ForesightFlow: An Information Leakage Score Framework for Prediction Markets

Anonymous Authors
Submission to ICML 2026 Workshop on AI Forecasting

Abstract

Decentralized prediction markets such as Polymarket aggregate dispersed beliefs into continuously updated price signals, but their on-chain transparency and pseudonymous participation also create unusually permissive conditions for trading on material non-public information. Recent empirical work has documented hundreds of millions of dollars in anomalous profits on Polymarket between 2024 and 2026; existing detection approaches are almost exclusively *post-hoc* and offer no actionable signal during the window when informed flow is moving prices. We propose an information-theoretic framework for quantifying informed flow on prediction markets. We introduce the *Information Leakage Score* (ILS), a label generator that quantifies how much of a market's terminal information move was priced in before the corresponding public news event, and show that ILS admits a clean interpretation in terms of the Murphy decomposition of the Brier score: high ILS corresponds to front-loaded resolution. We specify the score's resolution-typology and operational scope conditions and report a pilot empirical study on n=725 event-resolved Polymarket markets that produces structural negative findings: a resolution-anchored news-timestamp proxy does *not* separate event-resolved markets from a matched control population, and zero of 24 documented insider-trading cases satisfy the original ILS scope. The audit reveals that documented Polymarket insider cases are systematically *deadline-resolved* ("Will event X occur by date Y?"), falling outside the original scope. We accordingly extend the score to a deadline-ILS variant anchored at the underlying event timestamp, with a per-category constant-hazard baseline for the time-to-event distribution. The contribution is a methodologically transparent label generator with an explicit scoring-rule reading, scope conditions validated by negative findings, and an extension that closes the gap between methodology and the population in which informed trading has been empirically attested.[1]

Keywords: prediction markets, informed trading, information leakage, Murphy decomposition, proper scoring rules, market microstructure, blockchain forensics, Polymarket.

1 Introduction

Prediction markets aggregate dispersed private information into a publicly observable price signal [Wolfers and Zitzewitz, 2004, Manski, 2006]. Modern blockchain-based platforms have brought these mechanisms to scale: Polymarket alone processed over five hundred million dollars of contract volume in 2024, and weekly volume reached the billion-dollar range during peaks of the 2024 U.S. presidential election cycle. Their published probabilities have become reference points that journalists, donors, traders, and institutions cite when responding to political and economic uncertainty—a coordination role distinct from forecasting accuracy.

The same features that make these markets useful information aggregators—low-friction global access, on-chain transparency, pseudonymous participation—also create unusually permissive conditions for trading on material non-public information (MNPI). Mitts and Ofir [2026] estimate approximately $143 million in aggregate anomalous profit extracted across Polymarket during a two-year window, applying a five-feature composite screen to over 210,000 wallet–market pairs. Mamageishvili et al. [2025] independently document $40 million of arbitrage and quasi-arbitrage profit. Documented case studies include hours-before-event positioning around the February 2026 U.S.–Israeli strike on Iran, the December 2025 release of Google's proprietary "Year in Search" rankings, the U.S. operation in Venezuela, and several recent OpenAI product launches.

These findings establish that informed trading is a quantitatively significant phenomenon in decentralized prediction markets. They do not address what we view as the operationally relevant question: *can informed flow be detected while a market is still open*, in time for an outside observer to either act on the signal or, in a regulatory or platform-integrity setting, to intervene before resolution? Existing work is overwhelmingly post-hoc. Mitts and Ofir [2026] use a screen that explicitly conditions on profitability and proximity to resolution, both observable only after the fact. The gap between post-hoc identification and real-time detection is substantial: real-time detection requires features that can be computed from information available before resolution, and a label that is itself recovered from historical data without leakage.

Contribution. We separate *retrospective forensics* from *real-time detection* and develop the methodological core of a framework whose design goal is the latter. Our contributions are: (1) the Information Leakage Score (ILS), a label generator that quantifies the fraction of a market's terminal information move accomplished before public

news, with three explicit scope conditions (Section 2); (2) a Murphy-decomposition reading of ILS that connects the score to the proper-scoring-rule literature, identifying it with the *front-loaded resolution* component of the Brier decomposition (Section 3); (3) a three-way resolution typology (event-resolved, deadline-resolved, unclassifiable) on the 911,237-market Polymarket corpus, validated by the structural fact that markets the classifier identifies as deadline-resolved exhibit a 100% NO-resolution rate by construction (Section 2); (4) a pilot empirical study on n=725 event-resolved markets that produces negative findings substantively informative for the framework (Section 4); and (5) a deadline-ILS extension that closes the gap between methodology and the population in which insider trading has been documented (Section 5).

2 Information Leakage Score

Consider a resolved binary market M with three timestamps: T_{open} (market creation), T_{news} (first public mention of resolution-relevant information), and T_{res} (UMA Optimistic Oracle resolution). Let $p(t) \in [0,1]$ denote the YES-token mid-price at time t, and $p_{T_{\text{res}}} \in \{0,1\}$ the binary resolution outcome.

Definition. The pre-news drift, total information move, and Information Leakage Score are

$$\Delta_{\text{pre}}=p(T_{\text{news}})-p(T_{\text{open}}), \quad \Delta_{\text{tot}}=p_{T_{\text{res}}}-p(T_{\text{open}}), \quad (1)$$

$$\text{ILS}(M) \;=\; \frac{\Delta_{\text{pre}}}{\Delta_{\text{tot}}}, \quad (2)$$

defined whenever $|\Delta_{\text{tot}}| > \varepsilon$ for a small threshold (we use ε=0.05, corresponding to a 5-cent total move). The score takes value ILS≈1 when the full information move is priced in before the news (strong leakage), ILS≈0 when the market reacted only to the public announcement, and ILS<0 when pre-news price moved against the eventual outcome.

Scope conditions. ILS is interpretable only under three conditions: (i) edge-effect, $|p(T_{\text{open}}) - 0.5| \le 0.4$, ruling out near-degenerate openings; (ii) trivial-resolution, $|\Delta_{\text{tot}}| \ge \varepsilon$; (iii) anchor-sensitivity, requiring that the recovered value be robust across multiple anchor specifications (we evaluate this with the proxy-sensitivity analysis of Section 4). Reporting raw ILS values on markets that violate these conditions produces results that cannot be cleanly interpreted; we treat the conditions as mandatory.

Resolution typology. A principled application of ILS requires a structural distinction between markets in which the resolution is triggered by a publicly observable event (the news) and markets in which it is triggered by the elapsing of a contractual deadline. We classify the full Polymarket corpus into three types: *event-resolved*, *deadline-resolved*, and *unclassifiable*. The classification is empirically validated by the structural fact that markets identified as deadline-resolved exhibit a 100% NO-resolution rate by construction in resolved-and-expired form: "Will X happen by Y?" resolves NO iff the deadline elapsed without occurrence. ILS as defined by (2) is well-posed on event-resolved markets only; the deadline-resolved case is treated separately in Section 5.

3 Murphy Decomposition Reading

The classical Murphy decomposition [Murphy, 1973] expresses the Brier score of a probabilistic forecaster as

$$B \;=\; \text{UNC} \;+\; \text{REL} \;-\; \text{RES},$$

where $\text{UNC} = \bar{o}(1-\bar{o})$ is irreducible outcome uncertainty, REL is calibration error, and RES is the resolution component capturing squared gap between bin-conditional and unconditional outcome frequencies. Lower Brier requires higher resolution.

The market-implied probability $p(t)$ can itself be viewed as a sequence of forecasts. Let $\text{RES}(t)$ denote the resolution component computed using market predictions up to time t. Then $\text{RES}(T_{\text{res}})$ is the total resolution accomplished over the market's lifetime, and $\text{RES}(T_{\text{news}})$ is the portion already accomplished before public news.

Remark 1 (ILS as front-loaded resolution)**.** *Under regularity conditions on the price process—that pre-news price moves are unbiased estimators of the conditional outcome probability and that bin-conditional outcome frequencies converge to bin means—ILS is approximately the share of the market's total resolution component accumulated before* T_{news}*:*

$$\text{ILS}(M) \;\approx\; \frac{\text{RES}(T_{\text{news}})}{\text{RES}(T_{\text{res}})}.$$

A high-ILS market is one whose discriminative power was front-loaded: it had separated outcome from base rate before any public information justified that separation. Under the null of an efficient market with no private information, $\text{RES}(T_{\text{news}}) \approx 0$ and ILS ≈ 0, with all resolution accruing through the post-news price reaction.

This connection places ILS on the same theoretical footing as the proper-scoring-rule literature and clarifies the sense in which informed flow is detectable: it is the component of market discriminative power that arrives ahead of its public-information justification. Unlike post-hoc PnL-based screens, the score is well-defined on any resolved market with a recoverable T_{news} and admits a calibration-aware reading. The same UNC/REL/RES partition applies diagnostically to a downstream binary detector trained on ILS-derived labels.

4 Pilot Findings: Negative Results

We report a pilot study on the n=725 event-resolved markets in three target categories (military and geopolitical actions, corporate proprietary disclosures, regulatory decisions) for which ILS is computable after price-data and timestamp-validity filters. In the ab-

Table 1: Pilot vs. matched control under the resolution-anchored T_{news} proxy. Bootstrap 95% CI on the difference of medians (control minus pilot) is $[+0.023, +0.066]$, entirely positive.

Group	n	Med. ILS	Pos. %
Pilot (event-resolved, target)	725	−0.084	15.2%
Control (unclassifiable, matched)	683	−0.043	21.4%

Table 2: FFIC inventory audit. Reasons compound. The dominant exclusion—resolution-type—is structural, not implementation noise.

Exclusion reason	n	%
Resolution-type not event-resolved	21	87.5%
No trade history (n_{trades}<100)	16	66.7%
No `market_label` record	19	79.2%
Edge-effect violation ($p_{T_{\text{open}}}$>0.9)	2	8.3%
Eligible markets (all satisfied)	**0**	**0%**

sence of recoverable article-derived timestamps for admin-resolved markets, we adopt the resolution-anchored proxy $T_{\text{news}} \equiv T_{\text{res}} - 24\,\text{h}$ as a working hypothesis to be evaluated, not as a substitute for a real news timestamp.

Finding 1: The proxy does not separate target from control. Table 1 reports the central comparison. The pilot is matched against a control of n=683 unclassifiable markets (sports outcomes, behavioural-prediction markets) with identical volume and trade-coverage filters. The control population shows the higher positive-ILS rate (21.4% vs. 15.2%); the Mann–Whitney U test rejects equal distributions at p=1×10^{-6}, but in the wrong direction. The mechanism becomes legible once the resolution structure is examined: for the unclassifiable controls (sports, behavioural), the resolution event *is* the news event, and the resolution-anchored proxy legitimately captures a pre-news interval. For event-resolved political and regulatory markets, the actual news typically precedes T_{res} by hours to days, and $T_{\text{res}} - 24$h captures a window that is partially or entirely *post*-news.

Finding 2: Tighter proxy windows do not recover the signal. Recomputing ILS under three tighter resolution-anchored offsets ($T_{\text{res}}-6$h, -2h, -1h) collapses the positive-ILS share toward zero and inflates the share of $|\text{ILS}| > 1$ extreme values. Spearman rank correlation between the 24-hour and 1-hour proxies is ρ=0.542 on the 221 markets where both are defined—moderate, not robust. The 15.2% rate at the 24-hour offset does not survive tightening.

Finding 3: 0/24 documented insider cases are in scope. The structurally most consequential finding emerges from auditing the *ForesightFlow Insider Cases* (FFIC) inventory—24 markets across 8 publicly documented insider-trading episodes (Iran-strike, Maduro/Venezuela, FTX/SBF, Bitcoin ETF, 2024 U.S. Presidential top-line, Romanian election, Year-in-Search, OpenAI launches)—against the scope conditions of Section 2. Table 2 reports the result. Zero markets simultaneously satisfy all conditions. The dominant exclusion is structural: 21/24 (87.5%) markets are classified as deadline-resolved or unclassifiable rather than event-resolved.

Documented insider cases on Polymarket are systematically deadline-resolved. Inspection of the public-reporting record shows a consistent pattern: documented insider activity has been reported on contracts of the form "Will event X occur by date Y?"—an Iran strike on Israel by Nov. 8, U.S. entry into Iran by Apr. 30, Maduro in U.S. custody by Jan. 31, the Bitcoin ETF approved by Jan. 15. The reported trades are pre-event purchases of YES shares that pay out when the event occurs and the deadline is satisfied. This is not coincidence: the deadline structure is precisely what makes a market actionable for a participant with private information about *timing*—the contract has a fixed expiration, the YES leg is asymmetrically rewarded if the event happens on time, and the price during the pre-event window is a direct function of the market's aggregated belief about whether the event will occur.

The audit therefore exposes a structural gap between the score as defined in (2) and the population in which the phenomenon of interest is empirically attested. Closing this gap is not a question of better data or tighter proxy; it requires extending the score to deadline-resolved cases.

5 Deadline-ILS Extension

Consider a deadline market on the question "Will event E occur by deadline D?". Let $\theta_t \in [0,1]$ denote participants' marginal belief at time t that E will occur before D; under a calibrated market $p(t) \approx \theta_t$. Without informative news, θ_t is determined by survival of the deadline window:

$$\theta_t \;=\; \theta_{T_{\text{open}}} \cdot \frac{S(D-t \mid T_{\text{open}})}{S(D-T_{\text{open}} \mid T_{\text{open}})},$$

where $S(\tau \mid T_{\text{open}})$ is the survivor of the time-to-event distribution conditional on T_{open}-information. We adopt a constant-hazard parametric form $S(\tau) = e^{-\lambda\tau}$ with rate λ fitted on a per-category sample of resolved deadline markets in the same category that resolved YES, giving the closed form $\theta_t = \theta_{T_{\text{open}}}\,(1-e^{-\lambda(D-t)})/(1-e^{-\lambda(D-T_{\text{open}})})$. The constant-hazard assumption is a deliberately weak baseline: it asserts only that conditional probability of arrival in the next instant given non-arrival is constant. Tighter parametric forms (Weibull, time-varying λ) are natural refinements once empirical evaluation surfaces deviations.

Let T_{event} denote public observation of E (defined when E occurs, undefined when the deadline expires without occurrence). For a deadline-resolved market with $T_{\text{event}} \in$

$[T_{\text{open}}, D]$ we define

$$\Delta^{\text{dl}}_{\text{pre}} = p(T^{-}_{\text{event}}) - \theta_{T_{\text{open}}},$$
$$\Delta^{\text{dl}}_{\text{tot}} = p_{T_{\text{res}}} - \theta_{T_{\text{open}}},$$
$$\text{ILS}^{\text{dl}}(M) = \Delta^{\text{dl}}_{\text{pre}} / \Delta^{\text{dl}}_{\text{tot}}.$$

The extension preserves the scope conditions of Section 2 with one explicit modification: the trivial-resolution condition is evaluated on $\Delta^{\text{dl}}_{\text{tot}}$ relative to $\theta_{T_{\text{open}}}$, not $p(T_{\text{open}})$, to absorb the rational-decay component into the baseline. The Murphy-decomposition reading of Remark 1 carries through with T_{event} replacing T_{news}.

The extension closes the gap between the original methodology and the population in which informed trading has been empirically attested. The end-to-end empirical evaluation of ILS^{dl} on the 2026 U.S.–Iran conflict cluster—the largest documented Polymarket deadline-cluster in the public-reporting record—is reported in a companion paper, where the single-case computation produces $\text{ILS}^{\text{dl}} = +0.113$ at the article-derived event anchor, versus -0.331 at the legacy resolution-anchored proxy: a 0.444 shift on opposite sides of zero, establishing that the article-derived anchor materially changes the substantive reading of the score. A population-scale evaluation across 12,708 markets in three insider-relevant categories (October 2020–April 2026) is reported in a second companion paper.

6 Conclusion

We have introduced an information-theoretic framework for quantifying informed flow on prediction markets. The Information Leakage Score is a label generator with explicit scope conditions, a clean Murphy-decomposition reading as front-loaded resolution, and an extension that addresses the population in which insider trading on these markets has been documented. The pilot study's structural negative findings—that resolution-anchored proxies are misaligned with event timing for the target population, and that documented insider cases are systematically deadline-resolved—are themselves the empirical basis for the framework's eventual operational form: ILS is interpretable only on event-resolved markets passing all three scope conditions, and the deadline-ILS extension is the natural object on which to evaluate informed flow in the subpopulation where it has been most clearly attested.

The score's connection to proper scoring rules ties this work to the broader AI-forecasting evaluation literature and to parallel evaluation frameworks for AI forecasting agents. The framework is platform-agnostic in construction; the empirical findings are Polymarket-specific. The framework's principal use is as a label-generator stage in a real-time detection pipeline; it is not, on its own, an informed-flow detector. Subsequent work develops detector training on the joint feature space of ILS^{dl} and on-chain wallet features, with operational metrics evaluated under a longitudinal cumulative-evaluation protocol.

References

J. Mitts and M. Ofir. From Iran to Taylor Swift: Informed trading in prediction markets. Working paper, Columbia Law School / University of Haifa, 2026.

A. H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

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

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

A. Mamageishvili, A. Shcherbenko, P. Saldaña, and L. Kiffer. Strategic bidding wars in on-chain auctions. Preprint, IMDEA Networks Institute, 2025. arXiv:2502.16606.

D. Easley, N. M. Kiefer, M. O'Hara, and J. B. Paperman. Liquidity, information, and infrequently traded stocks. *Journal of Finance*, 51(4):1405–1436, 1996.

D. Easley, M. M. López de Prado, and M. O'Hara. Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5):1457–1493, 2012.

A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

R. Gómez-Cram, Y. Guo, T. I. Jensen, and H. Kung. Prediction market accuracy: Crowd wisdom or informed minority? Working paper, London Business School / Yale SOM, 2026.