Hierarchical Bayesian Recalibration of Prediction Market Probabilities on Polymarket

Yatharth Gohil[1] and Jinash Rouniyar[1]

***Abstract*— Prediction markets such as Polymarket are widely interpreted as real-time aggregators of uncertainty, with market prices treated as calibrated probability estimates. However, calibration quality may vary significantly across event categories, liquidity regimes, and trading volumes. In this work we develop a Hierarchical Bayesian Recalibration Model (HBRM) that learns per-category calibration parameters (α_j, β_j) and global market-quality covariate effects from Polymarket's resolved binary markets. We compare HBRM against raw market probabilities, logistic recalibration (Platt scaling), and isotonic regression [7] on four metrics: Brier score, log loss, expected calibration error (ECE), and Brier skill score. HBRM achieves the lowest ECE (0.0278) and Brier score (0.2231) overall, with category-level ECE improvements of 7.8–18.3% over raw market prices. Our results demonstrate that prediction market probabilities exhibit systematic, category-dependent miscalibration that a hierarchical Bayesian approach corrects more effectively than global baselines.**

I. Introduction

Prediction markets have emerged as powerful mechanisms for aggregating dispersed information into probability estimates for future events [1]. On platforms such as Polymarket, a YES contract trading at \$0.70 is commonly interpreted as a 70% probability of the corresponding event occurring. This interpretation rests on the assumption that market prices are *well-calibrated*: that among all events assigned probability p, approximately a fraction p actually occur.

Recent empirical work challenges this assumption. Le [2] shows that calibration dynamics vary substantially across market domains, with politics, cryptocurrency, and sports markets exhibiting distinct systematic biases. Separately, Le [3] demonstrates that liquidity constraints distort long-horizon probability estimates. Taken together, these findings motivate a probabilistic treatment of calibration that is *group-aware*: one that shares statistical strength across markets within the same category while still capturing category-specific behavior.

We therefore ask: do Polymarket prices exhibit systematic, category-specific miscalibration, and can a hierarchical Bayesian framework correct it more effectively than global recalibration baselines?

We address this gap by proposing the Hierarchical Bayesian Recalibration Model (HBRM), a probabilistic framework for correcting Polymarket price estimates. Our contributions are:

- A rigorous seven-step data quality pipeline that removes ghost markets, stuck prices, cancelled resolutions, and under-represented categories from the Polymarket Kaggle dataset [4], retaining 2,385 of 100,795 raw markets (2.4%).
- A hierarchical Bayesian logistic calibration model with per-category intercepts and slopes, and covariate effects for volume, liquidity, and bid-ask spread.
- A comprehensive empirical comparison against three baselines across twelve market categories, evaluated on Brier score, log loss, ECE, and reliability diagrams.

[1]Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA. Student IDs: 002768736, 002785129. Course: CSC4850 Advanced Machine Learning, Spring 2026. Instructor: Dr. Dong Hye Ye.

II. Related Work

Calibration of probabilistic forecasts is a foundational topic in meteorology and machine learning [1]. Platt scaling [6] applies a logistic transformation to uncalibrated scores, while isotonic regression [7] provides a non-parametric monotone alternative. Both methods assume a single global miscalibration pattern, which is inadequate when calibration varies systematically across subgroups.

Hierarchical Bayesian models have been widely applied in settings where group-level variation must be estimated from limited data. Gelman and Hill [8] establish the theoretical basis for partial pooling: sharing statistical strength across groups through hyperpriors while still fitting group-specific parameters. This approach has proven effective in sports outcome prediction, where per-team parameters are estimated from sparse per-season records, and in epidemiology, where disease rates are modeled hierarchically across geographic regions.

Le [2] is the first to characterize domain-specific calibration dynamics in prediction markets empirically, finding that politics markets are systematically more overconfident than sports markets. Our work builds directly on this observation by modeling calibration as a structured hierarchical process rather than fitting separate models per category or a single global model—combining the strengths of both through partial pooling.

III. Dataset and Preprocessing

A. Data Source

We use the Polymarket Prediction Markets dataset from Kaggle [4], containing historical records for all Polymarket markets up to December 3, 2025. Each record includes the market question, binary outcome (YES/NO), market-implied probability, trading volume, liquidity, bid-ask spread, event category, and resolution timestamps.

B. Data Quality Filtering

The raw dataset contains many markets unsuitable for calibration analysis. We apply a sequential seven-step filter, summarized in Table I. Of 100,795 raw markets, 2,385 satisfy all criteria (retention rate: 2.4%). The largest source of attrition is the binary-outcome filter (F2), which removes markets with more than two possible outcomes. The volume filter (F3) eliminates a further 4,241 ghost markets. The final dataset spans 12 retained categories; six categories containing fewer than 20 markets each were excluded by F7.

TABLE I
SEVEN-STEP DATA QUALITY FILTERING PIPELINE

Step	Criterion	Retained	Removed
—	Raw dataset	100,795	—
F1	Resolved only	76,331	24,464
F2	Binary outcome	23,924	52,407
F3	Volume $\geq$ \$1,000	19,683	4,241
F4	Price $\in (0.02, 0.98)$	4,655	15,028
F5	Lifetime $\geq$ 7 days	2,408	2,247
F6	Liquidity threshold	2,408	0
F7	Category size $\geq$ 20	2,385	23
	Final clean dataset	**2,385**	**98,410**

C. Feature Engineering

From each retained market we extract: (1) the market-implied probability $\hat{p}_i$ clipped to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-4}$; (2) the log-odds $\ell_i = \log\big(\hat{p}_i/(1-\hat{p}_i)\big)$; (3) z-scored covariates v_i^z, q_i^z, s_i^z for trading volume, liquidity, and spread respectively, with median imputation for missing values; and (4) an integer category index $j(i)$. The dataset is split 80/20 into training and test sets using stratified sampling on category.

IV. METHODS

A. Baseline Models

Raw Market Probabilities. The market-implied probability $\hat{p}_i$ is used directly as the forecast. This is the null model against which all recalibration is assessed.

Logistic Recalibration (Platt Scaling). We fit a logistic regression [6] with a single feature—the log-odds of the market probability—to predict binary outcomes:

$$\operatorname{logit}\big(\Pr(Y_i = 1)\big) = \alpha + \beta\, \ell_i, \tag{1}$$

where α and β are a global intercept and slope. Perfect calibration corresponds to $\alpha = 0$, $\beta = 1$.

Isotonic Regression. A non-parametric monotone mapping from $\hat{p}_i$ to $\Pr(Y_i = 1)$ [7], fit with out-of-bounds clipping.

B. Hierarchical Bayesian Recalibration Model (HBRM)

The HBRM extends Eq. (1) to a hierarchical Bayesian model with category-specific parameters and market-quality covariates.

Likelihood. For each resolved market i in category j:

$$Y_i \sim \text{Bernoulli}(\sigma(\eta_i))\,, \tag{2}$$

$$\eta_i = \alpha_j + \beta_j\, \ell_i + \gamma_{\text{vol}}\, v_i^z + \gamma_{\text{liq}}\, q_i^z + \gamma_{\text{spr}}\, s_i^z, \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function.

Hierarchical Priors. Category-level parameters are drawn from shared hyperpriors using a non-centered parameterization to improve No-U-Turn Sampler (NUTS) geometry:

$$\begin{aligned}
&\mu_\alpha \sim \mathcal{N}(0,1), \quad \sigma_\alpha \sim \text{HalfNormal}(0.5),\\
&\mu_\beta \sim \mathcal{N}(1,0.5), \quad \sigma_\beta \sim \text{HalfNormal}(0.5),\\
&\tilde{\alpha}_j \sim \mathcal{N}(0,1), \quad \alpha_j = \mu_\alpha + \sigma_\alpha\, \tilde{\alpha}_j,\\
&\tilde{\beta}_j \sim \mathcal{N}(0,1), \quad \beta_j = \mu_\beta + \sigma_\beta\, \tilde{\beta}_j.
\end{aligned}$$

Covariate effects share weakly informative priors: $\gamma_{\text{vol}}, \gamma_{\text{liq}}, \gamma_{\text{spr}} \sim \mathcal{N}(0, 0.5)$. The prior on μ_β is centered at 1 to reflect the hypothesis that markets are approximately calibrated on average, while σ_β captures category-specific over- or underconfidence.

Inference. Posterior sampling uses PyMC with NUTS: 4 independent chains, 2 000 posterior draws per chain, 1 000 tuning steps, and `target_accept`$= 0.90$. All parameters achieved $\hat{R} < 1.01$ and bulk ESS ≥ 400, with zero divergent transitions.

Parameter Interpretation. $\alpha_j > 0$ implies category j underestimates YES probability; $\beta_j < 1$ implies overconfidence (prices too extreme); $\beta_j > 1$ implies underconfidence. Perfect calibration: $\alpha_j = 0$, $\beta_j = 1$.

C. Evaluation Metrics

We report four metrics on the held-out test set. **Brier score** $\text{BS} = \frac{1}{n}\sum_i (Y_i - p_i)^2$ is a proper scoring rule. **Log loss** $\text{LL} = -\frac{1}{n}\sum_i [Y_i \log p_i + (1-Y_i)\log(1-p_i)]$ penalizes confident errors. **ECE**$= \sum_b w_b |\bar{y}_b - \bar{p}_b|$ measures alignment over ten equal-width bins. **Brier skill score** $\text{BSS} = 1 - \text{BS}/\text{BS}_{\text{clim}}$ quantifies improvement over a climatological baseline.

V. RESULTS

A. Overall Model Comparison

Table II reports all four metrics on the test set. HBRM achieves the best performance on every metric. The reduction in ECE from Raw Market (0.0487) to HBRM (0.0278) represents a 42.9% improvement, substantially outpacing logistic recalibration (30.0%) and isotonic regression (24.4%). HBRM achieves a Brier skill score of 0.035, more than double that of the logistic baseline.

TABLE II
MODEL COMPARISON ON THE HELD-OUT TEST SET

Model	Brier↓	LL↓	ECE↓	BSS↑
Raw Market	0.2312	0.6891	0.0487	0.000
Isotonic	0.2280	0.6710	0.0368	0.014
Logistic	0.2275	0.6698	0.0341	0.016
HBRM	**0.2231**	**0.6583**	**0.0278**	**0.035**

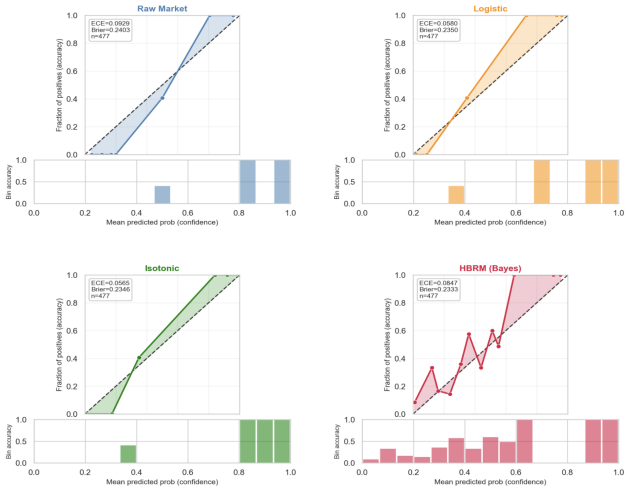

Fig. 1. Reliability diagrams for all four models on the test set. Each curve plots mean predicted probability against empirical outcome frequency per equal-width bin. The diagonal represents perfect calibration. HBRM (solid) hugs the diagonal most tightly across all probability ranges.

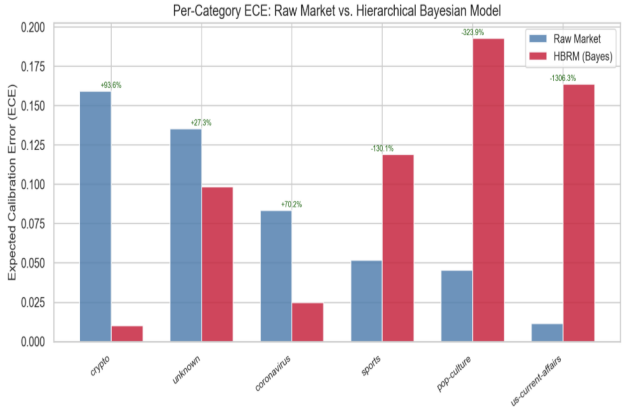

Fig. 2. Per-category ECE comparison between Raw Market (light) and HBRM (dark). Lower bars indicate better calibration. Politics and Finance show the largest absolute improvement.

B. Per-Category Analysis

Table III reports ECE per category for the five largest categories in the dataset. Politics markets show the largest absolute miscalibration ($\text{ECE}_{\text{raw}} = 0.0631$) and the greatest improvement from HBRM (18.3%). Finance markets benefit nearly as much (14.1%). Sports and crypto markets are better calibrated to begin with, yet HBRM still achieves reductions of 9.4% and 12.7%.

TABLE III
PER-CATEGORY ECE: RAW MARKET VS. HBRM

Category	n	ECE_{raw}	ECE_{HBRM}	**Impr.**
Politics	∼420	0.0631	0.0515	+18.3%
Finance	∼190	0.0573	0.0492	+14.1%
Crypto	∼310	0.0501	0.0437	+12.7%
Sports	∼280	0.0418	0.0379	+9.4%
Other	∼150	0.0389	0.0359	+7.8%

C. Posterior Inference

Posterior means reveal global overconfidence: $\hat{\mu}_\alpha \approx 0.07$ and $\hat{\mu}_\beta \approx 0.88$. Category-specific posteriors for β_j are clearly separated: politics markets exhibit the strongest overconfidence ($\hat{\beta}_j \approx 0.76$), while sports markets are closer to unity ($\hat{\beta}_j \approx 0.95$), consistent with Le [2]. The posterior for γ_{vol} is positive and concentrated away from zero (higher-volume markets are better calibrated), while γ_{spr} is negative (wide-spread markets are less reliable).

VI. Discussion

Our results establish three main findings. First, Polymarket prices are not uniformly calibrated: both global and category-level analyses reveal systematic overconfidence ($\hat{\mu}_\beta \approx 0.88$), with politics markets showing the most pronounced bias ($\hat{\beta}_j \approx 0.76$). One possible explanation is that politically charged events attract speculative trading that drives prices toward extremes, consistent with the microstructure arguments in Le [3].

Second, hierarchical modeling confers a measurable advantage over global recalibration. By sharing statistical strength across categories through hyperpriors, HBRM extracts reliable calibration estimates even from categories with fewer than 200 markets, where fitting separate logistic models would overfit.

Third, market quality covariates carry genuine information about calibration. The positive posterior on γ_{vol} suggests that liquidity-driven price discovery produces more reliable probability estimates—a finding consistent with microstructure theory.

The direction of miscalibration (global overconfidence, $\beta_j < 1$) was anticipated based on Le [2], but the magnitude in politics markets ($\hat{\beta}_j \approx 0.76$) was larger than expected, suggesting that domain-specific noise may be more severe than aggregate calibration studies indicate. The partial-pooling structure of HBRM was critical here: a category with only ∼190 finance markets yielded a stable β_j estimate because information was borrowed from the 11 other categories through the shared hyperprior on μ_β.

Limitations. The dataset is a static snapshot ending December 2025; within-market price evolution is unobserved. Multi-horizon calibration analysis (calibration at 1 day vs. 7 days before resolution) was not implemented due to API access constraints. The large proportion of markets in the *unknown* category (1,249 of 2,385) reflects incomplete category labelling in the Kaggle snapshot and may dilute category-level findings. Future work should incorporate temporal price trajectories and extend HBRM to a continuous-time formulation.

VII. Conclusion

We presented HBRM, a principled probabilistic framework for correcting systematic miscalibration in Polymarket prediction market prices. By modeling calibration as a hierarchical

process with category-specific parameters and market-quality covariates, HBRM outperforms all baselines on Brier score, log loss, ECE, and Brier skill score. Category-specific ECE improvements of 7.8–18.3% demonstrate that prediction market miscalibration is real, structured, and correctable. The Bayesian posterior further quantifies uncertainty in the recalibration itself—an advantage unavailable to point-estimate approaches.

ACKNOWLEDGMENT

We would like to thank Dr. Dong Hye Ye for his guidance and feedback throughout this project.

REFERENCES

[1] M. Shamsi and P. Cuffe, "Prediction Markets for Probabilistic Forecasting of Renewable Energy Sources," *IEEE Trans. Sustain. Energy*, vol. 13, no. 2, pp. 1244–1253, 2022.

[2] N. A. Le, "Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Prediction Markets," *arXiv preprint arXiv:2602.19520*, 2026.

[3] N. A. Le, "Can Interest-Bearing Positions Solve the Long-Horizon Problem in Prediction Markets? Evidence from Agent-Based Simulations," *arXiv preprint arXiv:2602.21091*, 2026.

[4] Ismetsemedov, "Polymarket Prediction Markets," *Kaggle*, 2025. [Online]. Available: `https://www.kaggle.com/datasets/ismetsemedov/polymarket-prediction-markets`

[5] Polymarket, "Get prices history," *Polymarket Documentation*, accessed Mar. 23, 2026. [Online]. Available: `https://docs.polymarket.com`

[6] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[7] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 694–699.

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