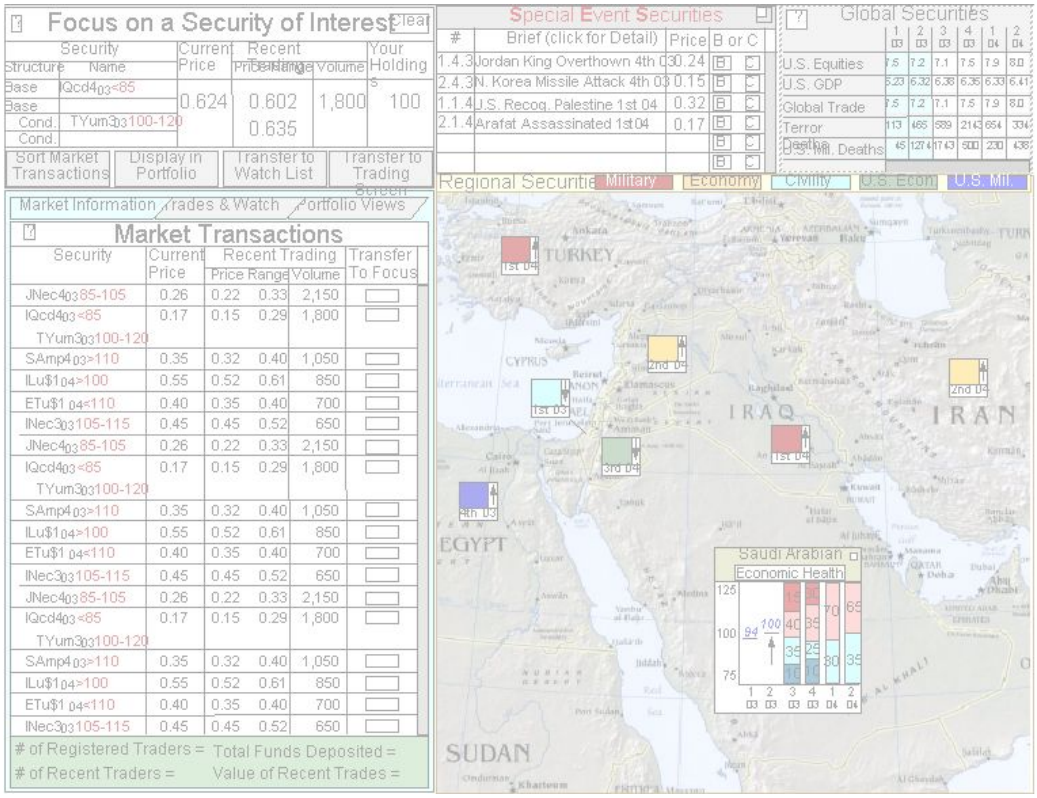

Figure 1.1: Policy Analysis Market (PAM) public website archive. Note the "Special Event Securities" section located in the upper middle that much of the negative media response centered around.

The following morning, news began to break of this plan. *"Pentagon Prepares A Futures Market On Terror Attacks,"* wrote the New York Times (Hulse 2003). *"Trading on the Future of Terror,"* the Los Angeles Times called it (Meyer 2003). Both outrage in the Senate and uproar from the public was swift, and by the end of the day, the plan had been scrapped. Shortly thereafter, the project's head was forced to resign. Bipartisan opinion concluded that the DARPA plan was "unbelievably stupid" (Meyer 2003) and that "this thing should be stopped" (Schoen 2003).

DARPA's Policy Assessment Market, although extremely short-lived, represented an attempt (albeit crude) at the creation of a "Prediction Market." Prediction markets have existed in various forms since the late 20th century, to varying degrees of success. Perhaps the oldest and most well-known of these (and certainly less controversial than DARPA's) is the Iowa Electronic Markets (IEM), an academic research project first

developed by the faculty at the University of Iowa Henry B. Tippie College of Business in 1988 (Berg, Nelson, and Rietz 2003). These markets established a controlled setting where vetted researchers and students could trade in financial contracts on real-world events ranging from economic indicators, quarterly earnings, stock price movements, movie box office receipts, and of course most famously, political outcomes.

Based on analysis of empirical evidence by researchers, it has been shown that the IEM was generally much better at predicting the outcome of elections than traditional polls (Berg, Nelson, and Rietz 2003). Both the relative success and perceived accuracy of the IEM in predictive power (at least for election-centered markets) have inspired a wide range of new prediction markets to spawn over the years. In addition to the many imitation academic and/or research-oriented markets that have been created by universities around the world, other prominent examples include but are not limited to:

- The Hollywood Stock Exchange to predict box office success (Wolfers and Zitzewitz 2004)
- Hewlett Packard's collaboration with Caltech to develop an "Information Aggregation Mechanism" for forecasting sales (Plott and Chen 2002)
- Google's internal prediction market to forecast the success of various corporate strategic aims (Yeh 2006)
- Eli Lilly's efforts to improve outcomes of drug research and development (Yeh 2006)

While many of these corporate projects achieved limited success utilizing the principles of prediction markets to improve internal operation and information-sharing, there is a separate swath of prediction markets that have arisen over the years in the hopes of capturing the public attention and operating as a for-profit financial

exchange. These include, but are not limited to, names such as Tradesports.com, PredictIt, InTrade, Newsfutures.com, Betfair.com, and more (Wolfers and Zitzewitz 2004). Many of these exchanges, due largely to an unfriendly U.S. regulatory environment at the Commodities Futures Trading Commission (CFTC) and the issuance of "no-action" letters, are now defunct (Berg, Nelson, and Rietz 2003).

With the recent resurgence in the popularity of prediction markets, led by dominant players in the form of Polymarket and Kalshi, the historical success of the IEM is once again being put to the test. But in order to discuss the modern state of these markets and its implications, this thesis first wishes to return to and thoroughly define the basic concept of a prediction market itself.

Broadly, prediction markets are exchange-traded markets in which individuals may trade on financial contracts with payoffs tied to the binary outcomes of real future events. Because of the crucial binary feature of these markets, most proponents of these markets claim that a probabilistic interpretation is justified: given the current trading price of any given event contract, this should theoretically be understood as the probability of that event occurring. What does this mean in practice?

As an illustrative example, consider the example below, as inspired by PAM:

> **Event**: Will Regime X of Nation Y fall this year?
>
> **Description**: This market will resolve to "Yes" if Nation Y's current ruling regime is overthrown, collapses, or otherwise ceases to assume governing duties by December 31st, 2026 at 11:59 PM GMT. Otherwise, this market will resolve to "No."

Suppose this market is structured as a winner-take-all approach: each "Yes" position must be exactly matched with one "No" position, these positions must sum to $1.00, and there can only be one winner (i.e., binary).[1] Based on these assump-

[1]The binary market constraint is the reason for the highly detailed market description in the above example; there is no room for ambiguity or "gray lines" in the outcome of a prediction market event contract due to the binary nature of the market structure.

tions of market structure, then, one might claim that the market currently places the probability of Regime X falling at p, where p is the current trading price of the market. If an individual believes that this market is *underestimating* (*overestimating*) the probability of a regime change, then it follows that the individual should purchase N contracts of "Yes" ("No"), where N is an arbitrary initial investment determined by the individual. Then, on December 31st, this market will either resolve in the individual's favor, or it will not.

For example, if the "Yes" price is currently trading at \$0.05 and "No" at \$0.95, but the individual believes "Yes" is much more likely, then they should purchase N shares of "Yes." If it turns out the individual is wrong and there is no regime change, the individual will receive nothing and their loss will be $N * \$0.05$. If it turns out the individual is right and there is in fact a regime change, the individual will receive a payout of $N * \$1.00$.

The underlying assumption in this example is clearly that the probability of a regime change is inherently tied closely to the current trading price of the regime change market. In modern prediction markets, it is often true that not only is the probability assumed to be closely aligned with the trading price, but that they are effectively interchangeable. For example, in Figure 1.2, it can be seen that the implied probability of a certain future event occurring is often even the *first* and most prominent numerical quantity presented to a market viewer. To a novice user, only by navigating to the trade window or by expanding the order book (see Figure 1.3) is one presented with live trading prices denoted in financial currency rather than a probability percentage. Additionally, a quick trip through the Polymarket platform documentation reveals many instances of the blanket, straightforward statement: "Prices are Probabilities" (*Polymarket 101* 2026).

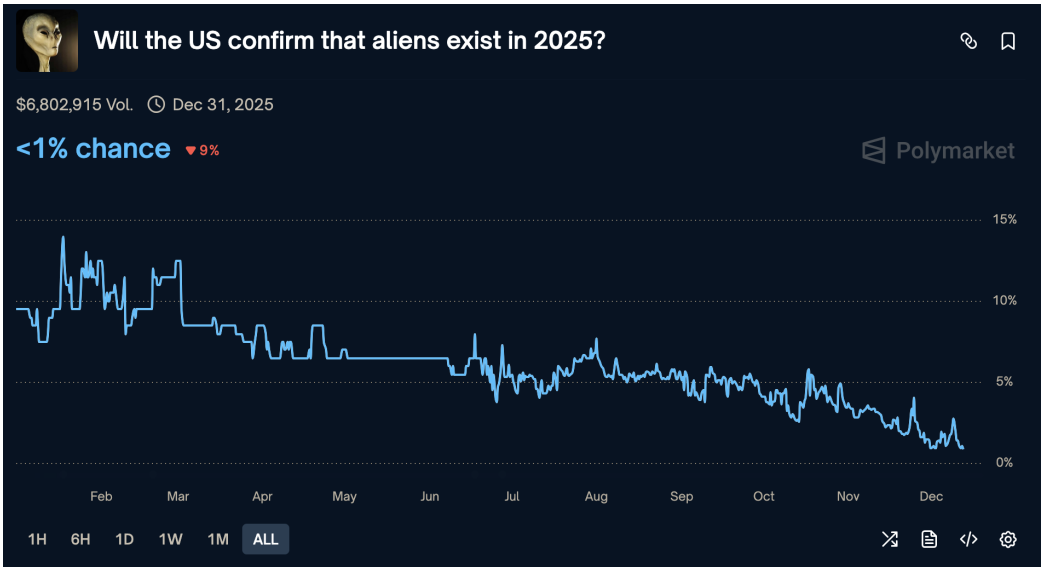

Figure 1.2: Historical "probability" chart for the Polymarket market: "Will the U.S. confirm that aliens exist in 2025?"

Figure 1.3: Open order book as of Friday, December 12th, 2025 for the Polymarket market: "Will the U.S. confirm that aliens exist in 2025?"

Naturally, this raises the question: How interchangeable are the two quantities (namely, the probability of a future real-world event occurring and the current prediction market trading price) in reality? Clearly, it is economically advantageous for a for-profit financial exchange to promote the accuracy and strength of its markets;

Polymarket and other modern prediction markets make this claim quite extensively (*How accurate is Polymarket?* 2025). However, the author would like to remind the reader of the famous appropriated Russian idiom that tells us to "доверяй, но проверяй," or "trust, but verify."

Prediction markets are based on the underlying fundamental assumption that prices accurately reflect the probability of some future event occurring. The basis for this assumption comes from an extensive history of classical theoretical arguments, namely on the subjects of wisdom of the crowds (Galton 1907; Surowiecki 2004), the Hayek hypothesis and information aggregation (Hayek 1945), and the efficient capital markets hypothesis (Fama 1970). Collectively, these suggest that prediction market prices should converge to true probabilities, and indeed, there is a substantial amount of empirical evidence to support the notion that prediction markets may be more accurate than alternative forecasting methods. For example, political election markets have been shown to outperform traditional expert polling mechanisms (Berg, Forsythe, et al. 2008).

However, the core belief in the accuracy of prediction market prices reflecting probabilities is not always supported in the research and empirical work. There is documented evidence of systematic deviations, including the favorite-longshot bias (Ottaviani and Sørensen 2008), temporal mispricing (Restocchi, McGroarty, and Gerding 2019), and domain-specific miscalibration (Le 2026), all of which serve to undermine trust and the informational value of these markets.

A key theme that frequently arises in the discussion of the above deviations is the question of **"calibration."** What do we mean when we discuss the concept of calibration as it relates to prediction markets? A prediction market is said to be "well-calibrated" if its observed trading prices correspond to the true outcome rate of future events. Without loss of generality, this means that "Yes" contracts currently trading at \$0.70 should ultimately resolve true (to \$1.00) approximately 70% of the

time. Conversely, a prediction market is "miscalibrated" if there is some systematic departure or bias from this ideal state in which trading prices do not correspond to true outcome probabilities.

Has increased liquidity and growing global interest in modern prediction markets (as compared to earlier generations of prediction markets) led to improved market efficiency and predictive power, or alternatively, is there evidence of systematic deviation? Specifically, the overarching question of this thesis is: **Are Polymarket prices reliably well-calibrated probability forecasts?**

Investigation into this question will come in the form of attempting to answer the following sub-questions:

1. Overall, does Polymarket as a prediction market support the notion of wisdom of the crowds, or is there evidence against this theory?
2. Does calibration quality vary across market domains?
3. Does calibration quality vary over time horizons (i.e., time to market resolution)?
4. How does accounting for the non-independence structure of markets affect calibration?

This thesis makes the following contributions. First, most prediction markets research predates entirely the arrival of modern global prediction markets (e.g., Polymarket, Kalshi) and is instead focused on much smaller-scale or limited projects (e.g. IEM, PAM). Polymarket is the world's largest prediction market, and represents a fundamentally different kind of market compared to those that came before—it is a decentralized, globally accessible, largely unregulated, cryptocurrency-based, dramatically higher liquidity market. To our knowledge, this work presents the first rigorous calibration study using historical data focused primarily on Polymarket across multiple domains and time horizons. Second, it tests whether the miscalibration patterns

found on previous platforms and in the relevant literature can be replicated in a new environment, and if so, to what extent. Third, it attempts to unify an understanding of how non-independent events may influence pricing.

Why does this research matter? The implications of this work have the most direct impact on prediction market consumers and traders who are using these platforms. The question of whether trading prices are accurately calibrated clearly has a large impact on those with financial "skin in the game," so to speak. In addition, these findings have implications for prediction markets themselves and the designers and developers responsible for improving market microstructure and their service offering. Finally, and perhaps most importantly, there are implications for those who fall into the broader category, namely those who may not be directly involved on the trading side or the service side (i.e., the general population). As prediction markets continue to gain global attention, millions of people now look to prediction market prices and interpret them as probabilities at face-value (due in part to prediction market advertising, but also by mainstream news reporting). Additionally, this research seeks to provide further evidence towards the question of under which conditions (if any) do crowds produce reliable probability forecasts. On the one hand, if miscalibration patterns are replicated, this strengthens the case that these such patterns may be fundamental features of how crowds reason and act under probabilistic uncertainty. On the other hand, if these patterns *don't* replicate, it suggests that some new combination of market microstructure, user demographic, trader interaction, or other factors present in prediction markets may affect price accuracy.

The remainder of this thesis is organized as follows: in Chapter 2, we review previous literature related to prediction markets and calibration. In Chapter 3, we discuss the methodology and data sources that will later be analyzed in Chapter 4. Finally, Chapter 5 will be used to discuss relevant findings and implications.

Chapter 2

Literature Review

2.1 Theoretical Foundations of Prediction Market Calibration

In the early 20th century, a weight-judging competition was held at the West of England Fat Stock and Poultry Exhibition. The rules of the competition were simple and were outlined as follows: attendees could sign up to estimate the weight of a slaughtered ox for a price of sixpence, with prizes available for those who guessed the closest. In Galton (1907)'s *Nature* publication, he investigated the sum total of all submitted estimates and came to the remarkable conclusion that the crowd's median estimate was within 1% of the true weight. This is the oldest documented case of "Vox Populi" (as coined by Galton), or what we more commonly refer to today as the "Wisdom of the Crowds"—the belief that aggregation mechanisms of crowds of non-experts can be surprisingly capable at producing accurate estimations amidst individually noisy judgments.

The conceptual foundation for this belief as it relates to financial markets comes to us from Hayek (1945) in his criticism of a centrally planned economy. Hayek argues that a central planner cannot possibly possess enough information to allocate

resources efficiently, and instead that a free market economy enables the efficient conveyance of decentralized information. This insight serves as the intellectual foundation of prediction markets: that if market prices are successful at aggregating dispersed information, then a standalone prediction market price can efficiently aggregate the beliefs of many traders about the true likelihood of a future event. In fact, Hayek's foundational hypothesis was explicitly cited by the creators of the IEM, the first instance of prediction markets as discussed above.

A further theoretical foundation of these markets is taken from Fama (1970)'s seminal 1970 paper in which he outlines his theory of Efficient Capital Markets: roughly, that all available information is reflected in financial asset prices. There are three forms given for the level of market efficiency. The first is *weak*, that an asset's current price is a reflection of all past prices; the second is *semi-strong*, that an asset's current price reflects all public information; and the third is *strong*, that an asset's current price reflects all public *and* private information (Fama 1970). The relevant form for prediction markets is Fama's "semi-strong" hypothesis: if a prediction market serves as a truly reliable aggregator of public information, then no publicly available poll, model, or analysis should be able to improve on the market-generated price as a forecast. There has been much debate on the subject of Fama's Efficient Capital Markets Hypothesis. Central to this debate is later work by Grossman and Stiglitz (1980) that presents a core paradox: perfectly efficient markets are a logical impossibility. On the one hand, if prices already perfectly reflected all information, then there would be no financial incentive to spend resources gathering information. On the other hand, if nobody collects information, then asset prices can't reflect it. This provides the beginnings of a theoretical opening for the later discussion on prediction markets miscalibration: Grossman and Stiglitz (1980) argue that some degree of pricing error is not only expected but necessary for markets to function.

The theoretical cases from Hayek (1945) and Fama (1970) suggest that prediction

market prices *should* reflect true probabilities. Wolfers and Zitzewitz (2004) provide a canonical overview of prediction markets in which they address directly the gap of *when* this interpretation holds. They identify three contract types—namely winner-take-all, index, and spread—and show that the price of a winner-take-all contract (the type Polymarket and other modern prediction markets use that is relevant to this work) is "essentially a state price, which will equal an estimate of the event's probability under the assumption of risk neutrality" (Wolfers and Zitzewitz 2004). They argue that traders' utility functions are linear over the relevant range of prices, and that the equilibrium price (i.e. the current trading price) approximates the mean of the belief distribution across dispersed traders. Furthermore, they formalize the methods through which prediction markets achieve this accurate aggregation: via incentivizing information seeking, via incentivizing truthful revelation of said information, and via an algorithm for collecting disparate information into a single market. These formalizations and assumptions are important to our work in that they clarify some of the conditions under which calibration can be expected to hold.

In the period immediately following this theoretical grounding, prediction markets began to attract wider attention due to their alleged power of accurate prediction (there will be no foolish claims made on the causality vs. correlation question here). Surowiecki (2004) published the now famous book *The Wisdom of Crowds*, in which Surowiecki laid out four conditions for crowd "wisdom:"

1. *Diversity of opinion*: Each person has access to private information (even if that private information is simply an eccentric interpretation of the known facts).

2. *Independence*: A person's opinions are not determined by the opinions of those around them.

3. *Decentralization*: People are able to specialize and draw on local knowledge.

4. *Aggregation*: Some mechanism exists for turning private judgments into a collective decision.

By design, prediction markets are designed to satisfy each of these four conditions; as such, this work served as an instrumental introduction of prediction markets to the popular conversation. In addition, also around this time, a hodgepodge of notable economics researchers released a letter titled "The Promise of Prediction Markets" arguing for broader adoption of prediction markets and certifying them as a "potent research tool " (Arrow et al. 2008).

This theoretical optimism surrounding the power of prediction markets is perhaps most apparently realized in research surrounding political prediction markets. In other words, according to multiple sources of empirical evidence, prediction markets have historically been strong at accurately predicting political elections. In fact, some would argue, this has been the singular most valuable contribution made by prediction markets to date. For example, Berg, Forsythe, et al. (2008), in a study spanning the 1990s and early 2000s on the Iowa Electronic Markets, find that elections markets accurately predict outcome over polling 74% of the time, with an average error of 1.33% on election eve. Rothschild (2009), in a study of 2008 U.S. elections on InTrade, finds that "debiased prediction market-based forecasts provide more accurate probabilities of victory and more information than debiased poll-based forecasts." More recently, Cutting et al. (2025), in a study of the 2024 U.S. Presidential election on Polymarket, find that prediction market prices inferred as probabilities were superior to polls (especially in swing states), showcasing the Wisdom of the Crowds in action with large volumes. Other studies with similar directional findings can be cited here too.

Importantly, while this empirical evidence supports the above outlined theoretical case of an accurate track record for prediction markets, this evidence tends to concentrate in narrow settings. For example, there is outsized research on political election markets as well as smaller and/or now defunct prediction markets. The empirical

evidence in this segment does *not* necessarily support the questions of interest to this study: whether prices are calibrated throughout the entire life of a market, across multiple domains, and on large modern prediction markets.

2.2 Empirical Evidence and Challenges to Prediction Market Calibration

While much of the theory and limited empirical evidence discussed above supports the case for the accurate calibration of prediction markets, there are some key points that these theoretical arguments assume, including but not limited to: approximately risk-neutral traders, an efficient market mechanism, and prices that can be interpreted as probabilities. While the theoretical case for calibration is strong, the reality is more complicated; there is more ambiguity surrounding the accuracy of prediction market pricing that calls into question this theoretical basis. New and more in-depth evidence, in collaboration with theoretical arguments that contrast with those outlined above, appear to reveal that miscalibration in prediction markets does not follow random noise patterns, but instead is better explained by a more systematic nature. Understanding these patterns is essential for the calibration analysis of the Polymarket prediction market that follows.

One of the most direct theoretical challenges to the perceived price-as-a-probability interpretation of prediction markets comes from Manski (2006). In particular, he opens by cautioning against this interpretation by stating: "the arguments for this interpretation have been imprecise" and continues by calling out the vague nature of Hayek (1945)'s hypothesis of information aggregation. Whereas Wolfers and Zitzewitz (2004) argue that prediction market prices successfully approximate the mean belief of traders (as described above), Manski (2006)'s key finding shows that even under the same standard risk-neutral trader preference assumption, "the equilibrium

price of a prediction-market contract is a particular quantile of the budget-weighted distribution of traders' beliefs" (i.e. not the mean). What this distinction means in practice is that prices are compressed to the \$0.50 threshold level. Near this threshold, prices are much less informative about traders' true beliefs: there is a tendency for traders to have true beliefs higher (lower) than the price when the price of a contract is above (below) \$0.50. Consequently, this suggests prices near the ends of the spectrum (i.e. \$0.00 or \$1.00) are in fact very informative about the average beliefs of traders. This is precisely the pattern of favorite-longshot bias that we will discuss below.

The Favorite-Longshot Bias (FLB) as predicted by Manski (2006)'s model has explanations in both market research and human psychology. But what is the FLB? The FLB refers to a well-studied phenomenon in betting markets (originally horse racing) in which "longshots"—those with true low probabilities of success—are regularly overpriced, while "favorites"—those with true high probabilities of success—are underpriced (Ali 1977). Ottaviani and Sørensen (2008) provide a comprehensive overview of seven theoretical mechanisms that may contribute to the FLB. Of these seven distinct explanations, the "heterogeneous beliefs" option is the most relevant to prediction markets: that market probabilities are less extreme than median beliefs. In other words, with classic constraints on traders (including risk-neutral preferences, heterogeneous beliefs, and limited budgets), the market cannot accurately reflect the strength of an individual trader's belief and so the market price ends up being below the true sum total of trader beliefs. This occurs because prices in a prediction market are determined not only by individual traders' beliefs, but also by their ability to act on these beliefs. As a result, wealth constraints on traders with high conviction may allow a price signal to be outweighed by a collective of traders with moderate beliefs, and therefore extreme beliefs are effectively "diluted" in the aggregation process. In addition to this market dynamics explanation for FLB, Kahneman

and Tversky (1979) offer a complementary explanation related to human psychology. Formalized in their "Prospect Theory" paper, they find that individuals will systematically overweight small probabilities and underweight large ones. This produces an S-shaped distortion of perceived probabilities, a finding particularly relevant to prediction markets calibration curves (where true outcome rate is plotted vs. price as an implied probability). Importantly, these two disjoint explanations present the same qualitative result: that prices on favorites will be too low, prices on longshots will be too high, and the calibration curve relating these discrepancies will exhibit a characteristic S-shape.

Page and Clemen (2013) provide a rigorous pre-modern study of prediction market calibration and encounter findings in support of the FLB. They analyze over 500,000 transactions across 1,787 markets on InTrade from 2002-2007 using both non-parametric and parametric methods to analyze the relationship between market prices and realized event outcomes. There are several key findings worthy of being addressed here that are directly related to the Research Questions outlined above. First, there is clear evidence in support of the FLB: "A price of \$0.20 is on average associated with a relative frequency of 15.3%; conversely, a price of \$0.80 is on average associated with a relative frequency of 87.4%." Second, there is clear evidence in support of miscalibration across different domains. In particular, "Market prices for political events show a very strong longshot bias, while the other markets are only slightly miscalibrated." For political markets, a price of \$0.20 corresponds to an even lower frequency of 10.9% (as compared to average across all domains), while a price of \$0.80 corresponds to an even higher frequency 92.8%. Third, Page and Clemen (2013) claim to be "the first to take into account the time dimension when modelling the equilibrium price in a prediction market." When investigating the novel time horizon, they find that prices are more miscalibrated as the time to contract expiration (i.e., event resolution) increases. Furthermore, the authors employ a key innovation

of "clustered bootstrapping" that will be instrumental to later developments in this thesis: they are the first to directly account for the the non-independence of contracts whose outcomes are linked when attempting to gauge market calibration (e.g. for the event "Who will win the presidential election?"). The study by Page and Clemen (2013) provides a strong overall empirical baseline for prediction market calibration research that our work extends. However, this research is outdated: the prediction market InTrade is now defunct, domain coverage was smaller, and volume was orders of magnitude smaller than Polymarket today.

Restocchi, McGroarty, and Gerding (2019) extend the work of Page and Clemen (2013), specifically on the time horizon front. They are the first to present a complete temporal study of prediction market mispricing as a function of market duration, using data of daily prices from $3,363$ political markets on PredictIt over the course of two years from 2014-2016 (Restocchi, McGroarty, and Gerding 2019). Their key findings include two important (and slightly counterintuitive) patterns. First, longer duration markets correspond to lower evidence of the FLB: markets lasting more than 50 days exhibit on average a Φ of 0.008 (where Φ is a cumulative FLB measure introduced by the authors), while this value jumps to 0.038 for markets lasting 10 days or fewer. Second, this dynamic *fundamentally changes* during the last 24 hours of trading, where the FLB becomes larger the longer the market duration (i.e., longer duration markets have bigger FLB spikes near resolution). This is a finding that directly contrasts with Page and Clemen (2013)'s previous conclusion that calibration improves monotonically as expiration approaches. These key findings, namely that the duration of a market can determine prediction market effiency patterns, are highly relevant to the later work of this thesis. However, similar to the limitations faced by Page and Clemen (2013) above, PredictIt is a small prediction market with very low volumes and only offering trading on a single political domain (politics).

The most comprehensive analysis of prediction market calibration to date, by

far, is undertaken by Le (2026) in February of this year. Le analyzed 292 million trades across over 327,000 contracts primarily on prediction market Kalshi (and using Polymarket for cross-validation). This study provides both a sound methodological template and the most direct point of comparison for the current study. Le (2026)'s primary contribution is a logistic recalibration approach, where a slope parameter b is estimated in logit space ($b = 1$ represents perfect calibration; $b > 1$ represents underconfidence in the market; $b < 1$ represents overconfidence). Upon doing so, the author finds that prediction market calibration can be decomposed into four components that together explain 87.3% of calibration variance:

1. *Universal Horizon Effect*: All domains share a tendency toward underconfidence (where prices are compressed towards 50%) at long time horizons.

2. *Domain-Specific Structural Biases*: Separate from the universal effect described above, this component reveals sharply different intercepts (biases) across domain area.

3. *Domain-by-Horizon Interactions*: This component constitutes the largest explanatory variable of explained variance, confirming that domains have fundamentally different calibration trajectories over time.

4. *Trade-Size Scale Effect*: For political markets, large trades cause greater underconfidence in the market. To explain, Le introduces the "bilateral cancellation hypothesis," whereby political markets attract partisan traders with strong large bets that partially cancel each other out.

Several limiting factors are important to consider when interpreting the results of Le's work and its relation to this thesis. First, this study showed that there must be some key distinctions between the prediction market platforms of Kalshi and Polymarket, due to the fact that Le's identification of the Trade-Size Scale Effect on Kalshi is

not replicated on Polymarket. Second, Le (2026)'s cross-validation with Polymarket only covers three of six domains that are available for comparison (namely Politics, Sports, and Crypto but excludes Weather, Entertainment, and Finance). Third, Le (2026) explicitly states that no consideration has been given to the non-independent contract structure on Polymarket. Fourth, the study uses a conservative price trimming approach, which potentially obscures the behavior at extreme prediction market prices, though previous research has shown that extreme prices are where the most documented biases exist (Rothschild 2009). Finally, and perhaps most importantly, Le treats Polymarket only as a secondary validation platform on the primary Kalshi analysis.

The calibration literature has progressed from first documenting the FLB in horse racing (Ali 1977), to providing formal theoretical explanations for it (Manski 2006; Ottaviani and Sørensen 2008), to establishing it empirically in prediction markets (Page and Clemen 2013), to characterizing its temporal structure (Restocchi, McGroarty, and Gerding 2019), and most recently to decomposing it into structured components that vary by domain, time horizon, and trade size (Le 2026). At each stage, the evidence has grown more detailed and the markets studied have grown larger. However, to our knowledge, no study has conducted a dedicated, multi-domain, temporal calibration analysis of Polymarket, one of the world's largest prediction market by volume. This thesis aims to address that gap.

Chapter 3

Methodology and Data

3.1 Methodology

This study asks whether Polymarket prices are well-calibrated probability forecasts. It has been established that the Favorite-Longshot Bias, domain heterogeneity, horizon dynamics, and market structure are the key empirical means that will enable a comprehensive review of the research questions posed in Chapter 1. In order to test these questions on Polymarket, it is necessary to (a) formalize a definition of calibration, (b) identify measures that can determine miscalibration across domains and time horizons, and (c) define analytical methods that account for the specific structure of Polymarket data. The rest of this section provides each in turn.

3.1.1 Defining Calibration

First, we define what it means to say that a prediction market price is "well-calibrated." Assume ω represents some unknown future state of the world, drawn from all possible future states of the world Ω. Let y_i represent the final outcome value of a given contract for a binary market i, i.e. $y_i \in \{0, 1\}$. Let p_i represent the market price of the same market i, with $p_i \in [0, 1]$. Then, for a market price to be "well-calibrated,"

we have that

$$p_i|\omega = \mathbb{E}[y_i|\omega]$$

This cannot be tested directly, since $\mathbb{E}[y_i|\omega]$ deals with future states of the world that cannot be known. However, because of prediction market design and that each contract is linked to the binary resolution of a future event, there is a probabilistic interpretation at play here. For a prediction market to be "well-calibrated," it follows that among all contracts trading at price p, it is expected that the event occurs with empirical frequency p. For example, a contract trading at \$0.70 resolves to "Yes" approximately 70% of the time, a contract at \$0.20 resolves to "Yes" approximately 20% of the time, and so on across the full price range. Alternatively, if prices depart systematically from the probability of the event occurring, this is evidence of prediction market "miscalibration."

3.1.2 Measuring Calibration

There are various methods employed to measure the level and shape of this calibration: (1) reliability diagrams, (2) Brier scores, and (3) logistic recalibration.

The first method relies on the basic concept of signed error. Signed error is the quantity $(p - y)$ that represents the difference between current market price p and actual outcome y. For example, if a market is currently trading at price $p = \$0.70$ and the actual outcome is $y = \$1.00$, the signed error is $(p - y) = -0.30$. By itself, this statistic has limited interpretive value on its own; however, the mean signed error MSE[1] is far more useful for analyzing probability forecasts and determining calibration. It is given by

$$Mean\ Signed\ Error\ (MSE) = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)$$

[1]MSE will refer to Mean Signed Error for the remainder of this paper, *not* the traditional Mean Squared Error.

where N is the number of price observations, p_i is the ith observed price, and y_i is the ith outcome. For an $MSE > 0$, prices systematically overstate true probabilities of outcomes (i.e. the market is overconfident). For an $MSE < 0$, the opposite is true. By using a non-parametric binning approach, daily prices can be grouped within a given price range (bin) that is a subset of the full price spectrum $[0, 1]$. Then, after computing the average realized outcome frequency within each bin, the actual outcome rate is plotted on the y-axis (y) against price on the x-axis (p), per bin. This creates "reliability diagrams" (also known as "calibration curves"), which provide a model-free estimate of outcome frequency across the full range of prices (i.e. the implied probability range). If the prediction market is well-calibrated, we expect the plotted points to fall along the 45-degree line $y = x$ of "perfect calibration." By comparing empirical data to this reference line, it provides meaningful insights into where miscalibration deviations may exist.

The second method relies on the concept of Brier scoring. The Brier score was originally devised as a method of verifying the accuracy of weather forecasts (Brier 1950), but in the context of this paper it is defined as follows

$$Brier\ Score\ (BS) = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

where N is once again the number of price observations, p_i is the ith observed price, and y_i is the ith outcome. This score is designed to measure the accuracy of probabilistic predictions, which for prediction markets is the mean squared error interpretation of pricing vs. outcome. As a simple example, the Brier score for predicting a fair coin flip would be 0.25. The Brier score is closely related to the mean signed error as discussed above, but differs in that the BS measures **magnitude** of error (with a squared penalty term) while the MSE measures **direction**. Importantly, the BS can be decomposed into three additive components, as defined by Murphy

(1973): (1) Reliability (REL), (2) Resolution (RES), and (3) Uncertainty (UNC), each of which can add valuable information to the score. The Brier score decomposition is defined as

$$BS = \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{p}_k - \bar{y}_k)^2}_{Reliability} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2}_{Resolution} + \underbrace{\bar{y}(1-\bar{y})}_{Uncertainty}$$

where N is the total number of price observations, K is the number of bins, n_k is the number of price observations in bin k, $\bar{p}_k$ is the mean price in bin k, $\bar{y}_k$ is the mean outcome in bin k (the "base-rate"), and $\bar{y}$ is the overall base rate. "Reliability" (also called "calibration loss") measures how far the average price (predicted probability) in each bin is from the average outcome. Counterintuitively, lower is better, where the floor of zero represents perfect calibration. "Resolution" measures how much outcomes vary across bins as compared to the overall base rate, i.e. how "sharp" the market is. Higher is better. Finally, "Uncertainty" is a property of the outcome distribution that represents an irreducible task difficulty; it is simply the variance of a Bernoulli random variable with a probability of success equal to the base rate. Taken alone, the reliability and resolution decompositions can be insightful into where the strength (or weakness) of forecasts lies. An additional final metric that is relevant to the Brier scoring methodology is the Brier Skill Score. Instead of simply knowing the mean squared error, it is often useful to know how well forecast performance is compared to the base rate benchmark. It is defined as

$$Brier\ Skill\ Score\ (BSS) = 1 - \frac{BS}{UNC}$$

where BS is the Brier Score as above and UNC is the Uncertainty component. The interpretation here is that a BSS of 0.20 represents a 20% reduction in the squared error relative to a naive forecaster who simply predicts the historical base

rate (or equivalently, a 20% improvement over the benchmark).

The third and final method is logistic recalibration. The theoretical background for this approach originally comes from the fields of behavioral economics and Kahneman and Tversky (1979)'s Prospect Theory (from above). Specifically, we adapt the Lattimore, Baker, and Witte (1992) probability weighting function, combining the approaches of Page and Clemen (2013) and Le (2026). The Lattimore function is a parametric function that describes how people psychologically weight probabilities when making risky decisions (whose key findings also provide support for the Favorite-Longshot Bias). Page and Clemen (2013) adapt this function in the context of prediction markets by assigning Lattimore, Baker, and Witte (1992)'s "objective probability" to true prediction market outcome and "subjective (perceived) probability" to observed prediction market price (assuming prediction market prices are reflective of the collective perceived probability of the event occurring). Page and Clemen (2013)'s adapted Lattimore function maps true probability π to observed market price p by the following:

$$p = \frac{\delta \pi^{\gamma}}{\delta \pi^{\gamma} + (1 - \pi)^{\gamma}}$$

The parameters γ and δ are estimated by maximum likelihood and correspond to the curvature and crossing point on a calibration curve, respectively. However, we follow the more natural approach of Le (2026) and invert the equation in logit space (because prices live on $[0, 1]$) to instead map observed price p to true probability π.

$$
\begin{aligned}
p &= \frac{\delta\pi^\gamma}{\delta\pi^\gamma + (1-\pi)^\gamma} \\
1-p &= \frac{\delta\pi^\gamma + (1-\pi)^\gamma - \delta\pi^\gamma}{\delta\pi^\gamma + (1-\pi)^\gamma} \\
1-p &= \frac{(1-\pi)^\gamma}{\delta\pi^\gamma + (1-\pi)^\gamma} \\
\frac{p}{1-p} &= \frac{\delta\pi^\gamma}{\delta\pi^\gamma + (1-\pi)^\gamma} \div \frac{(1-\pi)^\gamma}{\delta\pi^\gamma + (1-\pi)^\gamma} \\
\frac{p}{1-p} &= \frac{\delta\pi^\gamma}{(1-\pi)^\gamma} \\
\frac{p}{1-p} &= \delta\left(\frac{\pi}{1-\pi}\right)^\gamma \\
\log\left(\frac{p}{1-p}\right) &= \log\left(\delta\left(\frac{\pi}{1-\pi}\right)^\gamma\right) \\
\text{logit}(p) &= \log(\delta) + \gamma \cdot \log\left(\frac{\pi}{1-\pi}\right) \\
\text{logit}(p) &= \log(\delta) + \gamma \cdot \text{logit}(\pi) \\
\text{logit}(\pi) &= -\frac{\log(\delta)}{\gamma} + \frac{1}{\gamma} \cdot \text{logit}(p)
\end{aligned}
$$

Here, logit is defined as $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. This yields Le (2026)'s logistic recalibration model, where $P(y_i = 1)$ replaces π as the true outcome probability to be estimated from data:

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

with $a = -\frac{\log(\delta)}{\gamma}$ and $b = \frac{1}{\gamma}$. We fit this model by maximum likelihood (using Nelder-Mead and L-BFGS-B optimization) over all N price observations to estimate slope and intercept parameters b and a, which allows us to ask the question: given the market price p_i (expressed as log-odds), what does the data say the true log-odds of the binary outcome y_i is? If the market is perfectly calibrated, this corresponds to parameters of $a = 0$ and $b = 1$. The slope parameter b is the primary quantity of

interest: when b strays away from perfect calibration, we learn that systematic market deviations in calibration exist. When $b > 1$, the market systematically compresses prices towards \$0.50; longshots are overpriced and favorites are underpriced (i.e. the market is underconfident and this is the classic FLB). When $b < 1$, the market is overconfident and prices are reflecting more extreme probabilities than outcomes warrant. The intercept parameter a represents the directional bias present across markets.

3.1.3 Analysis Dimensions

To address the research questions outlined in the previous chapter and complete a comprehensive measurement of calibration signals from the previously outlined methods, calibration is estimated not just in aggregate but across two primary dimensions: domain category and time-to-resolution (horizon). First, domain category: markets and corresponding price observations are grouped into 7 domains, namely Sports, Crypto, Politics, Culture, Finance, Weather, and a catchall Other category. Applying the methods above on a more complete per-domain basis will allow us to test whether domain-specific calibration patterns identified on Kalshi also appear on Polymarket across the full range of available domains (Le 2026). Second, horizon: price observations are grouped into 8 time-to-resolution ranges (horizon bands), namely 1-3d, 4-7d, 8-15d, 16-30d, 31-60d, 61-90d, 91-180d, and >180d. Once again, applying the methods above across multiple horizon bands will reveal whether calibration improves as resolution approaches, i.e. if the temporal patterns previously documented on prediction markets Intrade (Page and Clemen 2013) and PredictIt (Restocchi, McGroarty, and Gerding 2019) or the universal horizon effect on Kalshi (Le 2026) are upheld, and to what extent.

3.1.4 Addressing Non-Independence

An important aspect of any regression and statistical inference, particularly when it comes to constructing confidence intervals and standard errors, is the assumption of independent observations. However, this does not always hold on Polymarket and other prediction markets; many events contain a multi-market structure. For example, consider the event: "Who will win the U.S. Presidential Election in 2028?" This event cannot be captured in the normal strict binary framework required by prediction markets. As a result, prediction markets usually resolve this issue by creating multiple *markets* that are all clustered under the larger *event* umbrella. Thus, an individual market within this event might be titled: "Will Person X win the U.S. Presidential Election in 2028?" Each of the markets located within this event are **not** independent: if Person X does indeed win and the market resolves with a final price of \$1.00, **all** other markets within this event **must** resolve to \$0.00. These are not independent observations; rather, they are based on the same underlying information.

Treating these correlated markets naively as purely independent price observations would understate standard errors and overstate the statistical significance of any detected miscalibration. To address this, this analysis employs an event-level "clustered bootstrap" for all confidence interval estimation and resampling taken in logistic recalibration, following the approach taken by Page and Clemen (2013). A significant limitation identified in Le (2026)'s paper is the lack of this clustered bootstrap (or similar) for Polymarket data analysis. The clustered bootstrap approach specifically sets the unit of resampling to be the overarching event—not individual observations—when quantifying uncertainty estimates. By doing so, this research ensures the non-independence introduced by multi-market events is reflected throughout all reported confidence intervals. By comparing the width of clustered bootstrap confidence intervals against their naive counterparts, we clarify the cost of ignoring this

market clustering in prediction markets.

3.2 Data Source

Polymarket is a global[2] prediction market platform that allows users to trade on the outcomes of a myriad of real-world events including, but not limited to: sports games, traditional financial and economic indicators, political elections, weather, and cultural moments. Founded in 2020, Polymarket has grown rapidly to become one of the largest prediction markets in the world, where it is consistently ranked among the top two by trading volume (alongside rival Kalshi). As of the time of this writing, Polymarket has a 30d trading volume of over $4.5 billion, accounting for roughly 45% of the market share (*Prediction Markets Dashboard* 2026). The key differentiator that separates Polymarket from other prediction market platforms—both past and present—is that it is based on cryptocurrency and blockchain technology; all trading happens in a decentralized manner on the Polygon blockchain network using smart contracts and USDC. While this does not necessarily change the nature of the prediction market itself, most of the intricacies of Polymarket's cryptocurrency-based operation are outside the scope of this thesis.

3.3 Data Collection and Filtering

Detailed code for data collection, data cleaning, outputting results, etc. for this and following chapters can be found at the author's GitHub page.

[2]Despite being advertised as a "global" prediction market, Polymarket (at the time of writing) is currently blocked from operating in 33 countries worldwide, including the United States, the Netherlands, North Korea, and others. Polymarket US, a separate legal entity from the global Polymarket.com, was recently granted permission to operate in the United States on December 2, 2025 following a settlement with the Commodity Futures Trading Commission (CFTC) for operating an unregistered derivatives trading platform.

3.3.1 Polymarket API

Real-time and historical Polymarket data is available for free via the public Polymarket API. Aligning with much of the ethos of the broader cryptocurrency industry, Polymarket's GitHub is open source and contains useful setup instructions and example code for using the Polymarket API. The relevant components of the public Polymarket API are (1) the Gamma API, and (2) the Polymarket Central Limit Order Book (CLOB) API. These are described as follows:

1. Polymarket Gamma API: Polymarket's Gamma service provides a comprehensive suite of endpoints for browsing and discovering cleaned market data. This service is designed to be used by researchers and developers for "things like non-profit research projects, alternative trading interfaces, automated trading systems, etc." The Gamma API is used in this project for collecting valuable metadata for historical markets.

2. Polymarket CLOB API: Polymarket's CLOB service provides full access to order book data, pricing history, and trading operations. Polymarket operates on the traditional central limit order book exchange model, where buy and sell orders are matched by price and bids and asks are aggregated into a transparent live order book for each market. The CLOB API is used in this project for collecting all historical price data.

3.3.2 Pipeline

The scope of the data collection initially contains all available Polymarket markets with listed endDate on or before December 31, 2025. First, the Gamma API /events endpoint was called repeatedly with pagination to fetch all markets that fall in this range. Key metadata was collected for each market: the market ID,[3] market slug,

[3]Polymarket differentiates between "events" and "markets." An event is a question you may ask about the state of the world. For example, "Will the US confirm that aliens exist in 2025?" In this

event ID, event slug, market start date, market end date, market close time, market resolution status, market outcomes and outcome prices, token IDs,[4] tag IDs, and tag labels. Upon completion, this data was saved to a market metadata parquet file that contains 259,542 rows of markets. Second, the CLOB API /prices-history endpoint was called once per market to fetch all historical price data for each market, on a daily frequency with prices floored to midnight. This price data was then stored in a local DuckDB database, where each row corresponds to one price observation for one market (e.g. a single market with 365 days of price data will contribute 365 rows to the database). A parallel ingestion log handled errors and enabled resumable collection at this step.

Before analysis of this price data, a series of filters are applied to the raw price data to ensure data integrity for the final calibration dataset. First, markets without a valid resolution timestamp are dropped entirely: without a valid resolution date, no horizon calculation is possible and the final outcome is ambiguous [- 468 markets]. Second, markets whose resolution date[5] falls after December 31, 2025 are excluded, as these fall outside the desired analysis window [- 2,711 markets]. Third, markets with ambiguous resolution outcomes that do not align with the required binary framework (e.g. ties) are dropped [- 574 markets]. Fourth, markets with total trading volume below $100 are excluded based on the understanding that markets with very low volume are wholly unreliable [- 38,802 markets]. Fifth, markets with no CLOB price data are trivially removed, as well as markets with no listed token ID to filter on [- 28,478 markets]. Sixth and finally, markets that have price observations that fall

case, this is already a binary event, so it has only one corresponding market (of the same name). However, for the event "Who will become President of the US in 2028?" which is not binary, this will have a collection of associated binary markets. For example, without loss of generality "Will Elon Musk become President in 2028?".

[4]Due to the cryptocurrency smart contract mechanics for Polymarket markets, Polymarket must differentiate within a market by two unique token IDs, where one corresponds to the price of a "Yes" token for any given binary market and the other corresponds to the NO token. All future data analysis and results are completed using exclusively the "Yes" token.

[5]In some markets, the resolution date is different from the market close date. This is why some markets have to be excluded again from the cutoff date despite filtering for end date above.

outside the binary range $[0, 1]$ are excluded [-2 markets]. As a result of these data filters, the final calibration dataset consists of 188,509 markets and 1,875,570 total price observations.

3.3.3 Domain Classification and Horizon Bands

In order to investigate Research Question 2 (domain analysis), a custom tag-based classification scheme is implemented for Polymarket markets. Polymarket's native method for handling domain tagging is an exhaustive list of over 5,000 individual tags, of which any quantity can be assigned to an individual market upon inception. In order to reduce this to a meaningful subset for analysis, we rank the top 100 tag IDs by count present in the finalized calibration dataset. After inspection, these tags were found to fall (roughly) into 6 domains—namely: Politics, Crypto, Finance, Sports, Culture, and Weather—each of which was merged with minor related tag IDs (e.g. minor tag ID 235 = "Bitcoin" was appended to major tag ID 21 = "Crypto") to create a curated mapping of tag IDs to domain. Each market is then assigned to the first matching domain according to a fixed priority ordering: Politics → Crypto → Finance → Sports → Culture → Weather. Markets with no matching tag in the curated mapping are assigned to a catchall Other category. This ordering is important because many markets carry overlapping tags, but only one is selected for the final domain bucket. The ordering reflects a key design decision: the most thematically distinct and research-relevant domains are assigned first, preventing their markets from being absorbed into broader categories. Collectively, this mapping selection accounts for 98.3% of all markets (i.e only 1.7% of markets are classified into a catchall final Other domain). The full tag-to-domain mapping can be found in Appendix Table A.1.

In order to investigate Research Question 3 (horizon analysis), a horizon band approach is implemented for price observation data. This relies on the key computation

of days-to-resolution, where the integer number of days between each price observation and its market resolution date is computed for each price row in the finalized DuckDB calibration dataset. Observations are grouped into 8 horizon bands—namely: 1-3d, 4-7d, 8-15d, 16-30d, 31-60d, 61-90d, 91-180d, and >180d—with finer granularity at shorter horizons where both price observations are concentrated and calibration dynamics are of interest.[6]

3.3.4 Descriptive Statistics

Table 3.1 summarizes the final calibration dataset across the seven categorical domains, reporting market distribution, median duration and volume, outcome base rates, and the distribution of price observations.

Table 3.1: Domain Summary Statistics

Domain	Markets	% Markets	N Multi	Med. Days	Med. Volume	Base Rate	Price Obs	% Obs
Sports	89,207	47.32	72,684	4	$5,631	0.39	817,789	43.60
Crypto	62,188	32.99	18,319	2	$18,242	0.43	234,032	12.48
Politics	16,873	8.95	13,877	8	$32,953	0.28	525,259	28.01
Weather	6,164	3.27	6,102	3	$8,191	0.15	25,556	1.36
Culture	5,655	3.00	5,168	8	$18,894	0.18	118,197	6.30
Finance	5,272	2.80	3,613	8	$6,046	0.42	79,120	4.22
Other	3,150	1.67	2,089	7	$12,313	0.32	75,617	4.03
Overall	188,509	100.00	121,852	3	$12,031	0.38	1,875,570	100.00

N Multi = number of markets belonging to multi-market events; Med. Days = median days to resolution; Med. Volume = median market volume (USD); Base Rate = fraction of markets resolving "Yes;" Price Obs = total price-day observations.

There are several key things to note on the above summary statistics. First, market composition. Sports and Crypto domains together account for over 80% of markets,

[6] Observations with days-to-resolution = 0 are excluded from all horizon analyses at runtime. This is an attempt at reducing the impact that very low time-to-resolution markets have on calibration analysis.

but have the shortest median durations. This is likely explained by the prevalence of short duration Sports games markets and even shorter duration Crypto markets (e.g. "Will the price of Bitcoin be up or down in 15 minutes?"). Despite making up under 9% of markets, the Politics domain contributes 28% of price observations, reflecting its much longer median duration. Second, multi-market structure. Weather stands out in this category, where 99% of markets have a multi-market structure. Politics also has a very high multi-market fraction, consistent with the election example presented previously. This sets up why clustering matters significantly, discussed later. Third, volume. Politics has the highest median volume by a significant margin, reflecting intense trader interest in political prediction markets and carrying implications for liquidity and price calibration. Fourth, base rates. All domains have a market-level base rate below the coin-flip reference of 0.50 (this is expected as a result of multi-market structure). Weather and Culture in particular have strikingly low base rates relative to Sports, Crypto, and Finance. This matters for cross-domain Brier score comparisons due to the lower irreducible uncertainty.

Figure 3.1 presents a six-panel exploratory overview of the final calibration dataset.

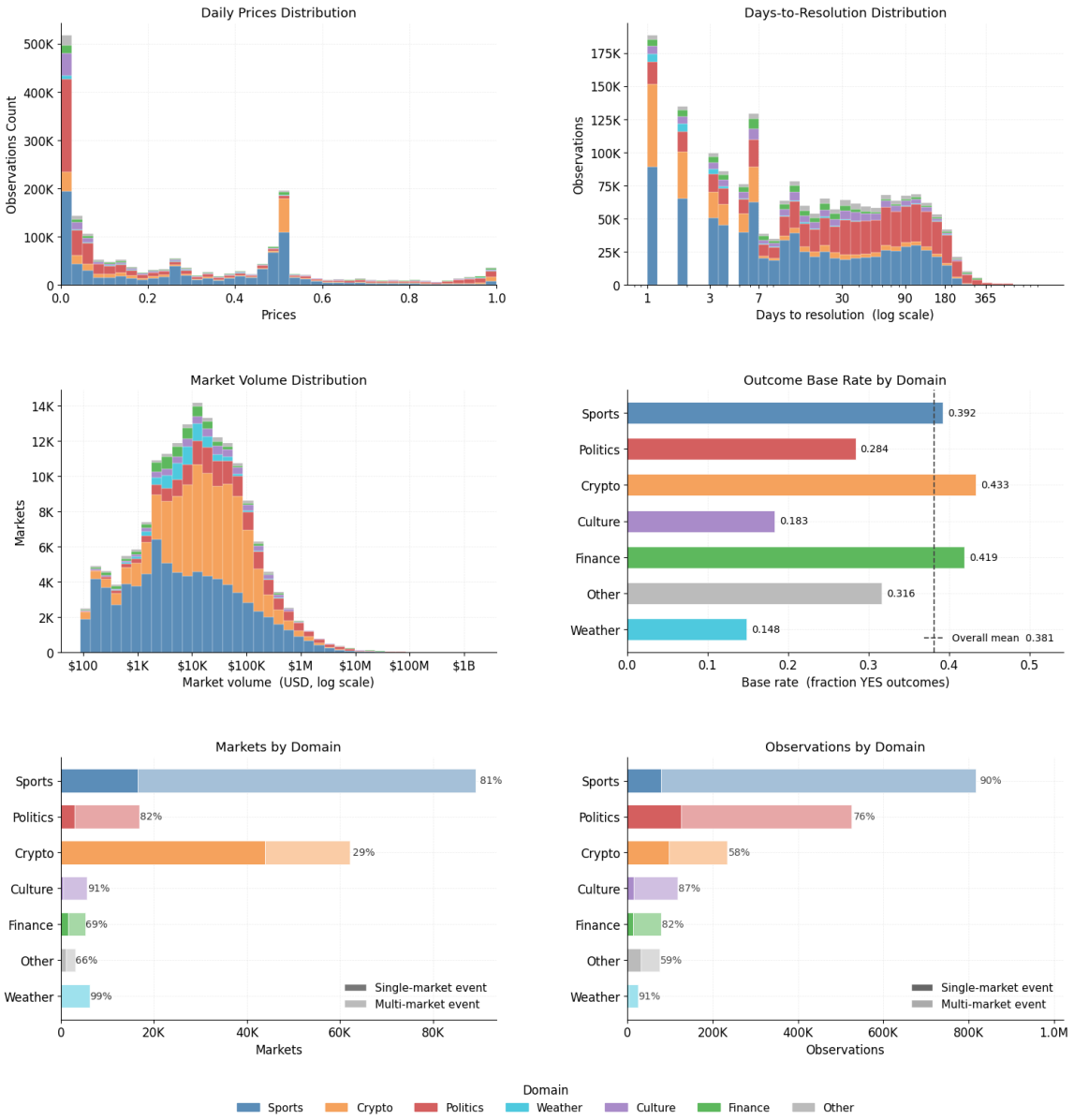

Figure 3.1: Exploratory overview of the final calibration dataset. Top row: distribution of daily closing prices (left) and days-to-resolution (right). Middle row: market volume distribution in USD (left) and outcome base rate by domain (right). Bottom row: market count (left) and price-day observation count (right) by domain, segmented by single-market and multi-market event structure. All panels are colored by domain.

From the figure, we identify key findings in the exploratory data analysis and present them panel by panel. For the daily price distribution chart (top left), there is an extremely heavy skew towards the 0.00 side of the range $[0, 1]$. There is also

evidence of a trimodal pattern, with peaks at 0.00, 1.00, and 0.50. This near-zero and near-one concentration reflects the dominance of near-resolution observations in the dataset, with an emphasis on multi-market events that can only resolve to 1.00 for one contained market. The sizable spike near the 0.50 mark is mostly made up of Sports and Crypto markets, and likely reflects the high concentration of Sports games or Crypto up/down markets that sit near the uninformative prior before information arrives.

For the days-to-resolution distribution chart (top right), the log scale shows that the majority of observations are within short ranges. In addition, Sports and Crypto price observations dominate at shorter horizons, while Politics becomes much more prevalent at longer horizons. Observations thin out sharply after about 180 days.

For the volume distribution chart (middle left), it appears that volume follows a roughly log-normal distribution with a peak in the $\$10K - \$50K$ range. There is a long right tail extending to $\$1B$ for a small but important number of very-high volume markets. Note that volume is clipped below the \$100 threshold, representing the design decision made earlier in data processing.

For the base rate by domain chart (middle right), there is substantial deviation across domains, with Weather and Culture on the low end and Sports, Crypto, and Finance on the high end. Note that the overall mean appears skewed to the right because it is a market-level mean, and the histogram bins do not represent relative sizes of the domain categories.

For the market and observations by domain charts (bottom), the bars represent the number of markets and observations, respectively, contributed to each count by domain. Sports, Politics, and Crypto represent the largest domains by both market count and price observation count. While Politics has a relatively fewer number of markets (left), it punches above its weight in daily prices contribution. The light-shaded region on each of these bars measures the concentration of multi-market events

within domains. Outside of the sole metric of single-market Crypto events, *all* other metrics show that multi-market events are not just expected, but the norm. The relative fraction of multi-market events has direct implications later for how much the clustered bootstrap will affect each domain.

Chapter 4

Results

This chapter presents the empirical results of the calibration analysis outlined in Chapter 3, organized sequentially by each of the four research questions identified in Chapter 1. Section 4.1 examines overall calibration of Polymarket across the full cleaned dataset of 188,509 resolved markets and 1,875,570 daily price observations. Section 4.2 partitions this calibration across market domains. Section 4.3 analyzes the temporal evolution of calibration as a function of time-to-resolution. Section 4.4 presents the clustered bootstrap analysis and its implications for statistical inference. Taken together, the results reveal that Polymarket prices are broadly well-calibrated, but also have systematic deviations that vary across domains and time horizons consistent with the Favorite-Longshot Bias documented in prior research.

4.1 Overall Calibration

Research Question 1 asks about Polymarket calibration as a whole: overall, does Polymarket support evidence for a well-calibrated prediction market, thereby supporting the notion of wisdom of the crowds? Figure 4.1 presents the overall "Reliability Diagram" (also known as a "Calibration Curve") for Polymarket price data. This is based on a simple but powerful graphic common in the probability forecasting lit-

erature, whereby "true probabilities" are plotted against "predicted probabilities" to determine the accuracy of probabilistic forecasts. To apply this same concept to the prediction market framework, we take "predicted probabilities" to be bins of daily price observations—which are crucially interpreted as implied probabilities due to the prediction market $[\$0.00, \$1.00]$ price structure—and "true probabilities" to be the actual outcome rates associated with these binned prices.[1]

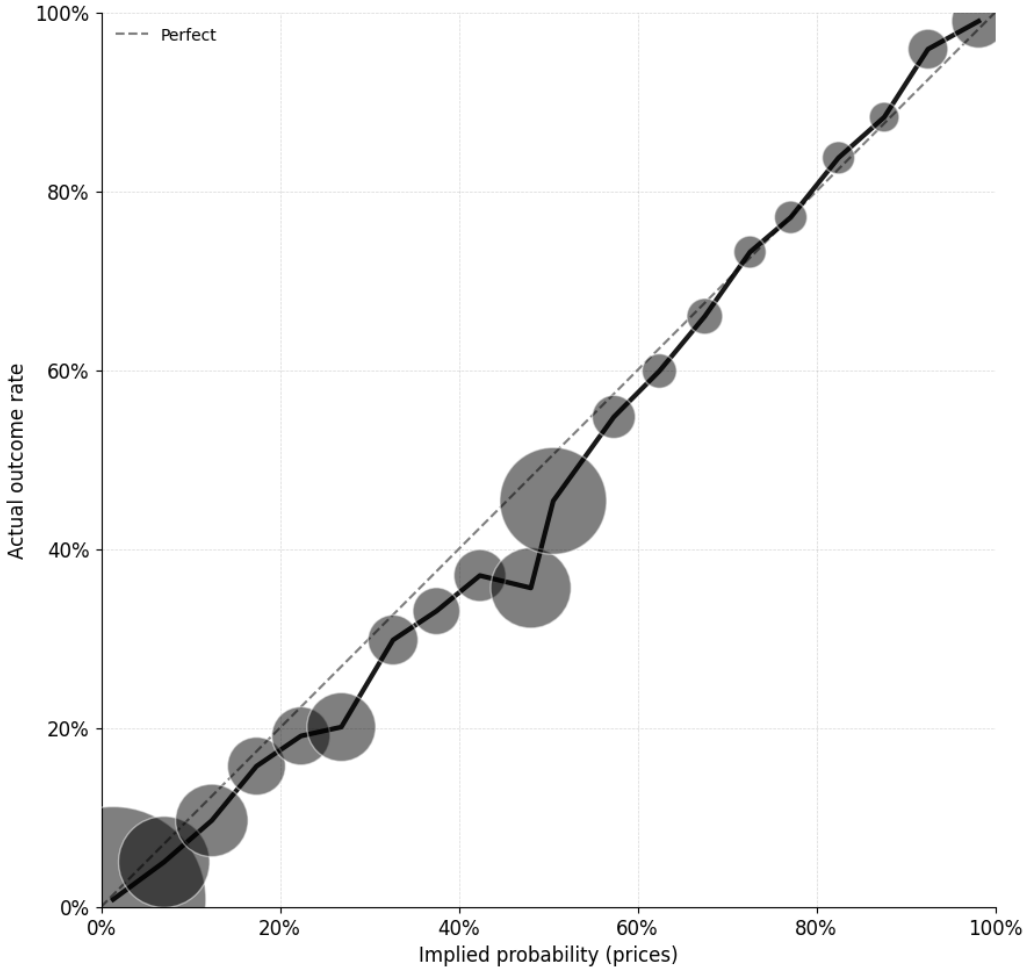

Figure 4.1: Overall Polymarket Reliability Diagram. All price observations are grouped into 20 bins, and mean outcome rates and prices are computed on a per-bin basis. Dot sizes encode the relative number of observation counts within each bin.

[1]The price observation database as a whole includes historical prices at every available day for every available market (e.g., a market with a 7-day duration contributes 7 individual rows and price points to the database, while a market with a 70-day duration contributes 70). Thus, longer duration markets are more heavily represented in the overall reliability diagram price bins.

The diagonal curve represents a perfectly calibrated forecast—that is, the observed price bin p corresponds to the same empirical frequency of occurrence p, as defined in Section 3.1.1. For the overall (i.e. *all* price observation data) Polymarket reliability diagram, we see that the calibration curve is shown to track the diagonal closely across the full range of implied probabilities, indicating Polymarket is broadly well-calibrated. Since dot sizes represent observation counts within each bin, we observe the skew in price observations toward the 0.0, 1.0, and 0.5 bins (as was found in Figure 3.1). Notably, there is a visible S-shape deviation within the diagram, where the curve sits below the diagonal at low prices and above the diagonal at high prices. This provides the first visual evidence of the FLB.

Figure 4.2 plots mean signed error per price bin, making the direction and magnitude of miscalibration explicit across the full price range.

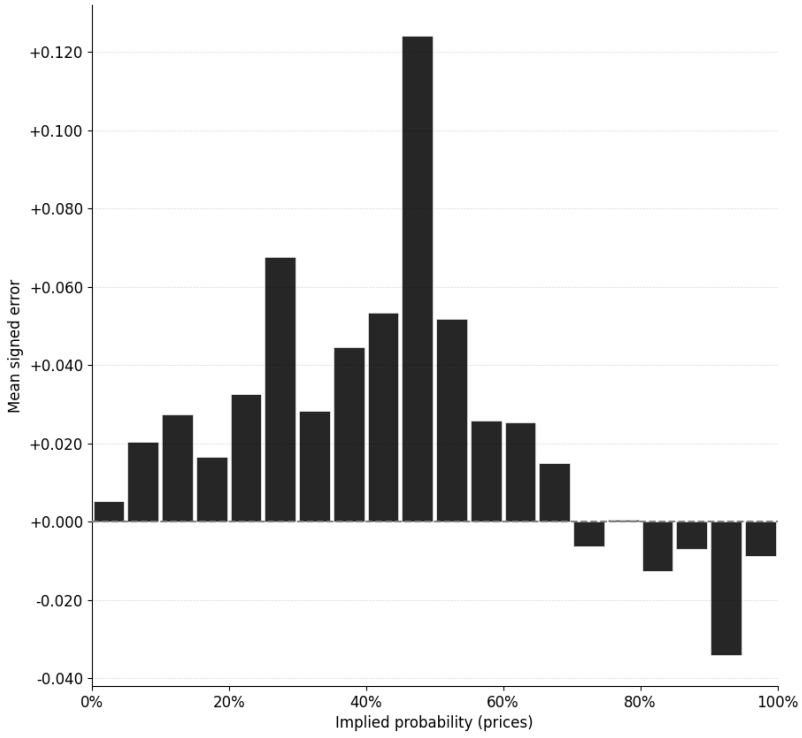

Figure 4.2: Polymarket Mean Signed Error by price bin. The height of the bar represents the difference between average observed price and actual outcome rate for a given bin.

Positive errors are a result of average price being **above** the actual outcome rate, or evidence of market overpricing. Negative errors are the inverse: the average price is **below** the actual outcome rate, which is evidence of market underpricing. There is a clear shape to the error bars, whereby positive errors dominate the $0.0 - 0.7$ range, and negative errors dominate $0.8 - 1.0$ range. There are a few anomalous spikes in the chart, namely around the 0.3, 0.5, and 0.9 price ranges, with the middle being the largest: this represents a high concentration of markets near the "coin-flip" threshold that systematically overprice outcomes on the order of 0.12.

Table 4.1 presents overall calibration statistics for Polymarket price data, including Brier Scores (BS), Brier Skill Scores (BSS), Brier Score decomposition, and logistic recalibration parameters.

Table 4.1: Overall calibration statistics for Polymarket (n = 1,875,570).

Block	**Statistic**	**Value**
Overall Accuracy	Brier Score (BS)	0.1100
	Base-Rate BS Reference	0.1827
	Brier Skill Score (BSS)	0.3980
	Mean Signed Error (MSE)	+0.0275
BS Decomposition	Reliability (REL) ↓	0.0019
	Resolution (RES) ↑	0.0744
	Uncertainty (UNC)	0.1827
Logistic Recalibration	Intercept (a)	−0.201
	Slope (b)	1.112

Investigation of the overall price forecasting accuracy of Polymarket reveals a Brier Score (the probabilistic interpretation of mean squared error) of 0.1100, which

represents a nearly 40% improvement over the naive base rate benchmark Brier Score (also the uncertainty in BS decomposition) of 0.1827. This is genuine forecasting skill. Looking more closely into the Brier Score decomposition, following the methodology set forth by Murphy (1973), we see that the reliability score of 0.0019 is extremely low (as a reminder—contrary to intuitive belief—lower is better, where REL = 0 represents perfect calibration) and thus the market is well-calibrated. For the resolution score (where higher is better), 0.0744 indicates that Polymarket meaningfully distinguishes outcomes into different probability levels. For the logistic recalibration fit, the parameter b is intended to be formal parametric evidence for market calibration (or a lack thereof). In particular, the slope $b = 1.112 > 1$ represents market underconfidence, following the classic FLB direction. The negative intercept $a = -0.201$ indicates a systematic upward bias in log-odds, meaning Polymarket slightly overestimates the probability of "Yes" outcomes on average. This aligns with the direction of the overall mean signed error (predicted - actual) being +0.0275.

Figure 4.3 shows observation counts for each bin in overall calibration analysis.

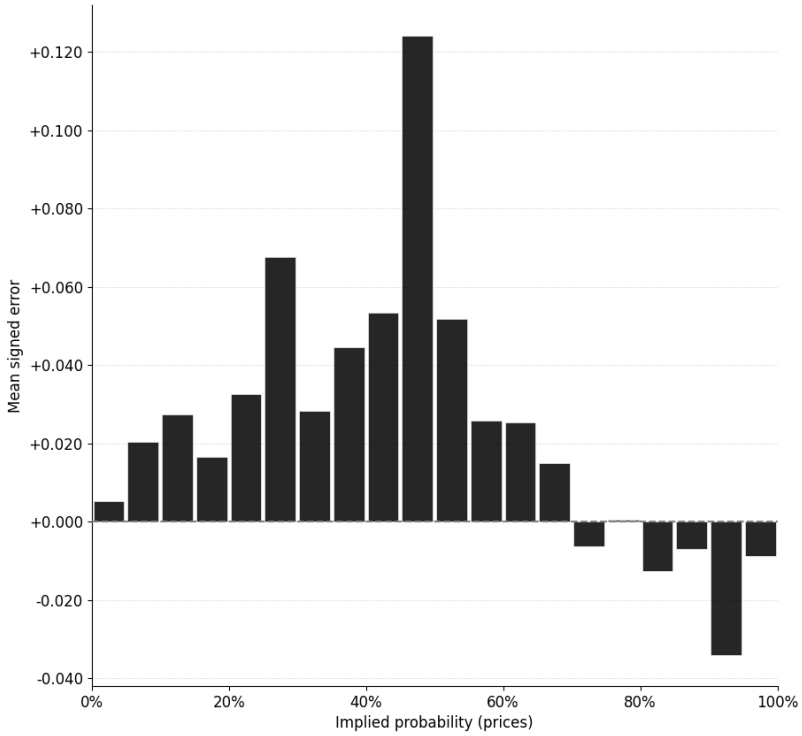

Figure 4.3: Number of observations per bin across the $[0, 1]$ price (predicted probability) range.

The key observation here is that the observation distribution is heavily skewed. This provides important context for the reliability diagrams, as well as for aggregate statistics like MSE and BSS which are disproportionately affected by (likely) near-resolution observations from multi-market events. This motivates the domain and horizon breakdowns that follow in the next sections.

4.2 Domain Analysis

Research Question 2 asks whether calibration dynamics change across market domain. In the following section, we test whether the aggregate calibration picture from Section 4.1 contains domain-level heterogeneity across seven identified domains: Sports, Crypto, Politics, Finance, Weather, and Other. Figure 4.4 presents the reliability diagrams for all seven domains (along with Overall for reference) overlaid on a single

plot for cross-domain comparison. Figure 4.5 then shows these reliability diagrams on per-domain subplots.

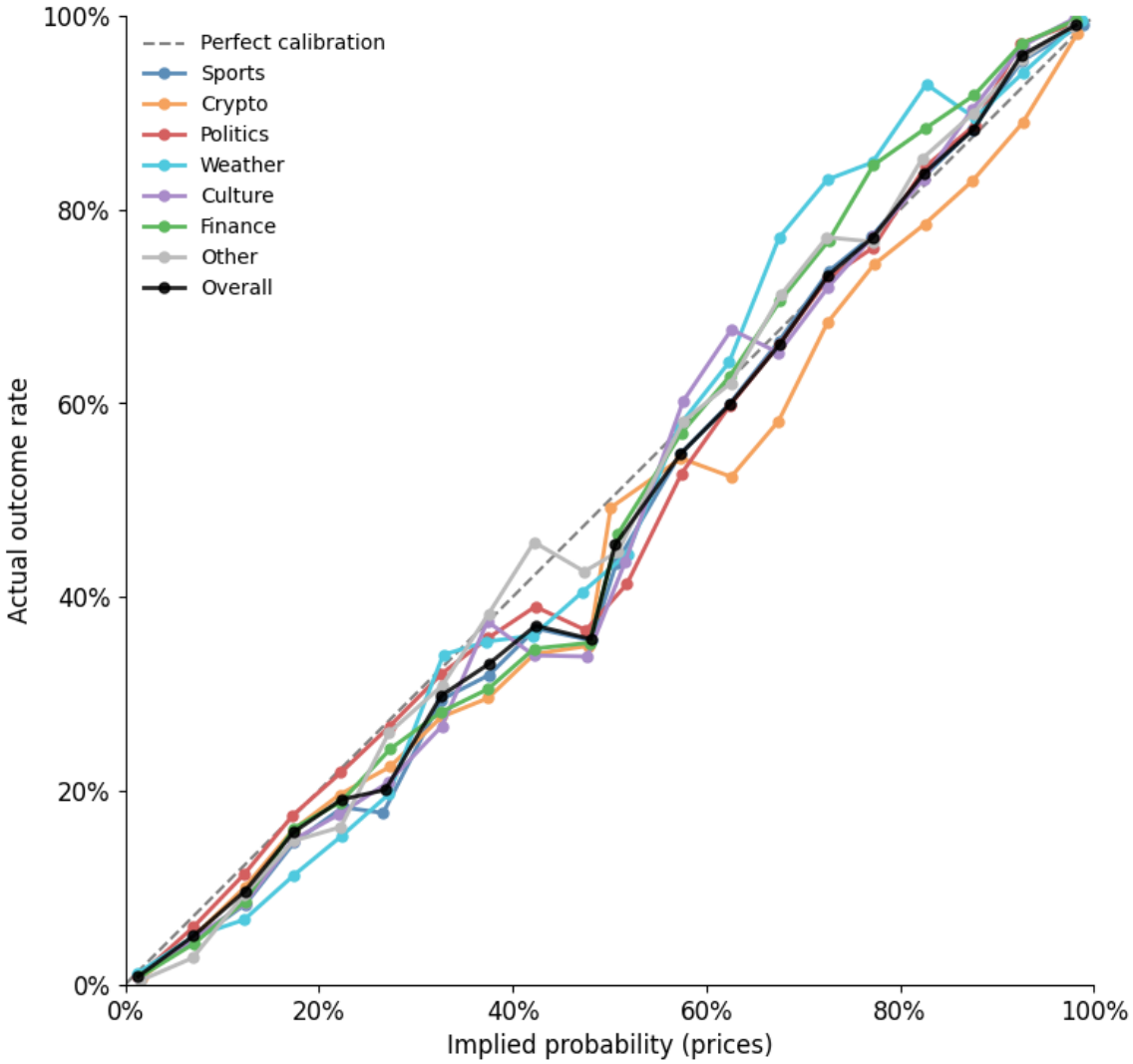

Figure 4.4: Overlaid reliability diagrams for all seven market domains and the overall curve. Each line connects the 20 binned mean outcome rates vs. average price. The dashed diagonal line represents perfect calibration.

.

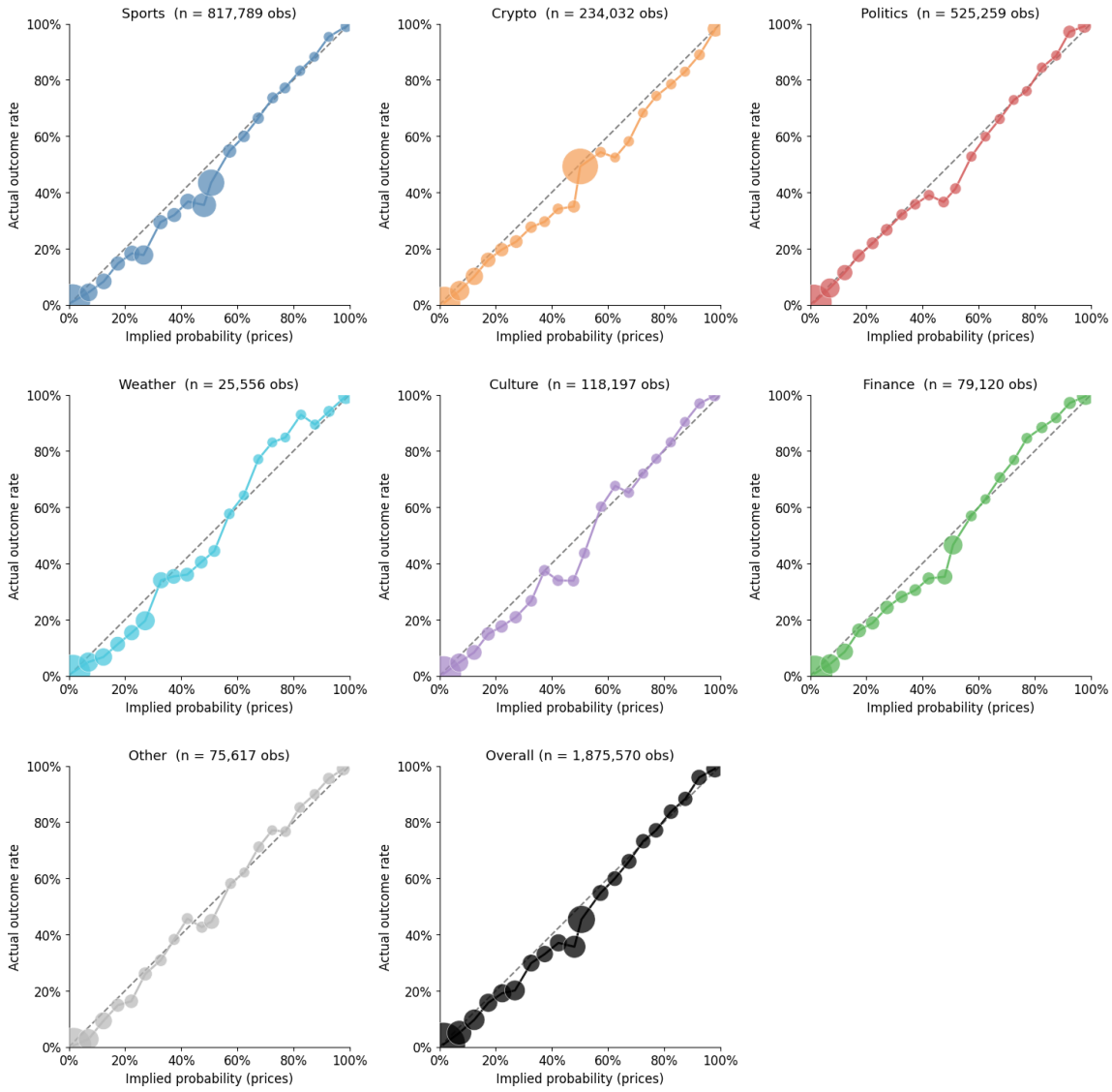

Figure 4.5: Per-domain reliability diagrams. Dot size is proportional to the number of observations in each bin.

.

There is clear visual spread across domains, confirming calibration is not uniform. For each market domain, there are clear abnormal patterns visible near the mid-range. Sports, Politics, and Culture markets appear most well-calibrated at the upper ranges of the price distribution. Besides Politics (which also hugs the diagonal at the lower ranges of the price distribution), the calibration curves of **all** other domains fall below the diagonal over approximately $[01 - 0.3]$. Crypto markets show

a brief spike near the mid-range, but otherwise are the only domain where prices (predicted probabilities) appear to be clearly universally overestimated across the full price range. The high concentration of observations near the mid-range is evident for Sports, Crypto, and to a lesser extent Finance domains. Finally, the Weather and Culture domains appear to be the largest culprits of the S-shaped calibration curve, where prices are overestimated for lower prices and underestimated for higher prices.

Table 4.2 summarizes the key calibration statistics for each domain. For a visual reference, Figure 4.6 compares the key statistics of mean signed error and Brier Skill Score, providing a visual ranking of directional bias and forecast skill, respectively.

Table 4.2: Domain-level calibration summary. Brier Score (BS) and Brier Skill Score (BSS) measure forecast accuracy. BSS is relative to a naive Base Rate Reference forecast (BR Ref.). Mean signed error (MSE) measures directional bias. Parameter b comes from logistic recalibration fitted by maximum likelihood.

Domain	**Base Rate**	**BS**	**BR Ref.**	**BSS**	**MSE**	***b***
Politics	0.211	0.073	0.167	0.565	+0.009	1.113
Crypto	0.312	0.137	0.215	0.361	+0.022	1.079
Finance	0.314	0.103	0.215	0.523	+0.018	1.221
Sports	0.245	0.136	0.185	0.264	+0.045	1.097
Culture	0.159	0.067	0.134	0.499	+0.019	1.180
Weather	0.206	0.089	0.164	0.459	+0.021	1.188
Other	0.244	0.087	0.184	0.526	+0.016	1.232
Overall	0.241	0.110	0.183	0.398	+0.028	1.112

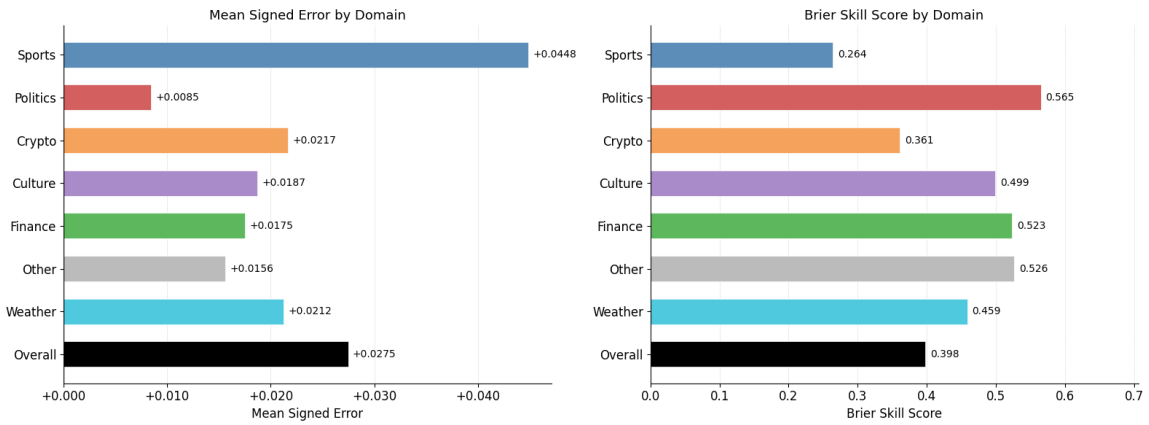

Figure 4.6: Left: mean signed error (predicted - actual) by domain; positive values represent systematic overpricing. Right: Brier Skill Score by domain; higher values indicate greater improvement over a naive per-domain base rate forecast. Both panels include the overall aggregate for reference.

.

Some key observations can be made from the table and figures above. Politics is the strongest performing domain, with an MSE of +0.0085 (which represents nearly zero bias on average) and BSS of 0.565 (which represents a 56% improvement over the base rate uncertainty reference). Sports, on the other hand, is the weakest performing domain (both MSE = +0.0448 and BSS = 0.264 well below the overall). It is important to note that raw Brier Scores are not the appropriate cross-domain reference statistic: they do not provide consideration of the differing base rates and therefore differing irreducible uncertainty quantification. By comparing to this metric, therefore, BSS provides the proper cross-domain metric for comparison. For example, though the Culture domain has the lowest Brier Score, it does not translate to the highest BSS due to the low base rate. Separately, the Finance and Other domains have the highest slope b parameter (representing the greatest S-shaped outcome to price distribution and therefore FLB), while Crypto and Sports have the lowest. The differences by domain here are not static; the next section will investigate how calibration evolves with time-to-resolution.

4.3 Horizon Analysis

Research Question 3 asks whether calibration dynamics change across horizons, i.e. by varying time-to-resolution. In the following section, we test whether miscalibration patterns identified in previous Sections 4.1 and 4.2 are stable across the market lifetime or concentrated at specific horizons.

Figure 4.7 presents reliability diagrams for each of the eight horizon bands (1-3d, 4-7d, 8-15d, 16-30d, 31-60d, 61-90d, 91-180d, and >180d). This provides a direct visual comparison of the calibration curve shape across the full range of market durations.

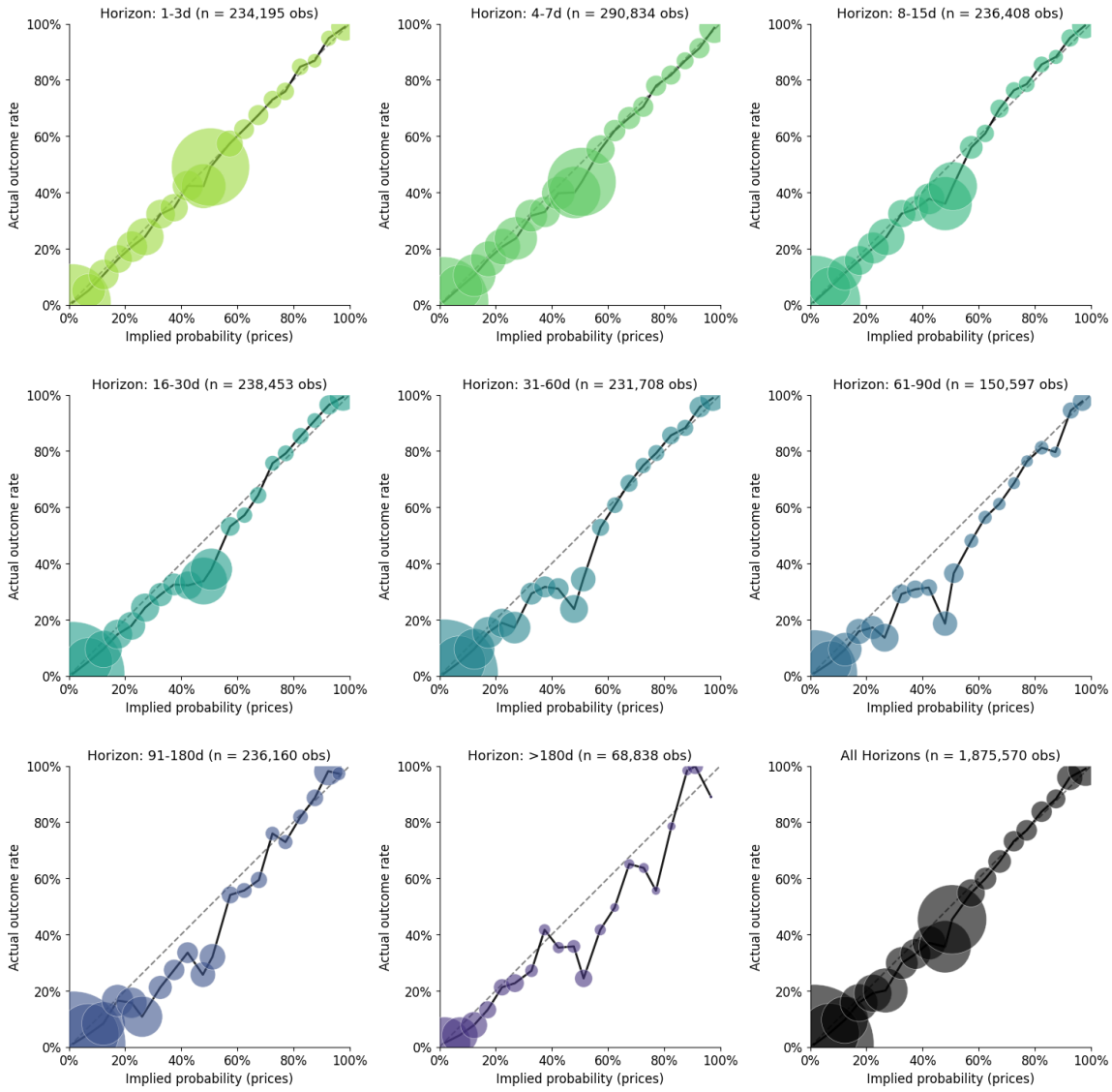

Figure 4.7: Reliability diagrams by horizon band, organized by nearest (top left) to furthest (bottom center) time to resolution. As before, dot size is proportional to bin observation count. The overall reliability diagram is shown for reference (bottom right).

Calibration is tightest in the short horizon bands (1-3d and 4-7d), as is to be expected for markets nearing resolution. As time-to-resolution increases, calibration gets progressively worse. However, there is clear evidence for the characteristic S-shape curve, even beginning at short horizon bands. There are large concentrations of price observations near the 0.50 mark for shorter horizon markets (see size of dots above), which gradually disappear as market time-to-resolution increases (which

based on previous domain analysis, likely comes largely from Sports and Crypto markets). For longer horizon bands (in particular 61-90d) prices appear to be chronically overpriced across a much larger range of the price spectrum. The >180d band is notably noisier, reflecting more sparse data available for observation at extremely long horizons (for prediction markets at least).

The following Table 4.3 provides key calibration statistics for each horizon band that will be used in later figures in this section.

Table 4.3: Horizon calibration summary. Obs is number of observations per horizon band. BS, BR, BR Ref., BSS, MSE, b, and a defined as before. FLB is defined as the difference in mean signed error between longshot ($p < 0.10$) and favorite ($p > 0.90$) groups.

Band	**Obs**	**Obs %**	**BS**	**BR**	**BR Ref.**	**BSS**	**MSE**	***b***	***a***	**FLB**
1-3d	234,195	12.5	0.152	0.332	0.222	0.315	+0.014	1.107	−0.066	0.017
4-7d	290,834	15.5	0.138	0.289	0.206	0.329	+0.027	1.042	−0.185	0.007
8-15d	236,408	12.6	0.121	0.249	0.187	0.353	+0.028	1.065	−0.211	0.022
16-30d	238,453	12.7	0.095	0.204	0.162	0.416	+0.034	1.114	−0.307	0.032
31-60d	231,708	12.4	0.072	0.168	0.140	0.486	+0.028	1.135	−0.289	0.030
61-90d	150,597	8.0	0.077	0.155	0.131	0.414	+0.038	1.048	−0.450	0.023
91-180d	236,160	12.6	0.073	0.152	0.129	0.432	+0.037	1.107	−0.390	0.059
>180d	68,838	3.7	0.088	0.176	0.145	0.393	+0.035	1.073	−0.334	0.094
Overall	1,875,570	100.0	0.110	0.241	0.183	0.398	+0.028	1.112	−0.201	0.026

The following figures measure the temporal evolution of mean signed error (i.e. calibration—as determined by mean signed error—as a function of time-to-resolution), coined "calibration convergence." First, Figure 4.8 visualizes Overall Polymarket calibration convergence. Second, Figure 4.9 breaks down calibration convergence by domain, using an overlaid plot. And finally, Figure 4.10 examines calibration convergence by domain more closely through subplots.

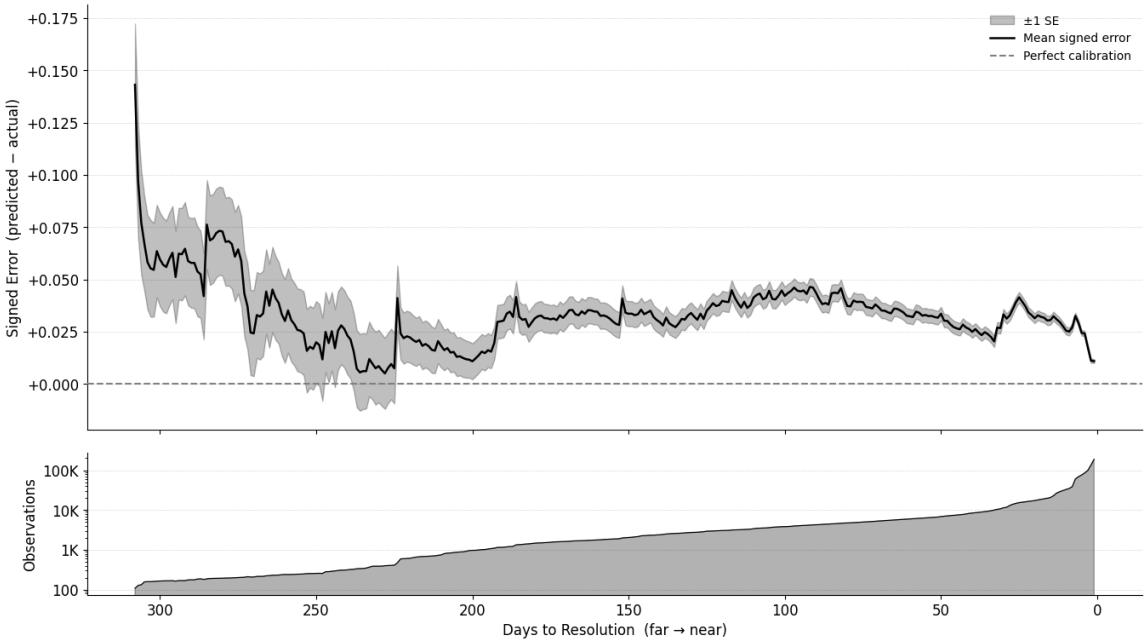

Figure 4.8: Upper: mean signed error by days-to-resolution (left to right: far to near), computed using all available price observations and true outcomes per day. Lower: price observation count per day (min = 100). Days-to-resolution is on a *linear* scale, while number of observations is on a *log* scale. Shaded band represents $\pm$ 1 standard error (SE).

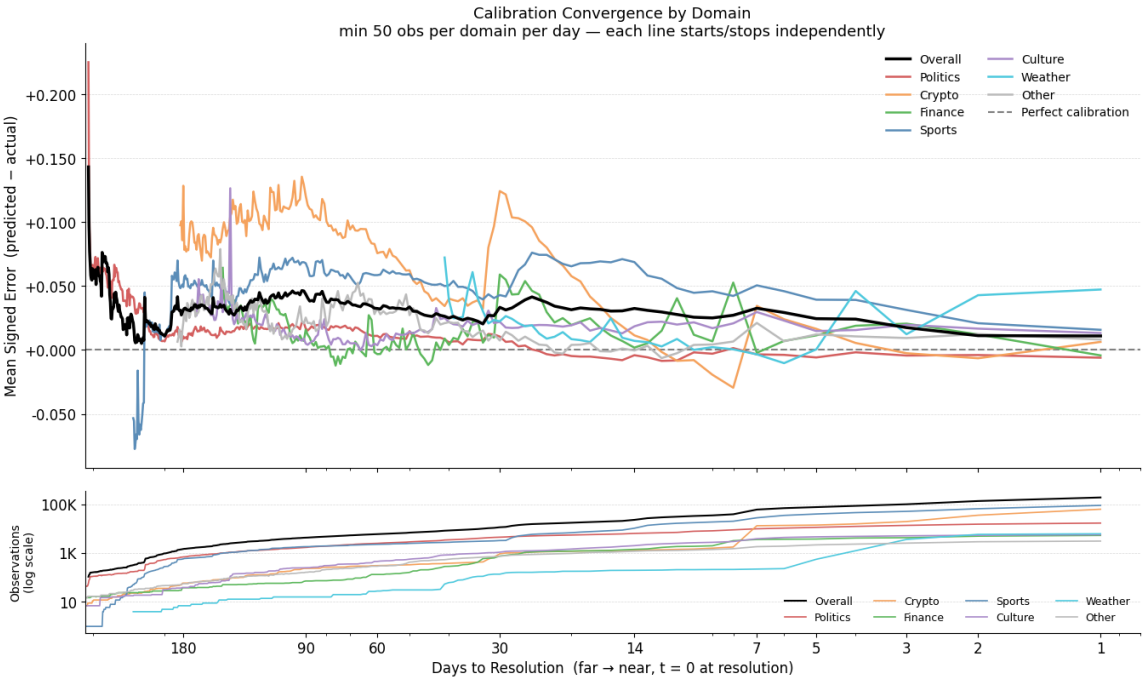

Figure 4.9: Upper: mean signed error by days-to-resolution (left to right: far to near), overlaid by domain. Lower: price observation count per day per domain (min = 50). Each domain line starts independently based on minimum observation threshold. Both days-to-resolution and number of observations are on *log* scales.

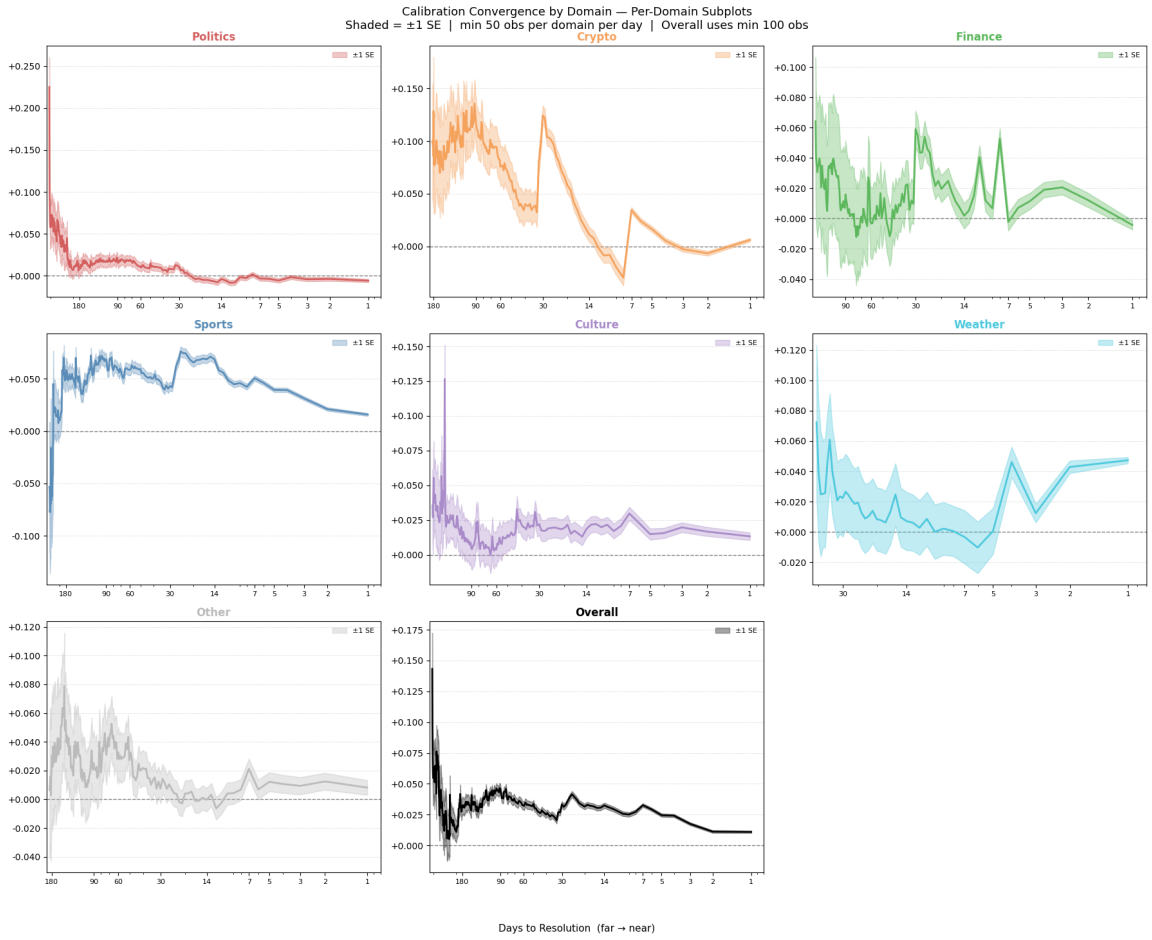

Figure 4.10: Per-domain mean signed error by days-to-resolution. Days-to-resolution is on a *log* scale, where each subplot's x-axis starts independently based on minimum observation threshold (min = 50). Shaded band represents $\pm$ 1 standard error (SE).

As can be seen, overall MSE is highest and most volatile far from resolution (where data is extremely sparse) and declines toward zero as resolution approaches. However, the decline is not linear: signed error increases over the approximate range $[200, 100]$, gradually declines from $[100, 30]$, and has a sharp spike around 25 before dropping off quickly in the final days. It remains above 0 for the entire horizon chart. Examining by domain, however, reveals dramatically different convergence behavior. Politics has the cleanest convergence, with a slight increase over the same range described above before reaching near-zero MSE by about 30 days and remaining there. Crypto is the most volatile, with large positive spikes around 30 and 7 days before terminating near

zero. Sports maintains positive MSE throughout and never fully converges. Finance and Weather have wide standard error bands and high volatility at far horizons, but dramatically different convergence patterns near resolution. Table 4.4 provides numerical binned data for these observations.

Table 4.4: Mean signed error by domain × horizon band. Positive values represent systematic overpricing; negative values represent underpricing.

Domain	**1-3d**	**4-7d**	**8-15d**	**16-30d**	**31-60d**	**61-90d**	**91-180d**	**>180d**
Politics	−0.004	−0.004	−0.004	−0.001	+0.010	+0.017	+0.016	+0.040
Crypto	−0.005	+0.019	−0.005	+0.075	+0.056	+0.098	+0.104	+0.067
Finance	+0.016	+0.009	+0.022	+0.032	+0.007	+0.003	+0.031	+0.018
Sports	+0.026	+0.043	+0.052	+0.067	+0.053	+0.060	+0.056	+0.021
Culture	+0.018	+0.020	+0.020	+0.018	+0.018	+0.008	+0.022	+0.053
Weather	+0.031	+0.024	+0.005	+0.016	+0.021	−0.072	−0.114	−0.061
Other	+0.011	+0.013	+0.001	+0.004	+0.021	+0.034	+0.030	+0.026
Overall	+0.014	+0.027	+0.028	+0.034	+0.028	+0.038	+0.037	+0.035

The following figures isolate the Favorite-Longshot Bias (FLB) magnitude component of miscalibration and examine its evolution. First, Figure 4.11 visualizes Overall Polymarket FLB by days-to-resolution. Second, Figure 4.12 breaks down FLB by days-to-resolution by domain, using an overlaid plot. And finally, Figure 4.13 examines FLB by days-to-resolution by domain more closely through subplots.

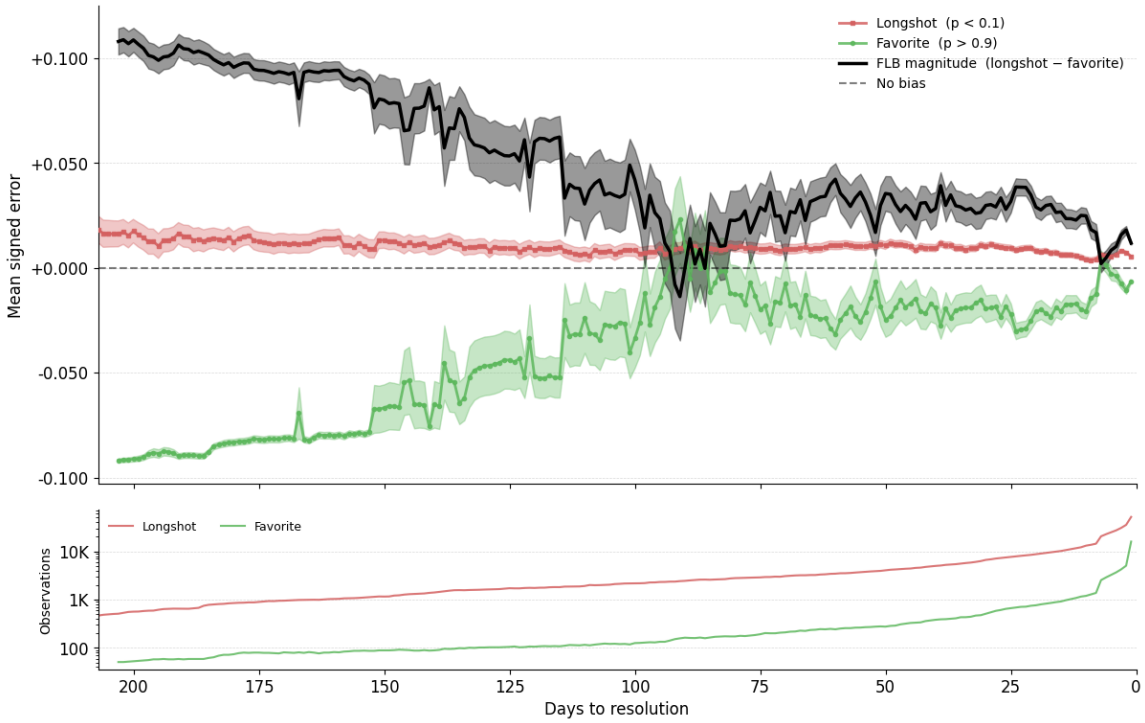

Figure 4.11: Red: signed error for longshot ($p < 0.10$) group as a function of days-to-resolution. Green: signed error for favorite ($p > 0.90$) group as a function of days-to-resolution. Black: FLB magnitude, defined as (longshot mean error - favorite mean error). Lower: price observation count per group (min = 50). Days-to-resolution is on a *linear* scale, while number of observations is on a *log* scale. Shaded band represents $\pm$ 1 standard error (SE).

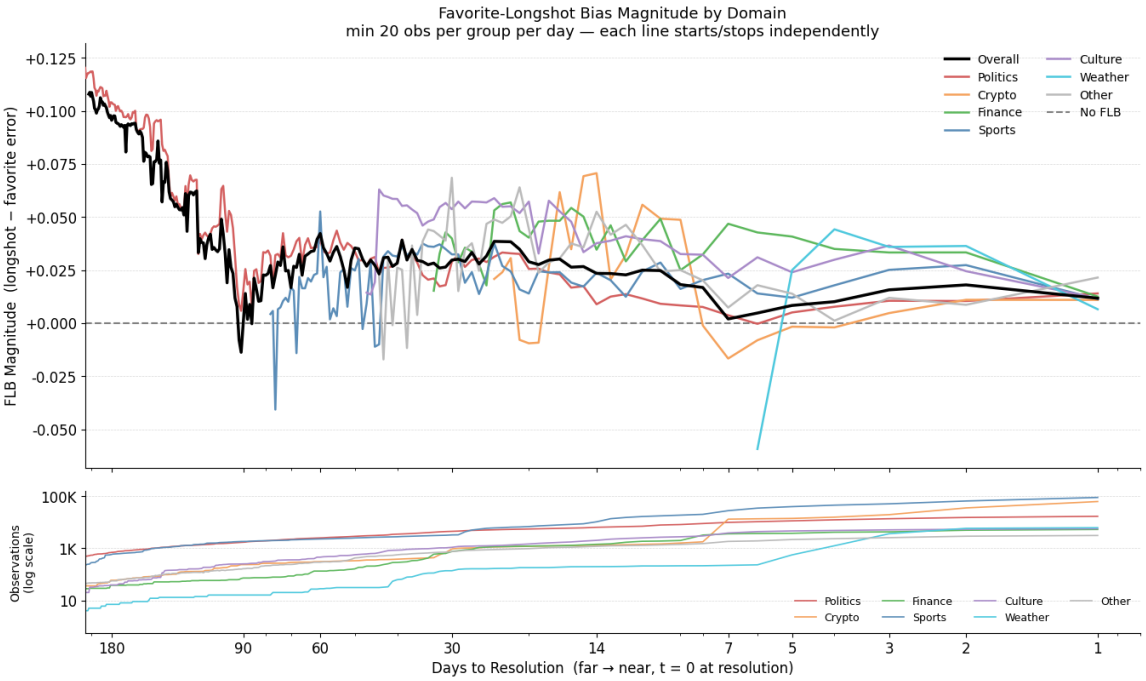

Figure 4.12: Upper: FLB magnitude (longshot mean error - favorite mean error) by days-to-resolution, overlaid by domain. Lower: price observation count per group per domain (min = 50). Each domain line starts independently based on minimum observation threshold. Both days-to-resolution and number of observations are on *log* scales.

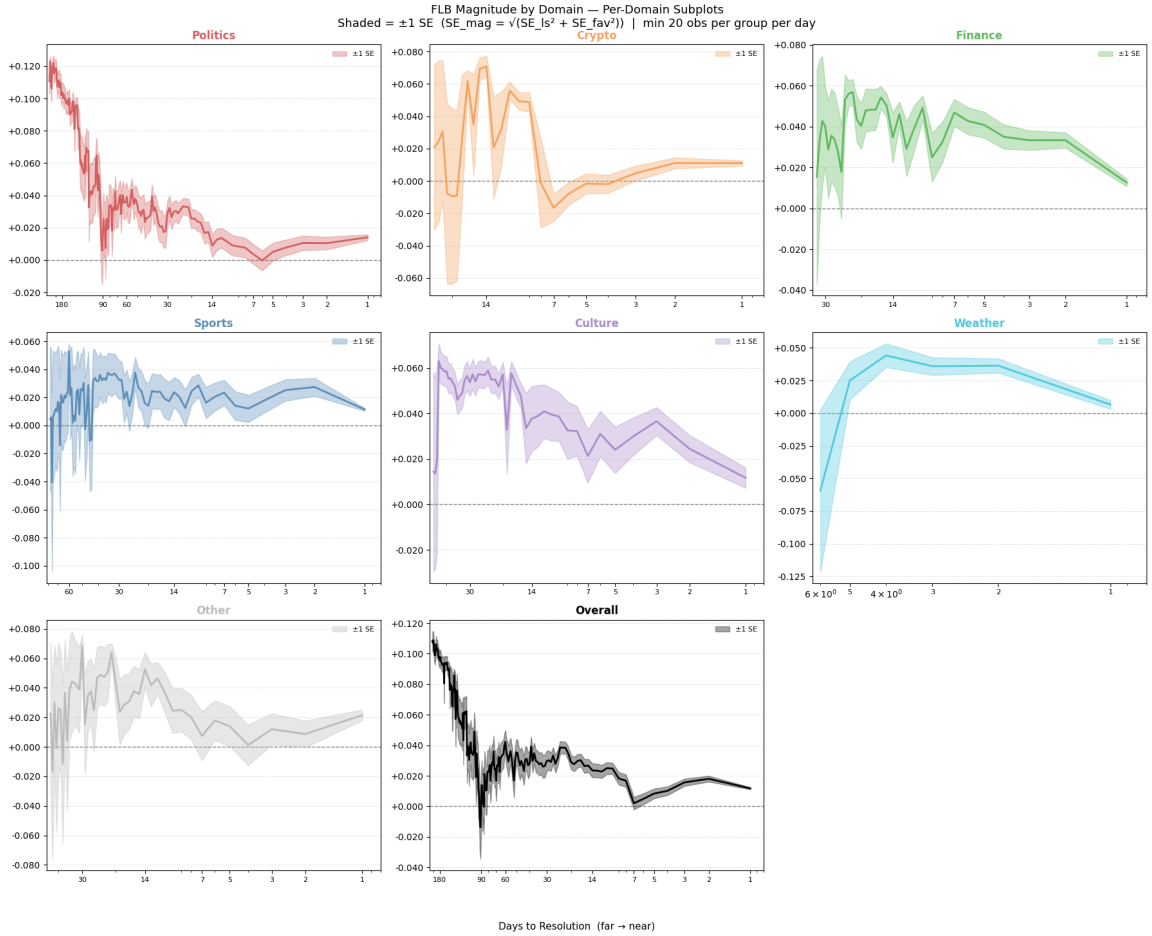

Figure 4.13: Per-domain FLB magnitude by days-to-resolution. Days-to-resolution is on a *log* scale, where each subplot's x-axis starts independently based on minimum observation threshold (min = 50). Shaded band represents $\pm$ 1 standard error (SE).

In the overall chart, we see that longshot error sits at a relatively constant positive level across the entire observable range: in other words, longshots are consistently overpriced at every horizon. Favorite error is much more volatile, but generally grows more negative as days-to-resolution increases: in other words, favorites are consistently underpriced at most horizons, and the effect is worse the further out the price observation is. However, there are some key exceptions: Favorite error enters the positive territory for a brief period around the $[90, 80]$ day range before dipping back down. Additionally, there is a sharp convergence pattern near resolution, beginning at about the 7 day mark. FLB magnitude is therefore largely dependent on the much

more volatile favorites (as shown by lower observation count and wider standard error bars): it is largest far from resolution and follows inverse dynamics as favorite error. Examining by domain, the first striking characteristic is that FLB magnitude is positive across nearly all domains and all horizons. There is a wide variety in data to reach minimum observation count thresholds, which results in dramatically different x-axes. Politics again shows the clearest FLB convergence.

Table 4.5 summarizes parametric calibration across both domains and horizon bands via the logistic recalibration slope b. Figure 4.14 presents logistic recalibration fitted parameters for Overall Polymarket data.

Table 4.5: Logistic recalibration slope b by domain and horizon band. Cells with fewer than 500 observations are ignored (–).

Domain	**1-3d**	**4-7d**	**8-15d**	**16-30d**	**31-60d**	**61-90d**	**91-180d**	**>180d**
Politics	1.071	1.029	1.041	1.105	1.105	1.090	1.203	1.192
Crypto	1.071	0.990	1.209	1.132	1.031	0.891	1.023	0.974
Finance	1.342	1.332	1.165	1.181	1.429	2.430	0.956	0.502
Sports	1.103	1.051	1.081	1.149	1.129	0.935	0.872	0.756
Culture	1.176	1.098	1.089	1.181	1.179	1.218	1.424	3.382
Weather	1.067	1.111	1.235	1.091	1.193	1.016	2.401	—
Other	1.119	1.107	1.182	1.265	1.289	1.165	1.250	1.733
Overall	1.107	1.042	1.065	1.114	1.135	1.048	1.107	1.073

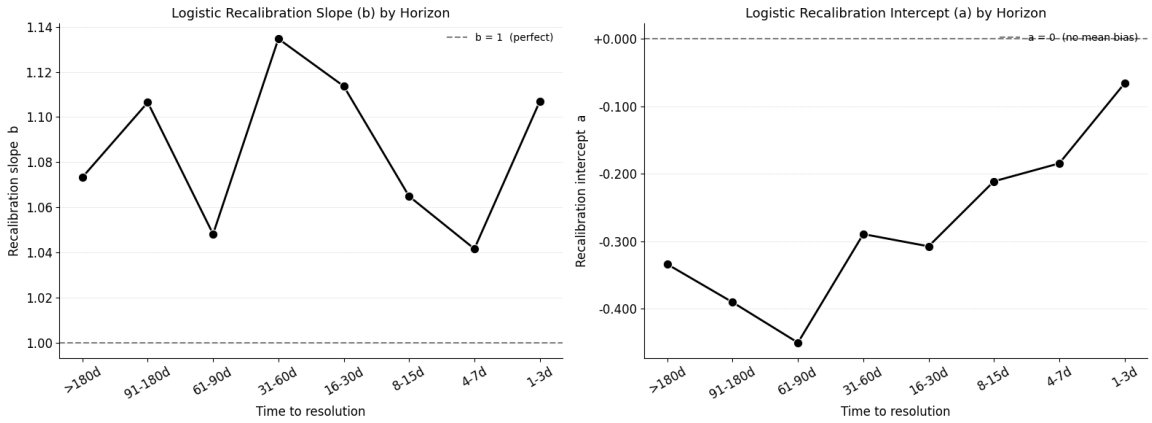

Figure 4.14: Left: logistic recalibration slope b by horizon. Right: logistic recalibration intercept a by horizon. Dashed lines indicate perfect calibration ($b = 1$, $a = 0$.

Looking at the overall structure first, we see that $b > 1$ across all horizon bands. This implies that market underconfidence and the FLB are universal across all horizons. Furthermore, this does not appear to follow any meaningful time-to-resolution pattern: the slope parameter is low at 61-90d and 4-7d, but peaks at 31-60d. As for the intercept parameter: a is negative across all horizons, and generally becomes more negative as time-to-resolution increases. This implies the systematic overpricing phenomenon worsens at longer horizons. From the domain table, almost all b entries across domains are above 1, with some notable exceptions. However, extreme values (such as those in Finance and Weather) should be interpreted with caution given sparse data.

The key patterns identified in the previous sections 4.1 through 4.3 have been market underconfidence, domain heterogeneity, and horizon-dependent calibration. However, calculation of these metrics have all been based on point estimates computed across the entire 1,875,570 price observation dataset (and filtering within when necessary). However, as discussed above in Section 3.1.4, Polymarket's multi-market event structure means many of those observations are *not* independent. This introduces correlation that naive standard errors do not accurately account for. This raises

the question of whether or not these estimates stand up to a more rigorous test of uncertainty.

4.4 Clustered Bootstrap

Research Question 4 asks whether calibration dynamics change after accounting for the non-independence structure of markets. The following section addresses this directly using the clustered bootstrap approach outlined in Section 3.1.4. The goal is to quantify how accounting for the non-independent nature of markets affects the conclusions of the previous sections. Importantly, the *estimates themselves do not change*: measuring the outcome rate of a market vs. its observed prices does not change the point estimate (the prices have not changed, and neither has the final outcome). However, the *uncertainty around those estimates does change*. When markets are clustered by event and have mutually exclusive outcomes, treating each historical price observation as an independent draw artificially inflates the sample size. The clustered bootstrap approach corrects for this by resampling (with replacement) at the *event* level rather than the *market* level to preserve the correlated structure. On Polymarket, only 21.1% of events fall into the multi-market category; however, this relatively small fraction of markets accounts for 80.5% of daily price observations. This cannot be ignored.

Figure 4.15 presents an updated overall reliability diagram, using new terminology of naive and clustered 95% confidence bands.

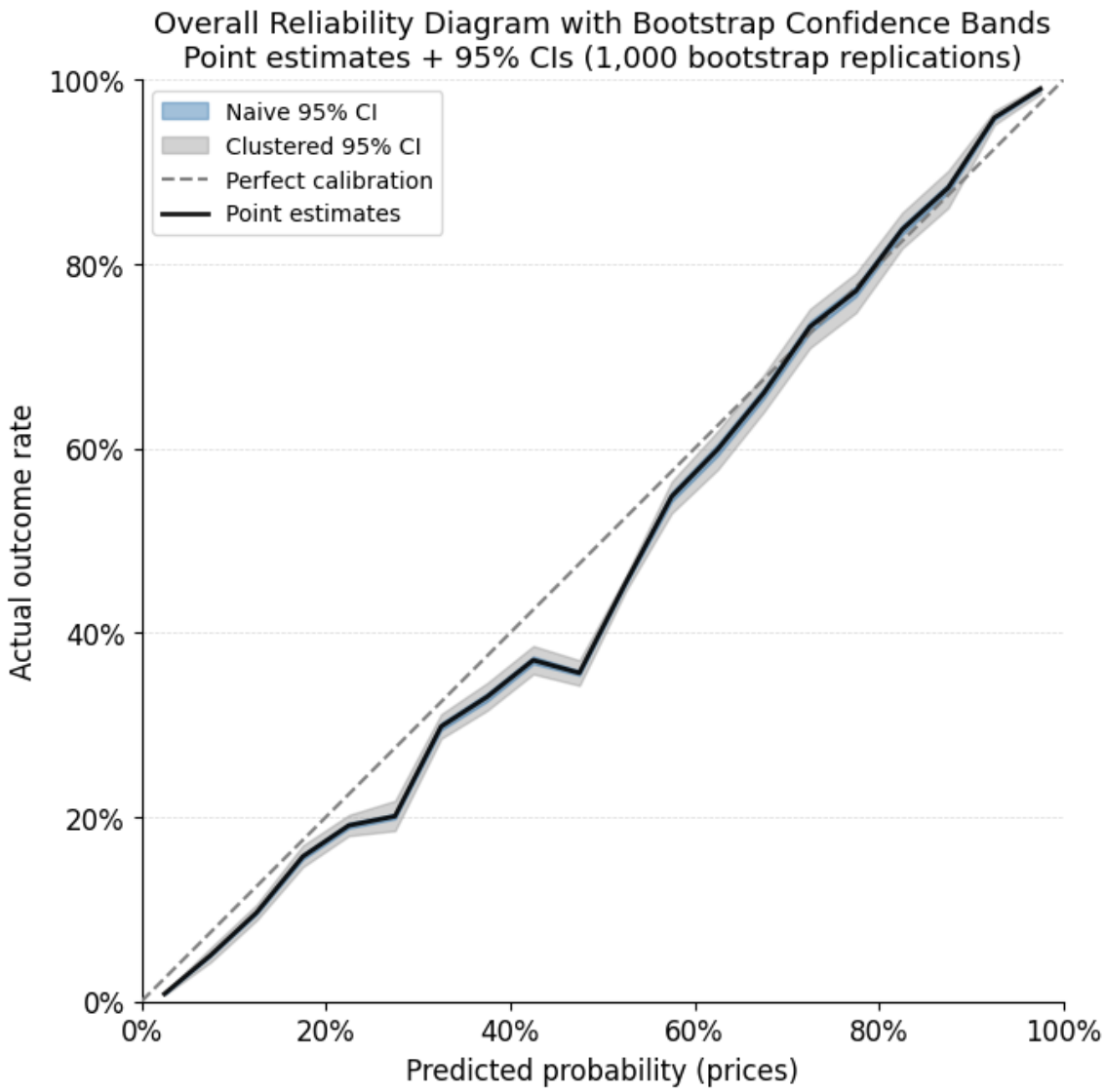

Figure 4.15: Overall reliability diagram with naive (blue) and clustered (gray) bootstrap confidence bands. Both bootstrap resampling methods use 1,000 replications, separated at the market- and event-level, respectively. Point estimates (black) and perfect calibration line (dashed) remain unchanged.

Note that the naive confidence band is barely visible at this scale, as it is almost entirely covered by the clustered band and calibration curve itself. The clustered (gray) is substantially wider throughout the curve, especially in the price ranges between the highly concentrated points of 0.0, 0.5, and 1.0. Despite this widening, the overall curve remains well outside the perfect calibration diagonal for the entire

lower range of prices. This confirms that overall miscalibration is robust to clustering. Table 4.6 shows overall statistics on Polymarket data when accounting for clustering. Importantly, mean signed error and recalibration slope b remain above the critical thresholds, confirming the miscalibration finding is robust to clustering.

Table 4.6: Overall calibration statistics. Naive vs. clustered 95% confidence intervals from 1,000 bootstrap replications are shown. Width Ratio is the clustered CI width divided by the naive CI.

Statistic	**Point Est.**	**Naive 95% CI**	**Clustered 95% CI**	**Width Ratio**
Brier Score	0.1100	[0.1098, 0.1102]	[0.1066, 0.1134]	14.4
Mean Signed Error	+0.0275	[0.0270, 0.0280]	[0.0239, 0.0314]	7.7
Recal. slope b	1.1125	[1.1085, 1.1163]	[1.0796, 1.1436]	8.1
Recal. intercept a	−0.2009	[−0.2058, −0.1966]	[−0.2367, −0.1687]	7.5

Figure 4.16 extends the confidence interval comparison to each domain.

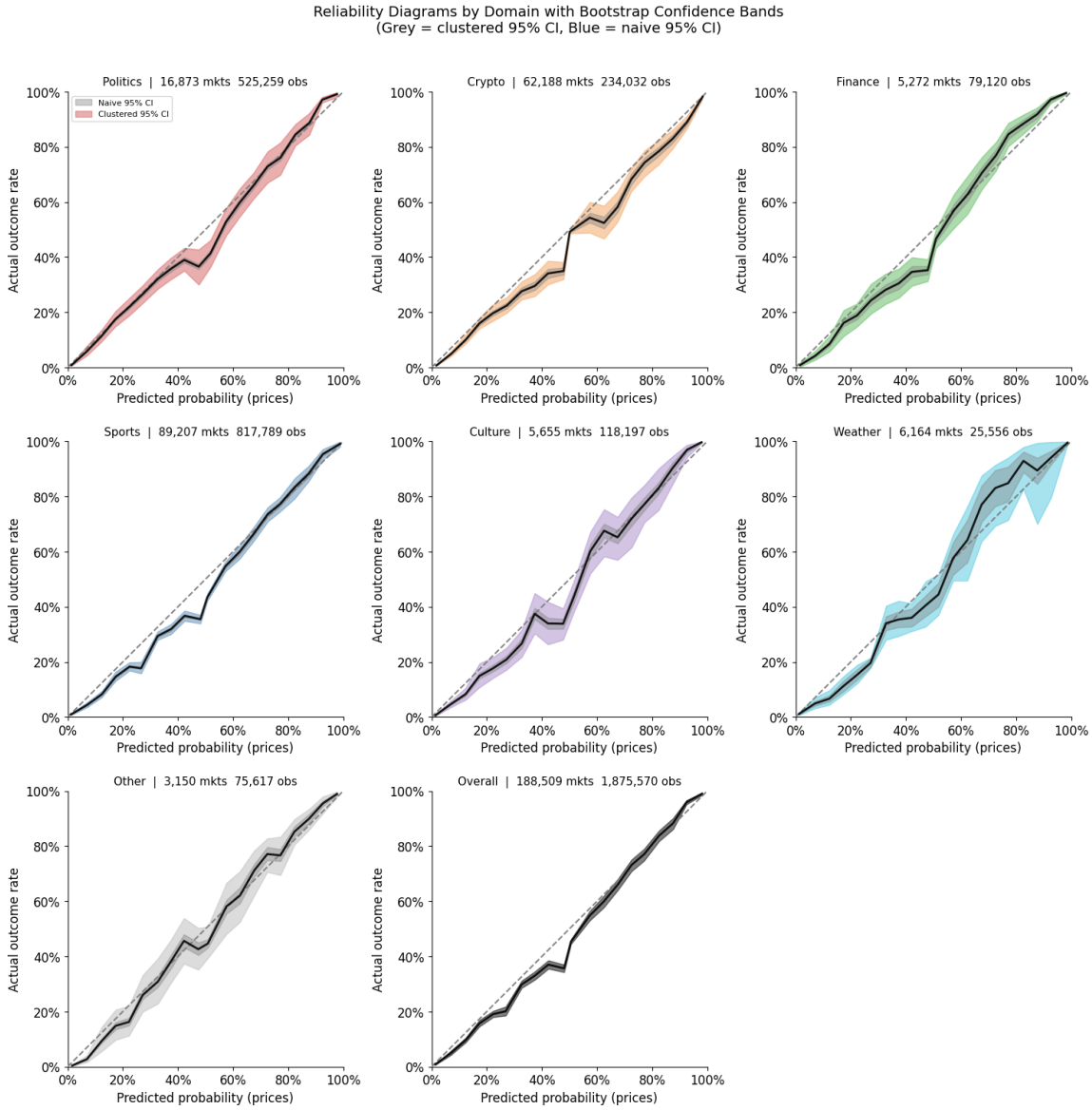

Figure 4.16: Reliability diagrams by domain with naive (gray) and clustered (domain color) bootstrap confidence bands.

The contrast between naive and clustered bands varies widely by domain. The Sports domain shows relatively little widening despite having the largest observation count, likely reflecting a lower concentration of multi-market events. On the other hand, Politics, Culture, and Weather domains show much wider bands, explained by their relatively higher concentration of multi-market events. Regardless, the calibration curves still largely appear to follow their previous S-patterns (although some are

further from the diagonal than others).

Similarly, Figure 4.17 further extends the confidence interval comparison to each horizon.

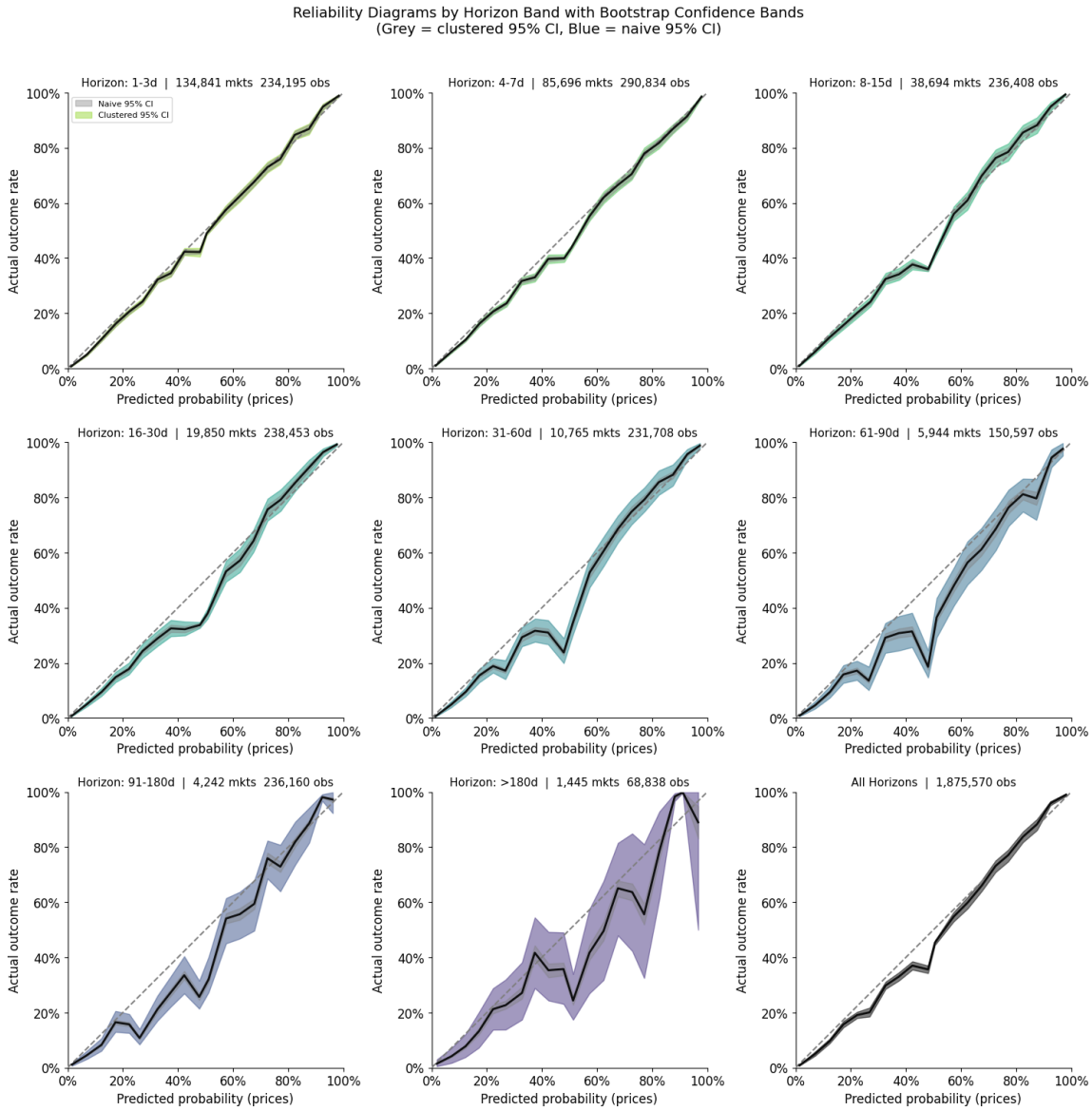

Figure 4.17: Reliability diagrams by horizon with naive (gray) and clustered (horizon color) bootstrap confidence bands.

The key difference here is that there is a clear pattern visible in the subplots: as time-to-resolution increases, clustered confidence band width grows considerably. At short horizons (1-3d, 4-7d), the clustered confidence bands are barely distinguishable

from the naive bands. On the opposite end of the spectrum, however, at longer horizons (91-180d, >180d), the clustered confidence bands are massive compared to the relatively small naive bands. Once again (and perhaps even more so than with the domain breakdown), a visible S-shape is still present.

Tables 4.7 and 4.8 summarize logistic recalibration slope b point estimates and compare them to naive and clustered 95% confidence intervals for each domain and horizon, respectively.

Table 4.7: Logistic recalibration slope b by domain with naive vs. clustered 95% confidence intervals. Width Ratio is the clustered CI width divided by the naive CI width. Sig. (Significant) indicates whether the clustered CI excludes $b = 1$.

Domain	**N_obs**	**Point b**	**Naive 95% CI**	**Clustered 95% CI**	**Width Ratio**	**Sig.**
Politics	525,259	1.113	[1.106, 1.119]	[1.050, 1.187]	10.9	Yes
Crypto	234,032	1.079	[1.069, 1.089]	[1.030, 1.127]	4.8	Yes
Finance	79,120	1.221	[1.203, 1.240]	[1.142, 1.316]	4.7	Yes
Sports	817,789	1.097	[1.090, 1.103]	[1.048, 1.148]	8.0	Yes
Culture	118,197	1.180	[1.166, 1.196]	[1.094, 1.267]	5.8	Yes
Weather	25,556	1.188	[1.151, 1.225]	[1.018, 1.368]	4.8	Yes
Other	75,617	1.232	[1.214, 1.250]	[1.127, 1.349]	6.0	Yes

Table 4.8: Logistic recalibration slope b by horizon with naive vs. clustered 95% confidence intervals. Width Ratio is the clustered CI width divided by the naive CI width. Sig. (Significant) indicates whether the clustered CI excludes $b = 1$.

Band	**N_obs**	**Point b**	**Naive 95% CI**	**Clustered 95% CI**	**Width Ratio**	**Sig.**
1-3d	234,195	1.107	[1.095, 1.118]	[1.089, 1.125]	1.6	Yes
4-7d	290,834	1.042	[1.032, 1.050]	[1.022, 1.063]	2.2	Yes
8-15d	236,408	1.065	[1.055, 1.075]	[1.033, 1.097]	3.3	Yes
16-30d	238,453	1.114	[1.104, 1.125]	[1.070, 1.157]	4.1	Yes
31-60d	231,708	1.135	[1.125, 1.146]	[1.083, 1.191]	5.1	Yes
61-90d	150,597	1.048	[1.035, 1.061]	[0.987, 1.126]	5.4	**No***
91-180d	236,160	1.107	[1.096, 1.117]	[1.021, 1.200]	8.6	Yes
>180d	68,838	1.073	[1.058, 1.090]	[0.901, 1.285]	11.9	**No***

On a pure breakdown by domain and horizon, all estimates for b are found to be significant, except for the horizon bands 61-90d and >180d. The significance test indicates whether the updated clustered confidence interval excludes the perfect calibration logit slope $b = 1$. If the clustered confidence interval on b continues to lie outside the critical tipping point where the interpretation of b changes (i.e. excludes $b = 1$), then the parameter is deemed significant and the FLB holds even with a clustered bootstrap approach. Most Width Ratios in both tables are very large, representing the prevalence of multi-market events and the importance of accounting for them.

Figure 4.18 presents the full logistic recalibration slope b, scoped by domain and horizon.

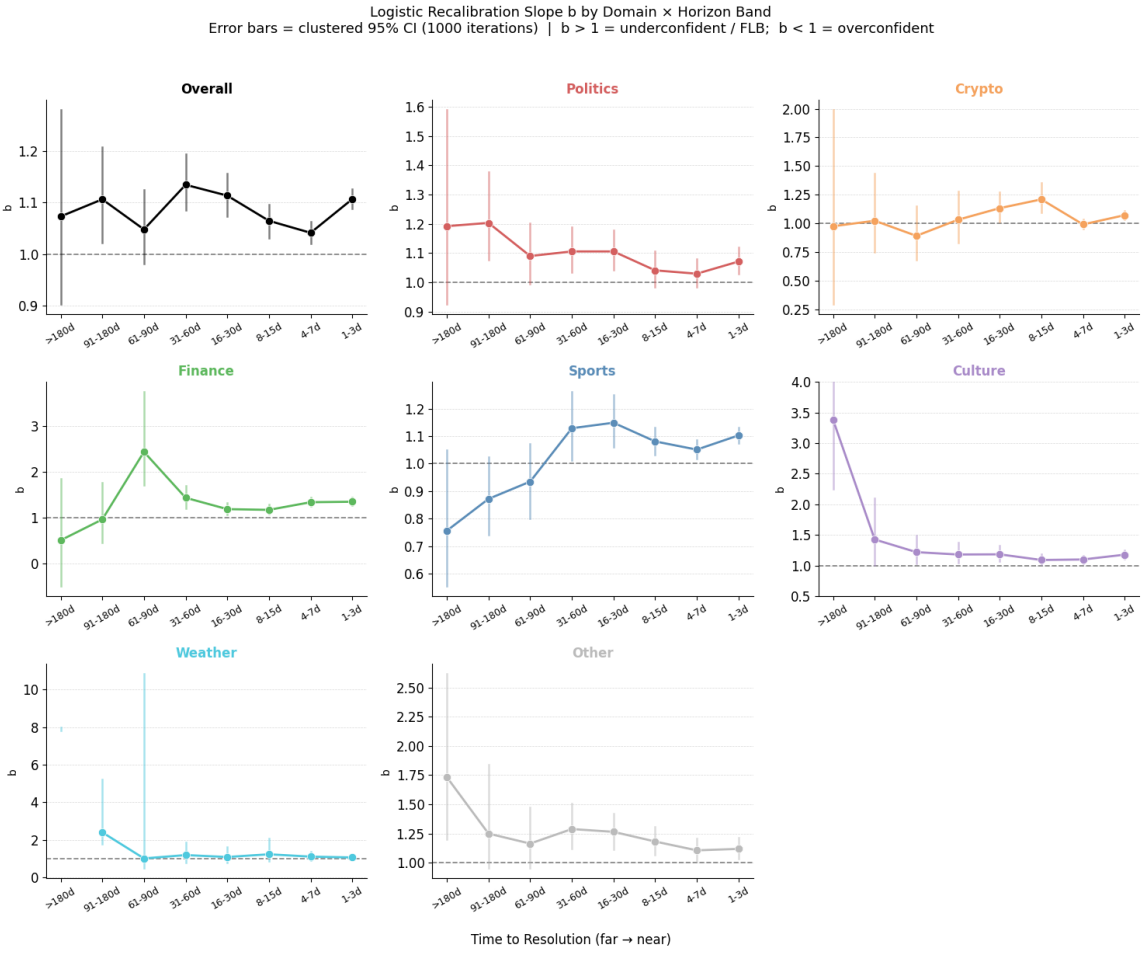

Figure 4.18: Logistic recalibration slope b by domain and horizon band with clustered 95% confidence interval vertical error bars. Reference line at $b = 1$ represents perfect calibration. Note that each subplot uses an independent y-axis adjusted for individual variance.

The extreme values identified in Section 4.3's logistic recalibration analysis by domain and horizon (e.g. Finance's $b = 2.430$ at the 61-90d horizon and Weather's $b = 2.401$ at the 91-180d horizon) are justified here: the clustered confidence intervals across all subplots are extremely wide at any high time-to-resolution. Politics shows the most consistent and stable trajectory above $b = 1$ across horizons, and Culture also shows consistent slope within the confidence interval after an initial steep rise at long horizons. For periods with more data (i.e. closer to resolution), all domains show a clear trend of lying above the line $b = 1$, most of which are also within confidence

intervals.

Taken together, the clustered bootstrap analysis confirms that naive methods significantly understate uncertainty in Polymarket calibration estimates. This is a direct result of the multi-market event structure that is so common among price observations. Despite this correction, the core calibration conclusions—tested through the logistic recalibration robustness to expanded confidence intervals—appear to remain unchanged. Two exceptions exist in the horizon analysis at longer durations with limited data. Further discussion of these findings, as well as their implications and relation to prior literature, follows below.

Chapter 5

Discussion

This thesis set out to answer a deceptively simple question: are prediction markets good at predicting real-world events? More concretely: do Polymarket prices reflect well-calibrated probability forecasts for the outcomes of real-world events? The short answer is yes, but there are important caveats. After analyzing 188,509 resolved markets and 1,875,570 individual daily price observations spanning November 2022 through December 2025, Polymarket prices indeed demonstrate true forecasting skill. Using the Brier Skill Score common in the forecasting literature, Polymarket prices as a whole achieve a BSS of 0.398, a 40% improvement over a base rate benchmark. Furthermore, the reliability diagram of Figure 4.15 closely tracks the 45-degree line of perfect calibration across the full implied probability range. These findings provide empirical evidence for prediction markets as reliable information aggregation mechanisms, as supported by the theoretical work of Galton (1907), Hayek (1945), Fama (1970), and Surowiecki (2004) and confirmed in empirical prediction market studies by Berg, Nelson, and Rietz (2003), Wolfers and Zitzewitz (2004), and Cutting et al. (2025). However, at the same time, the data reveals that this headline result conceals meaningful deviation and bias structure. The fitted logistic recalibration slope of $b = 1.112$ confirms the presence of the Favorite-Longshot Bias. This

key finding—that Polymarket systematically overprices longshots and underprices favorites—is a pattern that has been identified in theoretical work by Ali (1977), Manski (2006), Ottaviani and Sørensen (2008), Kahneman and Tversky (1979) and confirmed in empirical prediction market studies of InTrade (Page and Clemen 2013), PredictIt (Restocchi, McGroarty, and Gerding 2019), and Kalshi (Le 2026). This bias varies across market domains (Section 4.2), intensifies at longer time horizons (Section 4.3), and most importantly, proves robust to a rigorous uncertainty quantification (Section 4.4). The discussion that follows interprets these collective findings in the context of prior literature, examines key structural differences across domains and time horizons, and considers practical implications for the belief in Polymarket prices as probability estimates.

First, we return more in-depth to the original wisdom of the crowds and market efficiency discussion. From the Brier Score decomposition (as reported in Table 4.1), the "Reliability" score was found to be 0.0019, whose nearly negligible value suggests a very low calibration error at the overall level. Taken together with the BSS of 0.398 mentioned above, it is important not to understate the strong evidence in favor of Polymarket's price calibration. Indeed, this may be explained by Polymarket's fulfillment of each of the four necessary conditions for Wisdom of the Crowds set forth by Surowiecki (2004): diversity of opinion (Polymarket covers a wide range of domains and interest categories), independence (slightly more difficult to verify, but Polymarket has millions of users), decentralization (Polymarket's global footprint), and aggregation mechanism (the trivial case: the Polymarket markets themselves).

However, the empirical story does not end there; overall logistic recalibration slope $b = 1.112$ and mean signed error $+0.0275$ from Table 4.1 show that this apparent crowd "wisdom" is imperfect and directionally biased. According to Manski (2006), this is not a bug, but a feature: price compression toward 50% (i.e., the Favorite-Longshot Bias where $b > 1$) is expected even in well-functioning markets. Addi-

tionally, Grossman and Stiglitz (1980) suggest the remaining logical step: if prices perfectly reflected all information, there would be no incentive to collect information and trade. Thus, while Polymarket may broadly support a loose interpretation of the wisdom of the crowds thesis, the confirmation of structural bias calls into question the degree of this belief.

Analyzing the aggregate calibration of Polymarket is informative in and of itself, but it has been likened to only looking at the class average on a MAT201 exam—and failing to realize half the students failed and the other half aced it.[1] In other words, without context into the markets themselves, the aggregate obscures as much as it reveals. As shown in Table 4.2 and Figure 4.6, the BSS range varies dramatically by domain, ranging from 0.264 for Sports markets to 0.565 for Politics markets—a striking difference given that these domains also contain the two largest volumes of data. This reveals that domain is the first and most important stratification of the data. For the best performing domain Politics, this is likely explained by the higher concentration of liquidity and longer duration markets relative to peer domains (see Table 3.1), as well as a more informed or "wiser" (from Surowiecki's interpretation) trader base. This finding is consistent with decades of empirical evidence in support of political prediction markets outperforming polls, ranging from their inception in the late 20th century IEM (Berg, Nelson, and Rietz 2003) to the modern day (Cutting et al. 2025). For the worst performing domain Sports (and to a lesser extent Crypto), there is relatively limited informational advantage over the prior. This may be explained by the higher concentration of short-duration, near 50/50 markets that are prevalent in these domains (e.g. markets that are more akin to gambling in the traditional sense: "Which team will win the U.S.A. vs. Canada match tonight?" or "Will the price of Bitcoin be up or down in the next week?"). All domains exhibit a positive slope parameter b under logistic recalibration (where forecasts are

[1] This analogy was revealed to me in a dream.

modeled in log-odds space), implying prices are universally compressed toward 0.50 and providing empirical support for the heterogeneous beliefs framework of Manski (2006). Additionally, Weather and Finance represent the most egregious examples out of the named domains, as can be shown by the S-shape deviation patterns in Figure 4.5. Using the Prospect Theory of Kahneman and Tversky (1979), this suggests that collective traders in these domains have the worst probability weighting of extreme prices. Collectively, these results tell us that the wisdom of crowds and the calibration within Polymarket is not a domain-agnostic property but depends critically on the nature of the market conditions. Next, we discuss the next natural dimension of analysis after investigation of the domain-level: that is, how calibration evolves over the lifetime of prediction markets.

The intuition that price calibration should improve monotonically as a market approaches resolution is natural for any model where information accumulates over time. On Polymarket, however, this is only partially true. Figure 4.8 shows that mean signed error (in the aggregate view) generally reduces as price observations track closer to the resolution date (i.e., "calibration convergence"). This is more visually obvious in the reliability diagrams of Figure 4.7, where daily price observations are grouped by horizon bands: calibration curves track the diagonal line of perfect calibration much more closely at shorter horizons and gradually get worse at longer horizons. Looking more closely at the data, however, there are complications with this understanding. As information accumulates and markets approach resolution, we would expect that BSS improves and MSE declines. This is not supported in Table 4.3: BSS peaks at 0.486 in the 31-60d horizon band and counterintuively *drops* as the horizon approaches market resolution, bottoming out at 0.315 at the shortest 1-3d band.[2] While an interesting find, at the aggregate level this can be explained in part by an observation

[2]To be clear (as discussed above), even the low end of the observed BSS range represents genuine forecasting skill when interpreted as a snapshot. Here, we are interested in observing the *relative* BSS across horizons, rather than BSS as an *absolute* magnitude.

mix: shorter horizon bands are artificially depressed by the inclusion of more near-resolution, short term events that evidently do not have the same predictive power as that of a long term event nearing completion. Additionally, there is a distinct plateau in the MSE (as seen in Figure 4.8) from about the 180-50d range, which suggests a persistent and relatively stable mispricing at longer horizons. One explanation for this is that markets remain active (and therefore continue to contribute daily prices to our observation set), but information flow is slow and may not reflect materially new information at longer horizons, thus allowing mispricing to remain relatively stable. Furthermore, as seen from the parametric methodological approach in Figure 4.14, from the 31-60d band until resolution, the logistic recalibration parameters by-and-large behave as expected: the slope b (shape of the S-curve applied to logit space) decreases as days-to-resolution falls and the intercept a (directional bias of prices in logit space) approaches zero over the same timeline. Outside of this window, however, at longer horizons (and at least for the slope parameter, also at very short horizons), the parameters do *not* behave as would be expected by a monotonic intuition of calibration convergence. For example, the slope considerably *rises* at the shortest horizon band—which is associated with a *worsening* of the S-shaped deviation in the calibration curve—where we might expect minimum uncertainty and oscillates at longer horizons where we might expect strictly increasing maximum uncertainty. This non-monotonic story—which is evident in empirical evidence from MSE, BSS, and b slopes—calls for a closer look at the components driving these dynamics.

Before diving into the domain-level analysis of these horizon dynamics, we flag a special empirical finding on the Favorite-Longshot Bias. Table 4.3 confirms the existence of overall FLB across all time horizons—in line with the theoretical work of Kahneman and Tversky (1979), Manski (2006) and Page and Clemen (2013). Interestingly, Figure 4.11 reveals that that the magnitude of the FLB by horizon is driven almost entirely by the **favorite group**, rather than the longshot group. In this study,

we employ a relatively strict interpretation of favorites and longshots, whereby "longshots" are determined as observations with prices below 0.10 and favorites are those with prices above 0.90. The longshot signed error (in red) is approximately flat and positive throughout the entire observable price range, reflecting that longshots on Polymarket (in the aggregate) are consistently overpriced at a level that varies very little with days-to-resolution, even at very long horizons. This aligns directionally with the Prospect Theory proposed by Kahneman and Tversky (1979), but the static nature of longshot overpricing is surprising: at longer horizons where prices are more uncertain, there is minimal increase in probability weighting distortion as we might expect. The favorite signed error (in green), however, tells a dramatically different story: while favorites are nearly universally underpriced, the variance is much higher and does not follow any meaningful visual pattern besides a gradual reduction in error with a decrease in time-to-resolution. Interestingly, there is a brief inversion of the FLB at around the 90 day mark; more granular market-level investigation is needed before making any claims here. The extreme differences in favorite vs. longshot dynamics are almost certainly related to the multi-market event structure of Polymarket price observations: there are far more price observations that are near-zero and grouped into the longshot group than near-one observations grouped into the favorite group. Therefore, the FLB magnitude pattern seen in Table 4.3 is almost entirely the story of how *favorite* calibration evolves over time, not longshot. As a result, this qualifies the findings of Restocchi, McGroarty, and Gerding (2019), who used an extremely loose interpretation of favorites and longshots ($p < 0.49$ and $p > 0.51$, respectively) in their magnitude analysis: using a loose interpretation obscures much of the near-boundary variability in error between the two groups. In order to explain the specific shape of these curves, we must turn to a deeper domain-level examination.

Both the overall calibration convergence and FLB temporal evolution discussed above hide domain-level dynamics that vary substantially from one another. As shown

in Figure 4.10, Politics clearly shows the strongest signs of calibration convergence, even at very long horizons: MSE is low but positive from the 180d range and onwards (reflecting slight overpricing at longer horizons), but converges to near-zero by 30d and stays there. Other domains, such as Sports, Crypto, and Weather, exhibit far poorer calibration dynamics over time. We believe these results can be explained by a combination of compounding effects. First, Politics markets likely attract a much wider variety of traders than potential "domain-experts" interested in Sports, Crypto, or Weather markets. Two key factors identified by Surowiecki (2004) that contribute to wisdom of the crowds are diversity of opinion and decentralization; these factors are almost certainly more relevant to Politics markets. Second, this could be explained by the bilateral cancellation hypothesis of Le (2026), whereby politics markets in particular attract strongly partisan traders with opposing views that contribute to market efficiency. Third, Politics markets in particular (as opposed to other domains) may enjoy what we coin an outsized "snowball effect." This speculative hypothesis goes something like the following: global news agencies and social media have increasingly thrust prediction markets into the public view for their perceived accuracy, thereby attracting more public interest and traders, thereby leading to a more diverse and more informed crowd of traders, thereby improving price calibration. Looking further into the FLB breakdown by domain in Figure 4.13, we find that once again Politics shows the most dramatic FLB convergence over long horizons. The magnitude for each domain (notable exception: Crypto) remains positive outside of an initial high uncertainty state where price observations are thin, indicating the Favorite-Longshot Bias is present among *all* Polymarket domains and at almost *all* time horizons. Interestingly, we find that the FLB is not minimized at resolution (as would be expected), but rather at roughly the 7-day horizon across multiple domains including Politics, Crypto, Sports, and Other (and therefore strongly reflected in the aggregate view). This finding is unseen among all other literature.

Turning attention to the logistic recalibration parametric estimation (as shown in Table 4.5), we find that there are very few instances of market domains approaching perfect calibration (i.e., $b = 1$) across horizons. Once again, Politics, Crypto, and Sports appear to show the most promise, which may be a result of their relatively high price observation counts. However, while there may be an absence of perfect calibration reflected across the domain-level slope parameters, the direction is quite clear: $b > 1$ across almost all domains and at almost all time horizons. This indicates that systematic deviations in the price observation data are pervasive and that markets are almost universally underconfident (that is, prices are compressed towards 0.50 rather than the extremes). This aligns with the findings of FLB magnitude discussed above. Within domains, the pattern of these slope parameters is much less conclusive. Politics, Culture, Finance, and Other domains show a gradual approach towards $b = 1$ as time-to-resolution falls, but these approaches are not smooth (as would be expected for a market to become more well-calibrated as event resolution approaches). Taken together, Figure 4.10, Figure 4.13, and Table 4.5 represent a core takeaway: calibration trajectories and biases vary substantively by domain, and any complete calibration analysis of prediction markets should account for this. Le (2026)'s decomposition identified Domain-by-Horizon Interaction as the largest single component of calibration variance on Kalshi, a finding this study's Polymarket data is broadly consistent with even if the specific decomposition was not replicated here.

The domain and horizon results discussed above characterize the structure of miscalibration on Polymarket in detail. Before drawing final conclusions, however, it is vital to discuss the problem of non-independence referenced in Section 3.1.4. As a brief refresher, Polymarket prediction markets directly violate the independence assumption of standard errors and statistical inference: in fact, non-independent multi-market events are the norm (see Table 3.1 and the bottom panels of Figure 3.1). Despite multi-market events making up only 21.1% of the total events in the cleaned

dataset, 80.5% of total price observations are contributed by these events. Multi-market events mean price day observations within the same event are correlated (see the example from Section 3.1.4), so when naive standard errors treat correlated observations as independent, uncertainty is understated. To resolve this, we employ the clustered bootstrap model outlined by Page and Clemen (2013), which is devised for exactly this reason. The question is whether the previous findings stand up to a new, more rigorous uncertainty quantification.

In short, the answer is yes, the key findings presented by this study are robust to new confidence intervals defined by the clustered bootstrap approach. Using this method, we find that confidence intervals widen significantly depending on the context, often by several orders of magnitude (see Table 4.6, Width Ratio column). Despite the major corrections, the core conclusions survive. As a proxy, we examine the logistic recalibration slope b as both an aggregate metric and over all domains and horizon bands. We understand "remains significant" to mean that the slope parameter clustered CI continues to remain greater than 1 and exclude the critical $b = 1$ line of perfect calibration. In particular, the overall slope $b = 1.113$ remains significant on clustered CI [1.080,1.144] and the slope of all seven domains remain significant (see Table 4.7). For horizon bands, two exceptions exist for the 61-90d range and >180d range (see Table 4.8). What this means is that by using naive inference, we would wrongly conclude that these bands represent significant Polymarket miscalibration, however, the correct conclusion is that the evidence is insufficient due to the effective clustered sample size being far smaller (which also explains why the domain-horizon b point estimates in Table 4.5 are extreme and unreliable at long horizons). Despite the large study undertaken by Le (2026)—primarily on Kalshi prediction market data but also using Polymarket data—an explicit limitation stated by the study was the absence of a clustered bootstrap approach. This thesis fills that gap and demonstrates that it was consequential: Le (2026)'s Polymarket conclusions at long horizons should

be treated with additional caution. The broader implication here is that any calibration study for modern multi-market prediction markets should treat event-level clustered standard errors as a baseline requirement.

With the robustness of the core findings confirmed, we return to the original question that motivated this work: are Polymarket prices well-calibrated probability forecasts? Polymarket's own documentation states plainly: "Prices are Probabilities" (*How accurate is Polymarket?* 2025). As we caution in the Introduction, the appropriate response to this claim is to "trust, but verify." The data itself provides a specific, empirically grounded answer: "Mostly trust, but with a calibrated correction." A raw Polymarket price is not a tuned probability in the strict sense, but instead a probability that has been systematically compressed toward the mid-level 0.50 threshold. The degree of compression depends on domain category and horizon (time-to-resolution) that have been numerically characterized and rigorously tested. The bootstrapped reliability diagrams in Figure 4.15, Figure 4.16, and Figure 4.17 present a clean visual on how error materializes across the full Polymarket price range and additionally, how it varies across domains and horizons. Further, the logistic recalibration models fitted in this analysis (summarized in Table 4.5) provide a domain- and horizon-specific correction factor for any raw Polymarket price. Without loss of generality, a price of 0.30 in a Politics market at 90 days to resolution implies a meaningfully different probability than a price of 0.30 in a Crypto market at 15 days to resolution. Given any raw observed price (implied probability), domain, and days-to-resolution, it is possible to apply the fitted logistic recalibration model to return a corrected probability estimate with confidence bounds. With that being said, this correction is only as reliable as the historical calibration data it is trained on. As Polymarket and prediction markets as a whole continue to evolve, a new calibration study using updated data will be warranted.

The implications of these findings extend beyond individual traders correcting

their assumptions on raw price observations. Millions of people now encounter Polymarket prices in mainstream news coverage and may interpret them at face value as unbiased probabilities.[3] However, the systematic deviation in observed prices compared to actual outcome rates, which aligns with the well-documented Favorite-Longshot Bias, means this interpretation must be cautioned against. These deviations are most severe at long horizons and in certain domains, but they follow the general trend of market underconfidence: that prices are compressed toward 0.50, longshots are overpriced, and favorites are underpriced. The study of prediction market calibration has built toward this point over several decades: from the first documentation of the FLB in horse track betting Ali (1977), to the theoretical explanations of Kahneman and Tversky (1979), Manski (2006), and Ottaviani and Sørensen (2008), to the first studies of prediction markets by Berg, Nelson, and Rietz (2003) on the IEM, to the empirical foundation work by Page and Clemen (2013), to the temporal evolution study by Restocchi, McGroarty, and Gerding (2019), and finally to the most recent comprehensive study by Le (2026) on Kalshi. This thesis extends that history to Polymarket, with broader domain coverage, a longer data window, and the first event-level uncertainty quantification for Polymarket data. The core finding is that the Polymarket crowds are broadly wise, but imperfect. This imperfection follows a structure that theory predicts and empirical data confirms, in line with key findings that came before. Whether this imperfection will diminish as Polymarket matures or will persist as a fundamental feature of prediction market crowds reasoning under uncertainty remains an open and important question.

[3] And indeed, millions more may soon encounter them after the imminent launch of Polymarket US, which as of this writing continues to roll out to waitlist users (*USA Waitlist* 2026).

5.1 Limitations

Several limitations of this study exist. First, the cleaned data window is from November 2022 - December 2025, which happens to be the exact time under which Polymarket was "restricted" from operating in the United States. Along with the 32 other countries where Polymarket is currently restricted (see footnote in Section 3.2), this implies Polymarket is operating without access to an extremely large portion of its potential global user base, which could systematically affect trader composition, liquidity, and therefore calibration. Second, calibration is assessed only on resolved markets with publicly available Polymarket API CLOB data. Like any study that filters for resolution, there is an inherent survivorship bias here. Furthermore, the available Polymarket API data is incomplete: an automated market maker (AMM) mechanism was previously used for Polymarket pricing (rather than the current central limit order book mechanism), and historical pricing data from the previous AMM era is no longer available. Third, the study uses price data taken on a daily frequency: intraday price movement and continuous information arrival are invisible. This affects short-horizon calibration estimates, which may be noisier than they appear. Fourth, this study is a single-platform prediction market study of Polymarket. The author makes no claims on the generalizability of findings to other prediction markets with different market microstructure, user demographics, or regulatory environments. Fifth, the first-match priority rule of our domain classification means some markets with multiple tags are inevitably assigned to the wrong domain, which could bias domain-level interpretation in unknown directions. Sixth, our conclusions are limited by model dependence: the findings are only as good as the models we estimate them with. In particular, the parametric logistic recalibration model may not capture some domains or patterns well because of the assumptions it places on curves in logit space. Seventh and finally, the study is limited by time stationarity: the valid data window is treated as a single period. Early Polymarket calibration could be systematically

different from late calibration over the eligible data window, and pooling price data together homogenously could mask these discrepancies.

5.2 Future Work

There are several key areas for future work related to this study. First, trader-level or blockchain-level analysis. This is a unique possibility enabled by the open nature of Polymarket's cryptocurrency-based platform and the fact that all trades are publicly available on the blockchain. A potential research direction would investigate "whale" addresses—those blockchain addresses that are highly active, highly liquid, or both—and how they contribute to calibration. Second, an in-depth cross-platform comparison. This was one of the original motivations for this work, before encountering data issues. This is also motivated by Le (2026)'s approach of using Polymarket as a cross-validation approach for Kalshi analysis. The goal would be to apply the same methodology of rigorous calibration analysis across multiple platforms and compare findings. Third, as an extension of the above, a study with more granular price data could be an area for future work (e.g. minute-by-minute or transaction-level pricing data and order book analysis) that would enable a much more data-rich approach and potentially be far more insightful. Fourth, expanding the calibration analysis with another dimension represents a logical next progression: in addition to domain and horizon, a liquidity or volume dimension might add further clarity to Polymarket calibration. Fifth, a study that incorporates a more advanced domain classification tool. This is related to the first-match priority rule limitation identified above. A potential research direction would investigate whether the use of natural language processing (or another multi-domain classifier) could better group domains for accurate analysis. Sixth, a study that incorporates time period stratification. This is related to the time stationarity limitation identified above. A future direction could be splitting

the data window into distinct periods and testing whether calibration improved or systematically changed as Polymarket matured. Seventh, a study that includes more careful analysis of Polymarket's unique "negRisk" mechanism. negRisk is the tool that enables a user to swap their contracts in a multi-market event setting (e.g. enables swapping 1 "Yes" contract for all other "No" contracts, due to the linked nature of the event). This represents an interesting case study in market microstructure and how it may relate to price calibration. Eighth and finally, answering the question: so what does a trader do with all this information? Using fitted parameters (ideally on a more granular price observation dataset) from calibration analysis, can a user trade on this pricing miscalibration to create meaningful returns and help contribute to better price efficiency? The open-source nature of the Polymarket cryptocurrency-based architecture means that all historical trades are auditable, making this a tractable and interesting direction for future work.

Appendix A

Tag to Domain Mapping

Table A.1 presents the full tag-to-domain mapping used for market classification in this study. Polymarket assigns markets one or more tags from a pool of over 5,000 unique tag IDs. To reduce this to a meaningful set of domains, the top 100 tag IDs by frequency in the final cleaned calibration dataset were inspected and grouped into six named domains plus a catchall Other category. Each market is assigned to the *first* matching domain according to a fixed priority ordering: Politics → Crypto → Finance → Sports → Culture → Weather (e.g., a market tagged with both "Politics" and "Sports" is classified as Politics). Markets with no matching tag are assigned to Other. This mapping accounts for 98.3% of all markets in the final dataset.

Table A.1: Domain classification tag mapping. Tag IDs and Label correspond to Polymarket's internal tagging system. Markets are assigned to the first matching domain according to the ordering above.

Domain	Tag ID — Label	Domain	Tag ID — Label
Politics	2 — Politics	Sports	1 — Sports
	126 — Trump		100639 — Games
	1597 — Global Elections		28 — Basketball
	101191 — Trump Presidency		745 — NBA
	100265 — Geopolitics		100350 — Soccer
	144 — Elections		100396 — NCAA
	24 — USA Election		450 — NFL
	188 — U.S. Politics		102114 — NCAA Basketball
	1101 — US Election		64 — Esports
Crypto	21 — Crypto		100351 — CFB
	1312 — Crypto Prices		100381 — MLB
	235 — Bitcoin		864 — Tennis
	39 — Ethereum		100219 — Golf
	818 — Solana		899 — NHL
	101267 — XRP		435 — Formula 1
	101312 — Ripple		100389 — F1
Finance	120 — Finance		306 — EPL
	604 — Stocks		102780 — CWBB
	102676 — Equities		102112 — PGA Tour
	100328 — Economy		82 — Premier League
	102831 — Stock Prices		101178 — CBB
	102682 — Indices		279 — UFC
Culture	596 — Culture		100398 — Today's Sports
	53 — Movies		102643 — EFL Championship
	18 — Awards		1234 — Champions League
	100 — Music		102146 — Grand Prix
	100407 — Netflix		10 — Football
Weather	84 — Weather		780 — La Liga
Other	*No matching tag*		100088 — Hockey
			100100 — MLS
			100977 — UCL
			256 — Chess
			678 — Baseball
			101787 — UEL
			1494 — Bundesliga
			102070 — Ligue 1

Bibliography

Ali, Mukhtar M. (1977). "Probability and Utility Estimates for Racetrack Bettors". In: *Journal of Political Economy* 85.4, pp. 803–815.

Arrow, Kenneth J. et al. (2008). "The Promise of Prediction Markets". In: *Science* 320.5878, pp. 877–878.

Berg, Joyce, Robert Forsythe, et al. (2008). "Results from a Dozen Years of Election Futures Markets Research". en. In: *Handbook of Experimental Economics Results*. Vol. 1. Elsevier, pp. 742–751. ISBN: 9780444826428. URL: https://doi.org/10.1016/S1574-0722(07)00080-7 (visited on 10/30/2025).

Berg, Joyce, Forrest Nelson, and Thomas Rietz (2003). "Accuracy and Forecast Standard Error of Prediction Markets". Working Draft.

Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1, pp. 1–3. DOI: 10.1016/0167-2681(92)90015-U.

Cutting, Laurie E. et al. (2025). "Are Betting Markets Better than Polling in Predicting Political Elections?" In: arXiv: 2507.08921. URL: https://arxiv.org/abs/2507.08921.

Fama, Eugene F. (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2, pp. 383–417.

Galton, Francis (1907). "Vox Populi". In: *Nature* 75, pp. 450–451.

Grossman, Sanford J. and Joseph E. Stiglitz (1980). "On the Impossibility of Informationally Efficient Markets". In: *The American Economic Review* 70.3, pp. 393–408.

Hanson, Robin (2003). *The Policy Analysis Market (and FutureMAP) Archive*. PAM archive: https://cryptome.org/pam/pam-site.htm#Concept. URL: https://mason.gmu.edu/~rhanson/policyanalysismarket.html (visited on 10/31/2025).

Hayek, Friedrich A. (1945). "The Use of Knowledge in Society". In: *The American Economic Review* 35.4, pp. 519–530.

How accurate is Polymarket? (2025). en. URL: https://polymarket.com/accuracy (visited on 12/13/2025).

Hulse, Carl (July 2003). *THREATS AND RESPONSES: PLANS AND CRITICISMS; Pentagon Prepares A Futures Market On Terror Attacks*. URL: https://www.nytimes.com/2003/07/29/us/threats-responses-plans-criticisms-pentagon-prepares-futures-market-terror.html.

Kahneman, Daniel and Amos Tversky (1979). "Prospect Theory: An Analysis of Decision Under Risk". In: *Econometrica* 47.2, pp. 263–291.

Lattimore, Paul K., Jonathan R. Baker, and Ann D. Witte (1992). "The Influence of Probability on Risky Choice: A Test of Prospect Theory". In: *Journal of Economic Behavior & Organization* 17.3, pp. 377–400. DOI: 10.1016/0167-2681(92)90015-U.

Le, Naman H. (2026). "Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Prediction Markets". In: Working paper.

Looney, Robert E (2004). "DARPA's policy analysis market for intelligence: outside the box or off the wall?" In: *International Journal of Intelligence and Counterintelligence* 17.3, pp. 405–419.

Manski, Charles F. (2006). "Interpreting the Predictions of Prediction Markets". In: *Economics Letters* 91.3, pp. 425–429.

Meyer, Josh (July 2003). *Trading on the Future of Terror*. en-US. URL: https://www.latimes.com/archives/la-xpm-2003-jul-29-na-terror29-story.html (visited on 10/31/2025).

Murphy, Allan H. (1973). "A New Vector Partition of the Probability Score". In: *Journal of Applied Meteorology* 12.4, pp. 595–600. DOI: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Ottaviani, Marco and Peter Norman Sørensen (2008). "The Favorite-Longshot Bias: An Overview of the Main Explanations". en. In: *Handbook of Sports and Lottery Markets*. Elsevier, pp. 83–101. ISBN: 9780444507440. DOI: 10.1016/B978-044450744-0.50009-3. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780444507440500093 (visited on 10/31/2025).

Page, Lionel and Robert T. Clemen (2013). "Do Prediction Markets Produce Well-Calibrated Probability Forecasts?" In: *The Economic Journal* 123.568, pp. 491–513.

Plott, Charles R and Kay-Yut Chen (2002). "Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem". In.

Polymarket 101 (2026). en. URL: https://docs.polymarket.com/polymarket-101 (visited on 04/09/2026).

Prediction Markets Dashboard (2026). *Prediction Markets Dashboard: TVL, Volume and Share 2026*. en. URL: https://predictiontalk.org/prediction-markets/ (visited on 04/09/2026).

Restocchi, Valerio, Frank McGroarty, and Enrico Gerding (June 2019). "The temporal evolution of mispricing in prediction markets". en. In: *Finance Research Letters* 29, pp. 303–307. ISSN: 15446123. DOI: 10.1016/j.frl.2018.08.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1544612318303349 (visited on 10/30/2025).

Rothschild, David (2009). "Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases". In: *Public Opinion Quarterly* 73.5, pp. 895–916.

Schoen, John W. (July 2003). *Pentagon kills 'terror futures market'*. en. URL: https://www.nbcnews.com/id/wbna3072985 (visited on 10/31/2025).

Surowiecki, James (2004). *The Wisdom of Crowds*. New York: Doubleday. ISBN: 9780385503860.

USA Waitlist (2026). en. URL: https://polymarket.com/usa. (visited on 04/09/2026).

Wolfers, Justin and Eric Zitzewitz (2004). "Prediction markets". In: *Journal of economic perspectives* 18.2, pp. 107–126.

Yeh, Puong Fei (2006). "Using Prediction Markets to Enhance US Intelligence Capabilities". In: *Studies in Intelligence* 50.4.