Operating-Point Control for Delegated Evidence Acquisition in Prediction-Market Monitoring

Moin Mattar [1] **Taka Khoo** [1]

Abstract

We study delegated evidence acquisition for unresolved prediction markets under bounded verification cost. Our central claim is that useful delegation is an operating-point problem over evidence quality, coverage, and review load, rather than a push toward maximal autonomy. We design a rule-grounded, stagewise hybrid agent with persistent external memory and evaluate it along five axes: scale, calibration, ablations, lead time, and failure modes. The system screens more than 1,200 live markets and escalates only 1.92% for active tracking. At the default operating point, channel precision is 0.588 (N=3616), message precision is 0.526 (N=5000), and the evidence-positive market rate is 0.575. A stricter policy improves precision but sharply reduces downstream evidence yield, while removing persistent memory increases review load by 41.9%. These results support a simple systems conclusion: stagewise control plus memory yields materially more controllable behavior than collapsed or memory-free variants.

1. Introduction

Prediction markets are a strong benchmark for delegated monitoring because they aggregate scattered information into continuously updated prices while attaching real costs to false alerts and missed signals (Wolfers & Zitzewitz, 2004; Hanson et al., 2006; Bossaerts et al., 2024). The goal in this setting is not maximal autonomy. The goal is controllable autonomy.

This problem is also operationally real because market platforms expose machine-readable interfaces. Polymarket provides official REST, WebSocket, and SDK support for market access; Kalshi provides APIs and market data; and Telegram provides an official Bot API for downstream messaging and interaction (Polymarket, 2026; Kalshi, 2026; Telegram, 2026). Recent work has also begun to study agentic structure directly in prediction-market environments, including semantic relationship discovery on Polymarket and LLM-based semantic filtering in market settings (Capponi et al., 2025; Kim et al., 2026).

Our contribution is not a new monolithic model. It is a systems design for agentic monitoring under bounded review cost. We propose a stagewise hybrid agent with explicit thresholds and persistent external memory. We show that this architecture creates a tunable frontier over evidence quality, coverage, and review burden, and that this frontier degrades under collapsed or memory-free variants. This framing is consistent with recent work on AI delegation and monitoring (Tomašev et al., 2026; Ray, 2025).

2. Problem Formulation

Consider a market m, where q_m denotes market text, r_m denotes resolution context, and z_m denotes latent event state. The agent interacts with

$$\mathcal{G} = (\mathcal{S}, \mathcal{O}, \mathcal{A}, T, \Omega, U), \tag{1}$$

where observations consist of market snapshots, channel metadata, and message streams, and actions include triage, retrieval, filtering, matching, and activation.

The policy is stagewise:

$$m \xrightarrow{\phi_{\text{triage}}} \mathcal{Q}_m \xrightarrow{\phi_{\text{chan}}} \mathcal{C}_m \xrightarrow{\phi_{\text{msg}}} \mathcal{E}_m \xrightarrow{\pi_{\text{act}}} a_m, \tag{2}$$

where $\mathcal{Q}_m$ is a set of market-conditioned queries, $\mathcal{C}_m$ is the set of retained channels, $\mathcal{E}_m$ is the retained evidence, and $a_m \in \{0, 1\}$ is the downstream activation state.

Channel scoring is

$$s_{\text{chan}}(c, m) = w_e E(c, m) + w_t T(c, m) + w_p P(c, m) + w_\ell L(c, m), \tag{3}$$

where E is entity overlap, T is topical alignment, P is rule-triggered phrase evidence, and L is a locality/language prior. Channel retention requires $s_{\text{chan}} \geq \tau_c$.

[1]Dartmouth College, Hanover, NH, USA. Correspondence to: Moin Mattar <moin.a.mattar.28@dartmouth.edu>, Taka Khoo <taka.khoo.TH@dartmouth.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Table 1. Main results at scale and default operating point.

Metric	**Value**
Live markets screened	1200
Escalated for active tracking	1.92%
Mean screening throughput (markets/s)	6.84
Channel precision (τ_c=0.18, N=3616)	0.588
Message precision (default, N=5000)	0.526
Evidence-positive market rate	0.575
Review-load proxy (retained artifacts/market)	39.66
Lead-time mean / median (h, N=9)	0.84 / 0.71

Message retention is contract-conditioned rather than merely event-conditioned:

$$p_{\text{msg}}(x, m) = \mathbb{P}(x \in \mathcal{E}_m \mid x, m, r_m), \qquad p_{\text{msg}} \geq \tau_x. \tag{4}$$

Let $\theta = (\tau_c, \tau_x, \mu)$ denote the operating point, where μ captures message-stage policy strictness. The delegated objective is

$$\max_{\pi_\theta} \mathbb{E}\left[\sum_t \gamma^t \left(\alpha u_t - \beta \text{FA}_t - \eta \text{VC}_t - \kappa \text{OC}_t\right)\right], \tag{5}$$

where u_t is evidence utility, FA_t is false-activation cost, VC_t is verification cost, and OC_t is operational cost. The question is therefore not whether a single threshold is universally best, but whether the system exposes useful and stable operating regimes.

3. Agent Design and Evaluation

The system is built as a stagewise hybrid controller. Triage allocates attention across markets, source filtering limits noisy discovery, message matching enforces contract relevance, and activation happens only downstream of retained evidence. Persistent external memory stores market, source, message, and match state so the system can reduce rediscovery, avoid duplicate review, and support replay.

This design maps directly onto live infrastructure: market discovery and streaming are pulled from exchange APIs, and analyst-facing escalation is delivered through bot interfaces (Polymarket, 2026; Telegram, 2026). We evaluate one hypothesis from multiple angles: *stagewise hybrid control with memory yields better quality–coverage–verification controllability*. Experiments target scale, operating-point calibration, architecture and policy ablations, lead time, and failure modes.

4. Results

Scale. The agent screens 1,200 active markets and escalates only 1.92%, showing explicit monitoring-budget allocation rather than uniform polling.

Table 2. Architecture and policy ablations.

Mode	**Msg-P**	**Msg-R**	**Load**	**Evid+**
Full hybrid	0.526	0.638	39.66	0.575
No channel filter	0.524	0.642	39.83	0.578
No persistent memory	0.523	0.641	56.28	0.580
Planner-lite	0.503	0.611	18.26	0.548
Strict policy	0.746	0.138	31.60	0.063
Loose policy	0.425	0.950	44.60	0.675

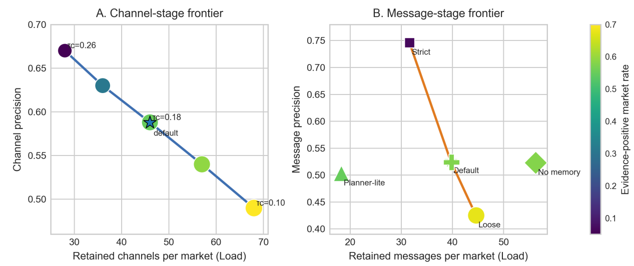

Figure 1. Two-panel operating frontier. **A**: channel-stage threshold sweep over retained channels per market and channel precision. **B**: message-stage policy sweep over retained messages per market and message precision. Marker size/color encode evidence-positive market rate.

Operating frontier. Figure 1 shows a stage-separated frontier. At the channel stage, increasing τ_c reduces retained channels per market while improving channel precision. At the message stage, stricter matching increases message precision from 0.425 to 0.746 while reducing message load from 44.60 to 31.60 and evidence-positive rate from 0.675 to 0.063. The point here is not a universal optimum, but exposed controllability.

Ablations. Table 2 separates architecture effects from policy effects. Policy strictness drives the expected precision–coverage trade-off. Removing persistent memory leaves precision nearly unchanged but increases review load by 41.9% and retained channels by 49.7%, indicating that external state materially improves efficiency.

Lead time and failures. Lead time is positive in the evaluated cases (mean 0.84h, range 0.18–1.92h), indicating a modest but usable early-signal margin. The dominant failures are semantically adjacent but contract-irrelevant channels and negated-evidence messages, pointing to contradiction-aware matching and stronger rule conditioning as the next bottlenecks.

5. Conclusion

Delegated evidence acquisition for prediction markets is best framed as operating-point control, not maximal autonomy. Our results show that a stagewise hybrid agent with persistent memory can screen broadly, escalate sparingly, and expose a usable frontier over quality, coverage, and review cost. The practical value is not just better precision at one point; it is better controllability across points. This is exactly the property a human supervisor needs in a noisy, high-volume monitoring setting (Tomašev et al., 2026; Ray, 2025).

References

Bossaerts, F., Yadav, N., Bossaerts, P., Nash, C., Todd, T., Rudolf, T., Hutchins, R., Ponsonby, A.-L., and Mattingly, K. Price formation in field prediction markets: The wisdom in the crowd. *Journal of Financial Markets*, 68: 100909, 2024. doi: 10.1016/j.finmar.2023.100909.

Capponi, A., Gliozzo, A., and Zhu, B. Semantic trading: Agentic ai for clustering and relationship discovery in prediction markets. *arXiv preprint arXiv:2512.02436*, 2025. doi: 10.48550/arXiv.2512.02436. URL `https://arxiv.org/abs/2512.02436`.

Hanson, R., Oprea, R., and Porter, D. Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior & Organization*, 60(4):449–459, 2006. doi: 10.1016/j.jebo.2005.12.009.

Kalshi. Kalshi api documentation. `https://docs.kalshi.com/welcome`, 2026. Official documentation for market data, trade execution, REST, and WebSocket APIs.

Kim, S., Kim, M., Kwon, J., Kim, Y., Kagan, N., Lee, J. W., Levy, O., Lopez-Lira, A., Lee, Y., and Choi, C. Llm as a risk manager: Llm semantic filtering for lead–lag trading in prediction markets. *arXiv preprint arXiv:2602.07048*, 2026. doi: 10.48550/arXiv.2602.07048. URL `https://arxiv.org/abs/2602.07048`.

Polymarket. Polymarket documentation. `https://docs.polymarket.com/`, 2026. Official documentation for REST, WebSocket, and SDK access.

Ray, K. Monitoring teams of ai agents. *Journal of Artificial Intelligence Research*, 84, 2025. doi: 10.1613/jair.1.19798. URL `https://www.jair.org/index.php/jair/article/view/19798`.

Telegram. Telegram bot api. `https://core.telegram.org/bots/api`, 2026. Official HTTP-based API for Telegram bots.

Tomašev, N., Franklin, M., and Osindero, S. Intelligent ai delegation. *arXiv preprint arXiv:2602.11865*, 2026. URL `https://arxiv.org/abs/2602.11865`.

Wolfers, J. and Zitzewitz, E. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004. doi: 10.1257/0895330041371321.