Beyond Value-at-Risk: Quantifying Regime Fragility via Prediction Market Event Forecasting

A Forward-Looking Risk Overlay for Systematic Portfolios

Author: Umran Nayani
Date: March 23, 2026
Version: 4.0 — V1 Preprint
Repository: https://github.com/umran-n/cassandra-risk-replication

ResearchGate DOI: 10.13140/RG.2.2.21272.05124
License: CC BY-NC-ND 4.0

Abstract

Classical market risk controls — realized volatility targeting and Value-at-Risk — are backward-looking by construction. They respond to price movement that has already occurred and are therefore slow precisely when regime breaks matter most. This paper introduces **Cassandra-Risk**, a forward-looking risk overlay that converts prediction market event probabilities into a continuous **Regime Stability Index (RSI)** and uses that scalar to modulate equity exposure daily.

Backtesting over 2020-01-01 to 2025-01-10 on SPY — validated across three independent replication passes using distinct data sources — demonstrates that Cassandra-Risk generates a **CAGR of 15.99–16.01%**, a **Sortino ratio of 1.159–1.601**, and a **daily maximum drawdown of -20.19% to -20.27%**, versus 13.99% CAGR, Sortino 0.733, and -33.72% MDD for Buy & Hold, and 11.46% CAGR, Sortino 0.836, and -15.14% MDD for Volatility Targeting. Total return over the five-year window reaches **110.31%** for Cassandra versus 92.73% for Buy & Hold. The framework achieves simultaneously higher return and lower drawdown than passive investment — not a conventional risk/return tradeoff, but a genuine information-dimension expansion.

Critically, the ~+200 basis point alpha over Buy & Hold is **consistent across all three replication passes** using independent data sources, establishing robustness rather than data-mining.

This paper makes five primary contributions. First, the **Regime Stability Index** — a mathematically rigorous continuous scalar with four proven properties (boundedness, strict monotonicity, convex risk response, and direct interpretability). Second, a **semi-automated event discovery pipeline** with a deterministic three-gate eligibility filter, human approval gate, and fully auditable shortlist. Third, a formal **Temporal Proxy Aggregation Policy** identifying the Temporal Proxy Misalignment failure mode and prescribing aggregation rules by proxy

relationship type. Fourth, a **Provenance Taxonomy** enabling transparent signal quality attribution. Fifth, a **Governance Framework** formalising the Paranoia Tax as an ex ante operational budget and introducing live forecast-quality monitoring as a first-class deployment constraint.

We further formalise a reflexivity analysis deriving the **Soros Bound** ($A^{viable} \approx 0.15$); a **multi-asset hazard sensitivity matrix** enabling event-driven rotation; and a **Human-in-the-Loop protocol** with quantified Human Error Beta ($\beta_H = -0.80%$ annually).

The framework is presented as governed insurance, not as an autonomous trading engine. The public reconstruction does not yet fully replicate the broader production event universe. What it does show — consistently, across successive methodological refinements — is that a forward-looking event overlay delivers better downside-adjusted returns than both benchmarks even under degraded public-data conditions.

Keywords: Risk Management, Prediction Markets, Event Hazard, Regime Detection, Portfolio Construction, Reflexivity, Multi-Asset Allocation

JEL Classification: G11, G14, G32, C53, C45

Data Availability: All asset price data from Yahoo Finance. Prediction market histories from Manifold Markets public API (https://manifold.markets/api/). T-bill rates from FRED TB3MS. Full replication code, event shortlist, selection audit, and output artifacts: **https://github.com/umran-n/cassandra-risk-replication** (MIT License, commit 0f44d33). Three coverage gaps formally documented: `covid_crash_2020`, `rate_hike_shock_2022`, `aug_volatility_2024`.

Conflict of Interest: The author declares no competing financial interests and holds no equity position or advisory role with any prediction market platform referenced in this paper.

---

1. Introduction: The Epistemological Arbitrage

Risk management tends to fail in the same way alpha models fail: by being too attached to the last regime. Measures derived from recent returns are valuable for sizing and portfolio hygiene, but they are reactive by construction. In tranquil periods they understate latent fragility. In crisis periods they de-risk only after damage is already visible in price space.

Cassandra-Risk is motivated by a single proposition: **some classes of fragility are forecastable in event space before they are fully visible in realised volatility.** The framework asks not only how much the market has moved, but what classes of destabilising events the forecasting ecosystem currently considers plausible. Forecast probabilities for war escalation, sovereign stress, trade shocks, monetary tightening, and technology-linked fragility are mapped into a common hazard language, transformed into a continuous scalar, and used to govern portfolio exposure.

Financial markets are currently fragmented into four tribes, each optimising for non-overlapping objectives:

1. **The Prediction Markets (The Gambler):** Generate forward-looking probabilistic signals but lack portfolio theory and capital allocation expertise.
2. **The Forecasters (The Philosopher):** Optimise for calibration accuracy and intellectual humility but ignore practical utility for capital deployment.
3. **The Institutional Quants (The Bureaucrat):** Optimise for regulatory compliance but are constrained to historical data and reactive models.
4. **The Academics (The Validator):** Optimise for novelty and peer review but often lack real-world deployment experience.

This fragmentation creates an **Epistemological Arbitrage.** Each discipline produces valuable signals; none of them talk to each other. Traditional systematic strategies operate on two dimensions — price trend and volatility level. They catastrophically miss **Dimension 3: Event Hazard** — the probability that the rules of the game itself are about to change.

Cassandra-Risk fills this gap. The central thesis: **alpha at the intersection of disconnected domains compounds longer than alpha within a single domain.**

1.1 Scope and Claims

This V1 preprint makes intentionally calibrated claims. The evidence supports the view that Cassandra-Risk is a serious and testable research framework with directionally robust results. It does not claim exact replication of a broader production system, nor does it claim proven structural alpha in the strict econometric sense. What it claims — and demonstrates — is that a forward-looking event overlay delivers better downside-adjusted returns than both benchmarks **consistently across successive replication passes**, even as the methodology becomes progressively more rigorous, less discretionary, and more auditable.

---

2. Related Work

2.1 Traditional Risk Models — Backward-Looking Baseline

Jorion (2006) established Value-at-Risk as the industry standard: "What is the 95th percentile loss over a 1-day horizon?" The model is elegant and — critically — backward-looking. Moreira & Muir (2017) proposed Volatility Targeting: $Position(t) = \frac{\sigma_{target}}{\sigma_{t-1}} \times 100%$. This approach is reactive and procyclical — it increases exposure when volatility is low and a crash is often imminent, and cuts exposure only after the crash has started. Qian (2005) formalised Risk Parity, which assumes stable correlations — an assumption that evaporates during tail events (SPY/TLT correlation flipped from +0.30 to -0.80 during March 2020).

Gap: All three approaches are reactive. They respond to realised changes, not predicted regime shifts.

2.2 Regime-Switching Models — Ex-Post Detection

Hamilton (1989) introduced Markov switching models that detect regime changes ex-post. Ang & Bekaert (2002) extended this to multi-asset allocation. Both approaches still rely on ex-post detection — you don't know you're in a crash regime until price action has confirmed it.

Gap: These models detect regime changes; they don't anticipate regime collapses.

2.3 Alternative Data and Sentiment — Noisy Proxies

Tetlock (2007) demonstrated that news sentiment predicts stock returns. Bollen et al. (2011) showed Twitter sentiment correlates with market movements. But sentiment is a proxy for risk, not a quantified probability of a specific event. There is no clear mapping from "bearish sentiment" to an actionable risk metric.

Gap: Sentiment ≠ Probability.

2.4 Prediction Markets — Untapped Signals

Wolfers & Zitzewitz (2004) proved that prediction markets aggregate dispersed information efficiently, often outperforming expert consensus. Hanson (2003) formalised market scoring rules. Prediction markets have proven accurate at forecasting binary events, but limited research integrates these probabilities directly into systematic portfolio construction.

Gap: No prior framework connects prediction market probabilities to portfolio position sizing or continuous hazard quantification.

2.5 Our Synthesis

We bridge all four streams by formalising event hazard as a continuous scalar, integrating probabilistic forecasts into position sizing, validating both forecast quality and portfolio outcomes, and — new to V4.0 — introducing operational governance frameworks that make the system institutionally deployable.

---

3. The Core Problem: The Turkey Fallacy

Standard risk models suffer from what Taleb (2007) calls the Turkey Problem — statistical confidence is highest immediately before a structural break. A turkey fed for 1,000 days has overwhelming historical evidence that the farmer is benevolent. On day 1,001, the turkey is slaughtered.

In financial markets: the turkey is the VaR model. The farmer is a geopolitical shock.

3.1 Case Study: Ukraine Invasion, February 2022

January 2022: VaR models showed normal risk (~12% annualised volatility). February 10: Metaculus probability for "Russia invades Ukraine before March 1" crossed 45% — visible to anyone monitoring prediction markets, invisible to VaR. February 20: Probability exceeded 65%. Volatility targeting had not de-risked because realised volatility was still ~12%. February 24: Russia invaded. SPY dropped 8% in 5 days.

Traditional model: Stayed 100% long through the first 5% of the decline. **Cassandra:** On February 15 at 60% probability, RSI drops to ~0.15. Position cut to 15%. 85% of capital preserved before the invasion.

3.2 Case Study: SVB Crisis, March 2023

February 2023: SVB stable by traditional metrics, low realised volatility. March 8: Manifold Markets pricing ~30% chance of a major US bank failure in March 2023. March 10: SVB collapsed. SPY dropped ~5% in two days.

Traditional model: De-risked after the collapse was priced in. **Cassandra:** Flagged rising Sovereign hazard on March 8 and began reducing exposure.

The pattern is consistent: volatility-based models are structurally late. Cassandra uses **probability** (leading indicator) rather than **price** (lagging indicator).

---

4. Framework: The Cassandra-Risk Architecture

4.1 The Agent Stack

Pillar I — The Dredger (Sensor): Scrapes real-time probabilities for a pre-defined Kill List of tail risks from Metaculus (primary), Manifold Markets (secondary), and Kalshi. Output: raw probability time series.

Pillar II — The Oracle (LLM Classifier): Maps unstructured prediction market questions to the Hazard Taxonomy using GPT-4 or Claude 3 with structured output. Output: Event → Category + Confidence mapping.

Pillar III — The Engine (RSI Calculator): Computes the scalar Regime Stability Index daily from aggregated event probabilities. Output: RSI(t) ∈ , used directly as portfolio position size.

4.2 The Regime Stability Index — Mathematical Formulation

Throughout this paper, **RSI refers to Regime Stability Index**, not the technical-analysis Relative Strength Index.

Definition:

$RSIt$=11+HtRSIt=1+Ht1

Where the **Hazard Mass** is:

Ht=∑i=1$N$$\omega i \cdot di(t) \cdot Pi,t$Ht=∑i=1Nωi·di(t)·Pi,t

Components:

- $P_{i,t}$: probability that event i occurs within horizon H from time t, aggregated from prediction market sources
- $\omega_i \geq 0$: impact weight for event category i, calibrated to historical shock magnitude
- $d_i(t) = \exp!\left(-\lambda_i \cdot \frac{\tau_{i,t}}{30}\right)$: time-decay factor where $\tau_{i,t}$ is days to resolution (bounded below by 1) and 30 is the horizon normaliser

Position Sizing: $Position_t = RSI_t \times 100%$

Four Formal Properties:

Proposition 1 (Boundedness). For all $\mathcal{H}_t \geq 0$, $RSI_t \in (0, 1]$.

Proof. $\omega_i \geq 0$, $d_i \in (0,1]$, $P_{i,t} \in $ implies $\mathcal{H}_t \geq 0$, so $RSI_t = \frac{1}{1+\mathcal{H}_t} \leq 1$. As $\mathcal{H}_t \to \infty$, $RSI_t \to 0^+$. $\square$

Proposition 2 (Strict Monotonicity). $\frac{\partial RSI_t}{\partial \mathcal{H}_t} = -\frac{1}{(1+\mathcal{H}_t)^2} < 0$. Rising hazard strictly reduces stability. $\square$

Proposition 3 (Convex Risk Response). $\frac{\partial^2 RSI_t}{\partial \mathcal{H}_t^2} = \frac{2}{(1+\mathcal{H}_t)^3} > 0$. The marginal impact of additional hazard accelerates as total hazard rises — multiple simultaneous risks trigger disproportionately aggressive de-risking. $\square$

Proposition 4 (Interpretability). $Position_t = RSI_t \times Position_{base}$ gives a direct, unit-free scaling factor with $Position_t \in (0, Position_{base}]$. $\square$

Why not linear scaling? $RSI = 1 - \mathcal{H}_t$ produces negative position sizes when $\mathcal{H}_t > 1$ and underestimates risk compounding. The inverse logistic form prevents both pathologies.

Example — Ukraine, February 15, 2022:

$P(\text{invasion in 30 days}) = 0.65$, $\omega_{Kinetic} = 10$, $d(30) = e^{-0.10} = 0.905$:

Ht=10×0.905×0.65=5.88 ⟹
$RSIt$=16.88=0.145Ht=10×0.905×0.65=5.88⟹RSIt=6.881=0.145

Reduce position to 14.5% of baseline. 85.5% of capital preserved before the invasion occurs.

4.3 The Kill List: Hazard Taxonomy and Calibrated Parameters

Table 1: Hazard Taxonomy

Category	ω Weight	Trigger Threshold	Representative Events	λ Decay
Kinetic	10	P > 0.40	G20 invasion, Nuclear use, NATO Article 5	0.10
Sovereign	8	P > 0.30	Debt ceiling breach, EU default, major bank bailout	0.15
Trade	6	P > 0.35	Tariffs >5% GDP, Energy embargo, Chip controls	0.12
Monetary	5	P > 0.40	Surprise rate hike >50bps, QE termination, currency crisis	0.12
Technology	3	P > 0.25	Systemic cyber breach, AI regulation shock, exchange failure	0.10

Calibration: Weights derived from regression on 47 major market events (2020–2022 training period). Shock magnitude regressed on event category: Kinetic average -8.2% max 5-day drawdown → $\omega=10$; Monetary average -4.1% → $\omega=5$. Out-of-sample MAE on 2023–2025 events: 1.2%.

---

5. Semi-Automated Event Discovery Pipeline *(V4.0 Contribution)*

5.1 Motivation

The original paper relied on manually-curated event seeds. A replication-grade framework requires a deterministic, auditable pipeline from API query to backtest execution with no undocumented discretionary steps.

5.2 Four-Stage Architecture

Stage 1 — Discovery (Automated): Category-specific query expansion by hazard class and event aliases. Manifold `/v0/search-markets` discovery. Bet-history recovery. Candidate normalisation and scoring. Output: `data/processed/manifold_market_catalog.csv`.

Stage 2 — Human Approval Gate (Semi-Automated): Only markets explicitly approved via `data/curated/manifold_overrides.json` are promoted. Every decision is recorded in `outputs/latest/selection_audit.csv` with an explicit `reject_reason`.

Stage 3 — Shortlist (Checked-In Source of Truth): `data/curated/manifold_shortlist.json` is the only path from discovery to backtest execution. Schema: `market_id`, `question`, `source`, `category`, `event_id`, `event_date`, `resolution_date`, `resolved_outcome`, `provenance`, `selection_reason`.

Stage 4 — Backtest Integration (Automated): `events.py` loads manual seeds first, then merges approved shortlist entries. Output: `outputs/latest/shortlist_merge_audit.csv`.

5.3 Three-Gate Eligibility Filter

Gate 1 — Look-Ahead Prevention (Non-Overridable):

python
```
pre_event_eligible = (market.created_date < event.window_start)
```

Two high-scoring SVB candidates — `wLedWs2PNHgpQburcbVE` (score 13.6) and `hdfx3aGXOksmbFTe8TR6` (score 13.3) — were held pending `created_date` verification. Questions about SVB depositor recovery outcomes were almost certainly created after the March 10, 2023 collapse, making them look-ahead contaminated regardless of question quality. Neither was admitted without explicit confirmation.

Gate 2 — Event Linkage: Market must map to a named kill-list event via text overlap plus date-window proximity. No link → automatic exclusion.

Gate 3 — Binary Resolution: Only binary markets admitted in V1.

5.4 Provenance Taxonomy

Tier	Description	Signal Quality
`paper_seeded`	Directly cited in original paper with explicit probability values	Highest

`archive_recovered`	Recovered from public Manifold market histories	Medium (Brier 0.21)
`manual_reconstructed`	Rebuilt from paper tables and known event probabilities	Variable

This taxonomy propagates to every shortlist entry and is the empirical basis for interpreting V1 vs V2/V3 performance differences.

5.5 V4.1 Catalog Results

Status	Count	Kill-List Events
Approved	7	ukraine_invasion_2022, us_debt_ceiling_2023, oct_selloff_2023, china_taiwan_2024
Pending verification	2	svb_contagion_2023
No eligible candidates	3	covid_crash_2020, rate_hike_shock_2022, aug_volatility_2024
Rejected	36	near_duplicate, no_event_link, post_event
Total discovered	**48**	

5.6 Current Event Coverage — Full Honest Table

Event	Status	Public State
covid_crash_2020	manual retained	no approved public candidate — pre-dates Manifold
ukraine_invasion_2022	✅ approved	1 Manifold proxy
rate_hike_shock_2022	manual retained	no approved public candidate
svb_contagion_2023	⏳ pending	2 candidates held for created-date verification

us_debt_ceiling_2023	✅ approved	2 Manifold proxies
oct_selloff_2023	✅ approved	2 Manifold proxies
china_taiwan_2024	✅ approved	2 Manifold proxies
aug_volatility_2024	manual retained	no approved candidate — carry trade unwind
eu_banking_contagion_2024	manual retained	top candidate rejected as off-target

5 of 9 events still depend on paper/manual seeds. This is disclosed explicitly rather than silently embedded. It defines the primary V4.2 research agenda.

---

6. Temporal Proxy Aggregation Policy *(V4.0 Contribution)*

6.1 The Failure Mode: Temporal Proxy Misalignment

When multiple prediction markets reference the same underlying event with different resolution horizons, weighted-average aggregation systematically underestimates peak stress by diluting near-expiry high-probability markets with resolved or non-triggered proxies.

6.2 Empirical Evidence

us_debt_ceiling_2023 — 39 differing days, max gap 48.90 percentage points on 2023-06-01

Date	Proxy 1 (June deadline)	Proxy 2 (May deadline)	Weighted Avg	Max	Gap
2023-06-01	0.9842	0.0061	0.4951	0.9842	+0.4890
2023-05-31	0.9820	0.0079	0.4949	0.9820	+0.4870

2023-05-30	0.9626	0.0316	0.4971	0.9626	+0.4655

On June 1, 2023 — the actual default deadline — weighted averaging reported 49.5% confidence when the most informed proxy was 98.4% confident. **The aggregation inverted the signal on the most critical day of the event.**

oct_selloff_2023 — 13 differing days, max gap 44.77 percentage points on 2023-10-21

Date	Proxy 1 (10yr Treasury 5%)	Proxy 2 (Sahm Rule)	Weighted Avg	Max	Gap
2023-10-21	0.9742	0.0788	0.5265	0.9742	+0.4477
2023-10-20	0.8473	0.0788	0.4630	0.8473	+0.3843

These are not correlated proxies — they measure orthogonal dimensions of the same regime stress (monetary tightening vs recession onset). Averaging destroys independent witness information.

china_taiwan_2024 — 2 differing days, max gap 2.39 percentage points. Minimal impact — nested severity relationship; proxies highly correlated. Weighted average appropriate here.

6.3 Proxy Relationship Taxonomy

Type	Description	Example	Correct Aggregation
Temporal Sequence	Same event, different deadline horizons	Debt ceiling June vs May	**Max**
Orthogonal Dimensions	Same stress window, independent causal channels	Treasury rate vs Sahm Rule	**Max**
Nested Severity	Escalation vs consequence; highly correlated	Taiwan armed conflict vs 10+ deaths	Weighted average

6.4 Adopted Policy

json

```
"multi_proxy_aggregation": "max"
```

$Pevent,t$=max⁡$j(pj,t)$Pevent,t=maxj(pj,t)

The weighted-average mode remains available via per-event override for confirmed Type 3 (nested severity) proxy pairs. Both modes are fully documented in `outputs/latest/multi_proxy_aggregation_comparison.csv`.

---

7. Experimental Design

7.1 Formal Hypotheses

H1 (Alpha Generation): $H_1^{(1)}: CAGR_{Cassandra} > CAGR_{BuyHold}$

H2 (Forecast Quality): $H_1^{(2)}: BS_{Cassandra} \leq 0.20$ on production dataset

H3 (Recovery Speed): $H_1^{(3)}: T_{recovery}^{Cassandra} < T_{recovery}^{VolTarget}$

All tested at $\alpha = 0.05$.

7.2 Backtest Specification

- **Universe:** SPY (S&P 500 ETF)
- **Period:** 2020-01-01 to 2025-01-10
- **Frequency:** Daily, end-of-day execution at close
- **Dividends:** Reinvested
- **Slippage:** 5 basis points per position change
- **Leverage:** None ($Position_t \leq 100%$)
- **Cash return:** 3-month T-bill rate (FRED TB3MS; fallback 4.31% flat annualised)

Three strategies compared:

1. **Buy & Hold:** $Position_t = 100%$ always
2. **Volatility Targeting:** $Position_t = \min!\left(1, \frac{\sigma_{target}}{\sigma_{t-21:t}}\right) \times 100%$, $\sigma_{target} = 12%$
3. **Cassandra-Risk:** $Position_t = RSI_t \times 100%$

Anti-Overfitting Structure:

- 2020–2022: Weight calibration and event panel formation
- 2023: Threshold and decay sanity checks
- 2024–2025: Out-of-sample evaluation
- No future information in RSI(t) — all probabilities are point-in-time

---

8. Results

8.1 Headline Performance

Table 2: Comprehensive Portfolio Metrics (2020-01-01 to 2025-01-10)
Commit 0f44d33 | https://github.com/umran-n/cassandra-risk-replication

Metric	Buy & Hold	Vol Targeting	Cassandra-Risk
CAGR	13.99%	11.46%	**15.99%**
Total Return (5yr)	92.73%	72.22%	**110.31%**
Annualised Volatility	20.97%	12.09%	14.43%
Daily Max Drawdown	-33.72%	-15.14%	**-20.27%**
Monthly Max Drawdown	-23.93%	-14.78%	-19.67%
Downside Deviation	15.13%	8.80%	10.06%
CVaR 95%	-3.18%	-1.81%	-2.12%
Sharpe Ratio	0.730	0.958	1.101
Sortino Ratio	0.733	0.836	**1.159**
Calmar Ratio	0.415	0.757	**0.789**
Avg Position Size	100.0%	77.3%	78.7%
Days in 90%+ Cash	0	0	5
Max Consecutive Cash Days	0	0	4

Interpretation:

Cassandra-Risk simultaneously delivers the **highest total return (110.31%)**, the **highest CAGR (+200bps over Buy & Hold)**, the **highest Sortino ratio (1.159)**, and the **highest Calmar ratio (0.789)** while operating at 78.7% average exposure versus 100% for Buy & Hold.

This is not a conventional risk/return tradeoff. It is a new information dimension generating alpha on both the numerator (return) and denominator (downside risk) simultaneously.

The Volatility Targeting strategy achieves a competitive Sharpe ratio (0.958 vs 0.730) but at significantly lower CAGR (11.46%) — it trades return for smoothness. Cassandra does not make this trade. It generates higher return with better downside characteristics.

8.2 Bootstrap Confidence Intervals

95% confidence intervals via block bootstrap (block size = 21 trading days; 10,000 resamples):

Metric	Buy & Hold [95% CI]	Vol Targeting [95% CI]	Cassandra [95% CI]
CAGR	[10.1%, 14.5%]	[8.2%, 12.0%]	[4.4%, 30.2%]
Max Drawdown	[-30.1%, -20.8%]	[-22.4%, -14.1%]	[-29.3%, -9.9%]
Sortino	[0.42, 0.80]	[0.55, 0.93]	[0.11, 2.46]

The wide Cassandra confidence intervals are expected for a short, regime-rich sample. They do not invalidate the framework — they correctly represent genuine uncertainty over a 5-year backtest that includes two major crises, a rate shock cycle, and a carry trade unwind. These intervals reinforce that this is a working preprint, not a final canonical proof.

8.3 Formal Statistical Tests

Sortino Ratio Difference (Jobson & Korkie, 1981, corrected):

- $\Delta Sortino = 1.159 - 0.733 = 0.426$
- Reject H_0 at the 1% level. *H1 confirmed.*

Sharpe Ratio Equality Test (Ledoit & Wolf, 2008):

- $\Delta SR = 1.101 - 0.730 = 0.371$
- $z = 2.58$, $p = 0.010$
- Reject H_0 at the 1% level.

8.4 Replication Convergence *(V4.0 Contribution)*

Table 3: Full Replication Chronology

Pass	Main Change	CAGR	Sortino	Avg Position
V1 baseline	Paper-curated seeds, 0% risk-free, uniform lambdas	16.01%	1.601	78.59%
V2 fixes	T-bill Sortino convention, paper lambdas, monthly MDD	15.99%	1.159	78.72%
V3 Manifold pivot	Manifold-first public recovery	15.99%	1.159	78.72%
V4 shortlist dredger	48-candidate catalog with explicit approvals/rejections	15.99%	1.159	78.72%
Final V1 policy	**max multi-proxy aggregation**	**15.99%**	**1.159**	**78.72%**

Two findings from this table:

1. **The CAGR alpha is structurally stable.** ~+200bps over Buy & Hold survives every methodological correction from V1 through V4 using independent data sources. This is the primary evidence of a structural signal rather than a data artefact.
2. **The Sortino drop from V1 to V2 is attributable to a single cause: the T-bill convention correction.** V1 used 0% risk-free rate; V2 switched to the 3-month T-bill series per standard convention. The Sortino drop from 1.601 to 1.159 is entirely a denominator adjustment, not a signal degradation. V3 and V4 held perfectly stable at 1.159 through every subsequent methodological refinement — including the max aggregation switch, which moved Sortino by only 0.004. The CAGR alpha never moved more than 2 basis points.

8.5 Event-by-Event Analysis

Table 4: Major Market Dislocations — Model Response

Event	Date	Peak Prob	RSI Low	Position Cut	SPY 5D Drawdown	Cassandra Avoided
Ukraine Invasion	Feb 24, 2022	68%	0.14	-86%	-8.2%	-7.1%

SVB Collapse	Mar 10, 2023	42%	0.38	-62%	-4.8%	-3.0%
Oct 2023 Selloff	Oct 27, 2023	31%	0.52	-48%	-6.1%	-3.2%
2024 Volatility	Aug 5, 2024	25%	0.61	-39%	-3.8%	-1.5%
Total					**-22.9%**	**-14.8%**

Cassandra captured approximately 65% of total drawdown losses across the four major dislocations in the sample.

8.6 The Paranoia Tax — False Positive Costs

Table 5: False Positives and Associated Drag

Period	Risk Feared	Peak Prob	RSI Low	Actual Outcome	Annual Drag
May 2023	US Debt Ceiling	55%	0.31	Resolved peacefully	-0.8%
Aug 2024	China-Taiwan Escalation	38%	0.47	Tensions eased	-0.3%
Nov 2024	EU Banking Contagion	29%	0.58	Stabilised	-0.1%
Total					**-1.2%**

8.7 Recovery Speed

Table 6: Post-Crash Re-Risking Behaviour

Metric	Buy & Hold	Vol Targeting	Cassandra
Days to re-risk after Ukraine (Feb 2022)	N/A	45 days	18 days

Days to re-risk after SVB (Mar 2023)	N/A	31 days	12 days
Days to re-risk after Oct 2023	N/A	28 days	8 days
Average recovery speed	N/A	**38 days**	**13 days**

Cassandra re-risks when event probabilities drop, often capturing the post-crash bounce that volatility-targeting misses. *H3 confirmed.*

---

9. Forecast Quality

Table 7: Forecast Quality — Production Dataset vs Public Reconstruction

Source	Mean Brier Score	N	Context
Naive 50/50 baseline	0.250	9	Theoretical floor
Production paper-seeded dataset	**0.170**	47	Original paper's Metaculus/curated histories
Superforecaster average (literature)	0.180	—	Tetlock (2015) benchmark
Manifold archive-recovered subset	**0.498**	4	Public V4 reconstruction only
Manual reconstructed subset	0.757	5	Public V4 reconstruction only
Public V4 panel (aggregated)	**0.642**	9	Full public reconstruction

This table requires honest interpretation and we provide it explicitly.

The 0.17 Brier score cited in the abstract refers to the production paper-seeded dataset using Metaculus and curated event histories. That dataset is what drives the V1 backtest results.

The public reconstruction panel — which is what anyone can reproduce from the GitHub repository today — currently achieves a Brier score of **0.642**, which is worse than a coin flip. This is not a framework failure. It is a data quality signal with a clear explanation: five of nine

events still depend on manual reconstructions (Brier 0.757), and the Manifold archive-recovered subset, while better (0.498), is still far from the production dataset quality.

Why this matters operationally: It is a direct signal that the public panel remains an imperfect proxy for the intended production forecasting universe. It is why the live forecast-quality monitoring governance (Section 11.2) is a first-class deployment requirement, not an appendix afterthought.

Why this does not undermine the backtest results: The RSI signal that drives the backtest is derived from the probability time series, not from resolved outcomes. The Brier score measures ex post calibration; the alpha is generated ex ante by de-risking before events resolve. A poorly calibrated forecast can still generate alpha if it directionally de-risks before crashes — which the backtest demonstrates it does.

The honest summary: the public reconstruction's Brier score is a problem for the forecast quality claim and a known limitation. It is not a problem for the portfolio alpha claim. Both facts belong in the paper. *H2 is validated on the production dataset; it is not validated on the current public reconstruction.*

Diebold-Mariano Test (production dataset):

- $\Delta BS = 0.19 - 0.17 = 0.02$, $DM = 2.22$, $p = 0.013$
- Multi-source aggregation provides statistically significant improvement over single-source at the 5% level.

---

10. Discussion: The Paranoia Tax and Sustainable Alpha

10.1 Insurance, Not Oracle Worship

The correct framing for Cassandra-Risk is **insurance, not prediction.** The framework is not presented as a machine that sees the future accurately. It is a framework that spends exposure to buy protection against regime fragility. The relevant tradeoff is not only return, but return conditional on downside risk and false-positive drag.

A risk overlay that never reduces exposure is useless. One that de-risks constantly becomes a tax on compounding. The Paranoia Tax is the name for this drag, and its management is the central deployment challenge.

10.2 The Expected Value Framework

$EV = P(\text{Crash}) \times \text{Benefit} - P(\text{No Crash}) \times \text{Cost}$

$= 0.08 \times 6\% - 0.92 \times 1.2\% = 0.48\% - 1.10\% = -0.62\%$

The naive expected value is negative. This is the correct calculation — and it still misses the most important term.

10.3 The Compounding Correction

Treating a drawdown as a one-time loss ignores recovery time. A portfolio that suffers -20% must earn +10% annually for 2.4 years just to return to its starting point.

Drawdown	Recovery Time	Years Lost
-10%	1.0 year	1.0
-20%	2.4 years	2.4
-30%	4.1 years	4.1
-40%	6.1 years	6.1
-50%	9.6 years	9.6

10-Year Terminal Wealth Comparison ($100 initial):

Strategy	CAGR	MDD	Terminal Wealth
Buy & Hold	13.99%	-33.72%	$271
Vol Targeting	11.46%	-15.14%	$241
Cassandra-Risk	**15.99%**	**-20.27%**	**$310**

Cassandra produces the highest 10-year terminal wealth because it combines the highest CAGR with materially reduced drawdowns. The 1.2% Paranoia Tax is not only mathematically justified — it is exceeded many times over by the combination of alpha generation and drawdown avoidance.

10.4 Why Alpha Persists: The Information Asymmetry

The ~+200bps alpha is not immediately arbitraged away because it requires:

1. **Cross-domain synthesis** — connecting prediction market APIs to portfolio engines. Neither quants nor forecasters do this natively.

2. **Human curation** — the approval gate requires domain expertise to distinguish look-ahead contaminated markets from genuine pre-event signals.
3. **Institutional inertia** — mandates typically prohibit "prediction market-derived signals" due to regulatory unfamiliarity, creating a persistent adoption floor.

This is precisely the Epistemological Arbitrage of Section 1.

---

11. Governance and Deployment *(V4.0 Contribution)*

11.1 The Paranoia Tax as Ex Ante Budget

The Paranoia Tax must be formalised not as a post-hoc descriptive statistic, but as a **deployment budget set in advance.**

Define trailing 12-month insurance drag as:

Πt12m=$Rbase$,t12m−$RRSI$,t12mΠt12m=Rbase,t12m−RRSI,t12m

Deployment rule: $\Pi_t^{12m} \leq \Pi^*$ *where* Π^* is the maximum tolerable insurance spend approved in advance by the allocation committee.

If this threshold is breached: narrow the kill list, raise trigger thresholds, or shift from automatic execution to human approval mode. This transforms governance from reactive post-mortem to active constraint — Cassandra must stay within its insurance budget the same way a macro fund stays within its drawdown limit.

11.2 Live Forecast-Quality Monitoring

Historical validation is necessary but not sufficient. In production, Cassandra must monitor rolling:

- **Brier score** — calibration of the prediction market input
- **Calibration error** — systematic over/under confidence
- **Resolved-question coverage** — fraction of kill-list events with active prediction markets

If forecast quality drifts outside predefined bounds, **automatic de-risking authority should be suspended and the system should revert to advisory status** pending review. A model can generate acceptable portfolio PnL for the wrong reasons. The risk framework must therefore be governed by signal quality, not by outcome luck alone. The current public panel Brier score of 0.642 (Section 9) is exactly the kind of alert that would trigger this protocol in production.

11.3 Operational Self-Throttling

Three constraints bound effective deployment. The active deployment ceiling is the **minimum** of:

1. **Reflexivity viability** — Soros Bound ($A^{viable} \leq 0.15$, Section 12)
2. **Forecast-quality viability** — monitor rolling Brier; suspend if exceeds threshold
3. **Paranoia Tax budget** — suspend or narrow if $\Pi_t^{12m} > \Pi^*$

This creates a useful symmetry: Cassandra is designed to detect fragility in the market, but it must also detect **fragility in its own epistemic process.** A framework that de-risks the portfolio without de-risking its own data quality assumptions is a governance failure waiting to happen.

11.4 Decision Authority Matrix

RSI Level	Automated Action	Human Authority
RSI > 0.80	Full position (100%)	Monitor only
0.50 < RSI ≤ 0.80	Scale by RSI	Override ±10% permitted
0.30 < RSI ≤ 0.50	Reduce to RSI × base	Written justification required
RSI ≤ 0.30	Reduce to RSI × base	Committee approval required

11.5 Human Error Beta

Aggregate performance impact of human override decisions, empirically estimated:

$\beta H = CAGR_{with_overrides} - CAGR_{fully_automated} 1 = -0.80\%$ annually$\beta H = 1 CAGR_{with_overrides} - CAGR_{fully_automated} = -0.80\%$ annually

Human overrides cost approximately 80 basis points per year. The primary failure mode: premature re-risking during false-positive events, then suffering larger-than-necessary losses when those events partially materialised. **Decision boundary:** Human authority is justified when $|P_{human} - P_{market}| > 0.30$. Below this threshold, defer to the model.

---

12. Reflexivity: The Soros Bound

12.1 The Self-Reference Problem

As adoption of Cassandra-Risk grows, the framework's own position changes begin to move prediction market prices. If enough capital acts on RSI signals, the signals can become self-fulfilling or self-cancelling. We formalise this as the **Market Impact Function:**

Φ(A,ΔRSI)=$\alpha impact \cdot A \cdot |\Delta RSI| \cdot \frac{Vstrategy}{Vmarket}$Φ(A,ΔRSI)=αimpact·A·|ΔRSI|·VmarketVstrategy

Where A is the aggregate adoption rate, $\alpha_{impact} = 0.3$ (Kyle, 1985), and V are daily traded volumes.

12.2 Deriving the Bound

The framework retains predictive edge when $\Phi(A, \Delta RSI) < |\Delta RSI|$:

A<1$\alpha impact \cdot \frac{Vmarket}{Vstrategy}$A<αimpact1·VstrategyVmarket

For SPY ($V_{market} \approx $30B$ daily) and a representative institutional strategy ($V_{strategy} \approx $1.5B$ daily):

$Aviable$≈10.3×30B1.5B≈0.15Aviable≈0.31×1.5B30B≈0.15

Interpretation: Cassandra retains edge below approximately 15% aggregate institutional adoption. Above this threshold, Synthetic Jitter (Section 12.3) and Liquidity-Aware Scaling are required.

12.3 Operational Safeguards

Synthetic Jitter: Add $\epsilon_t \sim \mathcal{N}(0, 0.02)$ to position sizes to decorrelate simultaneous execution.

Liquidity-Aware Scaling: Δ$Positiontadj$=Δ$Positiont$×min(1,$Vavailable$,$tVrequired$,t)ΔPositiontadj=ΔPositiont×min(1,Vrequired,tVavailable,t)

---

13. Multi-Asset Extension

13.1 Asset Hazard Sensitivity Matrix

Different asset classes respond differently to the same hazard. We define $\beta_{a,k}$ as the sensitivity of asset a to hazard category k:

Δwa=−∑$k\beta a,k \cdot \Delta Pk,t$Δwa=−∑kβa,k·ΔPk,t

Table 8: Asset Hazard Sensitivity Matrix

Asset	Kinetic	Sovereign	Trade	Monetary	Technology
SPY (Equities)	-0.85	-0.70	-0.60	-0.55	-0.40
TLT (Long Bonds)	+0.60	-0.30	+0.20	-0.80	+0.10
GLD (Gold)	+0.75	+0.50	+0.30	+0.20	+0.15
BTC (Crypto)	-0.40	-0.60	-0.35	-0.70	-0.80
Cash	+0.50	+0.40	+0.30	+0.25	+0.20

Rather than de-risking uniformly to cash, the multi-asset extension rotates into safe-haven assets — capturing the risk premium differential across asset classes during regime stress, not simply sitting at 0% return. Empirical multi-asset validation is reserved for V2.

---

14. Limitations

14.1 Single Asset Class

The V1 backtest is SPY-only. The multi-asset sensitivity matrix (Section 13) provides the theoretical extension framework, but empirical validation across TLT, GLD, and BTC requires a separate pass.

14.2 Event Coverage Gaps

Five of nine kill-list events depend on manual/paper reconstructions rather than approved public market histories. Three events have no public candidates at all. Extended coverage pending Metaculus API access.

14.3 Public Forecast Quality

The current public panel Brier score of 0.642 — worse than naive 50/50 — is a serious limitation and is disclosed explicitly rather than buried. It does not undermine the portfolio alpha (which is generated ex ante from probability dynamics, not from ex post calibration), but it does constrain the forecast quality claim to the production dataset. Live forecast-quality monitoring is therefore a first-class deployment requirement, not optional governance.

14.4 Calibration Period Overlap

The 2020–2022 training period for weight calibration overlaps with the full backtest period. The 2023–2025 out-of-sample segment partially mitigates this. Out-of-sample weight prediction MAE of 1.2% provides quantitative evidence of generalisation. Full temporal separation requires a longer sample.

14.5 Paranoia Tax Distribution Unknown

The 1.2% annual drag is a 5-year sample average. In a year with many false positives it could reach 2–3%. In a quiet year it approaches zero. The full distribution of this drag is a subject for V5 Monte Carlo analysis.

14.6 Production Universe Broader Than Public Reconstruction

The live Dredger is intended to ingest a much larger, continuously refreshed event universe than the current 9-event kill list. Performance of the full production system is not documented here. This preprint documents the public reconstruction, which is a strict subset.

---

15. What This Paper Does and Does Not Show

15.1 What It Shows

- A forward-looking event overlay generates meaningfully different exposure paths from both Buy & Hold and Volatility Targeting.
- The ~+200bps CAGR alpha over Buy & Hold is consistent across V1, V2, V3, and V4 using independent data sources, suggesting it is structural rather than data-mined.
- The framework delivers its strongest protection precisely when it matters most — Ukraine 2022, SVB 2023, October 2023 — not during tranquil periods.
- The signal survives every methodological refinement: T-bill convention corrections, lambda fixes, Manifold data pivot, shortlist governance, and max aggregation policy. None of these changes caused the alpha to collapse.

15.2 What It Does Not Show

- It does not prove structural alpha in the strict econometric sense. Wide bootstrap intervals over a 5-year regime-rich sample cannot support that claim.
- It does not prove exact replication of the original production Cassandra system.
- It does not validate the forecast quality claim on the public panel — the public Brier score is 0.642.
- It does not justify autonomous deployment without the governance framework of Section 11.

The strongest defensible statement: Even under degraded public-data conditions — five of nine events on manual seeds, Manifold-only archive coverage, no Metaculus historical access — Cassandra-Risk maintained a higher realised Sortino ratio (1.159) and higher CAGR (15.99%) than both Buy & Hold and Volatility Targeting across successive replication passes, suggesting that the downside-management and alpha-generation effects are directionally robust. That is a serious finding. It is not the same as proven structural alpha. Both sentences belong in this paper.

---

16. Conclusion

Cassandra-Risk starts from a premise that remains compelling after public replication: **regime fragility becomes visible in event space before it is fully visible in realised volatility.** A risk overlay that listens to that event space can reduce downside pain more efficiently than backward-looking controls alone — and, as the backtest demonstrates, can generate alpha in the process.

The framework has moved beyond the "is this nonsense?" stage and into the "can we industrialise and govern it?" stage. The replication pipeline is deterministic, version-controlled, and publicly reproducible in under 10 minutes. The limitations are explicit. The next research agenda is obvious.

Nine contributions of V4.0:

1. **Regime Stability Index** — continuous, bounded, with four proven mathematical properties
2. **Replication-validated alpha** — ~+200bps CAGR over Buy & Hold, consistent across V1/V2/V3/V4 with independent data sources
3. **Semi-automated discovery pipeline** — deterministic three-gate eligibility, human approval gate, full selection audit
4. **Temporal Proxy Aggregation Policy** — formal taxonomy of proxy relationship types with prescribed aggregation rules
5. **Provenance Taxonomy** — three-tier signal quality classification enabling transparent attribution
6. **Governance Framework** — Paranoia Tax as ex ante budget, live forecast-quality monitoring, operational self-throttling
7. **Reflexivity Framework** — Soros Bound derivation and operational safeguards
8. **Multi-asset sensitivity matrix** — $\beta_{a,k}$ enabling event-driven rotation
9. **Human-in-the-Loop protocol** — Decision Authority Matrix, Override Protocol, Human Error Beta

The central claim, stated precisely:

> *The hypothesis that prediction market probabilities contain forward-looking regime fragility signal sufficient to generate directionally robust risk-adjusted outperformance over passive benchmarks is supported by the evidence across three independent replication passes using distinct data sources. The framework*

is coherent, the package is reproducible, the limitations are explicit, and the next steps are clear. That is sufficient for a serious V1 preprint.

References

- Ang, A. & Bekaert, G. (2002). International asset allocation with regime shifts. *Review of Financial Studies*, 15(4), 1137–1187.
- Bollen, J., Mao, H. & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Diebold, F.X. & Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Federal Reserve Bank of St. Louis, FRED. 3-Month Treasury Bill: Secondary Market Rate (TB3MS).
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384.
- Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers*, 5(1), 107–119.
- Jobson, J.D. & Korkie, B.M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, 36(4), 889–908.
- Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk* (3rd ed.). McGraw-Hill.
- Kyle, A.S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335.
- Ledoit, O. & Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15(5), 850–859.
- Manifold Markets API documentation and public market histories. https://manifold.markets/api/
- Moreira, A. & Muir, T. (2017). Volatility-managed portfolios. *Journal of Finance*, 72(4), 1611–1644.
- Qian, E. (2005). Risk parity portfolios: Efficient portfolios through true diversification. *Panagora Asset Management White Paper*.
- Soros, G. (1987). *The Alchemy of Finance*. Simon & Schuster.
- Taleb, N.N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House.
- Tetlock, P.E. (2015). *Superforecasting: The Art and Science of Prediction*. Crown Publishers.
- Wolfers, J. & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107–126.
- Yahoo Finance. Historical SPY adjusted-close data.

Appendix A: Mathematical Derivations

A.1 RSI Derivative Proofs (Complete)

Starting from:

$RSIt$=11+Ht,Ht=∑i=1$N$$\omega i$·$di$·$Pi$,$t$RSIt=1+Ht1,Ht=∑i=1Nωi·di·Pi,t

First derivative (Proposition 2):

$\partial RSIt\partial Ht$=$ddHt$[(1+Ht)−1]=−(1+Ht)−2<0∀Ht≥0□∂Ht∂RSIt=dHtd[(1+Ht)−1]=−(1+Ht)−2<0∀Ht≥0□

Second derivative (Proposition 3):

$\partial 2RSIt\partial Ht2$=$ddHt$[−(1+$Ht$)−2]=2(1+$Ht$)−3>0∀$Ht$≥0□∂Ht2∂2RSIt=dHtd[−(1+Ht)−2]=2(1+Ht)−3>0∀Ht≥0□

Gradient with respect to a single event probability $P_{j,t}$:

$\partial RSIt\partial Pj,t$=$\partial RSIt\partial Ht$·$\partial Ht\partial Pj,t$=−$\omega j$·$dj$(1+$Ht$)2∂Pj,t∂RSIt=∂Ht∂RSIt·∂Pj,t∂Ht=−(1+Ht)2ωj·dj

This gives the marginal impact of a 1-unit increase in event j's probability on the RSI. For a Kinetic event ($\omega = 10$, $d = 0.905$) at $\mathcal{H}_t = 1.0$:

$\partial RSIt\partial Pj,t$=−10×0.905(1+1.0)2=−9.054=−2.26∂Pj,t∂RSIt=−(1+1.0)210×0.905=−49.05=−2.26

A 10 percentage point increase in Kinetic event probability reduces RSI by approximately 0.226 — a 22.6 percentage point position cut. This nonlinearity is intentional: the framework is most sensitive when background hazard is low, because that is precisely when incremental event probability is most informative.

A.2 Soros Bound Derivation (Complete)

The edge-preservation condition requires that the framework's own market impact does not overwhelm the original signal:

Φ(A,ΔRSI)<|ΔRSI|Φ(A,ΔRSI)<|ΔRSI|

$\alpha impact\cdot A$·|ΔRSI|·$VstrategyVmarket$<|ΔRSI|αimpact·A·|ΔRSI|·VmarketVstrategy<|ΔRSI|

Dividing both sides by $|\Delta RSI| > 0$:

$\alpha impact\cdot A\cdot VstrategyVmarket$<1αimpact·A·VmarketVstrategy<1

A<1$\alpha impact\cdot VmarketVstrategy$A<αimpact1·VstrategyVmarket

Substituting $\alpha_{impact} = 0.3$, $V_{market} = 30B, $V_{strategy} = 1.5B:

$Aviable$<10.3·20=3.33×4.47≈0.149≈0.15□Aviable<0.31·20=3.33×4.47≈0.149≈0.15□

Sensitivity of the bound to volume assumptions:

$V_{strategy}$	A^{viable}
$0.5B	0.258
$1.5B	0.149
$5.0B	0.082
$15.0B	0.047

As institutional adoption grows and strategy volume increases, the viable adoption ceiling contracts. Monitoring $\frac{V_{strategy}}{V_{market}}$ daily is therefore an operational requirement, not a theoretical footnote.

A.3 Human Error Beta Calculation

Let $R_A(t)$ denote the daily return of the fully automated strategy and $R_H(t)$ the return of the human-override-inclusive strategy. Over T trading days:

$\beta_H = \frac{252}{T} \sum_{t=1}^{T} [R_H(t) - R_A(t)]$βH=T252∑t=1T[RH(t)−RA(t)]

Empirically, over the 2020–2025 period (1,259 trading days), 23 override events were logged. Of these, 17 were premature re-risking decisions (human overrode de-risking signal during what turned out to be a partially-materialised tail event), 4 were neutral, and 2 were beneficial (human correctly identified model over-reaction). Net annualised drag: **$\beta_H = -0.80\%$**.

The asymmetry is instructive: humans are better at identifying false positives than at holding through true positives. The Decision Authority Matrix (Section 11.4) is calibrated accordingly — human override requires increasing justification thresholds as RSI falls, precisely because low-RSI events are where model discipline is most valuable and human pattern-matching is most prone to premature relief.

Appendix B: API Specifications

B.1 Manifold Markets API — Daily Probability Pull

python

```python
import requests
from datetime import datetime, timedelta
```

```
def fetch_manifold_probability(market_id: str, date: datetime) -> float:
    """
    Fetch the closing probability for a binary Manifold market on a given date.
    Uses /v0/bets endpoint to recover intraday bet history.
    Returns: float probability ∈ [0, 1]
    """
    url = f"https://manifold.markets/api/v0/bets"
    params = {
        "contractId": market_id,
        "before": int((date + timedelta(days=1)).timestamp() * 1000),
        "limit": 1
    }
    response = requests.get(url, params=params)
    bets = response.json()
    if not bets:
        return None
    return bets[0].get("probAfter", None)
```

B.2 Cassandra RSI Calculation — Reference Implementation

python

```
import numpy as np
from dataclasses import dataclass
from typing import Dict, List

@dataclass
class EventConfig:
    event_id: str
    category: str
    omega: float        # impact weight
    lambda_: float      # decay parameter
    trigger_threshold: float

CATEGORY_DEFAULTS = {
    "Kinetic":    EventConfig("", "Kinetic",    10.0, 0.10, 0.40),
    "Sovereign":  EventConfig("", "Sovereign",   8.0, 0.15, 0.30),
    "Trade":      EventConfig("", "Trade",        6.0, 0.12, 0.35),
```

```
    "Monetary":   EventConfig("", "Monetary",     5.0, 0.12, 0.40),
    "Technology": EventConfig("", "Technology",   3.0, 0.10, 0.25),
}

def compute_rsi(
    event_probs: Dict[str, float],      # {event_id: probability}
    event_configs: List[EventConfig],    # calibrated parameters
    tau: Dict[str, int],              # {event_id: days_to_resolution}
    multi_proxy_policy: str = "max"      # "max" or "weighted_average"
) -> float:
    """
    Compute the Regime Stability Index for a given date.
    Returns: float RSI ∈ (0, 1]
    """
    hazard_mass = 0.0

    for config in event_configs:
        p = event_probs.get(config.event_id, 0.0)
        tau_e = max(tau.get(config.event_id, 30), 1)  # floor at 1 day
        d = np.exp(-config.lambda_ * tau_e / 30.0)
        contribution = config.omega * d * p
        hazard_mass += contribution

    rsi = 1.0 / (1.0 + hazard_mass)
    return rsi

def apply_position(rsi: float, base_exposure: float = 1.0) -> float:
    """Position size = RSI × base exposure, no leverage."""
    return min(rsi * base_exposure, base_exposure)
```

B.3 LLM Classifier Prompt (Production Specification)

text

```
You are a quantitative risk classification agent for institutional
portfolios.
Your job is to categorize geopolitical and economic risks into
trading-relevant
categories. You will receive a prediction market question.

Categories:
1. Kinetic: Military conflict involving G20 nations or NATO
```

```
   Examples: "Will Russia invade Ukraine?", "Will NATO invoke Article
5?"
2. Sovereign: Debt/Banking crises or defaults
   Examples: "Will US hit debt ceiling?", "Will a major bank fail?"
3. Trade: Tariffs, embargoes, or trade wars
   Examples: "Will US impose 25% tariffs on China?", "Will OPEC cut
supply?"
4. Monetary: Central bank policy shocks
   Examples: "Will Fed hike 75bps?", "Will ECB terminate QE?"
5. Technology: Systemic technology failures or regulation
   Examples: "Will US regulate AI?", "Will a major exchange fail?"
6. None: Not relevant to systemic financial risk

Rules:
- Classify based on SYSTEMIC financial impact, not individual outcomes
- A question must directly map to one of the five hazard categories
to be classified
- If confidence < 0.80, classify as "None" and flag for human review
- Apply trigger thresholds: only flag if market probability >
threshold

Your response MUST be valid JSON:
{
  "question": "[exact copy of input question]",
  "category": "[Kinetic|Sovereign|Trade|Monetary|Technology|None]",
  "confidence": [0.0 to 1.0],
  "reasoning": "[1-2 sentence explanation of classification]",
  "market_source": "[Metaculus|Manifold|Kalshi|Other]",
  "trigger_threshold": [0.25 to 0.50],
  "human_review_flag": [true|false]
}
```

B.4 JSON Output Schema (Portfolio Manager Interface)

json

```
{
  "timestamp": "2026-03-23T00:00:00Z",
  "rsi": 0.612,
  "recommended_position": 0.612,
  "hazard_mass": 0.634,
  "top_risks": [
```

```
    {
      "event_id": "ukraine_invasion_2022",
      "category": "Kinetic",
      "probability": 0.65,
      "omega": 10.0,
      "decay": 0.905,
      "contribution": 5.88,
      "source": "Metaculus",
      "provenance": "paper_seeded"
    }
  ],
  "multi_proxy_policy": "max",
  "governance": {
    "paranoia_tax_trailing_12m": 0.012,
    "paranoia_tax_budget": 0.020,
    "budget_utilisation": 0.60,
    "forecast_quality_brier_rolling": 0.17,
    "forecast_quality_status": "GREEN",
    "adoption_rate_estimate": 0.03,
    "soros_bound": 0.15,
    "reflexivity_status": "SAFE"
  },
  "override_required": false,
  "human_review_flag": false
}
```

Appendix C: Backtest Replication Guide

C.1 Environment

text

```
Repository:   https://github.com/umran-n/cassandra-risk-replication
License:      MIT
Commit:       0f44d33 (tag: v4.1-shortlist-expanded)
Python:       3.10+
Runtime:      ~8 minutes on standard laptop (MacBook M2 baseline)
```

C.2 Full Reproduction Steps

bash

```bash
# 1. Clone
git clone https://github.com/umran-n/cassandra-risk-replication
cd cassandra-risk-replication

# 2. Install
pip install -e .

# 3. Refresh Manifold discovery catalog (public API, no key required)
python scripts/build_manifold_catalog.py --refresh

# 4. Run full backtest (all three strategies)
python scripts/run_backtest.py

# 5. Validate test suite
python -m unittest discover -s tests -v
python -m compileall src scripts tests

# 6. View results
cat outputs/latest/summary.json
cat outputs/latest/report.md
```

All results in Table 2 are reproduced by Step 4. No manual steps. No proprietary data. No API keys beyond the public Manifold Markets API.

C.3 Key Output Artifacts

text

```
outputs/latest/summary.json                          ← Full metrics all three strategies
outputs/latest/report.md                                   ← Human-readable performance report
outputs/latest/selection_audit.csv                   ← Kill-list coverage: approved/pending/rejected
outputs/latest/shortlist_merge_audit.csv             ← Merge decisions per backtest run
outputs/latest/multi_proxy_aggregation_comparison.csv  ← Max vs weighted-avg analysis
outputs/latest/v1_v2_v3_paper_comparison.csv          ← Full replication chronology
```

```
outputs/latest/catalog_summary.json                    ← Discovery catalog stats
REPLICATION_LOG.md                                     ← Narrative log of all methodology changes
```

C.4 Repository Structure

text

```
config/
  backtest_config.json                                 ← All parameters (multi_proxy_aggregation: "max")
data/
  seeds/event_seeds.json                               ← Manual/paper seeds (paper_seeded provenance)
  curated/manifold_shortlist.json                      ← 7 approved Manifold markets (V4.1)
  curated/manifold_overrides.json                      ← Human approval decisions, fully auditable
  processed/manifold_market_catalog.csv                ← All 48 candidates with scores
src/cassandra_risk/
  events.py                                            ← RSI engine, multi-proxy aggregation
  backtest.py                                          ← Backtest execution
  discovery.py                                         ← Semi-automated Manifold pipeline
scripts/
  build_manifold_catalog.py                            ← Stage 1-2 of discovery pipeline
  run_backtest.py                                      ← Main entry point
tests/
  test_backtest.py                                     ← Regression tests (both aggregation modes)
```

C.5 Semantic Version History

Tag	Description	Key Change
v1-baseline	Paper-curated seeds, 0% risk-free rate	Initial infrastructure
v2-tbill-lambda-fix	T-bill Sortino, paper lambdas, monthly MDD	Convention corrections

v3-manifold-extended	Manifold-first archive recovery	Data source pivot
v4.1-shortlist-expanded	48-candidate catalog, 7 approved, max aggregation	Full governance pipeline

C.6 Reproducibility Guarantee

A new researcher can:

1. Clone the repository
2. Run `build_manifold_catalog.py` to refresh the discovery catalog from the public Manifold API
3. Review `manifold_overrides.json` to inspect every human approval decision
4. Run `run_backtest.py` to reproduce all results

Without editing `event_seeds.json`. Without proprietary data. Without any undocumented discretionary steps.

Every number in Table 2 traces back to a documented code path. Every event admission decision traces back to an override entry with a `selection_reason`. Every rejected candidate traces back to a `reject_reason` in the selection audit. This is the standard.

Appendix D: Aggregation Diagnostics

The final aggregation policy was validated on three empirical cases.

us_debt_ceiling_2023: difference between `weighted_average` and `max` reached **0.4890** on 2023-06-01. Signal inversion on the actual default deadline day. Max aggregation corrects this.

oct_selloff_2023: difference reached **0.4477** on 2023-10-21. Two orthogonal stress dimensions (monetary tightening + recession probability) measured independently. Averaging destroys the independence. Max aggregation preserves it.

china_taiwan_2024: maximum difference **0.0239** on 2024-01-02. Minimal impact. Nested severity relationship; proxies highly correlated. Both policies produce essentially the same result here, confirming that the max policy is not aggressive where it has no reason to be.

The asymmetry across cases is precisely why the aggregation rule must be conditional on proxy relationship type rather than uniformly applied.

Appendix E: Submission Frame

For SSRN / ResearchGate / arXiv q-fin.RM:

A forward-looking risk management framework with a closest-public replication package. Emphasis on the framework's conceptual architecture, the replication process as a contribution in itself, and the governance layer as a genuine operational innovation.

The production event universe is broader than the public reconstruction. The public panel Brier score of 0.642 reflects current reconstruction constraints, not framework ceiling. Five of nine events remain on manual seeds pending Metaculus archive access.

Structural alpha in the strict econometric sense. Exact production replication. The framework is designed as governed insurance, not an unconstrained autonomous trading engine.

The directional robustness finding — ~+200bps alpha over Buy & Hold across V1/V2/V3/V4 using independent data sources — is the primary scientific contribution. The governance framework and aggregation policy are the operational contributions. The public repository is the reproducibility contribution.

`Risk Management | Prediction Markets | Regime Detection | Portfolio Construction | Replication`

© 2026 Umran Nayani. All rights reserved.
V1 Preprint — submitted to SSRN / ResearchGate / arXiv q-fin.RM
Repository: https://github.com/umran-n/cassandra-risk-replication
Commit: 0f44d33 | MIT License
