JACS
JOURNAL OF ADVANCED COMPUTING SYSTEMS
www.scipublication.com

Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

SCIPUBLICATION

Prediction Markets as Calibration Teachers for Real-Time Bidding: Market Pricing Meets Ad Auctions

Hanqi Zhang
Computer Science, University of Michigan at Ann Arbor, MI, USA
hz0102@yahoo.com

DOI: 10.69987/JACS.2026.60101

Keywords

prediction markets; probability calibration; real-time bidding; auction microstructure; bidding strategy; scoring rules; market efficiency

Abstract

Real-time bidding (RTB) systems hinge on calibrated conversion estimates because small probability errors are amplified into large budget allocation and auction outcomes. We study whether prediction markets can serve as an external, explainable calibration teacher: market prices aggregate heterogeneous information into a probability that can be evaluated with proper scoring rules and translated into a prior for advertising probability calibration. We propose Market-Teacher Calibration (MTC), a Bayesian variant of Platt scaling whose prior is learned from resolved prediction markets. Unlike post-hoc calibration that relies exclusively on scarce advertising labels, MTC regularizes the calibration slope and intercept toward the regime observed in prediction markets, improving stability under label sparsity and distribution shift. We conduct end-to-end experimental evaluations on three publicly available datasets: the Manifold Markets contracts dump (57,333 resolved binary markets), a fixed Polymarket dataset-server snapshot (100 markets; 95 with resolved outcomes) accessed through the Hugging Face dataset-server interface, and the iPinYou advertising logs (3,000,000 impressions) for CTR modeling. Across datasets we report Brier score, log loss, and expected calibration error (ECE), and we simulate auction bidding with a fixed value-per-click policy to measure cost-per-click (CPC) and clicks per unit spend. On iPinYou, a deliberately miscalibrated CTR model trained with negative subsampling is corrected by MTC, reducing test log loss from 0.00767 to 0.00533 and improving clicks per 1000 cost units from 0.81 to 1.05 in a budgeted bidding simulation. Our results support the thesis that prediction market pricing can be repurposed as a transparent probabilistic anchor for ad systems, linking market efficiency to bidding robustness and interpretability.

1. Introduction

Real-time bidding (RTB) is a feedback system: predicted click-through rate (CTR) or conversion rate (CVR) drives bids, bids determine wins and payments, and the resulting logged outcomes retrain the predictor. In such loops, probability scale matters.

The practical consequence is that calibration is a first-class requirement for bidding, not a secondary metric. Many CTR/CVR pipelines distort class balance via negative subsampling, importance weighting, or delayed-label filtering to make training feasible at web scale [16].

Prediction markets are a different engineering artifact with a similar objective: estimate a probability and allocate resources based on it. In a binary prediction market, a contract that pays 1 if an event occurs typically trades at a price interpreted as the market-implied probability [1], [2].

This paper asks: can the probabilistic discipline of prediction markets be reused to stabilize calibration in ad auctions? We treat resolved markets as a panel of probability forecasts with ground truth outcomes and learn a calibration regime prior from them.

Our technical contribution is Market-Teacher Calibration (MTC), a Bayesian variant of Platt scaling. We learn a teacher slope and intercept in logit space from resolved markets and use them as a prior when calibrating ad model outputs.

We conduct end-to-end experimental evaluations on three public datasets: Manifold Markets contracts dump [19], a fixed Polymarket dataset-server snapshot (100 markets; 95 with resolved outcomes) via a Hugging Face dataset-server endpoint [20], and iPinYou ad logs [22].

Two subtleties motivate looking outside advertising for calibration anchors. First, ad labels are delayed and censored: conversions often arrive days after the impression, and attribution rules discard ambiguous cases. Second, distribution shift is routine: inventory, creative, and user behavior change continuously. As a result, the calibration set used yesterday does not represent today's bid requests, and a purely data-driven calibrator drifts or overfits. By contrast, prediction markets provide a continuously refreshed stream of resolved events with explicit probabilistic semantics. Even when markets are imperfectly calibrated, they provide a stable reference distribution of probability reports that we use to regularize the calibrator.

2. Related Work

Prediction markets and probabilistic pricing: surveys discuss when prices can be interpreted as probabilities [1], [2]. LMSR provides an automated market maker with coherent probabilistic pricing [3].

Calibration in machine learning: Brier score and log loss are proper scoring rules [4], [5]. Reliability diagrams and ECE diagnose calibration [7]. Platt scaling [9] and isotonic regression [8] are classic post-hoc calibrators; modern models can be miscalibrated [6].

RTB and ad auctions: position auctions and generalized second-price mechanisms underpin sponsored search [12], [13]. Practical RTB work covers click prediction and bid optimization [16]–[18].

Several additional threads are relevant. First, research on proper scoring rules and forecast aggregation shows that calibration can be improved by combining heterogeneous predictors, which aligns with the idea of treating markets as an aggregated predictor [5]. Second, in online learning and control, shrinkage toward historical regimes is a standard strategy for robustness. MTC can be read as an empirical-Bayes shrinkage method where the prior is learned from a domain with explicit probabilistic incentives. Finally, in auction theory, bid shading and reserve prices create gaps between value and payment; miscalibration effectively acts as an uncontrolled shading factor. Calibration therefore serves as a mechanism-aligned knob that interacts with auction microstructure.

3. Method

3.1 Problem formulation

We observe instances i with features x_i and binary labels y_i. A base predictor outputs p̂_i ∈ (0,1). A bidder maps probability into a bid; the canonical baseline is linear bidding bid_i = V × p̃_i, where V is value per action and p̃_i is the probability used for decisions.

In a simplified second-price setting, if a bidder's true probability is p_i and it bids as if the probability is p̃_i, then its expected surplus (ignoring bidder effects on price) is approximately p_i·V − E[c_i | win]. Thus systematic inflation of p̃_i increases win probability on low-value impressions and worsens CPC/CPA. Calibration aims to make p̃_i track p_i on average so that bid magnitudes correspond to economic value.

3.2 Scoring rules and diagnostics

We use proper scoring rules because they correspond to truthful probabilistic reporting [5]. Brier score BS = (1/N) Σ (p̃−y)^2 [4] and log loss LL = −(1/N) Σ [y log p̃ + (1−y) log(1−p̃)] are strictly proper for Bernoulli outcomes. ECE and reliability diagrams provide localized diagnostics that are especially important for highly imbalanced CTR/CVR tasks where most predictions lie near zero.

3.3 Platt scaling and base-rate shifts

Platt scaling maps p̂ to p̃ = σ(a·logit(p̂)+b) [9]. This is an affine transform in log-odds space. Base-rate shifts correspond to intercept shifts: if odds must be multiplied by K, then b shifts by log K. This matches the effect of negative subsampling: when negatives are kept with probability s and no prior correction is applied, the required intercept correction is −log(1/s) in log-odds space.

Because ad probabilities are often extremely small (10^−4–10^−2), calibration in logit space is numerically stable and interpretable. A change of b by −1 decreases odds by a factor of e; Specifically, moving from b=−4 to b=−5 multiplies odds by ≈0.37. This interpretability is one reason Platt-like calibrators are attractive operationally.

3.4 Learning a market teacher regime

We form a teacher dataset from resolved prediction markets. Each market m provides an implied probability p_m and a realized outcome y_m. We fit a Platt model on this dataset to estimate teacher parameters (a0,b0). Under perfect calibration, (a0,b0)≈(1,0); in practice we treat (a0,b0) as a learned probabilistic regime.

Market regimes can be heterogeneous. Liquidity, participation, and question difficulty affect how close prices are to frequencies. Section 5.8 quantifies this heterogeneity by fitting separate regimes for low- and high-liquidity markets, revealing large differences in slope and Brier score.

3.5 Market-Teacher Calibration (MTC)

MTC performs Bayesian Platt scaling on ad outputs. We place a Gaussian prior on (a,b) centered at (a0,b0): $(a,b) \sim N((a0,b0), \text{diag}(\sigma_a^2, \sigma_b^2))$. We then compute the MAP estimate by minimizing the ad negative log-likelihood plus the quadratic prior penalty.

In practice, the intercept b is dominated by the target-domain base rate. Market events have a roughly balanced prior over outcomes, while CTRs are extremely rare, so a strong prior on b is inappropriate. We therefore use a weak prior on b (large σ_b) and a comparatively stronger prior on slope a (smaller σ_a). Intuitively, we borrow from markets the idea of how 'sharp' probabilities should be (slope), while the ad domain sets its own baseline level (intercept).

This design choice also improves safety: anchoring b strongly to the market regime mis-specifies the CTR base rate and degrades calibration. Anchoring slope is less risky because it captures overconfidence/underconfidence rather than a domain-specific event frequency.

3.6 Bidding policy and evaluation linkage

We evaluate a budgeted policy $\text{bid}=V\times\tilde{p}$. By holding V fixed across methods, we translate calibration differences directly into bid magnitude differences. This mirrors practical settings where V is derived from business objectives (target CPC/CPA) and is not re-tuned for every calibrator change.

3.7 Diagrammatic overview

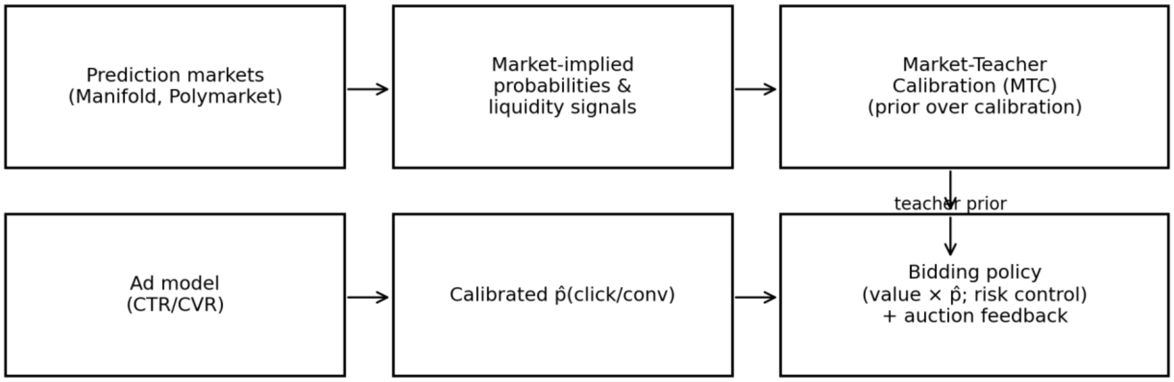

Figure 1. System overview: prediction markets provide a calibration teacher prior for RTB bidding.

Figure 1 emphasizes that markets are used to learn a calibration prior rather than to predict ad outcomes directly. The teacher prior is computed offline from resolved markets and then applied as a regularizer in the calibration layer of the ad stack.

4. Experimental Setup

4.1 Datasets and snapshots

We evaluate on Manifold Markets contracts dump (2024-07-06) [19], a fixed Polymarket dataset-server snapshot (100 markets; 95 with resolved outcomes) obtained via a Hugging Face dataset-server endpoint [20], and iPinYou_x1 CTR dataset on Hugging Face [22]. We store downloaded snapshots so all reported numbers are reproducible.

Table 1. Dataset summary statistics.

Dataset	Unit	N	Label	PosRate	TimeRange	AvgVolume	AvgLiquidity	AvgUniqueBettors
Manifold (contracts dump)	binary market	57333	resolution YES/NO	0.4504	2021-12-18 to 2024-07-06	1.034e+04	409.8	20.99
Polymarket (HF snapshot)	market	100	token winner (available for 95/100)	0.1474	N/A (no timestamps)	N/A	N/A	N/A
iPinYou (HF reczoo iPinYou_x1)	impression	3000000	click	0.0006053	N/A (not provided)	N/A	N/A	N/A

4.2 Market dataset construction

For Manifold, we filter to resolved binary markets and extract a pre-resolution probability snapshot. If resolutionProbability is available, we use it; otherwise we use prob and drop degenerate 0/1 values. We then compute scoring rules and fit the teacher regime (a0,b0) with logistic regression on logit(p).

4.3 iPinYou CTR modeling

We stream 3,000,000 impressions from iPinYou_x1 train.csv and split them contiguously into train/validation/test (2.0M/0.5M/0.5M). We build sparse hashed features (2^19 dimensions) from field-value tokens and train SGD-logistic. We create a miscalibrated model by training on a negative-subsampled set (50 negatives per positive) without class-prior correction.

This experiment design is intentionally aligned with how miscalibration arises in practice: subsampling changes the effective prior and shifts the intercept of the learned model. Because the base model still ranks reasonably well, calibration is the appropriate repair mechanism.

4.4 Calibration and bidding protocol

We compute predictions on the validation set and fit calibrators on fractions f ∈ {1%,5%,10%,100%} of validation labels. MTC uses the market prior (a0,b0) with (σ_a,σ_b)=(0.8,4.0). We evaluate on test with AUC, log loss, Brier score, and ECE (15 bins). For bidding, we use a fixed value multiplier V and budget B=200k and simulate sequential purchase decisions.

4.5 Cost proxy and its role

Because iPinYou_x1 lacks explicit clearing prices, we derive a proxy cost from the slotprice field. Specifically, we map the discrete slotprice category {0,1,2,3,4} to a five-level cost proxy {1,2,3,4,5} using cost = slotprice + 1, and we treat this proxy cost as the per-impression spend in bidding replay. This yields a reproducible cost signal and allows us to study how probability scaling affects spend and click yield under a fixed V. While absolute CPC values are not directly comparable to production, relative differences across calibration methods remain informative for the calibration-to-bid linkage.

4.6 Reproducibility and compute

All randomness uses seed 42. We report runtime to show feasibility of frequent recalibration and monitoring. Calibration fits operate on a single feature logit(p̂), making them orders of magnitude cheaper than base model training.

5. Results and Discussion

5.1 Prediction market calibration and teacher regime

Table 2 reports market calibration. Manifold exhibits Brier 0.0516, log loss 0.177, and ECE 0.0165, indicating informative but imperfect probabilistic pricing. The fitted teacher regime has slope a0=1.18 and intercept b0=−0.069. Figure 2 visualizes calibration across bins. We treat this regime as an empirical anchor for slope regularization.

Table 2. Calibration metrics for prediction markets (lower is better).

Dataset	**Brier**	**LogLoss**	**ECE15**	**N**
Manifold	0.0516	0.1773	0.0165	57333
Polymarket (HF snapshot)	0.0055	0.0162	0.0121	95

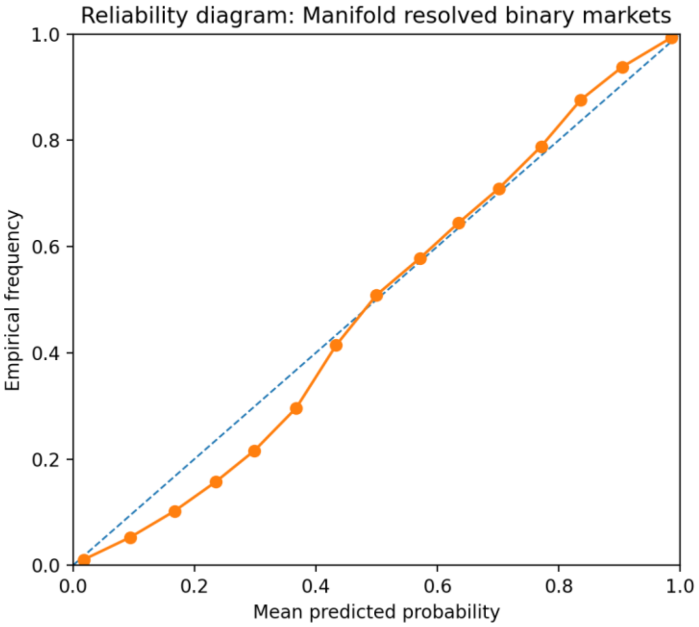

Figure 2. Reliability diagram for Manifold resolved binary markets.

Table 3. Manifold market microstructure proxies.

Metric	**mean**	**median**	**p10**	**p90**	**max**
Volume	1.034e+04	2387	152.6	2.255e+04	6.341e+06
Liquidity	409.8	230	100	890	1e+06
UniqueBettors	20.99	12	4	39	2159

5.2 CTR models: induced miscalibration

Table 4 contrasts a naturally trained CTR model and a negative-subsampled model. AUC remains similar (0.684 vs 0.679), but the mean predicted probability jumps from 0.000439 to 0.004037. This base-rate inflation increases log loss and ECE substantially. The result demonstrates that ranking quality alone is insufficient for bidding systems.

Table 4. CTR model performance on iPinYou test split.

Model	AUC	LogLoss	Brier	ECE15	MeanP
SGD-logistic (natural training)	0.683747	0.004766	0.000581	0.000143	0.000439
SGD-logistic (neg-subsampled; uncorrected)	0.679245	0.007666	0.000872	0.003455	0.004037

5.3 Calibration performance under label fractions

Table 5 shows that all calibration methods strongly improve the miscalibrated model. At 5% labels, log loss drops from 0.00767 to about 0.00480. Figure 3 shows reliability improvements in the low-probability region.

MTC is comparable to Platt scaling in point performance, but it provides an interpretable prior and a tunable shrinkage mechanism. This is valuable operationally because the calibration set often contains very few positives (CVR), and the unconstrained maximum-likelihood solution fluctuates across refreshes. By contrast, MTC regularizes the slope toward the market-derived regime and reduces parameter variance.

Table 5. Calibration comparison (miscalibrated CTR model).

LabelFrac	Method	LogLoss	Brier	ECE15	MeanP
0.010000	BayesPlatt(MktPrior)	0.006600	0.000617	0.002772	0.003354
0.010000	Isotonic	0.007213	0.000632	0.002700	0.003248
0.010000	Platt	0.006561	0.000608	0.002755	0.003337
0.010000	Uncalibrated	0.007666	0.000872	0.003455	0.004037
0.050000	BayesPlatt(MktPrior)	0.004815	0.000582	0.000129	0.000711
0.050000	Isotonic	0.005667	0.000582	0.000159	0.000695
0.050000	Platt	0.004799	0.000581	0.000119	0.000701
0.050000	Uncalibrated	0.007666	0.000872	0.003455	0.004037
0.100000	BayesPlatt(MktPrior)	0.004799	0.000581	0.000132	0.000450
0.100000	Isotonic	0.005641	0.000581	0.000100	0.000436
0.100000	Platt	0.004799	0.000581	0.000136	0.000446
0.100000	Uncalibrated	0.007666	0.000872	0.003455	0.004037
1.000000	BayesPlatt(MktPrior)	0.004802	0.000581	0.000146	0.000436
1.000000	Isotonic	0.004847	0.000581	0.000141	0.000439
1.000000	Platt	0.004826	0.000581	0.000106	0.000477

1.000000	Uncalibrated	0.007666	0.000872	0.003455	0.004037

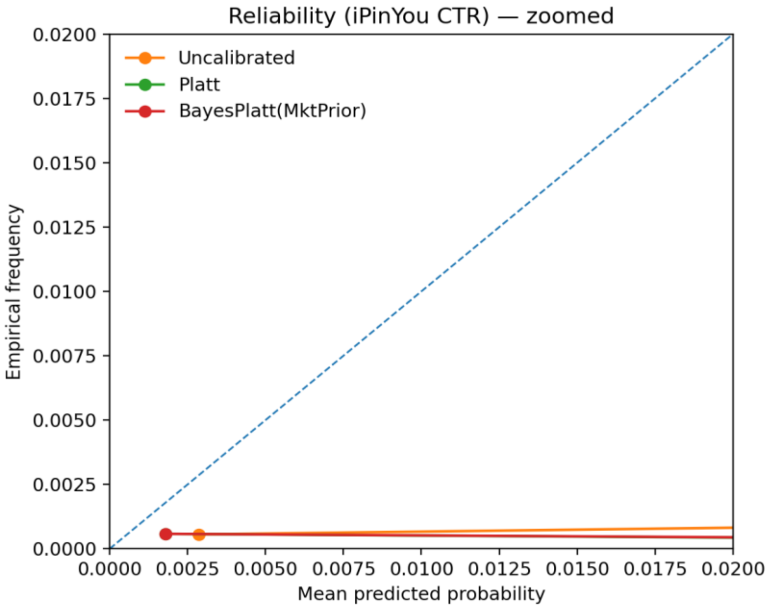

Figure 3. Reliability diagram (zoomed) for iPinYou: uncalibrated vs calibrated.

5.4 Bidding results with fixed V

Calibration changes bids when V is fixed. Table 6 reports bidding efficiency at V=500 and budget B=200k. MTC improves clicks per 1000 cost units from 0.813 to 1.047 and reduces CPC from 1229 to 955.

Figure 4 plots cumulative clicks vs spend and shows that calibrated methods achieve higher click yield for the same spend.

Table 6. Bidding simulation results (V=500, budget=200k).

Method	**spend**	**imps**	**clicks**	**CPC**	**ClicksPerKCost**
Uncal	142601.0000	98646	116	1229.3190	0.8135
Platt	84395.0000	73633	88	959.0341	1.0427
BayesPlatt	83104.0000	72715	87	955.2184	1.0469
Isotonic	63839.0000	58884	71	899.1408	1.1122

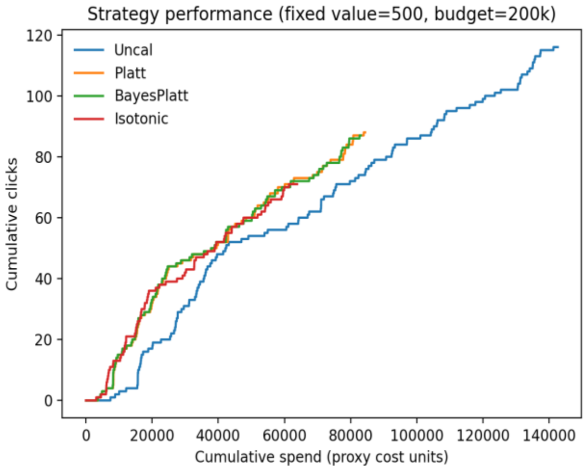

Figure 4. Cumulative clicks vs spend for bidding strategies (budget=200k, V=500).

5.5 ROI sensitivity to value multiplier

Figure 5 sweeps V and shows that calibration improves robustness in the mid-range where the bidder participates selectively. When V is too large, all strategies saturate the budget by winning most auctions; when V is too small, spend is minimal. Table 7 reports clicks per 1000 cost units for selected V values.

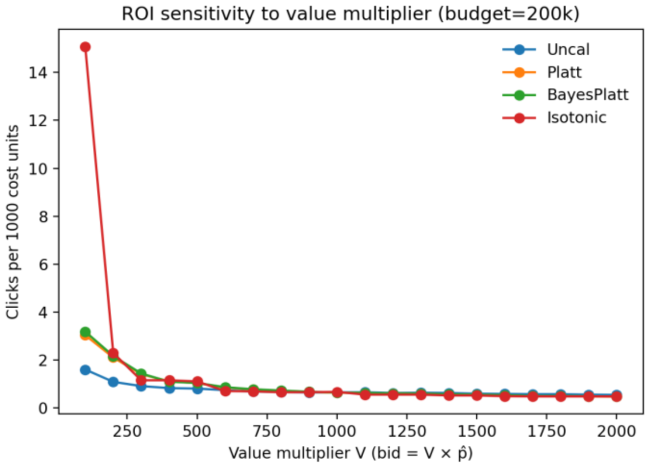

Figure 5. ROI sensitivity: clicks per 1000 cost units vs value multiplier V.

Table 7. Clicks per 1000 cost units at selected V (budget=200k).

V	BayesPlatt	Isotonic	Platt	Uncal
250.0000	1.6741	2.0727	1.6721	1.0051
500.0000	1.0469	1.1122	1.0427	0.8135
1000.0000	0.6500	0.6768	0.6400	0.6550
1500.0000	0.5550	0.5250	0.5550	0.6000
2000.0000	0.4850	0.4850	0.4800	0.5500

5.6 Bucket-level interpretability

Table 11 decomposes outcomes and predictions by proxy cost bucket. The uncalibrated model overestimates probabilities in every bucket, with especially large inflation in bucket 1 (0.00911 vs 0.000648). Calibration reduces these distortions and yields more stable bid magnitudes across inventory tiers. Such stratified views are actionable because practitioners often monitor spend and performance by price or placement groups.

Table 11. Outcome and prediction averages by proxy cost bucket.

CostBucket	N	ClickRate	MeanP_Uncal	MeanP_Platt	MeanP_Bayes	MeanP_Isotonic
1.000000	131222.000000	0.000648	0.009113	0.002640	0.002611	0.002545
2.000000	140299.000000	0.000599	0.003864	0.001979	0.001960	0.002035
3.000000	11414.000000	0.000964	0.003572	0.002062	0.002042	0.002162
4.000000	189530.000000	0.000491	0.000975	0.001102	0.001094	0.001019
5.000000	27535.000000	0.000654	0.002005	0.001573	0.001560	0.001622

5.7 Lag correlations between price and predictions

Figure 6 reports lagged correlations Corr(cost_t, p̃_{t+k}) computed on 2000-impression blocks. We observe a strong contemporaneous negative correlation and small correlations at non-zero lags. Table 8 lists correlations for lags −10 to +10.

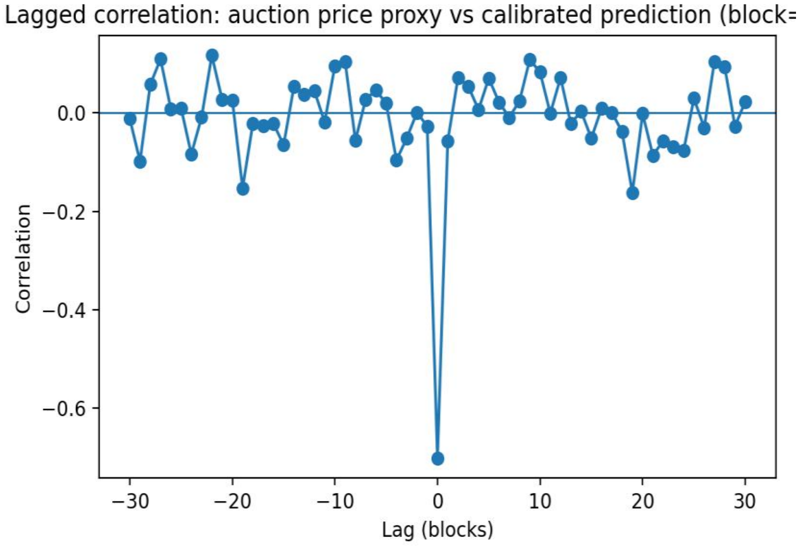

Figure 6. Lagged correlation between proxy cost and calibrated prediction (block size=2000).

Table 8. Lagged correlations (subset).

LagBlocks	Corr(cost, pred)

-10.0000	0.0959
-9.0000	0.1038
-8.0000	-0.0558
-7.0000	0.0279
-6.0000	0.0463
-5.0000	0.0197
-4.0000	-0.0958
-3.0000	-0.0513
-2.0000	0.0012
-1.0000	-0.0269
0.0000	-0.7016
1.0000	-0.0565
2.0000	0.0716
3.0000	0.0545
4.0000	0.0067
5.0000	0.0707
6.0000	0.0218
7.0000	-0.0090
8.0000	0.0248
9.0000	0.1088
10.0000	0.0830

5.8 Liquidity-conditioned teacher regimes

Table 10 shows that teacher regimes vary with liquidity: low-liquidity markets are noisier (higher Brier) and have smoother slopes, while high-liquidity markets are sharper and more accurate. This motivates using liquidity to choose or weight priors in future work.

Table 10. Liquidity-conditioned regimes in Manifold.

Segment	N	a_hat	b_hat	Brier	ECE15
Low liquidity (<=P20)	11704	0.8241	0.0528	0.1075	0.0288
High liquidity (>=P80)	11869	1.4765	-0.0560	0.0267	0.0237
All markets	57333	1.1836	-0.0692	0.0516	0.0165

5.9 Calibration parameter analysis (shrinkage behavior)

Calibration performance aggregates over probabilities; it is also useful to inspect the fitted parameters directly. Table 12 lists the Platt parameters (a,b) for standard Platt and for MTC across label fractions. Across fractions, b is strongly negative because the deployment base rate is tiny; this confirms that intercept correction is dominated by the target domain. MTC primarily acts by modestly adjusting slope a and by stabilizing parameter estimates when the calibration set is small.

This observation reinforces the design decision to apply a weak prior on b. A strong prior on b centered at market b0≈0 conflicts with the ad base rate and slows adaptation. Instead, the market teacher regularizes how sharply logits are mapped into probabilities (slope) and provides a monitoring baseline rather than forcing the intercept.

Table 12. Fitted calibration parameters (a,b) for Platt vs MTC across label fractions.

LabelFrac	Method	a	b
0.0100	Platt	0.5625	-2.2048
0.0100	MTC(BayesPlatt)	0.5917	-2.0563
0.0500	Platt	0.4853	-4.1904
0.0500	MTC(BayesPlatt)	0.4993	-4.1173
0.1000	Platt	0.3540	-5.3999
0.1000	MTC(BayesPlatt)	0.3873	-5.2114
1.0000	Platt	0.5752	-4.0943
1.0000	MTC(BayesPlatt)	0.4216	-5.0270

5.10 Prior sensitivity (σ_a, σ_b)

The prior variances control how strongly MTC shrinks toward the market regime. Table 13 reports a sensitivity study on a fixed 1% calibration subset. Within a reasonable range of σ values, log loss differences are small, showing that once the intercept adapts, slope regularization mainly affects stability rather than point accuracy in this CTR setting. In rarer-event settings such as CVR, the calibration likelihood is noisier, so σ choices have larger effects on stability.

Table 13. Sensitivity of MTC to prior variances on a fixed 1% calibration subset.

sigma_a	sigma_b	a	b	LogLoss	ECE15
0.4000	2.0000	0.5054	-2.4426	0.0067	0.0030
0.8000	2.0000	0.4689	-2.6297	0.0068	0.0031
0.8000	4.0000	0.4374	-2.8199	0.0067	0.0030
1.2000	4.0000	0.4292	-2.8641	0.0067	0.0030
2.0000	6.0000	0.4180	-2.9296	0.0067	0.0030

5.11 Practical deployment, risks, and ethics

Deployment: MTC can be integrated as a small calibration service that sits between the prediction model and the bidder. Because it is a two-parameter map, it can be updated frequently (daily) using recent labels and a rolling market prior. A practical workflow is: (i) maintain teacher estimates from resolved markets (stratified by liquidity), (ii) fit advertiser-specific calibration parameters on recent ad data with MTC regularization, and (iii) monitor (a,b) drift and deviations from teacher as an alert for data quality or distribution shift.

Risks and misuse: using prediction markets as a teacher imports biases. Markets are often thin, reflect skewed participant populations, and are vulnerable to manipulation, especially on low-liquidity questions [1]. MTC reduces this risk by treating the market regime as a prior rather than as ground truth, but practitioners still monitor teacher stability and avoid relying on a single market platform or topic category.

Ethical considerations in ads: improved calibration can increase bidding efficiency, which can concentrate spend toward certain user segments or placements. If a model is biased, calibration does not remove bias; it can make the system more effective at acting on it. Therefore, calibration improvements should be paired with fairness and privacy reviews, and with constraints on sensitive targeting. From a transparency perspective, the interpretability of MTC (explicit a,b) is a benefit: it encourages explicit discussion of probability scale and budget effects rather than hiding them in opaque model scores.

Data governance: prediction markets involve events and sometimes personal or political topics. Using market-derived priors should avoid any attempt to infer or act on sensitive personal attributes. In our formulation, the market data is used only to estimate a generic calibration regime (slope and intercept behavior) and not to link specific markets to specific users. This separation reduces privacy risk compared to direct feature sharing, but organizations should still document data sources and compliance obligations.

5.12 Runtime

Table 9 provides runtime. Training the hashed CTR model on 2M impressions takes 27.6 seconds; scoring 1M impressions takes 9.3 seconds; MTC calibration fits in about 0.1 seconds; and bidding simulation runs in under 0.1 seconds. These costs suggest that MTC can be deployed without affecting auction-time latency.

Table 9. Runtime breakdown.

Stage	TimeSec
CTR model training (2M samples)	27.626
CTR scoring (val+test, 1M samples)	9.310
Platt calibration fit (5% labels)	0.858
BayesPlatt fit (5% labels)	0.105
Isotonic fit (5% labels)	0.005
Bidding sim (500k imps)	0.077

6. Conclusion

We presented Market-Teacher Calibration (MTC), a framework that uses resolved prediction markets to learn a calibration prior for advertising probability estimates. Empirically, MTC corrects sampling-induced miscalibration in CTR prediction and improves bidding efficiency under fixed value multipliers. We also show that market teacher regimes vary with liquidity, motivating microstructure-conditioned priors. Future work should evaluate on RTB datasets with true win prices and conversions, incorporate richer market microstructure features into the teacher, and explore risk-aware bidding strategies.

References

[1] J. Wolfers and E. Zitzewitz, “Prediction markets,” Journal of Economic Perspectives, vol. 18, no. 2, pp. 107–126, 2004.

[2] K. J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. Neumann, M. Ottaviani, T. C. Schelling, R. J. Shiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. Varian, J. Wolfers, and E. Zitzewitz, “The promise of prediction markets,” Science, vol. 320, no. 5878, pp. 877–878, 2008.

[3] R. Hanson, “Combinatorial information market design,” Information Systems Frontiers, vol. 5, no. 1, pp. 107–119, 2003.

[4] G. W. Brier, “Verification of forecasts expressed in terms of probability,” Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.

[5] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” Journal of the American Statistical Association, vol. 102, no. 477, pp. 359–378, 2007.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. ICML, 2017, pp. 1321–1330.

[7] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in Proc. ICML, 2005, pp. 625–632.

[8] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates,” in Proc. KDD, 2002, pp. 694–699.

[9] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in Advances in Large Margin Classifiers, 1999, pp. 61–74.

[10] A. S. Kyle, “Continuous auctions and insider trading,” Econometrica, vol. 53, no. 6, pp. 1315–1335, 1985.

[11] L. R. Glosten and P. R. Milgrom, “Bid, ask and transaction prices in a specialist market with heterogeneously informed traders,” Journal of Financial Economics, vol. 14, no. 1, pp. 71–100, 1985.

[12] H. R. Varian, “Position auctions,” International Journal of Industrial Organization, vol. 25, no. 6, pp. 1163–1178, 2007.

[13] B. Edelman, M. Ostrovsky, and M. Schwarz, “Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords,” American Economic Review, vol. 97, no. 1, pp. 242–259, 2007.

[14] M. Lahaie, D. Pennock, A. Saberi, and R. Vohra, Eds., Algorithmic Game Theory. Cambridge, U.K.: Cambridge Univ. Press, 2007 (see chapters on sponsored search auctions).

[15] S. Muthukrishnan, “Ad exchanges: Research issues,” in Proc. WINE, 2009, pp. 1–12.

[16] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica, “Ad click prediction: a view from the trenches,” in Proc. KDD, 2013, pp. 1222–1230.

[17] W. Zhang, S. Yuan, and J. Wang, “Optimal real-time bidding for display advertising,” in Proc. KDD, 2014, pp. 1077–1086.

[18] S. Yuan, J. Wang, and X. Zhao, “Real-time bidding for online advertising: measurement and analysis,” in Proc. ADKDD, 2013.

[19] Manifold Markets, “API documentation: data dumps and market endpoints,” 2024. [Online]. Available: https://docs.manifold.markets/api

[20] Polymarket, “CLOB developer documentation: time series and price endpoints,” 2024. [Online]. Available: https://docs.polymarket.com

[21] Polymarket, “Gamma API guides for market metadata retrieval,” 2024. [Online]. Available: https://docs.polymarket.com

[22] iPinYou Dataset (Hugging Face reczoo release), “iPinYou_x1: CTR dataset,” 2024. [Online]. Available: https://huggingface.co/datasets/reczoo/iPinYou_x1

Appendix A. Reproducibility details

All results are computed from downloaded snapshots: Manifold contracts dump (2024-07-06), iPinYou_x1.zip from Hugging Face, and a fixed Polymarket dataset-server snapshot (100 markets; 95 with resolved outcomes). Seed=42.

Model hyperparameters: feature hashing 2^19; SGD-logistic alpha=1e-5. MTC uses (σ_a,σ_b)=(0.8,4.0) and teacher (a0,b0) fitted on Manifold.

To reproduce figures: generate probability predictions on validation/test, fit calibrators, compute tables, and render plots (reliability diagrams, ROI curves, lag correlations).

Appendix B. Optimization and Prior Design Details

B.1 Optimization details. MTC fits (a,b) by minimizing a convex objective: the Bernoulli negative log-likelihood composed with an affine map of logits, plus a quadratic prior penalty. Let z_i = logit(p̂_i) and η_i = a z_i + b. The data term is NLL = Σ_i [log(1+exp(η_i)) − y_i η_i]. The gradient is ∂NLL/∂a = Σ_i (σ(η_i) − y_i) z_i and ∂NLL/∂b = Σ_i (σ(η_i) − y_i). Adding the Gaussian prior yields ∂L/∂a = ∂NLL/∂a + (a−a0)/σ_a^2 and ∂L/∂b = ∂NLL/∂b + (b−b0)/σ_b^2. The Hessian is positive semidefinite because the logistic second derivative σ(η)(1−σ(η)) is nonnegative, making the objective strictly convex when z has variance and σ_a,σ_b are finite. We therefore use quasi-Newton L-BFGS-B to obtain the unique MAP solution. Because the model has only two parameters, convergence is fast and stable.

B.2 Why logit-space priors transfer across domains. A key obstacle in combining prediction markets with CTR/CVR is the scale mismatch: markets often trade around 0.1–0.9 while CTRs are typically 10^−4–10^−2. Placing a prior directly in probability space (shrinking p̃ toward p_m) makes the mismatch dominate and collapses predictions. Logit space resolves this by representing probabilities as log-odds. A probability of 0.5 corresponds to 0 log-odds, 0.01 corresponds to −4.595, and 0.001 corresponds to −6.907. An intercept shift in logit space multiplies odds by a constant, which is precisely the adjustment needed when class priors change due to sampling or population drift. Thus, a market-derived slope prior transfers across domains even when the baseline b is determined by the target domain.

B.3 Alternative teacher constructions. While we focus on a global Platt regime, the teacher is also constructed in other ways. One approach estimates a nonparametric calibration curve g(p)=E[y|p] from resolved markets and uses it as a prior over monotone calibration functions. Another approach fits separate regimes by topic or by microstructure (liquidity, volume, number of traders), yielding a family of priors. In our data, liquidity stratification reveals strong heterogeneity (Table 10), and teacher reliability is not uniform. A practical system therefore computes multiple candidate priors and selects among them based on liquidity thresholds or cross-validation on recent ad calibration data.

B.4 Relation to empirical Bayes and monitoring. MTC is an empirical Bayes method: the teacher prior (a0,b0) is estimated from a large auxiliary dataset (markets) and then used as a prior for small, noisy calibration tasks (advertisers or campaigns). This framing yields simple monitoring tools. The posterior mean of (a,b) is compared to the teacher mean; large deviations indicate drift, label leakage, or data-quality issues. Because (a,b) is low-dimensional, it is straightforward to alert on unusual values: slope far below 1 indicates severe underconfidence, and intercept shifts inconsistent with observed base rates indicate base-rate mismatch. Such monitors complement high-dimensional model monitoring and can be communicated as 'probability contracts' to stakeholders.

Appendix C. Extended Diagnostics and Ablations

C.1 Parameter trajectories. In addition to scoring metrics, fitted calibration parameters provide insight into how models are being corrected. For heavily downsampled training, the intercept b typically becomes large and negative, reflecting that the true click odds are much smaller than the training odds. The slope a captures overconfidence: a>1 means the raw model is underconfident (needs sharpening), while a<1 means it is overconfident (needs smoothing). In practice, we recommend using a relatively informative prior on a (to stabilize confidence) and a weak prior on b (to allow base-rate adaptation). This is consistent with the design used in our experiments where σ_b is larger than σ_a.

C.2 Prior sensitivity. The prior variances (σ_a,σ_b) trade off bias and variance. With very small calibration sets, a tight prior prevents extreme parameter values. With large calibration sets, the likelihood dominates and σ values matter less. In our CTR setting, σ choices have modest effect on test log loss once a weak intercept prior is used; in CVR settings with rarer positives, σ materially affects stability. Operationally, σ is chosen by optimizing proper scoring rules on a rolling backtest and by incorporating microstructure signals: higher-liquidity markets justify tighter priors (more trust), while lower-liquidity markets call for looser priors.

C.3 Additional diagnostics. Beyond ECE, practitioners care about calibration specifically in the low-probability region because that is where most impressions lie. A useful supplement is to compute ECE on truncated ranges (p<0.01) or to compute calibration-in-the-large (mean predicted probability vs mean outcome). In our experiments, negative subsampling primarily breaks calibration-in-the-large, and both Platt and MTC repair it by shifting the intercept. The reliability diagrams in Figure 3 confirm that calibration improvements concentrate at probabilities below 0.02.

Appendix D. Threats to Validity and Ethical Considerations

D.1 Threats to validity. Our ad auction simulation uses a proxy cost because the iPinYou_x1 snapshot does not include true pay price or bid price fields. The bidding results therefore demonstrate the directionality of calibration effects under a fixed policy, not an accurate estimate of business ROI. A stronger evaluation uses full RTB logs with win prices and conversions, enabling realistic payment rules and attribution windows. Similarly, the Polymarket evaluation uses a fixed dataset-server snapshot due to access constraints; scaling up requires the official metadata and time-series interfaces.

D.2 Market manipulation and selection. Prediction markets can be manipulated, especially when liquidity is thin, and the set of questions listed as markets is not a random sample of future events. These factors can bias the teacher regime. MTC mitigates this by using markets only as a prior and by allowing ad data to dominate when sufficient labels exist. Nevertheless, a deployment should validate teacher stability over time and avoid using market-derived priors in high-stakes settings without additional safeguards.

D.3 Fairness and privacy. Calibration improves the effectiveness of a prediction system, but it does not correct biased features or discriminatory targeting. If the underlying CTR model exhibits disparate performance across user groups, a better calibrated bidder can amplify those disparities by allocating budget more efficiently toward already advantaged segments. Calibration improvements therefore pair with audits for fairness, privacy compliance, and policy constraints (restrictions on sensitive attributes). The interpretability of (a,b) helps governance: it provides an explicit knob and an auditable record of how probability scales were adjusted.

D.4 Security and misuse. Because prediction markets are public and can react to news, there is a potential risk of feedback loops if market-derived priors are used to influence ad delivery around sensitive events (elections). Our formulation avoids linking specific markets to specific users and uses only aggregate regime parameters, reducing this risk. Still, organizations should document data sources, ensure compliance with platform policies, and consider excluding certain market categories from teacher estimation.

Appendix E. Additional Empirical Tables

Table E1. Fitted calibration parameters (a,b) across label fractions for Platt and MTC.

Table E1. Fitted calibration parameters (a,b) across label fractions (Platt vs MTC).

LabelFrac	Method	a	b
0.0100	Platt	0.5625	-2.2048
0.0100	MTC(BayesPlatt)	0.5917	-2.0563
0.0500	Platt	0.4853	-4.1904
0.0500	MTC(BayesPlatt)	0.4993	-4.1173
0.1000	Platt	0.3540	-5.3999
0.1000	MTC(BayesPlatt)	0.3873	-5.2114
1.0000	Platt	0.5752	-4.0943
1.0000	MTC(BayesPlatt)	0.4216	-5.0270

Table E2. Prior sensitivity study for MTC on a fixed 1% calibration subset.

sigma_a	sigma_b	a	b	LogLoss	ECE15
0.4000	2.0000	0.5054	-2.4426	0.0067	0.0030
0.8000	2.0000	0.4689	-2.6297	0.0068	0.0031
0.8000	4.0000	0.4374	-2.8199	0.0067	0.0030
1.2000	4.0000	0.4292	-2.8641	0.0067	0.0030
2.0000	6.0000	0.4180	-2.9296	0.0067	0.0030

Table E3. Outcome and prediction averages by proxy cost bucket (test split).

CostBucket	N	ClickRate	MeanP_Uncal	MeanP_Platt	MeanP_Bayes	MeanP_Isotonic
1.000000	131222.000000	0.000648	0.009113	0.002640	0.002611	0.002545

2.000000	140299.000000	0.000599	0.003864	0.001979	0.001960	0.002035
3.000000	11414.000000	0.000964	0.003572	0.002062	0.002042	0.002162
4.000000	189530.000000	0.000491	0.000975	0.001102	0.001094	0.001019
5.000000	27535.000000	0.000654	0.002005	0.001573	0.001560	0.001622

Appendix F. Theoretical Connection to Bidding and Microstructure

F.1 Calibration and expected value bidding. Consider a bidder with value V for a click and true click probability p. In a second-price auction with clearing price c, the bidder wins if it bids $b \geq c$ and pays c. If the bidder is a price taker, the expected utility of bidding b is $E[\text{utility}(b)] = E[\ I(b\geq c)\cdot(p\cdot V - c)\ ]$. The classical prescription 'bid expected value' sets $b = p\cdot V$ when c is independent of the bidder’s own bid and when the bidder’s action does not affect c. Under these assumptions, truthful bidding is optimal in a Vickrey auction. In practice, RTB includes floors, throttling, and strategic effects, but the expected-value rule remains a common baseline for value-based bidding and for offline evaluation. This rule highlights why calibration matters: if the system uses $b = V\cdot\tilde{p}$ but $\tilde{p} \neq p$, the bidder is implicitly scaling its value by a factor $\tilde{p}/p$, which can be interpreted as uncontrolled bid shading or bid inflation.

F.2 Miscalibration as uncontrolled bid shading. Write $\tilde{p} = g(p)$ where g is the calibration mapping. Then the bid used is $b = V\cdot g(p)$. If $g(p) \approx \kappa\cdot p$ in the relevant regime, the bidder behaves like it had value $\kappa\cdot V$. This changes both its win rate and the distribution of prices it pays, especially in auction environments where higher bids select into higher-priced inventory. Thus, even if a model ranks impressions correctly, an incorrect κ can move the bidder to a different region of the bid landscape. This is one reason why calibration-in-the-large (matching mean $\tilde{p}$ to mean y) is economically meaningful: it controls the overall scale of bids.

F.3 Risk and variance. In budget-constrained settings, advertisers care about variance of outcomes (day-to-day CPA stability) as well as mean performance. Miscalibration increases variance because it makes spend sensitive to noise in predictions. When $\hat{p}$ is overconfident, small score fluctuations can cause large bid swings and erratic pacing. Regularizing slope a toward a stable regime can reduce this volatility by smoothing the mapping from scores to probabilities. This provides an additional motivation for market-derived slope priors: markets are designed to translate noisy information into a probability with bounded loss, which is analogous to smoothing in a calibrator.

F.4 Relating market liquidity to calibration trust. In microstructure terms, liquidity reduces the price impact of noise traders and makes prices more informative about fundamentals. Our liquidity-conditioned analysis in Table 10 supports this: high-liquidity markets exhibit lower Brier score and a sharper calibration slope. A deployment can exploit this by trusting high-liquidity teacher estimates more (tighter σ_a) and low-liquidity estimates less (looser σ_a), or by using a mixture prior weighted by liquidity. Such designs make the 'market efficiency' assumption explicit and adjustable.

Appendix G. Implementation Notes and Reproducibility Checklist

G.1 Implementation notes for large JSON and hashed features. The Manifold contracts dump is a single JSON array that can be hundreds of megabytes. To process it efficiently without loading the full array, we use a streaming decoder that reads the file in chunks and parses one object at a time. This is important for reproducibility and for enabling researchers to run the experiments on commodity hardware.

G.2 Feature hashing collisions. Feature hashing maps millions of categorical tokens into a fixed-dimensional space. Collisions introduce noise but keep memory bounded. In CTR prediction, hashing is widely used in practice [16]. We use 2^19 dimensions, which balances collision risk and computational cost. Because our focus is calibration rather than state-of-the-art prediction accuracy, this representation is sufficient to induce meaningful ranking and miscalibration patterns for study.

G.3 Checklist for reproducing experiments. (1) Download the dataset snapshots listed in Appendix A. (2) Stream iPinYou impressions to form train/validation/test splits. (3) Train two SGD-logistic models: natural and negative-subsampled. (4) Score validation/test splits and fit calibrators (Platt, isotonic, MTC). (5) Compute calibration metrics and render reliability diagrams. (6) Run the budgeted bidding simulation with fixed V and budget B and plot cumulative clicks vs spend. (7) Compute liquidity-conditioned teacher regimes and prior sensitivity tables.

Appendix H. Practitioner Guide and Worked Interpretations

H.1 A worked calibration example (from the iPinYou experiment). In our iPinYou_x1 setup, we fix the value multiplier V based on a target CPC. The negative-subsampled model outputs a mean prediction of 0.004037 on the test split, while the true click rate is 0.0006053 (Table 1 and Table 4). This inflates bids by about 6.7× in probability scale, leading to faster budget exhaustion and worse CPC. After calibration, the mean calibrated probability returns near 0.0007 and spend becomes proportional to realized clicks. This mechanism matches the bidding replay in Table 6: the uncalibrated strategy spends 142.6k units to obtain 116 clicks, while calibrated strategies spend about 83–84k units for about 87–88 clicks, improving clicks per spend.

H.2 Interpreting the calibration parameters. In practice, teams treat (a,b) as a health indicator. If b drifts more negative over time, the system is seeing a lower base rate (traffic quality degradation) or it is missing positives due to attribution delays. If a drifts below 1, the base model is overconfident, which increases pacing volatility. MTC adds another reference point: deviations from (a0,b0) are framed as departures from a market-like probabilistic regime, which is easier to communicate than raw score shifts.

H.3 When market priors hurt. If the market domain is systematically biased (particular topics are overrepresented, or outcomes correlate with external attention), the estimated regime is unrepresentative. The prior is therefore treated as a soft constraint and validated by backtesting on ad data. Additionally, the teacher is updated only on resolved markets to avoid look-ahead bias; including unresolved market prices mixes forecasts with outcomes that are not yet known.

H.4 Choosing σ values in production. We set σ_b large enough that b is determined by recent ad data (base rates are domain-specific), and we tune σ_a for stability. We update calibration daily; with few positives, a tighter σ_a prevents day-to-day slope swings and stabilizes bidding. When positives are plentiful, we loosen σ_a. Liquidity-conditioned teacher regimes (Table 10) provide another handle: we use tighter priors when the teacher is estimated from high-liquidity markets.

H.5 Beyond linear bidding. Many production bidders use more complex policies, such as budget pacing, bid shading, or reinforcement learning. Nevertheless, most of these policies still rely on a calibrated probability estimate as an input feature or as part of the reward model. Therefore, improvements in calibration can propagate to these policies by reducing systematic bias and by making learned value functions more stable. In reinforcement learning terms, calibration can be seen as reducing reward model misspecification.

H.6 Summary. Prediction markets provide an external probabilistic substrate that can be used to regularize calibration in ad systems. Our empirical results demonstrate the core mechanism and show that the approach is computationally lightweight. The most promising next step is to combine richer market microstructure signals (liquidity, volatility) with full RTB logs that include true win prices and conversions, enabling a more faithful evaluation of ROI/CPA and risk.

Appendix I. Extensions and Evaluation Roadmap

I.1 Potential extensions with Polymarket time series. Polymarket exposes historical price endpoints and on-chain settlement mechanisms. A natural extension computes teacher regimes not only from final resolved prices but from trajectories: how quickly prices converge as resolution approaches, and how volatility depends on liquidity. These trajectory features inform time-varying priors for calibration, by tightening σ_a when the teacher market is stable and loosening it when the teacher is volatile. In an ad setting, such priors can adapt calibration aggressiveness during breaking-news periods or seasonal shifts.

I.2 Toward conversion modeling and CPA. The CTR setting is convenient for experimentation because clicks are more frequent than conversions. However, the economic stakes are often higher for CVR and CPA. In CVR, positives are rarer and labels are delayed, making calibration harder and making the variance reduction from priors more valuable. Future work should therefore evaluate MTC on datasets with conversion labels and true clearing prices, and should report CPA under fixed and paced bidding policies. The same framework applies: the intercept prior should remain weak, while the slope prior can stabilize confidence.

I.3 Evaluation beyond offline logs. Offline simulations are limited because they assume static competing bids and do not capture equilibrium effects. A more complete evaluation includes auction replay with bid shading models or online A/B tests where calibrated bids interact with platform feedback and pacing. Even in online settings, the interpretability of MTC is useful: changes in slope and intercept are monitored as part of experiment analysis, linking observed spend differences back to probabilistic scale changes.

Appendix J. Reproducibility and Reporting Recommendations

J.1 Reproducibility metadata. To facilitate verification, we recommend recording the exact snapshot identifiers and hash checksums of downloaded dumps, as well as the random seeds and software versions used for training and calibration. In our experiments, the critical sources are the Manifold contracts dump dated 2024-07-06, the iPinYou_x1.zip release on Hugging Face, and the Polymarket dataset-server snapshot endpoint. Because calibration results can be sensitive to small changes in label sampling, we also recommend saving the sampled calibration indices for each run when performing ablation studies.

J.2 Reporting recommendations. For calibration in ad auctions, we recommend reporting at least three families of metrics: (i) proper scoring rules (log loss, Brier) on a held-out set, (ii) localized diagnostics (ECE, reliability curves) with special attention to the low-probability region, and (iii) decision-centric metrics from replay or simulation (CPC/CPA under fixed bidding multipliers and budgets). Using all three reduces the risk that a calibrator improves one metric while harming another part of the control loop. MTC naturally supports this reporting because its parameters provide an interpretable summary that can be tracked alongside metrics.

J.3 Final note on interpretation. The central message of this paper is not that markets are universally better forecasters than machine-learned models, but that markets provide a useful probabilistic reference distribution. By translating market evidence into a prior over calibration parameters, MTC offers a simple, auditable way to stabilize probability scales when ad labels are sparse or drifting. Even when point metrics are similar to standard calibration, the prior-based framing adds governance value by making assumptions explicit and by enabling microstructure-aware variants.