Recovering Pre-Event Crowd Beliefs in Prediction Markets with a Risk-Calibrated Mixture of Inverse Experts

Idrees Aziz
Middle East Technical University
Ankara, Türkiye
idrees.mustafa.aziz@gmail.com

Abstract

Prediction-market event studies require the crowd's belief immediately before an event, yet that state may never appear in the transaction tape. Sequential arrival, social feedback, belief revision, and repeat trading can produce the same observed path from different latent beliefs. We introduce Adaptive Path Decomposition with Safe Routing (APD-Safe), a risk-calibrated mixture of inverse experts that reconstructs the fixed-cohort pre-event belief from trader-identified probability paths. Four experts reverse stable, feedback, shift, and joint mechanisms; a constrained correction and risk gate combine their estimates. On frozen tests from previously unseen mechanism families, APD-Safe cuts maximum family regret by 79.5% relative to direct black-box prediction. In leave-one-project-out tests on real prediction markets with pre-trading survey beliefs, it reduces RMSE by 10.2% relative to pooling. Unlike direct prediction, every output is fully traceable through named expert estimates and explicit mixture weights. APD-Safe recovers the missing baseline required to measure how events change crowd beliefs.

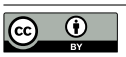
CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Economics*; • **Mathematics of computing** → *Time series analysis*.

Keywords

prediction markets, event studies, mixture of experts, latent-state reconstruction, model risk, financial AI

ACM Reference Format:
Idrees Aziz. 2026. Recovering Pre-Event Crowd Beliefs in Prediction Markets with a Risk-Calibrated Mixture of Inverse Experts. In *7th ACM International Conference on AI in Finance (ICAIF '26), November 14–17, 2026, Milan, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3858063.3858729



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICAIF '26, Milan, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3087-0/2026/11
https://doi.org/10.1145/3858063.3858729

1 Introduction

Financial event studies compare a market state before and after an event [43, 53]. Prediction markets appear to provide that state directly because a binary contract trades as a probability. The price, however, is an equilibrium object shaped by wealth, utilities, information, risk attitudes, and market design; it need not equal the equally weighted belief of the crowd an analyst intends to study [22, 30, 44, 59, 60]. The required pre-event belief may therefore be absent from every observed trade.

The transaction tape records an endogenous sequence rather than independent reports. Traders arrive sequentially, observe earlier prices, learn from or react against the path, revise beliefs when information arrives, and trade more than once [5, 8, 17, 41, 47, 50]. First-entry pooling removes repeat-trader weighting but not feedback or revision. Change-point correction removes one fitted shift but not endogenous path attachment. Graph inversion removes path attachment but can distort genuine information updates. Several mechanisms can explain the same path, so committing to one correction creates model risk.

We extend structural path correction with APD-Safe, a finance-structured mixture of inverse experts. The target is the pre-event mean belief of one fixed, equally weighted trader cohort. Four experts reconstruct that target under stable, feedback, persistent-shift, and joint mechanisms. A learned residual improves the structural mixture but is clipped to the range supported by the candidate reconstructions. A second gate predicts candidate-specific reconstruction risk and returns a soft mixture. The architecture therefore learns *which financial explanation to trust*, rather than mapping the price tape directly to the target. Every reconstruction retains an exact audit trail through its candidate estimates and mixture weights.

We test the method in two settings that observe the same target. First, the complete model is frozen and evaluated on 1,500 exact-ground-truth markets from ten mechanism families absent from training and validation. APD-Safe lowers RMSE by 12.8% relative to identity-balanced first-entry pooling. Against direct black-box prediction, it improves average RMSE by 3.1% and cuts maximum family regret from 8.8% to 1.8%, a 79.5% reduction. Second, four real replication-market projects link pre-market survey probabilities to trader identities and transaction paths [26]. Leave-one-project-out field calibration lowers market-balanced RMSE by 10.2% relative to pooling and by 6.6% relative to graph inversion across 98 real markets.

The paper makes three contributions:

(1) A fixed-cohort estimand for the pre-event crowd belief, separating belief change from trader turnover and repeated activity.
(2) APD-Safe, a risk-calibrated mixture of mechanism-matched inverse experts whose candidate estimates and mixture weights expose the complete reconstruction path.
(3) Controlled and real-market tests against simple aggregation, structural correction, generic stacking, and direct prediction: APD-Safe reduces maximum family regret by 79.5% on unseen mechanisms and beats graph inversion and pooling in all four held-out projects.

2 Related Work

Prediction markets and event baselines. Prediction markets can aggregate dispersed information and often forecast accurately [3, 59], but price depends on the trading mechanism, participant composition, information distribution, and preferences [4, 14, 31, 34, 44, 45, 60]. Event studies require a defensible pre-event state and window [39, 43, 53]. Announcement markets and high-frequency evidence show that prices update before and around observed events, while many jumps cannot be assigned to one recorded news item [2, 40]. We address the missing-baseline step directly.

Path dependence and structural correction. Sequential actions can create cascades, herding, contrarian responses, and wealth-dependent price reactions [5, 8, 17, 50]. Experiments show that social information can reduce crowd independence and that traders learn from realized prices [41, 47]. Change-point methods detect abrupt, multiple, dependent, or contaminated shifts [1, 7, 15, 18, 21, 36]; Hawkes and microstructure models represent endogenous path dependence and order-flow effects [6, 13, 19, 27, 29]. APD-Safe extends these structural corrections by treating them as competing inverse explanations of the same missing baseline.

Risk routing and mechanism generalization. Mixture-of-experts, model averaging, and stacking combine specialized predictors to reduce selection risk [9, 10, 28, 32, 33, 35, 48, 55, 61]. Selective prediction learns where a predictor should be trusted [20, 23, 49]. Controlled financial simulation provides exact latent states, but simulator misspecification must be tested explicitly [11, 16, 25, 42, 54, 56, 57]. APD-Safe extends risk routing from prediction to reconstruction: its gate chooses among the structural explanations above by predicted error on the fixed-cohort baseline. The final test freezes every model choice and evaluates complete mechanism families absent from development.

3 Problem Formulation

3.1 Fixed-cohort target

Fix the crowd whose belief change the event study intends to measure, $\mathcal{U}_\star = \{1, \dots, U\}$. Trader u has latent log-odds belief b_u^- before the persistent response and b_u^+ after the response stabilizes. The two crowd probabilities are

$$C^s = \frac{1}{U} \sum_{u \in \mathcal{U}_\star} \mathrm{sigmoid}(b_u^s), \qquad s \in \{-, +\}, \tag{1}$$

and the event effect is $\Delta = C^+ - C^-$. Fixed identities and equal weights prevent trader turnover or repeated activity from changing the estimand [4, 44, 45].

The observed pre-event history is

$$\mathcal{H} = \{(t_i, u_i, p_i)\}_{i=1}^N, \tag{2}$$

where p_i is the post-trade contract probability. We clip p_i to $[.005, .995]$, transform it to the log-odds mark $m_i = \mathrm{logit}(p_i)$, and normalize time to $z_i \in [0, 1]$.

3.2 Why the target can be missing

A compact observation model isolates the reconstruction problem:

$$m_i = \delta q_i + (1 - x_i) b_{u_i}^- + x_i G_i + \varepsilon_i. \tag{3}$$

Here q_i marks a persistent information revision, δ is its size, G_i summarizes the earlier path, x_i is the trade's attachment to that path, and ε_i is residual noise. Identity-balanced first-entry pooling removes repeat-trader weighting,

$$\widehat{C}^-_{\mathrm{IBFE}} = \frac{1}{U} \sum_{i : u_i \text{ first appears}} \mathrm{sigmoid}(m_i), \tag{4}$$

but it does not remove δq_i or $x_i G_i$.

Suppose an inverse expert subtracts $\widehat{\delta}\widehat{q}_i$, recomputes a path message $\widehat{G}_i$, and inverts feedback:

$$\widehat{b}_i = \frac{m_i - \widehat{\delta}\widehat{q}_i - x_i \widehat{G}_i}{1 - x_i}. \tag{5}$$

Substitution of Eq. (3) gives

$$\widehat{b}_i - b_{u_i}^- = \frac{\delta q_i - \widehat{\delta}\widehat{q}_i}{1 - x_i} + \frac{x_i (G_i - \widehat{G}_i)}{1 - x_i} + \frac{\varepsilon_i}{1 - x_i}. \tag{6}$$

The reconstruction error is also the event-study error. Let $\widehat{\Delta} = C^+ - \widehat{C}^-$ when the stabilized post-event state is observed. Then

$$|\widehat{\Delta} - \Delta| = |\widehat{C}^- - C^-|. \tag{7}$$

A lower pre-event reconstruction RMSE therefore produces the same reduction in event-effect RMSE. A correct expert recovers the latent belief; a wrong correction creates error. Because the tape does not reveal which mechanism is active, reconstruction is a conditional model-selection problem.

4 Method: APD-Safe

APD-Safe separates candidate construction from risk-based routing (Figure 1A). Every candidate estimates the same target C^-. The final gate predicts which candidate will reconstruct that target most accurately for the observed path.

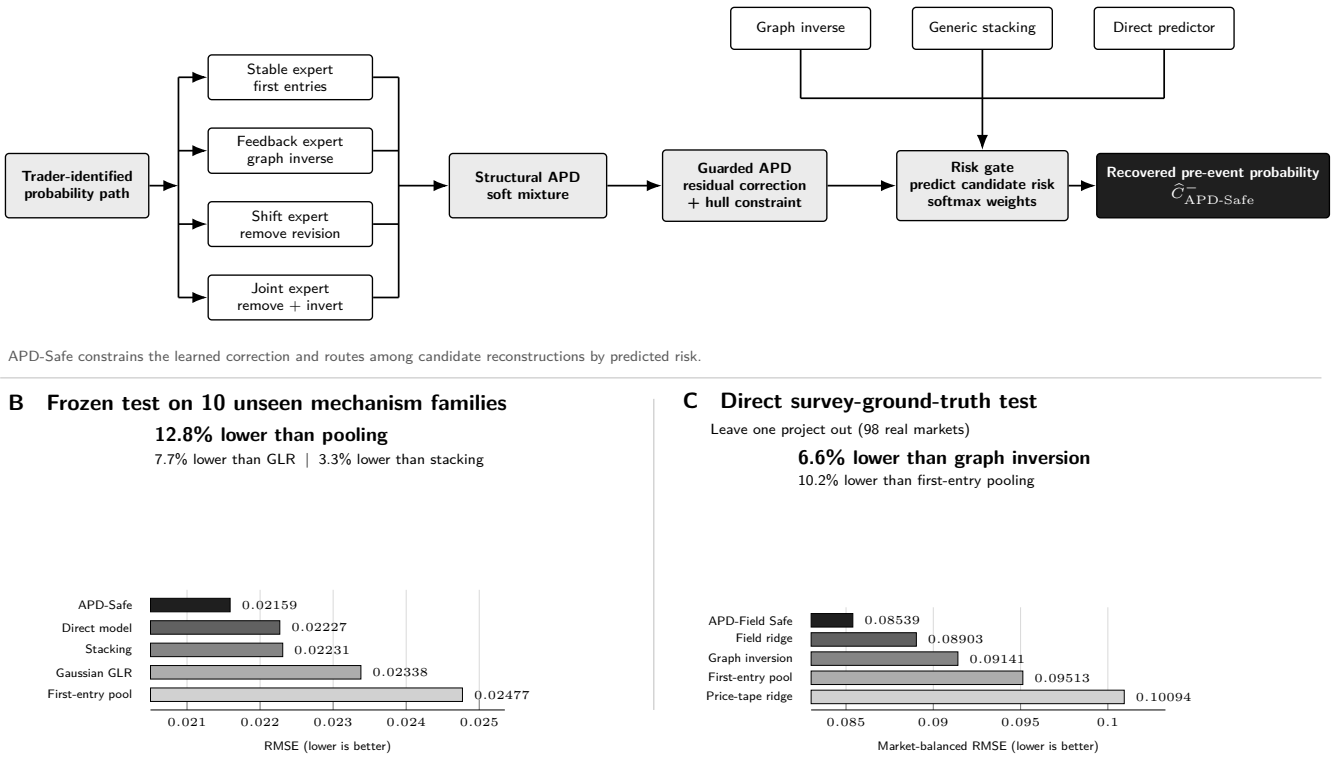


Figure 1: APD-Safe architecture and evaluation. Panel A shows the inverse-expert bank and risk-calibrated route. Panel B reports the frozen unseen-mechanism test. Panel C reports leave-one-project-out reconstruction against pre-market survey beliefs in 98 real markets.

4.1 Inverse expert bank

The expert bank contains four estimators: stable pooling, feedback inversion, persistent-shift removal, and their joint correction. For a corrected log-odds path y, the causal graph message is

$$a^G_{ij}(y) = \frac{\exp\{-(z_i - z_j)/2 - 5|y_i - y_j|\}}{\sum_{\ell<i}\exp\{-(z_i - z_\ell)/2 - 5|y_i - y_\ell|\}}, \quad j < i,$$
$$G_i(y) = \sum_{j<i} a^G_{ij}(y) y_j. \tag{8}$$

Set $G_1(y) = 0$; because $e_1 = x_1 = 0$, this convention does not affect the first reconstruction. The feedback exposure is $x_i = .5e_i$, where e_i is the fraction of the target cohort observed before trade i. Reconstructed logits are clipped to $[-5.3, 5.3]$ before the probability transform.

Each regime $r \in \{0, E, S, ES\}$ is fit over fixed time-kernel and change-point dictionaries. Following information-criterion logic [37, 52], its score combines fit with a parameter-count penalty:

$$Q_r = \min_{\lambda\in\Lambda_r}\{2\,\mathrm{NLL}_{r,\lambda} + k_r \log N\}. \tag{9}$$

With $\tau \in \{.1, .4, 1.2\}$, change points $c \in \{.35, .5, .65\}$, and dimensions $(k_0, k_E, k_S, k_{ES}) = (2, 3, 4, 5)$, the structural weights are

$$w_r \propto \exp\{-\tfrac{1}{2}(Q_r - \min_s Q_s) - 4\mathbb{1}[r \neq 0]\},$$
$$\widehat{C}^{-}_{\text{struct}} = \sum_r w_r \widehat{C}^{-}_r. \tag{10}$$

We add the fixed offset $-4\mathbb{1}[r \neq 0]$ so the stable expert remains the default until the path supplies evidence for correction.

4.2 Guarded correction and risk routing

Structural scores measure path fit, not reconstruction loss. A depth-5 CatBoost regressor [51] therefore learns a residual from 20 observable path summaries and the fixed structural-bank estimates. Let $h_\phi(x)$ be the predicted residual. The guarded estimate is

$$\widehat{C}^{-}_{\text{guard}} = \text{clip}_{[L(x),H(x)]}\left(\widehat{C}^{-}_{\text{struct}} + .9h_\phi(x)\right), \tag{11}$$

where $L(x)$ and $H(x)$ are the minimum and maximum candidate estimates. To block unsupported extrapolation, we

constrain the residual to the range of the candidate reconstructions, applying the broader robustness principle of aggregation under misspecification [46, 58].

The final candidate set is

$$\mathcal{C} = \{\widehat{C}^-_{\text{guard}}, \widehat{C}^-_{\text{graph}}, \widehat{C}^-_{\text{stack}}, \widehat{C}^-_{\text{direct}}\}. \tag{12}$$

The stacking candidate is an Extra Trees regressor on ten path summaries and the four inverse experts; a separate Extra Trees direct predictor uses 56 path features [24]. Four depth-4 CatBoost models predict candidate log squared error,

$$\widehat{\ell}_k(x) \approx \log\left[(\widehat{C}^-_k - C^-)^2 + 10^{-7}\right]. \tag{13}$$

Predicted risks become soft weights with fixed temperature $T = .5$:

$$\alpha_k(x) = \frac{\exp\{-\widehat{\ell}_k(x)/T\}}{\sum_j \exp\{-\widehat{\ell}_j(x)/T\}}, \qquad \widehat{C}^-_{\text{APD-Safe}} = \sum_{k\in\mathcal{C}} \alpha_k(x)\widehat{C}^-_k. \tag{14}$$

The compact gate uses ten causal path features, the four candidate estimates, their span, and two candidate gaps. Every summary and direct-feature dictionary is fixed before final testing and uses only $\mathcal{H}$; no feature uses C^- or a future transaction. Soft routing preserves candidate uncertainty and diversification [28, 32, 38, 48].

4.3 Cross-project field deployment

When earlier survey-linked projects are available, a ridge head maps the APD candidates and compact causal path summaries to the observed fixed-cohort survey target:

$$\widehat{C}^-_{\text{field}} = \text{clip}_{[0,1]}\Big(\beta_0 + \beta^\top z(x)\Big). \tag{15}$$

A final convex blend combines the field head, graph inversion, and first-entry pooling. For each held-out project, coefficients, regularization, and blend weights are selected only from the other projects using inner leave-one-project-out predictions. The held-out project contributes no survey response or tuning decision. With fixed dictionaries, expert construction uses $O(N^2)$ time and $O(N)$ memory; a single-threaded Python implementation evaluates one market in about 0.11 seconds on an AMD EPYC 9V74 core.

5 Experiments

5.1 Controlled exact-ground-truth benchmark

Model development uses 6,330 markets from 38 mechanism families. The source block contains stable, feedback, repeat-trading, news, mixed, and one-step paths. Two disjoint stress blocks add gradual and multiple shifts, drift, heavy-tailed and heteroskedastic noise, nonlinear and heterogeneous feedback, delayed and contrarian response, network copying, strategic repeat trading, identity corruption, temporary shocks, endogenous arrival, quantization, missing events, and variance regimes. Disjoint seed ranges separate candidate training, gate training, validation, and architecture selection.

After every model choice is fixed, the final test draws 1,500 markets, 150 from each of ten additional families: asymmetric subgroup shifts, belief-dependent noise, cohort block shocks, decaying shocks, identity merging, mixed probability/logit feedback, periodic feedback, price limits, selective censoring of repeat trades, and trend reversal. None appears in training or validation, and none is generated by Eq. (3) with one constant shift and one fixed feedback coefficient.

5.2 Real-market survey benchmark

Gordon et al. release four independent replication-market projects containing pre-market survey probabilities, transaction paths, and matching participant identifiers [26]. The fixed preprocessing extracts 426 causal prefixes from the 98 markets that yield a usable trader-linked prefix. Prefix endpoints are chosen before model fitting from transaction order and linkage availability, and the same prefixes are used for every method. For each prefix, the target is the equal-weight mean pre-market survey probability of the traders who have appeared by that decision. The median prefix contains 23 traders; 400 of 426 prefixes match every trader to a survey response, and every prefix matches at least 80%.

Evaluation leaves one complete project out. The held-out project supplies no information for field-head fitting, regularization, or blend selection. This is direct target validation: beliefs are recorded before trading, rather than inferred from a future price.

Ethics and data handling. This experiment is a secondary analysis of the publicly released, pseudonymized survey and transaction records of Gordon et al. [26]. No new participants were recruited or contacted. We used only the identifiers contained in the public release to link pre-market survey responses to transactions, and we made no attempt to re-identify any participant.

5.3 Baselines, metrics, and inference

Controlled baselines are identity-balanced first-entry pooling, Gaussian GLR, structural APD, generic stacking, and direct black-box prediction. Gaussian GLR fits one Gaussian mean shift over the fixed change-point dictionary and removes the selected shift before identity-balanced pooling. Field baselines are pooling, graph inversion, current price, a full-feature ridge model, and a price-tape-only ridge model. All learned baselines receive the same development data or the same leave-one-project-out privilege.

The primary controlled metric is pooled RMSE. Family-average RMSE weights each mechanism equally, and maximum family regret is the largest percentage gap from the best tested method in a family. Inference uses a family-balanced bootstrap. The field test uses market-balanced RMSE, project-level comparisons, and market-cluster bootstrap inference [12].

Table 1: Controlled and real-market reconstruction. Panel A freezes the complete model before testing 1,500 markets from ten unseen mechanism families. Panel B leaves one real-market project out at a time and scores 426 causal prefixes against directly observed pre-market survey beliefs. Lower is better.

Method	RMSE	Reduction vs. pooling	Mean family RMSE	Max family regret	Family wins
Panel A: controlled test on ten unseen mechanism families					
APD-Safe (proposed)	**.02159**	**12.8%**	**.02100**	**1.8%**	**6/10**
Direct black-box prediction	.02227	10.1%	.02174	8.8%	0/10
Generic stacking	.02231	9.9%	.02183	14.5%	3/10
Structural APD	.02324	6.2%	.02274	18.4%	0/10
Gaussian GLR	.02338	5.6%	.02284	17.4%	1/10
Identity-balanced pooling (IBFE)	.02477	–	.02426	39.4%	0/10

Method	Market-balanced RMSE	Reduction vs. pooling	Pearson	Below graph	Below pooling
Panel B: 98 real markets across four held-out projects					
APD-Safe (field-calibrated; proposed)	**.08539**	**10.2%**	**.902**	**4/4**	**4/4**
Full-feature field ridge	.08903	6.4%	.890	2/4	2/4
Graph inversion	.09141	3.9%	.896	–	3/4
First-entry pooling	.09513	–	.889	1/4	–
Current price	.09870	−3.7%	.874	2/4	2/4
Price-tape ridge	.10094	−6.1%	.854	1/4	2/4

5.4 Model selection and leakage control

The controlled pipeline uses four disjoint operations: candidate fitting, safety-gate fitting, architecture selection, and final testing. Mechanism families and seed ranges do not cross these boundaries. The ten final families are generated only after the four-candidate gate, temperature, shrinkage, clipping rule, feature set, and hyperparameters are fixed. The five-batch replication reuses the same frozen estimator without re-selection.

The field pipeline applies the same principle at project level. For each outer fold, all coefficients and blend weights are selected by inner leave-one-project-out predictions among the remaining projects. Prefixes are causal: every feature uses only transactions observed by that decision. Market-balanced scoring prevents markets with many prefixes from dominating the result, and the market-cluster bootstrap treats each market as the independent sampling unit.

5.5 Coding and manuscript assistance

ChatGPT was used to implement experimental scripts, produce figure code, and refine the manuscript. The APD-Safe method, evaluation protocols, and analysis plan were specified by the authors, who ran and verified the experiments and interpreted the findings.

6 Results

6.1 Controlled generalization

APD-Safe reaches RMSE 0.02159, compared with 0.02477 for pooling and 0.02338 for Gaussian GLR (Table 1A; Figure 1B). These are reductions of 12.8% and 7.7%. It also improves on generic stacking by 3.3% and direct prediction by 3.1%. Family-balanced gains are significant against pooling, GLR, direct prediction, and stacking ($p \leq .004$).

The sharper advantage over direct prediction is robustness to mechanism shift. Maximum family regret falls from 8.8% for the black-box model to 1.8% for APD-Safe, a 79.5% reduction. APD-Safe beats pooling, direct prediction, and structural APD in all ten families, generic stacking in seven, and GLR in nine. It is best in six families; maximum regret is 14.5% for stacking, 17.4% for GLR, and 39.4% for pooling.

6.2 Real-market reconstruction

The field estimator reaches market-balanced RMSE 0.08539 and Pearson correlation 0.902 (Table 1B; Figure 1C). Graph inversion reaches 0.09141, pooling 0.09513, and the price-tape learner 0.10094. The RMSE reductions are 6.6%, 10.2%, and 15.4%, respectively. APD-Safe (field-calibrated) beats graph inversion and pooling in all four held-out projects.

Market-cluster inference confirms the ranking. Relative to graph inversion, the mean market MSE gain is 1.06×10^{-3} with 95% bootstrap interval $[0.45, 1.67] \times 10^{-3}$ and $p < .001$. Relative to pooling, the gain is 1.76×10^{-3} with interval $[0.94, 2.59] \times 10^{-3}$ and $p < .001$. Relative to the price-tape learner, the gain is 2.90×10^{-3} with interval $[1.37, 4.62] \times 10^{-3}$ and $p < .001$.

The project breakdown shows that the gain is not created by one dataset (Table 2). The largest improvements occur on ML2 and EERP, where the raw cross-project error is highest. On RPP and SSRP the safety blend retains the strength of graph inversion while remaining best overall.

Table 2: Held-out project RMSE. Each row is produced by a model that did not use that project's surveys or transactions for selection.

Project	APD-Safe (field-calibrated)	Graph	Pooling
EERP	**.06653**	.07363	.07767
ML2	**.11345**	.12368	.13288
RPP	**.07529**	.07938	.07951
SSRP	**.07808**	.07981	.07969

Table 3: Routing ablation on the fixed architecture test set.

Variant	RMSE
Guarded correction alone	.04662
Uniform candidate average	.02843
Hard risk routing	.02344
Soft risk routing (APD-Safe)	**.02238**

6.3 Robustness and ablation

Five independent frozen batches add 4,200 markets from fourteen broad stress mechanisms. APD-Safe obtains mean RMSE 0.02305 versus 0.02902 for pooling, a 20.6% reduction. Every batch is positive, and 64 of 70 family-by-batch gains are positive; the cluster bootstrap interval for mean MSE gain is $[2.12, 4.18] \times 10^{-4}$.

The ablation isolates the architectural gain (Table 3). The learned correction alone fails under shift. Candidate averaging is safer but inaccurate. Hard routing improves sharply. Soft risk routing is best because it preserves uncertainty over competing financial explanations. The gain is not a feature-count effect: the compact gate also beats the full 56-feature risk router on the final family holdout.

7 Discussion

APD-Safe constructs financially meaningful inverse candidates and routes among them using predicted reconstruction risk. Each output retains an exact audit trail: the candidate reconstructions, their risk weights, and the resulting convex mixture. This exposes when the estimate is supported by structural inversion and when learned alternatives receive weight, rather than hiding the reconstruction inside one direct mapping. The guarded residual adds flexibility without leaving the range supported by the candidates. The outer gate evaluates candidate risk rather than path fit, which is why it avoids the large failures of the correction alone and the mechanism-specific failures of hard selection.

The controlled and field benchmarks assess different forms of generalization. The controlled benchmark asks whether the architecture recovers a known latent state when the market mechanism changes. The field benchmark asks whether the same candidate structure supports deployment on observed markets when earlier survey-linked projects are available. Exact latent-state labels answer the first question; pre-trading survey beliefs answer the second. Equation (7) connects both tests to the financial application: every unit of baseline error is one unit of event-effect error.

The method is directly applicable to binary, trader-identified markets. New contract structures or additional path mechanisms enter through new inverse candidates and candidate-specific risk models; the target remains the fixed-cohort belief required by the event study.

8 Conclusion

Prediction-market event studies need a pre-event crowd belief absent from the transaction tape. APD-Safe reconstructs it with mechanism-matched inverse experts and risk-calibrated routing, retaining full traceability through named candidates and explicit mixture weights. Compared with direct black-box prediction, APD-Safe lowers average RMSE by 3.1% and maximum family regret from 8.8% to 1.8%, a 79.5% reduction. On 98 real markets with pre-trading survey beliefs, leave-one-project-out calibration reduces RMSE by 10.2% relative to pooling. Together, these results show that competing structural explanations recover pre-event crowd beliefs more robustly than direct prediction while retaining a complete audit trail.

References

[1] Ryan Prescott Adams and David J. C. MacKay. 2007. Bayesian Online Changepoint Detection. arXiv:0710.3742.

[2] Yacine Aït-Sahalia, Chen Xu Li, and Chenxu Li. 2024. *So Many Jumps, So Few News.* Technical Report 32746. National Bureau of Economic Research. doi:10.3386/w32746

[3] Kenneth J. Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers, and Eric Zitzewitz. 2008. The Promise of Prediction Markets. *Science* 320, 5878 (2008), 877–878. doi:10.1126/science.1157679

[4] Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2017. Distilling the Wisdom of Crowds: Prediction Markets versus Prediction Polls. *Management Science* 63, 3 (2017), 691–706. doi:10.1287/mnsc.2015.2374

[5] Christopher Avery and Peter Zemsky. 1998. Multidimensional Uncertainty and Herd Behavior in Financial Markets. *American Economic Review* 88, 4 (1998), 724–748.

[6] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. 2015. Hawkes Processes in Finance. *Market Microstructure and Liquidity* 1, 1 (2015), 1550005. doi:10.1142/S2382626615500057

[7] Sujay Bhatt, Guanhua Fang, and Ping Li. 2022. Offline Change Detection under Contamination. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 180).* 191–201.

[8] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy* 100, 5 (1992), 992–1026. doi:10.1086/261849

[9] Leo Breiman. 1996. Stacked Regressions. *Machine Learning* 24 (1996), 49–64. doi:10.1007/BF00117832

[10] Stephen T. Buckland, Kenneth P. Burnham, and Nicole H. Augustin. 1997. Model Selection: An Integral Part of Inference. *Biometrics* 53, 2 (1997), 603–618. doi:10.2307/2533961

[11] David Byrd, Maria Hybinette, and Tucker H. Balch. 2020. ABIDES: Towards High-Fidelity Multi-Agent Market Simulation. In *Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. Association for Computing Machinery, New York, NY, USA, 11–22. doi:10.1145/3384441.3395986

[12] A. Colin Cameron and Douglas L. Miller. 2015. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources* 50, 2 (2015), 317–372. doi:10.3368/jhr.50.2.317

[13] Rama Cont, Arseniy Kukanov, and Sasha Stoikov. 2014. The Price Impact of Order Book Events. *Journal of Financial Econometrics* 12, 1 (2014), 47–88. doi:10.1093/jjfinec/nbt003

[14] Brice Corgnet, Cary Deck, Mark DeSantis, Kyle Hampton, and Erik O. Kimbrough. 2023. When Do Security Markets Aggregate Dispersed Information? *Management Science* 69, 6 (2023), 3697–3729. doi:10.1287/mnsc.2022.4463

[15] Herold Dehling, Roland Fried, and Martin Wendler. 2020. A Robust Method for Shift Detection in Time Series. *Biometrika* 107, 3 (2020), 647–660. doi:10.1093/biomet/asaa004

[16] Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and Francois-Xavier Briol. 2022. Robust Bayesian Inference for Simulator-Based Models via the MMD Posterior Bootstrap. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*. 943–970.

[17] Mathias Drehmann, Jörg Oechssler, and Andreas Roider. 2005. Herding and Contrarian Behavior in Financial Markets: An Internet Experiment. *American Economic Review* 95, 5 (2005), 1403–1426. doi:10.1257/000282805775014317

[18] Paul Fearnhead. 2006. Exact and Efficient Bayesian Inference for Multiple Changepoint Problems. *Statistics and Computing* 16 (2006), 203–213. doi:10.1007/s11222-006-8450-8

[19] Vladimir Filimonov and Didier Sornette. 2012. Quantifying Reflexivity in Financial Markets: Toward a Prediction of Flash Crashes. *Physical Review E* 85, 5 (2012), 056108. doi:10.1103/PhysRevE.85.056108

[20] Vojtech Franc and Daniel Prusa. 2019. On Discriminative Learning of Prediction Uncertainty. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*. 1963–1971.

[21] Piotr Fryzlewicz. 2014. Wild Binary Segmentation for Multiple Change-Point Detection. *The Annals of Statistics* 42, 6 (2014), 2243–2281. doi:10.1214/14-AOS1245

[22] Jianjun Gao, Zizhuo Wang, Weiping Wu, and Dian Yu. 2024. Price Interpretability of Prediction Markets: A Convergence Analysis. *Operations Research* 73, 1 (2024), 157–177. doi:10.1287/opre.2022.0417

[23] Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*. 2151–2159.

[24] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely Randomized Trees. *Machine Learning* 63, 1 (2006), 3–42. doi:10.1007/s10994-006-6226-1

[25] Dhananjay K. Gode and Shyam Sunder. 1993. Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality. *Journal of Political Economy* 101, 1 (1993), 119–137. doi:10.1086/261868

[26] Michael Gordon, Domenico Viganola, Anna Dreber, Magnus Johannesson, and Thomas Pfeiffer. 2021. Predicting Replicability—Analysis of Survey and Prediction Market Data from Large-Scale Forecasting Projects. *PLOS ONE* 16, 4 (2021), e0248780. doi:10.1371/journal.pone.0248780

[27] Martin D. Gould, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. 2013. Limit Order Books. *Quantitative Finance* 13, 11 (2013), 1709–1742. doi:10.1080/14697688.2013.803148

[28] Bruce E. Hansen. 2007. Least Squares Model Averaging. *Econometrica* 75, 4 (2007), 1175–1189. doi:10.1111/j.1468-0262.2007.00785.x

[29] Alan G. Hawkes. 1971. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika* 58, 1 (1971), 83–90. doi:10.1093/biomet/58.1.83

[30] Xue-Zhong He and Nicolas Treich. 2017. Prediction Market Prices under Risk Aversion and Heterogeneous Beliefs. *Journal of Mathematical Economics* 70 (2017), 105–114. doi:10.1016/j.jmateco.2017.02.005

[31] Paul J. Healy, Sera Linardi, J. Richard Lowery, and John O. Ledyard. 2010. Prediction Markets: Alternative Mechanisms for Complex Environments with Few Traders. *Management Science* 56, 11 (2010), 1977–1996. doi:10.1287/mnsc.1100.1226

[32] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. Bayesian Model Averaging: A Tutorial. *Statist. Sci.* 14, 4 (1999), 382–417. doi:10.1214/ss/1009212519

[33] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive Mixtures of Local Experts. *Neural Computation* 3, 1 (1991), 79–87. doi:10.1162/neco.1991.3.1.79

[34] Lian Jian and Rahul Sami. 2012. Aggregation and Manipulation in Prediction Markets: Effects of Trading Mechanism and Information Distribution. *Management Science* 58, 1 (2012), 123–140. doi:10.1287/mnsc.1110.1404

[35] Michael I. Jordan and Robert A. Jacobs. 1994. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation* 6, 2 (1994), 181–214. doi:10.1162/neco.1994.6.2.181

[36] Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. 2012. Optimal Detection of Changepoints With a Linear Computational Cost. *J. Amer. Statist. Assoc.* 107, 500 (2012), 1590–1598. doi:10.1080/01621459.2012.737745

[37] Sadanori Konishi and Genshiro Kitagawa. 1996. Generalised Information Criteria in Model Selection. *Biometrika* 83, 4 (1996), 875–890. doi:10.1093/biomet/83.4.875

[38] Anders Krogh and Jesper Vedelsby. 1994. Neural Network Ensembles, Cross Validation, and Active Learning. In *Advances in Neural Information Processing Systems 7*. MIT Press, Cambridge, MA, 231–238.

[39] Ashley Langer and Derek Lemoine. 2020. *What Were the Odds? Estimating the Market's Probability of Uncertain Events*. Technical Report 28265. National Bureau of Economic Research. doi:10.3386/w28265

[40] David S. Lee and Enrico Moretti. 2009. Bayesian Learning and the Pricing of New Information: Evidence from Prediction Markets. *American Economic Review* 99, 2 (2009), 330–336. doi:10.1257/aer.99.2.330

[41] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How Social Influence Can Undermine the Wisdom of Crowd Effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025. doi:10.1073/pnas.1008636108

[42] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob H. Macke. 2021. Benchmarking Simulation-Based Inference. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*. 343–351.

[43] A. Craig MacKinlay. 1997. Event Studies in Economics and Finance. *Journal of Economic Literature* 35, 1 (1997), 13–39.

[44] Charles F. Manski. 2006. Interpreting the Predictions of Prediction Markets. *Economics Letters* 91, 3 (2006), 425–429. doi:10.1016/j.econlet.2006.01.004

[45] Marco Mantovani and Antonio Filippin. 2026. When Do Prediction Markets Return Average Beliefs? Experimental Evidence. *Games and Economic Behavior* 156 (2026), 135–148. doi:10.1016/j.geb.2025.12.004

[46] Annie Marsden, John Duchi, and Gregory Valiant. 2021. Misspecification in Prediction Problems and Robustness via Improper Learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*. 2161–2169.

[47] M. Kathleen Ngangoué and Georg Weizsäcker. 2021. Learning from Unrealized versus Realized Prices. *American Economic Journal: Microeconomics* 13, 2 (2021), 174–201. doi:10.1257/mic.20180268

[48] Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. 2024. On Least Square Estimation in Softmax Gating Mixture of Experts. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. 37707–37735.

[49] Fedor Noskov, Alexander Fishkov, and Maxim Panov. 2024. Selective Nonparametric Regression via Testing. In *Proceedings of the 15th Asian Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 222)*. 1023–1038.

[50] Marco Ottaviani and Peter Norman Sørensen. 2015. Price Reaction to Information with Heterogeneous Beliefs and Wealth Effects: Underreaction, Momentum, and Reversal. *American Economic Review* 105, 1 (2015), 1–34. doi:10.1257/aer.20120881

[51] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost:

Unbiased Boosting with Categorical Features. In *Advances in Neural Information Processing Systems*, Vol. 31.

[52] Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464. doi:10.1214/aos/1176344136

[53] Erik Snowberg, Justin Wolfers, and Eric Zitzewitz. 2011. How Prediction Markets Can Save Event Studies. In *Prediction Markets: Theory and Applications*, Leighton Vaughan Williams (Ed.). Routledge, London, UK, 18–34. doi:10.4324/9780203815526-3

[54] Namid R. Stillman, Rory Baggott, Justin Lyon, Jianfei Zhang, Dingqiu Zhu, Tao Chen, and Perukrishnen Vytelingum. 2023. Deep Calibration of Market Simulations Using Neural Density Estimators and Embedding Networks. In *Proceedings of the 4th ACM International Conference on AI in Finance*. 46–54. doi:10.1145/3604237.3626881

[55] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. 2007. Super Learner. *Statistical Applications in Genetics and Molecular Biology* 6, 1 (2007), Article 25. doi:10.2202/1544-6115.1309

[56] Svitlana Vyetrenko, David Byrd, Nicholas Petosa, Mahmoud Mahfouz, Danial Dervovic, Manuela Veloso, and Tucker Balch. 2020. Get Real: Realism Metrics for Robust Limit Order Book Market Simulations. In *Proceedings of the First ACM International Conference on AI in Finance*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3383455.3422561

[57] Antoine Wehenkel, Juan L. Gamella, Ozan Sener, Jens Behrmann, Guillermo Sapiro, Joern-Henrik Jacobsen, and Marco Cuturi. 2025. Addressing Misspecification in Simulation-Based Inference through Data-Driven Calibration. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*. 65949–65980.

[58] Halbert White. 1982. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* 50, 1 (1982), 1–25. doi:10.2307/1912526

[59] Justin Wolfers and Eric Zitzewitz. 2004. Prediction Markets. *Journal of Economic Perspectives* 18, 2 (2004), 107–126. doi:10.1257/0895330041371321

[60] Justin Wolfers and Eric Zitzewitz. 2006. *Interpreting Prediction Market Prices as Probabilities*. Technical Report 12200. National Bureau of Economic Research. doi:10.3386/w12200

[61] David H. Wolpert. 1992. Stacked Generalization. *Neural Networks* 5, 2 (1992), 241–259. doi:10.1016/S0893-6080(05)80023-1