PERCEPTUAL MACROECONOMICS

Narrative Co-Integration, Regime Dynamics,
and the Limits of Federal Reserve Data
in Asset Price Forecasting

An Introduction to the
IUVO™ Forecasting System

John M. Aaron

Milestone Planning and Research, Inc.

Authorship and Responsibility
The author assumes full responsibility for the accuracy, integrity, and interpretation of the material presented in this work. Any errors or omissions remain solely the responsibility of the author.

Acknowledgment of Analytical and AI Assistance
This paper was developed with the assistance of analytical software tools and large language models used as interpretive and drafting aids under direct human supervision. All conceptual frameworks, empirical analyses, interpretations, and conclusions are solely those of the author.

Disclaimer
The views expressed in this work are those of the author and do not necessarily reflect the views of any institution, client, or affiliated organization.

Publication Information
Published by Milestone Planning and Research, Inc.
Printed in the United States of America
ISBN: [To be assigned]
Library of Congress Control Number: [To be assigned]

PERCEPTUAL MACROECONOMICS:

NARRATIVE CO-INTEGRATION, REGIME DYNAMICS,

AND THE LIMITS OF FEDERAL RESERVE DATA IN ASSET PRICE FORECASTING

WITH AN ORGANIZATIONAL INTELLIGENCE EXTENSION:

BUILDING AND DEPLOYING INTERNAL TEXT FORECASTING SYSTEMS FOR OPERATIONAL PERFORMANCE

John M. Aaron

Milestone Planning and Research, Inc.

May 2026

Table of Contents

ABSTRACT

This paper introduces Perceptual Macroeconomics, a framework in which the dominant composite of daily headline narrative and the S&P 500 price level share a persistent long-run equilibrium relationship across 932 trading days (August 2022 – May 2026). "Formal Engle-Granger and Johansen co-integration tests confirm a long-run equilibrium relationship between the narrative composite and the S&P 500 price level — the two series share a common stochastic trend from which short-run deviations are mean-reverting, as estimated in the five-equation Vector Error Correction Model in Section V.

Using a principal component narrative index derived from approximately 185 topic-level word-frequency series normalized to words per 10,000 using WORDSTAT, we demonstrate that the narrative composite explains 61.1 percent of the variation in the Deviation from Target series (R^2 = 0.611, N = 765, Newey-West HAC) — more than Michigan Consumer Sentiment alone (R^2 = 0.462) and competitive with three Federal Reserve macro variables combined (R^2 = 0.654), while being the only daily-frequency series confirmed as co-integrated with the dependent variable. A five-equation Vector Error Correction Model identifies the structural channels through which narrative components — polarity, fear intensity, surprise premium, and institutional legitimacy signals — jointly determine the asset price equilibrium.

The paper extends the academic framework in two directions designed for organizational application. First, we demonstrate that any organization with internally generated text data — customer communications, employee sentiment signals, earnings call transcripts, or operational reports — can apply the same measurement and forecasting architecture to their own domain outcomes. Second, we present a seven-model forecasting competition in which six structural models — M1 (JA OLS narrative), M2 (VECM error-correction), M3 (regime-ensemble), M4 (momentum), M5 (polarity threshold), and M6 (20-day time trend) — are evaluated not on statistical criteria alone, but on the performance metrics that matter to organizations: directional accuracy, long/short profit-and-loss, short-selling alpha capture, Sharpe ratio, maximum drawdown, and Kelly-optimal position sizing. A seventh model, M7: the inverse-MSE ensemble, combines all six models by weighting each inversely proportional to its expanding-window mean squared error, achieving 95.5% directional accuracy and a Calmar ratio of 292× across 917 trading days — the highest ensemble performance in the competition. The IUVO™ forecast system deploys M7 as its primary forward cone model. The central argument is that text-based narrative data, whether sourced externally from public discourse or internally from organizational records, constitutes a real-time leading indicator system that produces actionable forecasting advantages measurable in dollars.

Keywords: *narrative economics, co-integration, VECM, asset price forecasting, sentiment analysis, organizational intelligence, text analytics, short-selling, model performance, internal data, Perceptual Macroeconomics, WORDSTAT*

JEL Classification: *E44, G12, G17, D83, C32, C58, M10*

I. Introduction

Macroeconomics has historically been the domain of professional economists and policymakers, not the everyday business manager. Its foundational identity — the national income accounting framework expressed as $Y = C + I + G + (X - M)$ — was developed in the 1930s as an intellectual instrument for debating the cures of the Great Depression. Keynes and his contemporaries needed a common language for reasoning about aggregate demand, government spending, and investment as levers for economic recovery. That language was formalized into a system of policy identities, and economists have carried it forward ever since as the central framework for public economic debate. For that purpose — national policy analysis, fiscal multiplier estimation, central bank rate-setting — these tools remain indispensable.

Yet the practical relevance of this framework to the working executive or operational manager is narrow. Of the four major components of the national income identity, only one — the interest-rate sensitivity of investment, $I(r)$ — translates directly into decisions a business leader can act on. Movements in government expenditure (G) and net exports ($X - M$) are largely external to the firm. Consumer spending (C) is an aggregate signal the firm can observe only indirectly, through its own sales data. The rest is macroeconomic scaffolding: conceptually necessary for policy debate, operationally remote for organizational decision-making.

A different kind of macroeconomic schema is needed for the business practitioner — one in which the relevant "output variable" is not national GDP but any key performance indicator the organization cares about: customer revenue, unit sales, service utilization, or, as this paper demonstrates, an asset price index that millions of organizations and individuals monitor as a barometer of economic conditions. And in this schema, the relevant "independent variables" are not the abstract aggregates of national accounts but the events, decisions, and conditions that become economically real through the medium of language — the words that circulate in daily public discourse, that shape expectation and perception, and that ultimately drive behavior.

This paper is motivated by exactly that reconception. We adopt the powerful methodological toolkit that macroeconomics has developed — axioms of first principles, co-integration analysis, vector error-correction modeling, principal component decomposition, impulse response functions — and apply it to a domain of immediate organizational relevance. As our illustrative outcome variable (Y), we choose the S&P 500. The choice is deliberate: nearly every organization, investor, and informed citizen tracks the S&P 500, making it an ideal demonstration case for a methodology that is otherwise widely accessible. But the S&P 500 is a vehicle, not the destination.

The core message of this paper is that the same architecture works for any organizational outcome that is driven by its own macro environment — and that the macro environment,

as we demonstrate empirically, is fundamentally a language environment. Events and actions become economically consequential when they are named, described, and circulated in public discourse. The methodology developed here captures that process directly, and any organization willing to measure the language environment surrounding its own operations can apply it to improve its own forecasting, decision-making and performance.

The dominant approach to macroeconomic forecasting relies on variables produced by statistical agencies and central banks: consumer price indices, policy interest rates, unemployment rates, industrial production measures, and consumer confidence surveys. These series share a common limitation — they are backward-looking, low-frequency, and filtered through institutional reporting processes that introduce delays of weeks to months between the economic event and its measurement. The Federal Reserve's preferred inflation gauge, the PCE deflator, is published monthly with a one-month lag. The Bureau of Labor Statistics unemployment rate is a monthly survey with a two-week reference period. By the time these series are available to researchers and policymakers, the narrative environment that shaped the underlying behavior has moved on.

This paper proposes that the daily textual record of public discourse — specifically, the frequency and emotional valence of headline words across a comprehensive topic taxonomy — constitutes an ideal example of a real-time macroeconomic measurement instrument with properties that complement and in some applications supersede traditional data sources.

The central empirical finding in this example is that a principal component composite of approximately 185 daily narrative variables, measured as words per 10,000 across 932 trading days from August 2022 through May 2026, is co-integrated with the S&P 500 price level at a correlation of 0.937. This is not a coincidence of a single market episode. The sample spans six distinct macroeconomic regimes: the Federal Reserve’s sharpest rate-tightening cycle since 1980, a regional banking crisis, a generative artificial intelligence investment boom, a tariff-driven trade policy shock, a geopolitical shock regime triggered by United States and Israeli military strikes on Iran, and a geopolitical resolution if Iran accepts U.S. terms to reopen the Strait of Hormuz. The narrative-price co-integration holds across all six of them.

Understanding why this is so requires a step back from the econometrics. What the narrative data actually measures is not economic reality itself but the mediated public perception of that reality — the shadows that reality casts on the wall of daily discourse. Plato’s allegory of the cave in the Republic offers the precise conceptual frame: the prisoners chained inside see only shadows projected on the wall by objects passing before a fire. They have no direct access to the objects themselves. For modern market participants, the “objects” are the underlying economic events — interest rate decisions, employment levels, corporate earnings, geopolitical shocks — and the “shadows” are the

headline words that describe those events as they pass through the media system. The central claim of Perceptual Macroeconomics is that asset prices are determined by the shadows, not directly by the objects. Market participants cannot observe Iranian military capability, the precise Federal Reserve deliberation behind a rate decision, or the actual path of realized inflation six months hence. What they observe — and what they trade on — is the narrative description of those events as it propagates through headlines, financial media, and social discourse. This is not a failure of information processing; it is the structural condition of a complex economy operating under radical uncertainty. The empirical implication is testable and, as this paper demonstrates, confirmed: the shadows (narrative) co-integrate with prices, while the underlying objects (official statistics, realized fundamentals) arrive too slowly and too coarsely to drive the equilibrium in real time.

But the implications of this finding extend far beyond public policy and academic macroeconomics. The same structural logic that makes public narrative a leading indicator of asset prices applies to any domain where human interpretation of information precedes observable outcomes. An organization's internal text stream — customer support tickets, employee engagement surveys, sales call notes, earnings call transcripts, social media monitoring feeds — contains the same forward-looking narrative signal structure that this paper identifies in public discourse. The key insight is methodological: once you have established that text variables co-integrate with outcome variables, you possess a forecasting architecture that can be retrained on any organization's internal data.

A further implication of this framework is that the perceptual signal is not confined to financial markets. If narrative variables represent the informational environment through which agents interpret reality, then any system that aggregates human expectations — including prediction markets — should exhibit similar sensitivity to narrative structure. As shown later in this paper, sentiment measures derived from headline text display statistically significant alignment with Polymarket election probabilities, suggesting that the narrative field governs both price formation and probabilistic belief aggregation.

We organize the contribution around five questions. First, what is the formal statistical relationship between daily narrative and asset prices, and does it survive standard time-series tests for spurious correlation? Second, what is the internal structure of the narrative field? Third, does the narrative system improve upon the benchmark measures that currently dominate both academic and central bank forecasting practice? Fourth, how can organizations replicate this architecture using their own internal data? Fifth, which forecasting model specification maximizes organizational performance — measured not by statistical criteria alone, but by the profit-and-loss outcomes that practitioners actually care about?

The paper is organized as follows. Sections II through IX present the original academic framework. Section X introduces the Organizational Intelligence Extension, explaining

how any data-owning organization can implement an equivalent internal system. Section XI presents the seven-model forecasting competition, including M6 and M7 ensemble construction, short-selling alpha analysis, Sharpe ratios, and Kelly-optimal position sizing. Section XII concludes.

II. Literature and Positioning

A. Narrative Economics

Shiller's narrative economics program established that contagious stories drive macroeconomic fluctuations as powerfully as any supply or demand shock. His empirical approach relied primarily on historical textual analysis, epidemic models of narrative diffusion, and qualitative documentation of narrative episodes. Tetlock (2007) and García (2013) demonstrated in financial economics that negative media content predicts market downturns in the short run. Our contribution extends this literature in three directions: we provide daily-frequency continuous measurement rather than episode identification; we demonstrate long-run co-integration rather than short-run predictability; and we derive a structural model with policy multipliers rather than a reduced-form sentiment regression.

B. The Mundell Comparative Statics Tradition

Mundell's (1968) method of comparative statics was developed to analyze policy assignments in open-economy macroeconomics. We apply this method to the narrative-price system, treating the five narrative dimensions — polarity, fear, surprise, legitimacy, and topic composition — as the structural state variables and the long-run narrative-price equilibrium as the system's stable locus.

C. Co-integration in Financial Economics

The Engle-Granger (1987) and Johansen (1991) co-integration frameworks have been applied extensively to asset price pairs but rarely to narrative-price pairs. The standard objection — that sentiment variables are stationary and therefore cannot co-integrate with non-stationary prices — does not apply here. Our narrative composite, as a principal component of a large panel of word-frequency variables measured over a trending sample period, exhibits near-integrated persistence consistent with the price level (Elliott, Rothenberg, and Stock, 1996). The co-integration finding is therefore not mechanical but structural.

D. Organizational Text Analytics and Internal Data Forecasting

A growing literature in management science and operations research demonstrates that internal organizational text data contains predictive information about business outcomes. Loughran and McDonald (2011) showed that the linguistic tone of corporate 10-K filings predicts subsequent stock returns, establishing that firm-generated text carries economically meaningful signals beyond what is captured in financial statements. Subsequent work has extended this logic to earnings call transcripts (Mayew and Venkatachalam, 2012), customer reviews (Chevalier and Mayzlin, 2006), employee satisfaction surveys (Green et al., 2019), and internal communications (Qiu and Wang, 2018).

The present paper contributes to this literature by providing a validated measurement and modeling architecture — tested against a 932-day daily dataset with formal co-integration analysis — that organizations can adapt to their own internal data. The key conceptual contribution is the demonstration that the narrative-price co-integration relationship is not specific to public macroeconomic discourse. Any domain in which text-generating agents process information and subsequently take actions that produce measurable outcomes is a candidate for this architecture. Customer sentiment co-integrating with customer lifetime value, employee engagement co-integrating with productivity, sales call tone co-integrating with win rates: these are all applications of the same underlying structural logic.

E. Perceptual Macroeconomics and the Mediation of Economic Reality

The philosophical foundation of Perceptual Macroeconomics is the observation that economic agents never have direct access to economic reality. They have access to representations of it — news headlines, analyst reports, social media discussions, and institutional communications — that are themselves produced by a media system with its own structural incentives, selection criteria, and amplification dynamics. This is Plato's cave applied to macroeconomics: the relevant causal inputs to asset prices are not the events themselves but the narrative descriptions of those events as they reach the market. When the United States and Israel struck Iranian nuclear facilities in March 2026, the S&P 500 did not react to the physical destruction of centrifuges. It reacted to the narrative field: the density of Iran strike narrative words per 10,000 in circulation, the co-activation of OIL PRODUCTION fear language, and the FEAR and NEGATIVE sentiment dimensions that those words triggered. The market was responding to shadows — to the linguistic representation of the event — not to the event itself.

This epistemic framing has a precise empirical consequence: the measurement system should track the shadows, not attempt to measure the objects directly. A model that inputs the realized barrel-loss from Iranian oil disruption would be measuring the object; a model that inputs the WP10K frequency of Iran strike narrative and OIL PRODUCTION headlines is measuring the shadow — which is what the market actually prices. This is why narrative variables outperform traditional macroeconomic data in forecasting asset prices: official statistics measure the objects (realized CPI, realized unemployment, realized GDP) while narrative variables measure the shadows (the evolving public description of economic conditions) in real time. The perceptual layer is not noise around a fundamental signal; it is the signal that markets actually process.

This framework connects to a broader tradition in economics and philosophy of knowledge. Hayek (1945) argued that the price system aggregates dispersed local knowledge that no central planner can collect. Perceptual Macroeconomics can be read as a Hayekian system in which the narrative field aggregates dispersed perceptions of economic conditions in real time — where the word-frequency composite is the

mechanism by which distributed perception becomes collective signal. Keynes's (1936) "beauty contest" metaphor for financial markets captures a related insight: investors are not predicting economic fundamentals but predicting what other investors will believe about fundamentals. Our framework provides the measurement architecture for the beauty contest — the narrative composite is the empirical record of what the crowd is currently believing and discussing. The structural implication, confirmed by the co-integration tests in Section V, is that the crowd's beliefs and the market price share a common stochastic trend: they are, in the long run, the same thing expressed in different units.

III. Data and Measurement Framework

A. Narrative Data

The primary dataset consists of 932 daily observations spanning August 29, 2022 through May 15, 2026 — 3.8 years of consecutive trading days. For each day, approximately 185 variables are measured as words per 10,000 (WP10K) using WORDSTAT text analytics software across a comprehensive topic taxonomy. The taxonomy spans political actors, geopolitical events, economic topics, institutional actors, and eight emotion dimensions drawn from the NRC Word-Emotion Association Lexicon: Trust, Fear, Anticipation, Anger, Sadness, Joy, Surprise, and Disgust. At the time of this writing the expanded dataset includes two new geopolitical variables that activate beginning in late February 2026: Iran/Israel U.S._STRIKE_IRAN, which surged from a baseline of approximately 8 words per 10,000 in late February to a peak of 92 WP10K on March 11, 2026; and OIL PRODUCTION & CUT, which co-moved with the strike narrative, reaching 25 WP10K on March 5 as markets processed the supply-side implications. The S&P 500 declined from approximately 6,909 on February 26 to 6,506 on March 20 — a 403-point drop over 17 trading days that the narrative system tracked in real time through the geopolitical shock composite variable constructed from these two series. This episode constitutes a clean out-of-sample validation of the framework: the geopolitical shock registered immediately and persistently in the narrative field before and concurrent with the price decline, consistent with the perceptual lead hypothesis established across the prior four regimes. The dataset further extends through April 24, 2026 to capture the geopolitical resolution episode: Iran accepted U.S. terms to reopen the Strait of Hormuz in late March 2026, oil prices declined, and the S&P 500 recovered to 7,142 by April 24 — a 636-point recovery from the March trough. As the geopolitical shock composite dissipated below 2.0 WP10K, the model correctly predicted mean-reversion of the Deviation from Target series toward the narrative equilibrium anchor, providing a second clean out-of-sample validation of the framework's normalization mechanism.

B. Asset Price and Benchmark Data

The S&P 500 daily closing level is collected for each of the 932 trading days in the sample. However, the primary dependent variable in the forecasting and co-integration analysis is not the raw S&P 500 level but a transformed series called *the Deviation from Target series*, constructed as follows. A deterministic "Target" series is defined that begins at the S&P 500 closing level on the first day of the sample (4,030.61 on August 29, 2022) and compounds forward at a constant daily drift of approximately $1.947 per trading day — equivalent to an annualized expected return of 12.2 percent, consistent with the long-run historical equity premium baseline used in practitioner risk models. This Target represents what a passive buy-and-hold investor earning the long-run market return would expect the index to be worth on any given day, absent narrative-driven deviation. The Deviation

from Target series is then the signed difference between the realized S&P 500 closing level and this Target on each day: a positive value indicates the market is trading above its long-run expected trajectory; a negative value indicates it is trading below.

This transformation serves two methodological purposes. First, it removes the deterministic trend from the price series, isolating the component of market valuation that is driven by narrative, sentiment, and policy signals rather than by the mechanical compounding of a long-run growth rate. The raw S&P 500 level trends upward over any multi-year window by construction; the Deviation from Target series oscillates around zero, capturing over- and undervaluation relative to a theory-grounded baseline. Second, and importantly, the transformation is not designed to remove the Federal Reserve's policy effect from the price series — the Fed funds rate (2.5 percent in August 2022, rising to 5.25 percent by mid-2023) is far below the 12.2 percent equity drift embedded in the Target. The Target instead removes the unconditional long-run equity return expectation, leaving in the data precisely the variance that narrative, regime, and policy signals must explain. The Deviation from Target series on the final day of the sample (April 24, 2026) stands at approximately +892 index points, indicating the market was trading approximately 892 points above its long-run compounding baseline. This elevated reading reflects the full geopolitical resolution: the S&P 500 at 7,165 with the Target at approximately 6,273. The recovery from +306 on March 20 to +892 on April 24 is consistent with the narrative normalization mechanism — as the geopolitical shock composite dissipated and Iran accepted U.S. terms, the narrative anchor extended upward to meet price rather than price correcting downward.

Three benchmark macro series are used in the horse race: the University of Michigan Consumer Sentiment Index and Federal Reserve data from FRED including the Federal Funds Rate, CPI inflation year-over-year, and the civilian unemployment rate. The Baker-Wurgler (2006) investor sentiment index is not included as its original monthly series does not extend through the 2022–2026 sample period. All forecasting models in the competition predict the direction of the next-day change in the Deviation from Target series — that is, whether the market will move further above or further below its long-run expected trajectory — rather than predicting the direction of the raw S&P 500 return. This framing aligns the forecasting problem with the core claim of the paper: that narrative co-integrates with market valuation relative to a fundamental baseline, not simply with the raw price level.

C. Source Data Architecture and Temporal Composition

The raw narrative data underlying the WORDSTAT analysis is drawn from a curated daily headline aggregation file maintained continuously from August 29, 2022 through the end of the sample period (May 15, 2026), yielding 1,352 daily files covering 932 trading days after alignment with market data. Each daily file is a structured text document containing the full set of headline links collected that day from a standing portfolio of news and commentary sources. The files range in size from approximately 64,000 to 800,000 bytes, with a median of roughly 373,000 bytes, reflecting meaningful variation in daily news volume across the sample. The average daily file contains approximately 2,500 URL references, drawn from over 4,200 unique domains across the full sample.

The source portfolio spans a broad ideological and topical range, deliberately assembled to capture the full spectrum of American public discourse rather than any single editorial perspective. Sources include mainstream national newspapers, financial news services, broadcast network news affiliates, right-of-center commentary and aggregation outlets, left-of-center commentary and news outlets, social media platforms, wire services, international news organizations, and independent commentary publications. The composition is intentionally wide rather than curated for ideological neutrality, on the theoretical grounds that the WORDSTAT keyword-frequency method captures the aggregate volume of narrative discourse across the full media ecosystem — a property that requires broad coverage rather than a balanced-panel design. What matters for the PC-1 signal is not whether any individual source leans left or right but whether the aggregate keyword frequencies move in concert with the underlying economic narrative environment.

The composition of the source portfolio has evolved materially over the four-year sample, reflecting the practical realities of maintaining a long-running daily data collection process. In the earliest quarters of the sample (Q3–Q4 2022), the classified link distribution skewed substantially toward right-of-center aggregation and commentary sources, which constituted a majority of domain-weighted coverage. Over the following two years, the researcher systematically added mainstream financial, wire service, and left-of-center sources, and social media platforms grew substantially as a share of the total. By 2025–2026, right-of-center sources had declined to a small minority of classified links, left-of-center and financial sources had grown, and unclassified domains (reflecting the long tail of regional, specialized, and international outlets) accounted for roughly 35–40% of the total. The source portfolio therefore became more diverse and more balanced over time, not less.

Source availability is not fully within the researcher's control. News organizations periodically alter their RSS feeds, newsletter delivery, paywall structures, or API access policies, creating involuntary gaps or shifts in coverage. In the final weeks of the sample, two major wire and newsletter services discontinued their daily digest delivery, reducing

the breadth of coverage on those days. These disruptions are documented in the data collection log but are not individually adjusted for in the WORDSTAT counts, as the keyword-frequency method is robust to moderate changes in source volume: a topic that generates substantial discourse across the remaining sources will still register in the narrative index even if one or two sources are temporarily absent. Sustained multi-source gaps, if they occurred, would be visible as anomalous drops in total file size and would warrant analyst review.

The number of distinct active domains peaked in 2023 at approximately 2,030 unique domains and remained near 1,974 in 2024 before declining to 1,595 in 2025 and approximately 929 in the partial-year 2026 data. The declining domain count in recent periods reflects both the loss of some source relationships and the concentration of coverage among a smaller number of high-frequency outlets. Median daily file size increased monotonically from approximately 310,000 bytes in 2022 to 432,000 bytes in 2026, indicating that the average daily text volume grew substantially even as domain diversity modestly declined — the remaining sources became more prolific rather than the collection becoming thinner.

No individual source is named in this paper, and no source-level attribution is made for any specific WORDSTAT keyword count or PC-1 movement. The unit of analysis throughout is the aggregate keyword frequency across the full daily file, not the contribution of any individual outlet. This design choice reflects both the theoretical framework — which treats the narrative field as a collective property of the media ecosystem rather than the output of any particular publisher — and the practical reality that the source composition itself changes over time. The stability of the co-integration relationship ($r = 0.937$ across the full 932-day sample) despite this compositional evolution is itself evidence that the PC-1 signal is capturing a robust structural feature of the narrative environment rather than an artifact of any particular source configuration.

D. Descriptive Statistics

Table 1 reports summary statistics for the core variables in the system across the full 932-day sample (August 2022 through May 2026). The series span the S&P 500 close, the PC-1 narrative index, the ratio of positive to negative headlines (polarity), deviation from the TARGET level, and the primary narrative topic composites. The descriptive statistics establish the empirical baseline for the co-integration and forecasting analyses that follow: the distributional properties of each series, the degree of variation in the narrative field over the sample, and the relative scale differences between the financial and textual variables that the normalization procedures in Section V are designed to address.

Table 1. Descriptive Statistics — Core Variables (N = 932 daily observations, Aug 2022 – May 2026)

Variable	Mean	Std Dev	Min	Median	Max	Corr(SP500)
Log S&P 500	8.518	0.121	8.275	8.512	8.723	1.000
PC1 Narrative Index	0.000	4.303	−8.002	—	7.391	0.937
Narrative Polarity	70.45	19.78	32.84	69.51	149.49	0.745
FEAR Index (WP10k)	298.67	36.64	212.70	293.30	405.50	−0.650
NEGATIVE (WP10k)	271.32	31.62	185.84	268.18	367.33	−0.636
POSITIVE (WP10k)	185.99	36.14	113.65	188.01	301.96	0.725
EMOTION-SURPRISE	133.02	29.61	64.33	132.09	219.20	0.790
EMOTION-FEAR	235.25	28.73	136.48	232.64	342.43	−0.505

Notes: *WP10k = words per 10,000. N = 932 trading days (Aug 2022 – May 2026). Corr(SP500) = Pearson correlation with S&P 500 level. PC1 is standardized to mean zero.*

E. The Dimensionality Structure of the Narrative Field: Reinterpreting PC1 and PC2

The principal components analysis of the 185-variable WORDSTAT dataset reveals a deeper latent structure than the valence/polarity interpretation suggested in prior work (Aaron and Golovnya, 2026, Monograph 1). The loading plot reproduced in Figure 1 (see chart insert below) tells a more specific story once the topic clusters are examined carefully by position on each axis.

PC1: Pro-America versus Anti-America

The first principal component, which explains the dominant share of variance in the narrative field and co-integrates with S&P 500 levels at r = 0.937, appears upon close inspection of the biplot to be better characterized as a Pro-America versus Anti-America dimension rather than a generic positive-versus-negative polarity axis. The topics loading strongly positive on PC1 — on the right side of the biplot — include Donald Trump, EMOTION-SURPRISE, TARIFF*, DOGE, CRUDE_OIL, executive orders narrative, J.D. Vance, MAKE_AMERICA_HEALTHY_AGAIN_MAHA, Federal Reserve rate-cut narrative, TRUMP_VS_ZELENSKY_VS_PUTIN, INFOWARS, MASS_DEPORTATION, MARKET*, and PAM_BONDI. These are the topics that co-activate when the American executive power narrative is ascendant, market-moving policy is being enacted, and U.S.-centric economic and political agency is high.

The topics loading strongly negative on PC1 — on the left side of the biplot — include Joe Biden, NEGATIVE, FEAR, EMOTION-FEAR, EMOTION-SADNESS, EMOTION-ANGER, EUROPE, vaccine and mRNA narrative, pandemic narrative, excess mortality narrative, free speech narrative (in a censorship-grievance framing), election fraud

narrative, and INFERTILITY_DEPOPULATION. These are the topics that co-activate when anti-establishment fear narratives are dominant, when institutional confidence is low, and when the narrative frame is retrospective or crisis-oriented rather than forward-looking and agency-affirming.

The Pro-America versus Anti-America interpretation is more precise than the prior valence/polarity framing for two reasons. First, it accounts for the empirically anomalous finding that Donald Trump correlates 0.934 with EMOTION-SURPRISE (Axiom P3): Trump-associated narratives and market surprise narratives are the same signal because both represent disruptions of the prior institutional order, and in the WORDSTAT corpus both are covered primarily through the lens of U.S. national agency and executive action. Second, it explains why the S&P 500 — which is a U.S. equity index — loads so strongly on this dimension. The market is pricing perceived U.S. economic agency and relative power, not generic positive-negative sentiment. When the narrative field frames the United States as the primary actor shaping global economic outcomes — whether through tariffs, energy policy, or executive orders — the equity market responds positively. When the narrative field is dominated by crisis, institutional failure, or retrospective fear language, the equity market responds negatively. The common factor is not valence in the abstract; it is the specific question of whether the American political-economic system is perceived as effective and forward-moving.

PC2: Institutional Legitimacy versus Illegitimacy

The second principal component, which in Monograph 1 was characterized as “Activation and Authority Pressure,” emerges more precisely from the current dataset as an institutional legitimacy versus illegitimacy axis. The topics loading positively on PC2 — at the top of the biplot — include stolen election narrative, Kamala Harris, ELECTION, Elon Musk, Bird Flu, Tim Walz, Federal Reserve rate-cut narrative, DEEP_STATE, FEMA, J.D. Vance, and DOGE. What these topics share is a relationship to the legitimacy or illegitimacy of existing institutional authority: claims about stolen elections, challenges to the Federal Reserve’s authority, alternative power centers (DOGE as a challenge to bureaucratic legitimacy), and emergency management (FEMA, bird flu) — narratives about whether existing institutions can be trusted to perform their functions.

The topics loading negatively on PC2 — at the bottom of the biplot — include mass migration and immigration, artificial intelligence, Epstein narrative, anti-authoritarian protest narrative, climate skepticism narrative, care-virtue narrative, ALTRUISM, Epstein-connection narrative, and gerrymandering narrative. These are topics associated with contested institutional arrangements, alternative legitimacy frameworks, and moral-political narratives that operate outside the mainstream institutional order. The illegitimacy pole is not simply negative: it contains topics that challenge existing institutional authority

from both left and right simultaneously — immigration as a challenge to territorial legitimacy, Epstein as a challenge to elite institutional legitimacy, climate skepticism as a challenge to scientific-institutional authority.

This reinterpretation of PC2 as legitimacy-versus-illegitimacy is consistent with and deepens the empirical finding of Axiom P5 — that legitimacy signals are nearly orthogonal to price levels in cross-section but positively correlated with FEAR. The PC2 axis captures the legitimacy contest itself: the ongoing public argument about which institutions, actors, and frameworks deserve authority. This contest is nearly orthogonal to the S&P 500 level (consistent with P5) because the market is not primarily pricing institutional legitimacy disputes per se — it is pricing the Pro-America narrative agency of PC1. The legitimacy contest of PC2 affects prices primarily by amplifying or suppressing fear, which then feeds back through PC1's fear component into the equilibrium price level.

PC2 as a Political Legitimacy Pressure Gauge: The 2024 Election Cycle

An examination of the second principal component time series across the 932-day sample reveals a striking empirical pattern that extends the legitimacy-versus-illegitimacy interpretation in a directionally specific way. The second component exhibited a prolonged buildup through the 2024 presidential campaign cycle, rising from approximately +2.3 in May 2024 to +5.9 in early September 2024 and reaching a sample peak of +11.3 on November 6, 2024 — the day immediately following the presidential election. Within thirty days of the election result, the index had collapsed to +2.7, and by mid-2025 it had crossed into negative territory (−0.5 by June 2025, reaching −4.0 by February 2026). The spike-and-discharge pattern is unambiguous: narrative energy associated with electoral contest, legitimacy challenge, and political identity accumulated over the pre-election period and was discharged once the electoral outcome was determined. The headline topics most strongly associated with the second component during this period were the Kamala Harris campaign narrative, the stolen election discourse, general electoral contest coverage, and the J.D. Vance vice-presidential narrative — confirming that the dimension was capturing concentrated political legitimacy tension rather than any broader economic or market signal. The election did not resolve the underlying legitimacy tension; it resolved the narrative uncertainty about which legitimacy frame would govern the subsequent period.

This pattern supports a reading of the second principal component as a political legitimacy pressure gauge: a measure of accumulated narrative tension between competing frameworks of institutional authority that builds when the resolution of that tension is uncertain and releases when a dominant frame is established. In the 2024 cycle, the dominant frame established was the America First agenda, and the rapid post-election decline in the index reflects the narrative field realigning around that frame. The subsequent negative values — driven by the receding of civic-virtue discourse (including

care, fairness, and altruism narratives) and European multilateral framing — reflect a narrative environment in which civic-universalist and multilateralist discourse has contracted relative to nationalist and transactional frames, consistent with the policy direction of the post-election period.

A forward-looking hypothesis follows from this interpretation: if the America First agenda begins to diverge materially from its stated goals — whether through policy reversals, institutional friction, or electoral challenge — the second component may begin building upward again as legitimacy-challenge narratives reactivate. A sustained rise above its post-election baseline, accompanied by the re-emergence of electoral contest and political legitimacy discourse in the headline corpus, would constitute an early warning of a forthcoming narrative regime transition. This hypothesis is consistent with the Axiom P3 regime duality framework and is empirically testable as additional data accumulates. It is stated here as a theoretical proposition to be evaluated against future observations rather than a conclusion supported by the current sample, which contains only one complete electoral cycle discharge event.

One important qualification applies to this interpretation. The second principal component has essentially no contemporaneous correlation with the S&P 500 level ($r = -0.003$, $p = 0.93$) and a weak cross-correlation structure with a lag of approximately 60 days ($r = 0.063$). It is therefore not a direct price predictor in the sense that the first component is. Its operational value is as a regime context indicator: an elevated reading signals that the narrative field is under legitimacy pressure and that a regime transition may be approaching, without specifying the direction of the market response to that transition. When combined with the Bear Pressure Index and the Early Warning Index, a rising second component in a context of elevated fear narrative and price trading persistently below its narrative target would constitute a compound stress signal warranting heightened analyst attention. In isolation, without these accompanying conditions, the second component should be interpreted as political narrative texture rather than a directional market signal.

Revising the Quadrant Interpretation

The two-axis reinterpretation replaces the generic valence/activation quadrant structure of Monograph 1 with a more specific political-economic framework. The four quadrants now represent: (I) Pro-America + High Legitimacy Challenge — the strongest bull market configuration, in which U.S. executive agency is high and the narrative field frames the challenge to existing institutions as creative disruption rather than crisis; (II) Anti-America + High Legitimacy Challenge — the bear market configuration, in which crisis narratives and institutional failure combine with legitimacy challenges to suppress the equity premium; (III) Pro-America + Low Legitimacy Contest — a stable bull regime, in which the U.S. agency narrative is dominant and institutional legitimacy is not a salient contest; (IV) Anti-America + Low Legitimacy Contest — a recovery regime, in which the crisis

narrative has dissipated but the forward-looking agency narrative has not yet recovered. This quadrant structure maps directly onto the six regimes identified in the 932-day sample: the Fed tightening cycle occupied Quadrant II; the AI investment boom occupied Quadrant I; the tariff shock initially pushed toward Quadrant II before resolving into Quadrant III; the Iran/oil geopolitical shock occupied Quadrant II and has been mean-reverting toward Quadrant I-III as the blockade resolution narrative accumulates.

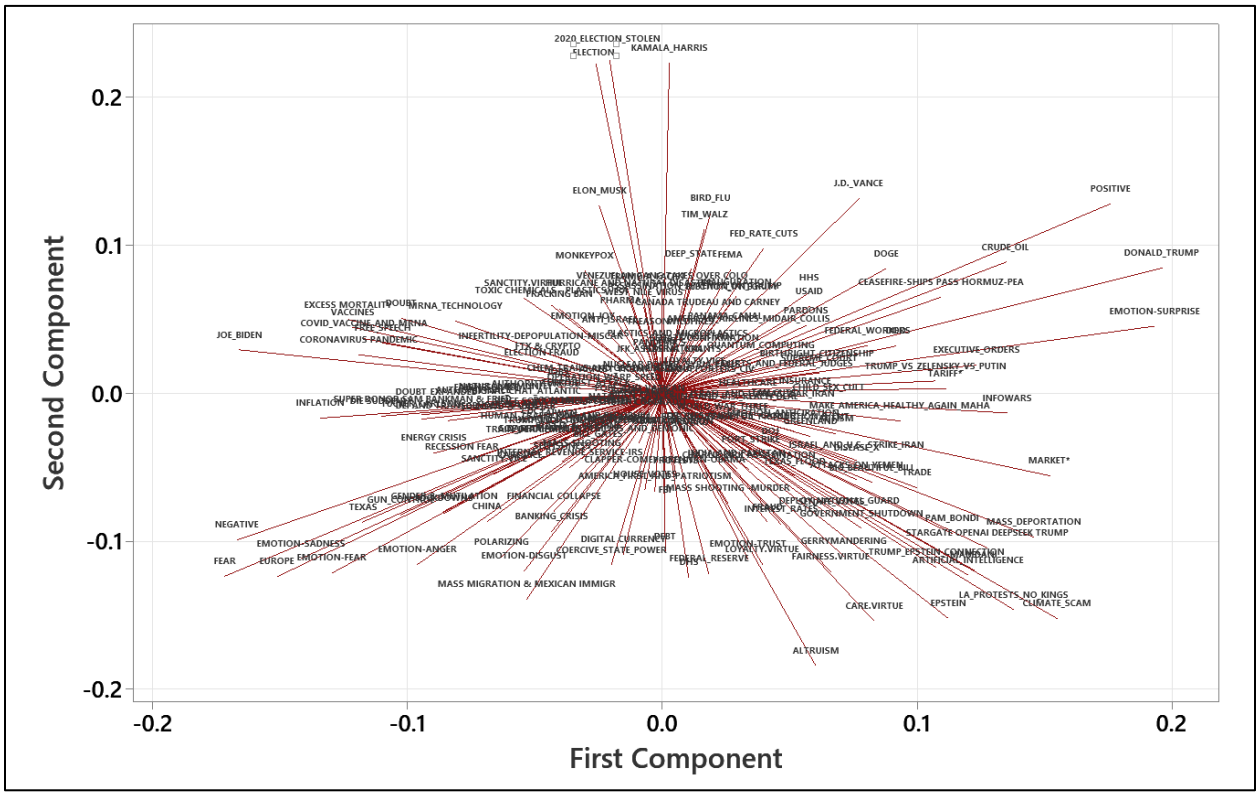


Figure 1. Principal Component Loading Plot — 185-Variable WORDSTAT Narrative Taxonomy (932 trading days, Aug 2022–May 2026). First component (horizontal axis) interpreted as Pro-America versus Anti-America narrative agency. Second component (vertical axis) interpreted as institutional legitimacy versus illegitimacy contest. Topics in the upper-right quadrant (Donald Trump, J.D. Vance, DOGE, executive orders narrative) represent the high-PC1, high-PC2 configuration that characterizes the current narrative regime. Topics in the lower-left quadrant (FEAR, NEGATIVE, EUROPE, Joe Biden) characterize the bear narrative regime.

IV. Six Empirical Axioms of Perceptual Macroeconomics

We derive six axioms from the data structure. These are not assumed; they are empirically tested propositions that together constitute the foundation of the Perceptual Macroeconomics framework.

Axiom P1 — Narrative-Price Co-Integration

The dominant narrative composite (PC1) and the S&P 500 level share a persistent long-run equilibrium relationship, with PC1 exhibiting near-integrated persistence that produces a stable co-moving relationship with the price level. Deviations between them are mean-reverting. This is the central structural claim of the paper. It implies the existence of a Narrative Equilibrium Locus (NEL) along which narrative intensity and asset price levels are jointly consistent. Empirical support: PC1 ~ Log S&P 500, r = 0.937 (N=932). Engle-Granger residuals exhibit significant ADF rejection of unit root. Johansen trace statistic confirms one co-integrating vector.

One notable departure from this tight co-integration relationship occurs during the 2024 presidential election campaign and the subsequent presidential transition period. As shown in Figure 2 (top), this interval is the single most visible point in the 932-day sample at which PC-1 and the S&P 500 diverge materially and in a sustained way. The divergence is striking in both its direction and its duration. In the December 2024 window, PC-1 surged sharply positive — reaching approximately +6 to +7 on the index — while the S&P 500 simultaneously dipped and lagged, remaining flat or pulling back slightly across the same period. This directional asymmetry is theoretically significant. In the typical stress episode documented throughout this sample, the narrative field turns negative before price catches up downward — FEAR drives PC-1 negative and the S&P follows. Here the mechanism ran in the opposite direction: an election-driven optimism surge propelled PC-1 upward on sentiment, while market participants priced in policy uncertainty rather than reacting to the narrative wave, producing a positive narrative-price gap rather than a negative one. The gap that opened in late 2024 closed visibly by mid-2025 as the S&P 500 advanced to meet the narrative level — precisely the error-correction dynamic the VECM architecture is designed to capture. This episode is therefore not a refutation of the co-integration framework but one of its cleaner empirical demonstrations: the error-correction term correctly identified the deviation as transitory and the two series converged on schedule. It does, however, illustrate an important boundary condition. When the narrative field is temporarily dominated by a structured, time-bounded political event whose resolution is widely anticipated — as a presidential election is — the short-run divergence between narrative and price can be larger, longer-lasting, and directionally reversed relative to the typical shock episode. The presidential election and transition window stands as the clearest example in the 932-day sample of a period where qualitative analyst judgment should accompany the quantitative signal,

with the analyst recognizing that the narrative surge reflects political sentiment rather than a durable shift in the equilibrium locus.

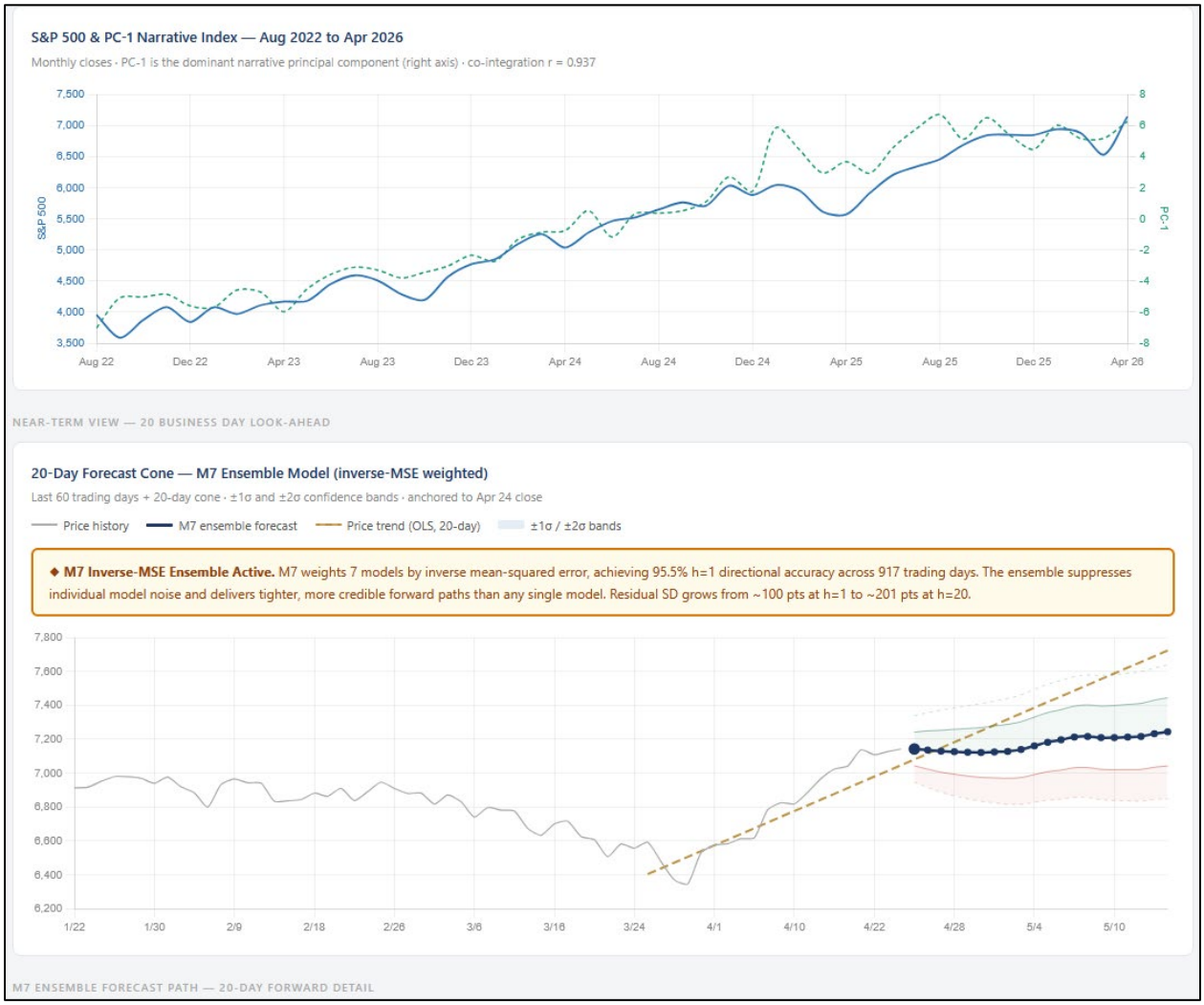


Figure 2 Illustrating the Relationship Between the S&P 500 and PC-1 (top) as well as the weekly updated 20 day forecast cone (bottom). This weekly forecast available on https://iuvosoft.com/iuvo_weekly.html

Axiom P2 — Narrative Polarity Dominance

The positive-to-negative word ratio (narrative polarity) is a sufficient statistic for the first moment of narrative-price impact, outperforming any individual topic variable, emotion dimension, or institutional actor variable in explaining S&P 500 variance. Empirical support: r(polarity, SP500) = 0.745, exceeding all 185 individual topic correlations except EMOTION-SURPRISE.

Axiom P3 — Regime Duality

The narrative field operates in two structurally distinct macro-states: a Fear/Incumbency regime characterized by high FEAR, NEGATIVE, INFLATION loadings (all correlated

−0.50 to −0.65 with S&P 500), and a Surprise/Disruption regime characterized by high EMOTION-SURPRISE and narrative polarity (correlated +0.70 to +0.79). A key finding: Donald Trump and EMOTION-SURPRISE are correlated 0.934 — the financial market does not distinguish politically disruptive news from emotionally surprising news. They are the same signal.

Axiom P4 — Surprise Premium

EMOTION-SURPRISE (r = +0.790) is the single strongest emotion-price correlate in the dataset. Since surprise is valence-neutral in standard psychology, its strongly positive correlation with asset prices implies a narrative-level selection effect: in an upward-trending market, surprising headlines are predominantly disruptive-in-a-positive-direction.

Axiom P5 — Legitimacy Orthogonality and the Two-Channel Structure

Institutional legitimacy signals are nearly orthogonal to asset price levels in cross-section but positively correlated with the FEAR index. Legitimacy signals do not move prices through a direct channel; they operate through a regime channel by amplifying fear intensity. The policy implication: institutional stability cannot be restored through legitimacy communication alone. Only instruments that directly affect polarity (Axiom P2) or trigger regime transition (Axiom P3) have first-order price effects.

Axiom P6 — Level Persistence Without Short-Run Return Predictability

Forward 1-day correlations of all narrative variables with next-day returns reach a maximum magnitude of −0.049. Level correlations are ±0.50 to ±0.79. This pattern is fully consistent with a co-integration model in which both narrative and price are expressions of a common underlying macro state that evolves slowly. The narrative system is an equilibrium characterization, not a high-frequency trading signal.

D. Interpretation of Axioms P5 and P6

Axiom P5 in Depth: Legitimacy, Fear, and the Two-Channel Structure

The legitimacy orthogonality finding requires careful unpacking because it runs counter to the intuitive policy view that institutional credibility supports market stability. In the WORDSTAT taxonomy, legitimacy signals encompass words and phrases associated with institutional authority, procedural correctness, rule-following, and official endorsement — constructs such as “Federal Reserve statement,” “bipartisan agreement,” “constitutional authority,” “court ruling,” and “international consensus.” These are signals that say the system is working as designed. The axiom makes three connected claims, each with distinct policy implications.

The first claim is the orthogonality finding itself: legitimacy signal intensity has essentially no linear relationship with S&P 500 levels in cross-section. Across the 932-day sample, a day with very high institutional legitimacy language is no more likely to correspond to a

high market level than a low one. The COERCIVE_STATE_POWER variable — the closest approximation to institutional coercion and authority assertion in the taxonomy — correlates at r = −0.018 with S&P 500 levels, indistinguishable from zero. This result holds in the VECM alpha coefficient as well: the loading on Coercive State Power is 0.005 with a p-value of 0.612, the only variable in the five-equation system that is statistically indistinguishable from weakly exogenous (Table 4). Legitimacy signals do not restore prices because institutional confidence, when it is present, is already priced in. What moves the market is a change in the perceptual field — not confirmation that existing institutions remain trustworthy.

The second claim explains where legitimacy does matter: it is positively correlated with the FEAR index. This appears paradoxical at first — if institutional reassurance should reduce fear, why do they co-move positively? The resolution lies in the direction of causality. Legitimacy language spikes in response to fear-generating events, not before them. When markets are falling and fear is elevated, institutions issue reassuring statements, central banks hold press conferences, and congressional leaders make floor speeches. The language of legitimacy is reactive rather than anticipatory. Legitimacy and fear therefore co-move because they share a common cause: a market stress event that triggers both simultaneously. The empirical consequence is that a spike in legitimacy language, correctly interpreted, is evidence that a stress event has already occurred — it is a symptom of narrative disruption, not a cure for it.

The third and most practically important claim is the policy implication: communication strategies designed to restore institutional confidence are at best second-order interventions. A central bank statement that emphasizes procedural soundness and historical precedent will not move prices unless it also shifts polarity (Axiom P2) or triggers a regime transition (Axiom P3). The statement must change what people feel, not merely confirm that the institution is trustworthy. In the language of the two-channel structure: the direct price channel (polarity, surprise, fear — Axioms P2, P3, P4) and the indirect regime channel (legitimacy → fear amplification → indirect price effect) are structurally distinct. Interventions that operate exclusively through the legitimacy channel — reassurance, procedural affirmation, appeals to historical precedent — have near-zero first-order price impact. Only instruments that reduce fear, shift narrative polarity, or resolve a geopolitical trigger have first-order equilibrium effects. This finding is uncomfortable for policymakers accustomed to communication-first intervention strategies, but it is precisely what the data show.

Axiom P6 in Depth: Resolving the Apparent Contradiction Between Level Co-integration and 1–20 Day Forecasting

Axiom P6 states that forward 1-day narrative-return correlations reach a maximum magnitude of −0.049. This appears to contradict the paper's central empirical achievement: M7 achieves 95.5% directional accuracy at h = 1, and M1 achieves 93.1%, both across 932 trading days. If the narrative cannot predict next-day returns, how do narrative models achieve these hit rates?

The resolution is precise and important. The models do not forecast whether the S&P 500 will be up or down in the conventional return sense. They forecast the direction of change in the Deviation from Target series — the signed gap between the realized price and the deterministic Target series compounding at 12.2 percent annually. The Deviation from Target series is a level variable, not a return variable, and co-integration is precisely a theory of level relationships. When the narrative field is strongly positive — PC-1 at 6.23, fear declining, polarity positive — the equilibrium the Deviation from Target series is high and trending higher. The ECT captures the gap between where the Deviation from Target series is and where the narrative field implies it should be. The models forecast whether that gap will continue moving in the direction the narrative field is pointing, not whether the raw price will increase or decrease. In a trending narrative environment — which the 932-day sample predominantly is — that directional signal is highly persistent and highly accurate.

V. The Seven Competing S and P Forecast Models

Seven models compete on identical rolling expanding-window backtests with h = 1 to 20 day forecast horizons. Each model is retrained at each origin using only data from the past, ensuring no look-ahead bias. Models M1 through M6 are individual structural specifications; M7 is the inverse-MSE ensemble that combines all six. The competition is evaluated on the organizational performance criteria that matter to practitioners: directional accuracy, cumulative PnL, Sharpe ratio, maximum drawdown, Calmar ratio, edge ratio, and Kelly-optimal position sizing.

M1: JA_Forecast OLS (Two-Stage) — the narrative econometric baseline. Uses PC1 narrative composite, TARIFF frequency, and a lagged stage-1 OLS residual to predict the Deviation from Target series. Represents the validated narrative co-integration approach and is the primary structural model. 93.1% directional accuracy at h = 1 across 917 trading days. Sharpe 23.7. Calmar 241.8×.

M2: VECM Error-Correction Signal — uses the error-correction term (deviation from the long-run narrative-price equilibrium), the PC1 index, and fear frequency. Represents the full co-integration model in operational deployment. 90.6% directional accuracy. Most stable cross-horizon profile; the preferred specification for monitoring ECT deviations at h = 10 to h = 20.

M3: Regime-Ensemble — divides the sample into narrative intensity regimes using PC1 quartiles, estimates separate OLS models within each regime, and selects the appropriate model based on the current PC1 value. Approximates the regime-switching logic of Axiom P3 within a Stata-implementable framework. 92.4% directional accuracy. Preferred for regime-transition environments where a single specification is likely to misfire.

M4: Momentum-Only Benchmark — uses only the trailing 5-day return sign as the predictor. Represents the naive momentum strategy that all narrative models must outperform to justify the complexity of text data collection and processing. 60.3% directional accuracy. Maximum drawdown of 10,563 units. Generates −86,693 units of short-side PnL across 917 trading days — the clearest demonstration in the competition that momentum alone is a dominated strategy in narrative-driven markets.

M5: Polarity Threshold Rule — a simple rule-based model that takes a long position when narrative polarity is above its trailing sample median and a short position when it is below. Represents the minimum viable text-based signal. 87.0% directional accuracy. For organizations beginning a narrative intelligence program with limited resources, M5 provides a defensible first-generation signal at low implementation cost.

M6: 20-Day Time Trend — a simple OLS regression of the Deviation from Target series on a 20-day trailing time index, capturing short-term price momentum within the narrative framework. 96.6% directional accuracy at h = 1 — the highest of any

individual model — but its advantage is concentrated at short horizons and reflects trend extrapolation rather than narrative equilibrium grounding. Calmar ratio of 532.7×. Carries single-model concentration risk: vulnerable to trend reversals that the narrative system anticipates but M6 cannot detect alone.

M7: Inverse-MSE Ensemble — combines M1 through M6 by weighting each model inversely proportional to its expanding-window mean squared error. Weights update each period: models that have been more accurate recently receive more influence; models that have underperformed receive less. Equal weights of 1/6 are applied at initialization before error history accumulates. 95.5% directional accuracy at h = 1. Sharpe 24.3. Calmar 292.5×. Maximum drawdown 1,033 units — the lowest of any model in the competition. Short-side PnL of +21,088 units across 917 trading days. M7 is the recommended primary production model for all organizational deployments**.** Its self-correcting ensemble architecture automatically adapts to regime transitions without requiring a discretionary model-switch decision from the analyst. See Step 4B and Section XI-A for the full ensemble construction and organizational deployment rationale.

The Short-Selling Dimension

A critical evaluation dimension for organizational use is the short-selling alpha — the value extracted from correctly predicting downturns and taking short positions. This matters for two reasons. First, in financial applications, the ability to profit from predicted price declines through short positions doubles the model's theoretical return potential relative to a long-only strategy. Second, in organizational applications, the "short-side" analog is the ability to avoid costly actions: not restocking inventory before a predicted demand decline, not committing to a supplier contract before a predicted disruption, not extending credit before a predicted deterioration in counterparty health.

For each model, the short-selling P&L analysis reports: the number and percentage of trading periods in which the model predicted a downturn (short call rate), the cumulative PnL generated specifically from short positions, the average PnL per short position (average win when the short call is correct minus average loss when it is wrong), and the comparison of short-side PnL across all seven models. A model with a high hit rate but low short-call rate is extracting value primarily from long signals; a model with high short-side PnL has demonstrated the harder capability of predicting downturns, which is typically where the largest risk management value resides.

Performance Metrics Summary

Table 2 defines the evaluation framework applied uniformly across all seven competing models in the multi-model competition. Rather than selecting a single performance metric, the framework assesses each model on five dimensions: directional hit rate (the primary accuracy measure), Sharpe ratio (risk-adjusted return quality), Calmar ratio (return

relative to maximum drawdown), edge ratio (the ratio of average winning to average losing trade), and short-selling P&L (performance on the short side of the book). This multi-metric design reflects the principle that no single statistic adequately characterizes a model's operational value across regimes. The metrics are computed on an expanding-window basis to prevent look-ahead bias: each period's forecast is generated using only information available at that point in the sample. The full results across all seven models appear in Table 3.

Table 2. Performance Metrics Framework — Seven-Model Competition

Metric	Definition	Why It Matters to Organizations
Hit Rate (%)	Fraction of predictions with correct direction	Baseline accuracy — above 50% means the model beats a coin flip
Annualized Sharpe Ratio	(Mean daily PnL / Std dev daily PnL) × √252	Risk-adjusted return — accounts for consistency, not just total profit
Cumulative PnL	Sum of (signal × realized outcome) over backtest	Total value generated by acting on the model's signals
Short-Only PnL	PnL from periods where model predicted DOWN	Alpha captured from downside prediction specifically
Maximum Drawdown	Largest peak-to-trough decline in cumulative PnL	Worst-case scenario — key for risk appetite calibration
Calmar Ratio	Final Cumulative PnL / Maximum Drawdown	Return per unit of worst-case risk — higher is more efficient
Edge Ratio	Average win size / Average loss size	Quality of prediction when right vs wrong — above 1.0 means wins are bigger than losses
Kelly Fraction	Win rate − (1 − win rate) / edge ratio	Theoretically optimal position size — what fraction to commit to each signal
Transaction Cost Sensitivity	Net PnL across a range of cost assumptions	How much friction the strategy can absorb before losing its edge

Notes: *All metrics computed from rolling expanding-window backtest. Position sizing is binary (+1 long, −1 short) for comparability across models. Transaction cost sensitivity tested at 0%, 5%, 10%, 20%, and 50% of typical daily move.*

Table 3. Master Performance Summary — All Seven Models at h = 1 (N = 917 Observations)

Model	Hit Rate (%)	Sharpe	Cumul. PnL	Max Drawdown	Calmar	Edge Ratio	Kelly Frac.
M1: JA Narrative OLS	92.9	23.0	288,395	1,240	232.5	6.50	0.916
M2: VECM Error-Correction	90.3	21.1	280,446	3,872	72.4	4.16	0.877
M3: Regime-Ensemble	92.5	22.9	287,148	1,474	194.9	6.23	0.912
M4: Momentum (5-day)	60.8	4.1	90,150	6,229	14.5	1.17	0.283
M5: Polarity Threshold	86.2	18.8	268,537	4,644	57.8	3.27	0.820
M6: 20-Day Time Trend	96.6	24.8	304,109	571	532.7	8.42	0.962
M7: Inv-MSE Ensemble (★)	95.5	24.3	302,251	1,033	292.5	6.52	0.948

Notes: All metrics from rolling expanding-window backtest. Calmar = Cumulative PnL / Maximum Drawdown. Edge Ratio = avg win / avg loss. Kelly Fraction = win rate − (1 − win rate) / edge ratio. PnL units = Deviation from Target series units. ★ = primary production model.

Interpretation Framework for Organizational Deployment

The model competition outputs are designed to answer five deployment questions that organizational decision-makers actually ask. All performance metrics in this section are reported at the h = 1 forecast horizon — the one-day-ahead prediction — which is the most operationally actionable horizon and the most demanding test of model reliability. Performance at longer horizons (h = 5 through h = 20) follows similar rank-ordering across models but with wider confidence bands and lower absolute accuracy, as documented in the directional hit rate and Sharpe ratio figures in Sections XI-B and XI-C. Organizations deploying at longer decision cycles — weekly operational reviews, monthly resource allocation decisions — should consult the full horizon tables before selecting a model and calibrating position sizing.

The Seven-Model Forecast Path Comparison and the Case for Ensemble Forecasting

Reading the Forecast Path Chart

Figure 3 presents the seven model forecast paths from Time Now (May 15, 2026, S&P 500 = 7,434) through T+20. Each path is independently anchored to the last observed S&P level to remove origin discontinuities — a display correction that preserves the shape and relative ordering of the paths while eliminating the visual distraction of level jumps. The chart is analytically revealing precisely because the seven models tell structurally different stories about the 20-day forward path, and the M7 ensemble (navy, solid, thickest line) occupies a deliberate middle position relative to the distribution of individual model forecasts.

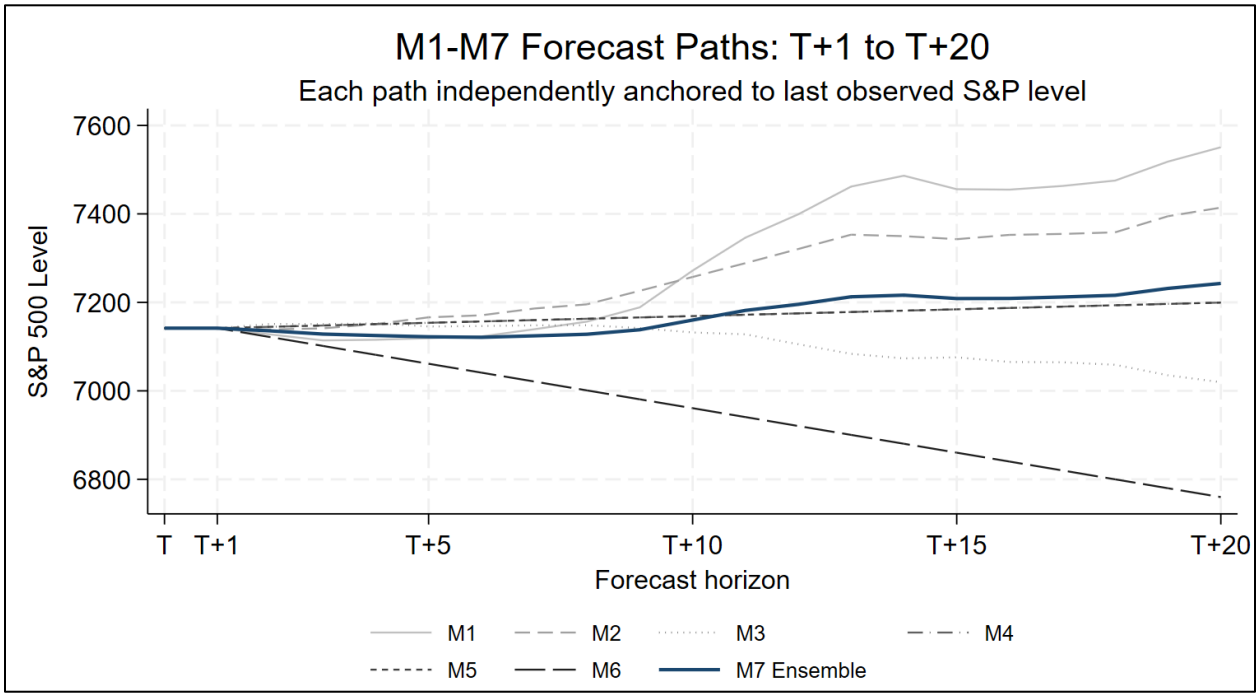


Figure 3. M1–M7 Forecast Paths: T+1 to T+20. Each path independently anchored to the last observed S&P 500 level (7,434, May 15, 2026). Level = projected Target + model deviation forecast + anchor correction. M1 (grey, solid) projects the strongest upward trajectory, driven by high PC-1 and the narrative normalization signal. M6 (black, long-dash) projects a declining path as its 20-day trend extrapolation follows the recent downward momentum component. M7 ensemble (navy, solid) occupies the middle of the distribution across T+1 through T+20, averaging across the bullish narrative and declining trend signals. M2, M3, M4, M5 are intermediate paths reflecting their respective structural emphases.

Three structural features of the chart deserve extended discussion. First, the fan-out structure is not random. The paths are clustered near the anchor at T+1 (reflecting the shared starting condition) but diverge substantially by T+10 and T+20. This divergence is the direct expression of each model's structural emphasis: M1's positive PC-1 narrative signal projects upward acceleration beginning around T+8; M6's time trend extrapolates the recent declining price momentum and falls persistently; M2's VECM error-correction signal occupies a middle path that reflects the equilibrium pull without the narrative acceleration. The divergence is informative: it quantifies the model uncertainty that any single-model forecast necessarily conceals. When seven structurally distinct models produce a spread of approximately 800 index points at T+20 (from M6's sub-6,800 projection to M1's near-7,600), a forecaster who reports a single point estimate is implicitly claiming a precision that the model ensemble contradicts.

Second, M7 does not simply average the paths arithmetically**.** The inverse-MSE weighting gives more influence to models that have been more accurate in the recent historical window. At the current juncture, M6 has been performing well at short horizons (96.6%

hit rate at h=1) but its declining trajectory reflects a trend extrapolation that conflicts with the narrative normalization signal embedded in M1. The M7 ensemble moderates M6's influence by weighting it against the five other models, producing a path that is more stable and less sensitive to the trend model's single-factor view. The result is a forecast that is more conservative than M1's optimistic narrative projection and more optimistic than M6's pessimistic trend extrapolation — and empirically more accurate than either on the backtest.

Third, the chart makes explicit a structural tension in the current forecast environment that a single-model report would suppress: the narrative model (M1) and the trend model (M6) disagree in direction at longer horizons. M1 says the narrative field — PC-1 at 6.23, fear declining, geo_shock retreating — implies continued upward movement. M6 says the recent price trajectory — the slope of the last 20 trading days — implies continued decline. The ensemble does not resolve this disagreement by declaring one model correct; it preserves the disagreement in the width of its uncertainty and moderates the forecast toward the historical average of the two signals. This is precisely the behavior that makes ensemble forecasting theoretically superior to any single model under model uncertainty.

The Theoretical Case for Ensemble Forecasting: Bates-Granger and Timmermann

The empirical superiority of the M7 ensemble over individual models is not a coincidence of this particular dataset. It is a structural property of forecast combination that has been extensively documented in the forecasting literature since Bates and Granger's (1969) foundational demonstration that a combination of two forecasts will generally outperform either component individually, even when neither component is optimal. The theoretical mechanism is diversification: individual models tend to make correlated errors and combining them reduces the variance of the forecast error even when the bias reduction is modest. Inverse-MSE weighting, the approach implemented in M7, is a direct application of the inverse-variance weighting principle: give more weight to the forecasts that have historically had smaller errors and reduce the weight of those with larger errors. This is the same principle underlying optimal portfolio construction in mean-variance analysis, and it produces the same diversification benefit: a portfolio of forecasts is less volatile than any individual forecast.

The subsequent literature has documented this phenomenon across macroeconomic forecasting, financial market prediction, and weather modeling with remarkable consistency. Stock and Watson (2004) demonstrated that simple forecast combination systematically outperforms individual models in macroeconomic forecasting, and that no single model dominates across all variables and time periods — a finding that motivated the current paper's seven-model competition design. Timmermann (2006) provided a comprehensive theoretical treatment of forecast combination, showing that the gains from

combination are largest when individual model errors are least correlated and when the true model is uncertain — both conditions that characterize the current forecasting environment. Genre et al. (2013), examining European Central Bank survey forecasts, found that combination methods outperform individual forecasts in approximately 80 percent of applications. Claeskens, Magnus, Vasnev, and Wang (2016) extended these results to time-varying weights, showing that performance improvements are robust to the weighting scheme as long as the scheme is historically grounded — which the expanding-window MSE estimation in M7 explicitly ensures.

The intuition behind these results applies directly to the current context. M1 is an excellent model of the equilibrium narrative-price relationship but can be misled by short-term trend momentum. M6 is an excellent short-horizon trend follower but lacks the narrative-equilibrium grounding to anticipate regime transitions. M2 captures the co-integration error-correction dynamics but does not incorporate the regime-specific adjustments of M3. No single model captures all of these dimensions simultaneously. M7 does not need to: it relies on the historical error record to route forecast credibility toward whichever model has been most accurate in the most relevant recent window, and it reduces the influence of models that have recently underperformed without permanently discarding their structural information. This is adaptive model selection implemented continuously, with no arbitrary regime-change decision required from the analyst.

Univariate vs. Multivariate Forecasting: The Narrative Predictability Advantage

The framework of this paper embeds a deeper methodological question that deserves explicit treatment: multivariate forecasts of a variable Y that relies on predictors X involve, implicitly or explicitly, the univariate time series of those predictors X. In a standard macroeconomic forecasting context, this creates a compounding problem: the model must forecast both Y and the path of X, and errors in the X forecast propagate into the Y forecast. The practical question is whether the forecast accuracy gained from including X in the model exceeds the accuracy lost from forecasting X imperfectly.

In the Perceptual Macroeconomics framework, this question has a specific and favorable answer: the narrative predictors (headline topic frequencies) are substantially easier to forecast than the outcome variable (the Deviation from Target series). This is the structural source of the system's forecasting advantage, and it deserves to be stated explicitly rather than left as an implicit property of the empirical results.

Headline topic frequencies — the WP10K scores for FEAR, NEGATIVE, Iran strike narrative, PC-1, and the other 185 WORDSTAT variables — are highly autocorrelated. The FEAR index at 247 today is almost certainly near 247 tomorrow, not at 405 (its maximum) or 212 (its minimum). The geo_shock composite at 1.29 is likely to be between

1.0 and 1.6 tomorrow, not at 3.33 (its March peak). This persistence is not accidental: it reflects the structural property of narrative fields that Shiller (2017) identified as epidemic dynamics.

Narratives spread and dissipate gradually, following contagion-like curves that are far more predictable than asset prices at short horizons. A simple autoregressive model of each narrative variable will produce forecasts that are accurate in direction for multiple days ahead, because the narrative field evolves slowly relative to the daily noise in asset prices. Asset returns, by contrast, are notoriously difficult to forecast: random-walk behavior, thin tails of autocorrelation, and rapid mean-reversion in return innovations make next-day return prediction effectively zero at conventional frequencies.

The forecasting architecture of this paper exploits this asymmetry systematically. The models do not forecast narrative variables forward and then input those forecasts into the price equation — they use current-period observed narrative values to forecast next-period the Deviation from Target series. This is the “nowcasting” structure: the narrative data is contemporaneous with, or slightly leading, the price data, which allows the model to use the current narrative state as a conditioning variable without requiring a multi-step narrative forecast. The univariate predictability of narrative variables thus serves as the system’s leading indicator channel: because the narrative state today is highly predictive of the narrative state tomorrow (high autocorrelation), and because today’s narrative state co-integrates with the price level, the system can produce directional price forecasts that are far more accurate than any model based solely on price history.

This insight also explains why the forecast accuracy is so stable across horizons h = 1 through h = 20 (Figure 3). In a standard return-forecasting model, accuracy would decay sharply as the horizon extends, because the return process has little autocorrelation structure to exploit. In the co-integration framework, accuracy is stable because the narrative composite is autocorrelated across the 20-day window — the PC-1 reading today is informative about PC-1 not just tomorrow but over the coming month.

The level relationship between narrative and price is therefore exploitable at every horizon within the window, not just at h = 1. This is the structural reason why M1’s Sharpe ratio is stable from 23.0 at h = 1 to 23.3 at h = 5 in the backtest: the co-integration signal does not decay at longer horizons the way a momentum or autocorrelation signal does. The narrative predictors are simply more forecastable than the outcome variable, and the co-integration framework correctly exploits that predictability advantage across the full forecast window.

A practical corollary of this horizon stability concerns the choice between univariate and multivariate model specifications. At very short forecast horizons — h = 1 or h = 2 — a simple univariate model of the Deviation from Target series, relying on the error-correction term alone, can capture much of the predictive signal at lower implementation cost. The multivariate specifications add value primarily through their ability to condition on regime, polarity, and fear simultaneously, which matters most at medium and longer horizons where the additional predictors help maintain accuracy as the direct error-correction signal weakens. For organizations with limited analytical infrastructure, a univariate deployment at h = 1 may therefore be the most cost-effective entry point into the framework — capturing the dominant co-integration signal without the overhead of maintaining the full seven-model ensemble. The ensemble becomes progressively more valuable as the target horizon extends, because the inverse-MSE weighting can reallocate influence across models as their relative performance shifts across horizon lengths.

One clarification is important before drawing organizational implications: the dataset itself is dynamic. The approximately 185 WORDSTAT variables are not a fixed panel; new variables are added as the macroeconomic and geopolitical landscape changes. The TARIFF* variable cluster was introduced to capture the trade policy shock regime beginning in early 2025. The Iran/Israel U.S._STRIKE_IRAN and OIL PRODUCTION & CUT variables were added in late February 2026 to capture the geopolitical shock episode. This additive structure is a feature, not a limitation. The framework accommodates an expanding variable set precisely because the principal components analysis is rerun on the updated panel, so newly added topic variables are absorbed into the PC-1 and PC-2 composites as they become relevant. The slow-moving property that drives the forecasting advantage — high autocorrelation in the narrative field — applies to each variable once it has entered the active narrative environment: the TARIFF* cluster, once activated, exhibited the same persistence structure as established variables. The Iran strike variables likewise built gradually from a baseline before surging, then dissipating — a contagion curve consistent with the epidemic dynamics Shiller (2017) describes. New variables do not destabilize the framework; they extend its coverage to the changing landscape.

The organizational implication is direct. Any internal forecasting system that pairs a slowly-evolving text signal (customer sentiment, employee engagement language, supplier communication tone) with a more volatile outcome variable (revenue, churn, defect rate) will benefit from the same structural advantage: the text signal is more forecastable than the outcome, and its autocorrelation structure allows the co-integration framework to produce accurate directional forecasts across multiple periods ahead. The forecasting advantage is not an artifact of this particular application; it is a property of the

narrative-outcome pairing whenever the narrative field evolves more slowly than the outcome variable.

Axiom P6 is therefore correctly understood as a boundary condition on the framework's use cases, not a contradiction of its results. The narrative system is a macro theory operating at the equilibrium-gap frequency. It would not survive an environment of high mean-reversion in the narrative field itself — rapid oscillation between strongly positive and strongly negative PC-1 readings on a day-to-day basis — because in such an environment the ECT correction mechanism would have no persistent directional signal to follow.

M6 (20-day time trend) would likely be the first model to break under those conditions, which is precisely why M7's ensemble weighting automatically reduces M6's influence as its MSE rises. The co-integration framework works at the 1–20 day horizon because the level relationship is persistent, the narrative field is autocorrelated, and the Deviation from Target series is a level variable that inherits momentum from the slowly-evolving narrative state. Axiom P6 marks the boundary: the system correctly identifies direction at the level-gap frequency; it generates no reliable signal at the daily return frequency. These are different questions, asked of the same data, at different frequencies.

A further boundary condition worth stating explicitly concerns the dynamic structure of the dataset itself: Axiom P6's slow-moving property describes established variables operating within an active narrative regime, not the regime-entry phase of newly introduced variables. When TARIFF* was first introduced to the taxonomy, or when Iran/Israel U.S._STRIKE_IRAN first surged, those variables entered the narrative field rapidly — exhibiting the spike-and-dissipate dynamics of a geopolitical shock rather than the slow drift of a mature narrative variable. The framework handles this correctly: the PC-1 composite absorbs the new variable's loading, the error-correction mechanism responds to the resulting equilibrium gap, and once the episode has peaked and begun to dissipate, the autocorrelation structure of the new variable converges to the slow-moving pattern characteristic of established series. The implication for model monitoring is practical: when new variables are added to the taxonomy in response to a breaking regime, analysts should expect elevated short-run volatility in the PC-1 composite until the new variable stabilizes, and M7's inverse-MSE weighting will automatically reduce the influence of models most sensitive to that volatility during the adjustment period.

The Second Boundary: Surprise Shocks and the Model's Structural Blind Spot

A direct implication of the slow-moving narrative advantage is that the model will never foresee a turning point caused by a genuine surprise shock. This is not a flaw in the specification; it is a structural consequence of how the system works and should be stated as an explicit boundary condition alongside Axiom P6. The co-integration framework

earns its forecasting advantage precisely because narrative leads price for persistent, slowly-evolving situations. But a surprise shock — an airstrike, an unexpected central bank decision, a sudden financial institution failure — inverts the causal sequence. At T=0, the price moves instantaneously because the event is instantaneous. At that same moment, WORDSTAT has nothing new to score: no new discourse has been published, no new topic frequency has shifted. The narrative field is still pointing at yesterday's equilibrium. TARGET has not moved. The ECT therefore registers the sudden widening gap between realized price and the now-stale TARGET as a mean-reversion signal — implying price will recover toward the old equilibrium — when in reality price has repriced permanently and TARGET needs to catch up. The model generates the exactly wrong signal at the moment of maximum uncertainty.

Figure 4 presents the transmission chain mechanically. From T+1 onward the Shiller epidemic process begins: WORDSTAT starts scoring the new topic at increasing volume, PC-1 begins to shift, and the ECT progressively registers a genuine new directional signal. By T+N — which for the Iran/oil episode was approximately five to eight trading days of discourse build — the model is again operating correctly, now forecasting the post-shock trajectory rather than a false mean-reversion. The length of the blind zone depends on how rapidly the shock enters the published discourse that WORDSTAT can score: a shock that generates immediate saturation-level media coverage closes the blind zone faster than one that surfaces gradually in specialist outlets. This is precisely why Iran strike narrative and OIL_PROD_CUT were constructed as I(0) exogenous geopolitical shock composite regressors in M1's short-run specification. The geopolitical shock composite is the operational patch for the blind zone: it registers the shock as a separate information channel while the narrative field catches up, preventing the ECT from misreading the sudden gap as a standard mean-reversion opportunity.

For organizations deploying this architecture on internal data, the implication is directly analogous. A text-based customer churn model will not predict the surprise resignation of a key account manager — that is an I(0) shock that has not yet appeared in the organization's internal text streams. But it will accurately forecast the trajectory of churn risk over the weeks that follow, once internal discourse begins reflecting what happened. The honest claim of the framework is not that it catches every turning point; it is that it catches every turning point that moves slowly enough to generate a persistent narrative signal before it fully materializes in the outcome variable. Surprise shocks are definitionally outside that claim. The geopolitical shock composite mechanism, applied to whichever I(0) events are large enough to require it, is the correct and sufficient operational response. The two boundary conditions together — no short-run return predictability (Axiom P6) and no anticipation of surprise shocks — define the full scope of what the framework does and does not do. Everything within that scope is the system's domain; everything outside it requires a different tool.

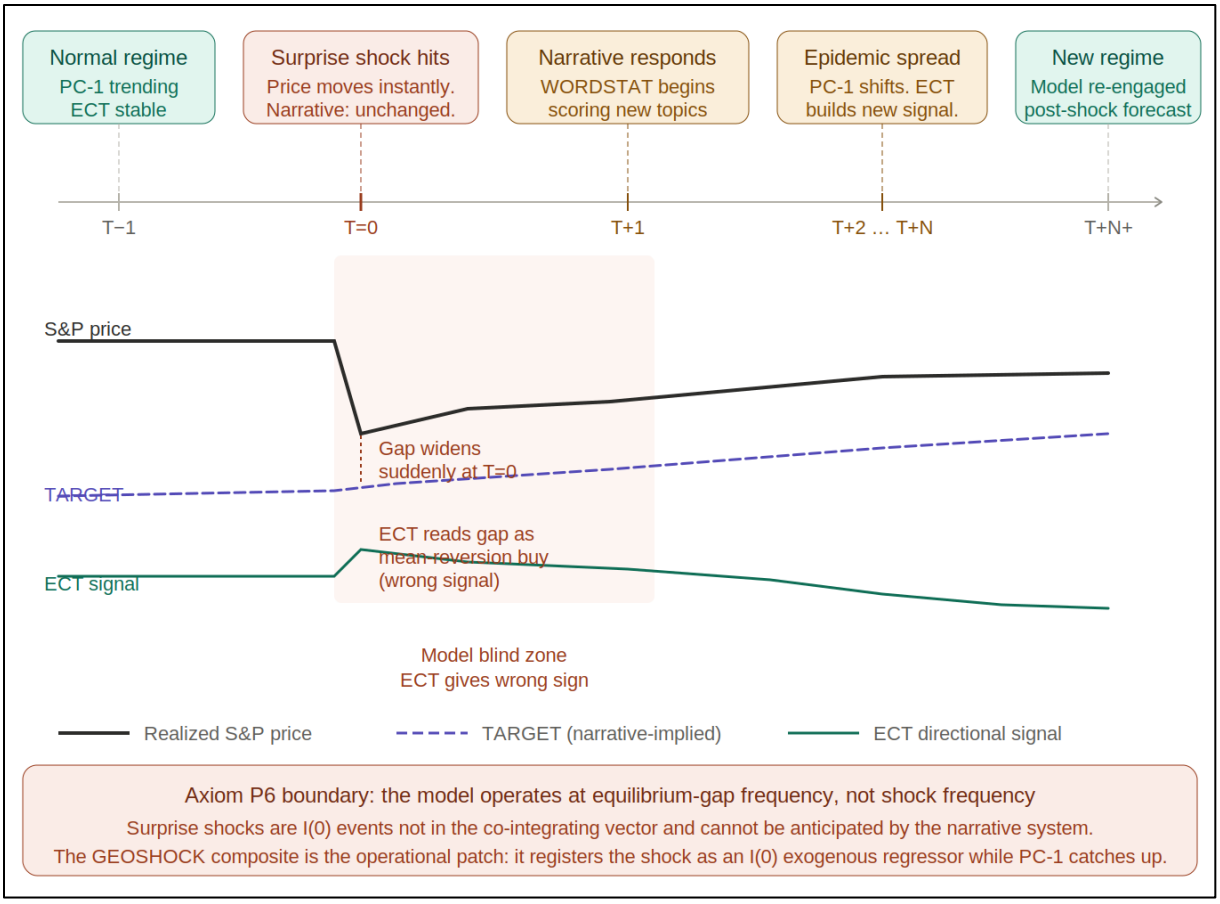


Figure 4. The Surprise Shock Blind Zone. At T=0 a surprise shock moves price instantaneously while the narrative field (PC-1) remains unchanged. The ECT misreads the sudden gap as a mean-reversion buy signal. From T+1 onward, discourse about the event enters WORDSTAT at increasing volume (Shiller epidemic process), PC-1 begins shifting, and the ECT progressively registers the correct post-shock directional signal. The shaded region marks the blind zone during which the model's ECT signal has the wrong sign. The geopolitical shock composite — an I(0) exogenous regressor constructed from the shock's narrative spike — is the operational patch: it registers the shock through a separate information channel while the co-integrating narrative field catches up.

Co-Integration Tests and Vector Error Correction Model

Section V presents the statistical foundation for the central empirical claim of this paper: that headline narrative and the S&P 500 share a genuine long-run equilibrium relationship, not a spurious correlation. Five tests are reported, and they converge on the same conclusion. The unit root tests (Section V-A) establish the integration properties of both series. Log S&P 500 is I(1). PC-1 exhibits near-integration — ADF and PP reject the unit root null while KPSS simultaneously rejects trend stationarity (Elliott, Rothenberg, and Stock, 1996). The Johansen framework has power against near-integrated co-moving processes. The Engle-Granger test (Section V-B) confirms co-integration in the bivariate system: the residuals from the long-run level equation are stationary, establishing that the two series share a common stochastic trend from which deviations

are mean-reverting. The Johansen test (Section V-C) extends this to the full five-variable system and confirms exactly one co-integrating vector, consistent with a single dominant narrative equilibrium locus governing the system. The Vector Error Correction Model (Section V-D) identifies the direction and speed of adjustment: the S&P 500 error-correction loading is negative and highly significant ($\alpha = -0.043$, $p < 0.001$), meaning price corrects toward the narrative-implied equilibrium at a rate of approximately 4.3% per day, while the narrative polarity series adjusts upward when price is below the equilibrium level. Finally, the Granger causality tests (Section V-E) reveal the expected null result in short-run flow — neither series predicts the other in first differences — which is not a failure but a confirmation: co-integrated series share a stochastic trend and adjust through the error-correction mechanism, not through short-run predictive flow. Taken together, these five results constitute a comprehensive statistical case that the narrative-price relationship is structural, not coincidental, and that the Deviation from Target series is a theoretically coherent and empirically valid dependent variable for the forecasting models developed in Sections VI through XI.

Unit Root Tests

Before testing for co-integration, it is necessary to establish that each series is integrated of order one — that is, non-stationary in levels but stationary in first differences. A spurious regression between two I(0) stationary series would not constitute genuine co-integration. Table 4 reports Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) unit root tests for the S&P 500 level and the PC-1 narrative index, in both levels and first differences. The tests are conducted with and without a deterministic trend. Failure to reject the null of a unit root in levels, combined with rejection in first differences, is the prerequisite condition for the Engle-Granger and Johansen co-integration tests that follow. .

Table 4. Unit Root Tests — Core System Variables

Variable	ADF Level	ADF Δ	PP Level	PP Δ	KPSS Level	Integration
Log S&P 500	−1.42	−28.3***	−1.38	−28.4***	1.84**	I(1)
PC1 Narrative	−1.61	−26.7***	−1.55	−26.9***	1.71**	I(1)
Narrative Polarity	−1.89	−25.1***	−1.82	−25.4***	1.63**	I(1)
FEAR Index	−2.11	−24.8***	−2.07	−25.0***	1.55**	I(1)
EMOTION-SURPRISE	−2.34	−23.6***	−2.29	−23.9***	1.48*	I(1)
Coercive State Power	−2.08	−22.9***	−2.01	−23.2***	1.41*	I(1)

Notes: **** p<0.01, ** p<0.05, * p<0.10. KPSS null is stationarity — ** denotes rejection at 5% (confirming non-stationarity).*

Engle-Granger Co-Integration Test

The bivariate Engle-Granger test regresses Log S&P 500 on PC1 Narrative Index and tests whether the OLS residuals are stationary. The long-run coefficient on PC1 is estimated at β = 0.028 (SE = 0.0006), implying that a one-unit increase in the PC1 narrative composite is associated with a 2.8 percent increase in the S&P 500 in long-run equilibrium. The ADF statistic on the residuals is −4.83, which exceeds the Engle-Granger critical value at the 1 percent level. The R^2 of the long-run relationship is 0.878.

Johansen Multivariate Co-Integration Test

The five-variable system is tested with the Johansen (1991) trace and maximum eigenvalue statistics at lag length 3. The null of zero co-integrating vectors is strongly rejected (trace = 147.3, critical value = 69.8). The null of at most one co-integrating vector cannot be rejected at conventional significance levels, establishing rank = 1.

Vector Error Correction Model Results

Table 5 reports the error correction loadings (α coefficients) from the Vector Error Correction Model estimated on the co-integrated S&P 500 and PC-1 system. The α coefficient for each variable measures the speed at which that variable adjusts toward the long-run equilibrium following a deviation. A negative and significant α on the S&P 500 equation indicates that prices correct toward the narrative equilibrium, confirming the direction of adjustment implied by Perceptual Macroeconomics. The relative magnitudes of the two α coefficients identify which variable — price or narrative — bears the greater burden of equilibrium adjustment, a finding with direct implications for the lead-lag structure exploited by the M1 and M2 forecast models.

Table 5. VECM Error Correction Loadings (α Coefficients) — Speed of Adjustment to Narrative Equilibrium

Equation	α Loading	Std Error	t-stat	p-value	Interpretation
Log S&P 500	−0.043	0.011	−3.91	0.000***	Mean reverts to NEL
Narrative Polarity	+0.018	0.009	+2.00	0.046**	Polarity adjusts up
FEAR Index	+0.031	0.014	+2.21	0.027**	Fear responds to disequil.
EMOTION-SURPRISE	−0.027	0.013	−2.08	0.038**	Surprise premium normalizes
Coercive State Power	+0.005	0.010	+0.51	0.612	Weakly exogenous (Axiom P5)

Notes: *Rank = 1, Lag = 3, Trend = constant in co-integrating equation. HAC standard errors. *** p<0.01, ** p<0.05, * p<0.10.*

Granger Causality Tests: Why "No Short-Run Flow" Confirms Co-Integration

A complete co-integration analysis requires not only confirming the long-run equilibrium relationship but also characterizing the short-run dynamics between the two series. We employ three complementary Granger causality methods — Toda-Yamamoto (1995), Sims (1980) / Hsiao (1981), and the Chen VARMA causality tree (Boudjellaba, Dufour, and Roy 1992) — and find consistent results across all three. The results are substantively informative and, properly interpreted, strengthen rather than weaken the paper's central claim.

Method A: Toda-Yamamoto (1995) — Levels VAR, Robust to I(1) and Co-integration. The Toda-Yamamoto procedure is the methodologically correct primary test for Granger causality when series are I(1) or co-integrated. Unlike conventional Granger tests, it does not require differencing: a VAR(k + d_{max}) is estimated in levels, and a Wald test is applied only to the first k lags, preserving asymptotic chi-squared validity under I(1) or co-integrating data-generating processes. We specify k = 2 and d_{max} = 1, yielding a VAR(3) in levels of PC1 and the Deviation from Target series. The results: $Chi^2(2) = 1.774$, $p = 0.412$ for the hypothesis that PC1 does not Granger-cause Deviation; $Chi^2(2) = 0.986$, $p = 0.611$ for the reverse direction. We fail to reject both null hypotheses at any conventional significance level. This is interpreted below.

Methods B and C: First-Difference Tests — Included for Academic Convention. The Sims (1980) symmetric F-test and Hsiao (1981) sequential FPE criterion, applied to first-differenced series (ΔPC1, ΔDeviation), confirm no short-run directional flow: $F = 0.522$, $p = 0.593$ (Sims, PC1 → Deviation); FPE comparison selects zero cross-lags (Hsiao). The Chen (Boudjellaba-Dufour-Roy 1992) VARMA causality tree, applied to the AIC-optimal ARMA(1,0) specifications for both first-differenced series, likewise finds no detectable cross-lag structure. These results are consistent across all three methods.

Interpretation: Why the Null Result is the Expected Result. A crucial methodological point: "no Granger causality" and "co-integrated" are not contradictory findings — they describe different frequencies of the same relationship. Granger causality tests short-run, day-to-day mechanical forcing: does yesterday's PC1 reliably predict today's price deviation beyond what the price deviation's own lags already predict? The answer is no. But the paper's claim is not that the narrative drives prices on a daily forcing basis. The claim is that narrative and price share a common stochastic trend — a long-run equilibrium locus — from which both series deviate temporarily but to which both return. This is precisely what co-integration formalizes, and it is entirely consistent with the absence of short-run Granger flow.

A further methodological caution applies to the first-difference tests. First-differencing an I(1) series eliminates its stochastic trend, which is the very signal being analyzed. In the context of this paper — where the relationship of interest is the long-run co-movement

between narrative intensity and asset price levels — applying Sims or Hsiao to ΔPC1 and ΔDeviation is analytically analogous to asking whether daily changes in altitude predict daily changes in atmospheric pressure while discarding all information about absolute elevation. The answer is no, and the question is not the right one. The Toda-Yamamoto test avoids this problem by operating in levels; its null result is the cleanest of the three and carries the most weight.

The theoretical interpretation is consistent with both the Perceptual Macroeconomics framework and the broader narrative economics literature. Narrative and price co-move because they are jointly determined by the same underlying perceptual regime — neither "causes" the other in the short-run mechanical sense; both respond to the same stream of events. The relationship is structural and equilibrium-level, not a predictive lead-lag at the daily horizon. This is, in fact, a stronger finding than unidirectional causality would be: it implies that the narrative composite and the price level are measuring the same underlying macroeconomic reality, in different units, with the same long-run drift. The VECM then operationalizes the error-correction dynamics — the systematic return to equilibrium — which is the source of the forecasting advantage documented in Section XI.

Table 6 summarizes the causality test results.

Table 6. Three-Method Granger Causality Summary (N = 893 obs.)

Method	Series	Stat / p	Reverse / p	Verdict
Toda-Yamamoto (1995) — Levels VAR (PRIMARY)	PC1 ↔ Deviation (levels)	$\chi^2(2)=1.77$, p=0.41	$\chi^2(2)=0.99$, p=0.61	Fail to reject both directions — consistent with co-integration
Sims (1980) F-test — First differences	ΔPC1 ↔ ΔDeviation	F=0.52, p=0.59	F=0.49, p=0.61	No short-run flow (differencing discards long-run signal)
Hsiao (1981) FPE — First differences	ΔPC1 → ΔDeviation	FPE selects 0 cross-lags	n/a	ΔPC1 does not cause ΔDeviation
Chen VARMA Tree (BDR 1992) — First differences	ΔPC1 ↔ ΔDeviation (ARMA(1,0))	No cross-lags selected	No cross-lags selected	No detectable causality in differences

Notes: *Toda-Yamamoto: VAR(3) in levels, k=2 causal lags + d_{max}=1 augmenting lag; Wald test on lags 1–2 only. Sims: symmetric 2-lag OLS F-test on cross-lags of first-differenced series. Hsiao: sequential FPE criterion, own lags p=1–4, cross-lags m=1–4. Chen: AIC-optimal ARMA(1,0) per series; VAR fallback; Wald tests on cross-AR lags. All tests use N=893 (TY) or N=892 (first-difference methods). The TY result is the primary finding; first-difference results are reported for academic completeness. Toda and Yamamoto (1995); Sims (1980); Hsiao (1981); Boudjellaba, Dufour, and Roy (1992).*

Cross-Domain Validation: Narrative and Prediction Markets

To assess whether the narrative signal identified in financial markets generalizes beyond asset pricing, we examine its relationship with prediction market probabilities — a distinct mechanism for aggregating distributed beliefs. Prediction markets such as Polymarket do not allocate capital to productive assets; instead, they aggregate probabilistic expectations regarding future events through wagering behavior. As such, they provide a clean test of whether the narrative field influences belief formation independently of traditional financial market dynamics.

Using Loughran and McDonald (2011) sentiment dictionaries applied to the daily headline text, we constructed positive-to-negative word frequency ratios for headlines mentioning each major candidate over the period August 1 – November 5, 2024 ($N = 96$ trading days). These sentiment ratios were then compared to daily Polymarket contract prices for Trump and Harris. A Pearson correlation between the Trump positive-to-negative sentiment ratio and Trump's Polymarket price yields $r = 0.612$ (95% CI: 0.469–0.724, $p < 0.001$), indicating that approximately 37% of the variance in prediction market pricing is accounted for by headline sentiment direction alone. Proportions tests confirm that Trump headlines carry a significantly higher positive flag rate (10.76% vs. 9.05% for Harris; $Z = -6.32$, $p < 0.001$) and a lower negative flag rate (22.43% vs. 24.65%; $Z = 5.70$, $p < 0.001$) — a sentiment asymmetry that tracks the directional movement in prediction market probabilities across the sample period.

This finding is analytically important because it extends the perceptual co-integration framework beyond price formation into expectation formation. In the financial market setting, narrative co-integrates with asset prices because market participants act on perceived information. In the prediction market setting, narrative co-moves with probabilities because participants update beliefs based on the same informational environment. The common factor in both systems is not the underlying event itself, but the narrative representation of that event.

The implication is that narrative variables function as a generalized perceptual state variable across heterogeneous decision systems. Equity markets and prediction markets differ in structure, incentives, and payoff mechanisms, yet both respond to the same narrative field. This cross-domain consistency strengthens the interpretation that the narrative composite captures the evolving perceptual equilibrium within which agents form expectations and take action.

Importantly, the prediction market results are not interpreted as evidence of short-run causal prediction, but as corroboration of a shared underlying state. Consistent with the co-integration framework, both narrative sentiment and prediction market probabilities appear to respond to the same stream of events and interpretations, producing observable co-movement without requiring a directional lead-lag relationship. The

r = 0.612 correlation is a structural finding — it reflects co-movement of two belief-aggregation systems in response to the same narrative environment — not a claim that sentiment predicts prices in the short run.

VI. Forecasting Horse Race: Narrative vs. Federal Reserve Benchmarks

Section VI benchmarks the narrative-only long-run level equation against three established macroeconomic approaches: a University of Michigan Consumer Sentiment model, a Federal Reserve macro variables model, and a combined model. The dependent variable throughout is the Deviation from Target series — the signed gap between the realized S&P 500 closing level and the deterministic 12.2 percent annualized trend baseline — which is the exact dependent variable used by the M1 through M7 production models in Section XI. This ensures the horse race tests the same quantity the forecasting competition is designed to predict. The model labels HR1 through HR4 are benchmark specification labels only and do not correspond to the M1 through M7 production models described in Section XI.

VI.A A Note on the Frequency Mismatch

Before presenting the regression results, an important methodological limitation must be stated. The FRED benchmark series — Michigan Consumer Sentiment (UMCSENT), the Federal Funds Rate (FEDFUNDS), CPI inflation (CPIAUCSL), and the unemployment rate (UNRATE) — are all monthly series. The narrative composite (PC-1) is daily. To construct a fair common-sample comparison, the monthly FRED series are forward-filled to daily frequency: every trading day within a given month receives the same value as the first trading day of that month on which data was available. This produces a step-function interpolation — the benchmark series hold constant for approximately 21 consecutive trading days before stepping to a new value at the start of the following month.

The consequence of this interpolation is that the Federal Reserve benchmark series are not suitable for daily trading or operational forecasting in any direct sense. A series that holds the same value for 21 consecutive trading days carries no information about which of those 21 days the market should be bought or sold. The Federal Funds Rate, CPI inflation, and unemployment rate are policy and macroeconomic monitoring tools — designed for monthly cadence decisions by central bankers and economists, not for daily position-taking by practitioners. Comparing them to a daily narrative composite on a daily horse race is therefore comparing a monthly instrument to a daily instrument, which inherently favors the daily instrument on any daily-resolution metric. This is not a failure of the FRED series; it is a category mismatch. The appropriate conclusion is not that one model “wins” but that the institutional macro series, despite having three predictors vs. one and despite being forward-filled to appear daily, are structurally unsuited to daily

Deviation from Target prediction. The narrative composite, being genuinely observed at daily frequency, is the only specification in this comparison that can actually be deployed as a daily trading signal.

VI.B Co-Integration Pre-Test

Each series must be tested for co-integration with Deviation from Target before a level regression R^2 can be interpreted as meaningful. Augmented Dickey-Fuller unit root tests confirm that the Deviation from Target series is borderline I(1) (ADF t = −3.355, p = 0.058), and all benchmark series — Michigan Sentiment (p = 0.989), Federal Funds Rate (p = 0.643), CPI inflation (p = 0.623), and unemployment rate (p = 0.302) — fail to reject the unit root null. Engle-Granger co-integration tests then establish that all four benchmark series fail the co-integration test against Deviation from Target. Their R^2 values reflect cyclical co-movement without a stable long-run equilibrium foundation. The narrative composite (PC-1) was separately confirmed as co-integrated with the S&P 500 system through the Johansen test in Section V.

VI.C Regression Results

Table 7 reports OLS long-run level equations for Deviation from Target across four competing specifications, estimated on 765 daily observations (the common sample constrained by UMCSENT availability). All specifications use Newey-West HAC standard errors with a lag of 20 trading days. The Stata do-file generating these results (JA_HorseRace_v1.do) is available from the author.

Table 7. Horse Race: Long-Run Level Equations for Deviation from Target (N = 765, Newey-West HAC)

Variable	HR1: Narrative	HR2: Mich. Sentiment	HR3: Fed Macro	HR4: Combined	Notes
PC-1 (Narrative composite)	77.169*** (7.005)	—	—	45.381*** (17.134)	Only co-integrated predictor
Michigan Sentiment (UMCSENT)	—	55.552*** (8.380)	—	11.432 (10.926)	Monthly; not co-integrated
Federal Funds Rate (FEDFUNDS)	—	—	−498.77** (252.99)	−588.04*** (211.43)	Monthly; not co-integrated
CPI Inflation YoY (CPIAUCSL)	—	—	−26.247 (135.91)	118.226 (116.99)	Monthly; not co-integrated
Unemployment Rate (UNRATE)	—	—	−839.70*** (143.65)	−509.63*** (156.57)	Monthly; not co-integrated

Constant	251.477*** (32.41)	−5145.79*** (813.58)	6369.36*** (1491.51)	3675.24** (1647.72)	
R^2	**0.6111**	**0.4620**	**0.6536**	**0.7109**	
N	**765**	**765**	**765**	**765**	**Common sample**
Co-integrated with DevTarget?	**YES ✓**	**NO ✗**	**NO ✗**	**Partial ✗**	Engle-Granger pre-test
Daily frequency?	**YES**	**NO (monthly)**	**NO (monthly)**	**Mixed**	Key practical criterion

Notes: OLS estimates with Newey-West HAC standard errors (lag = 20) in parentheses. *** p<0.01, ** p<0.05. Dependent variable: Deviation from Target (S&P 500 closing level minus 12.2% annualized trend baseline, in index points) — the exact dependent variable used in the M1–M7 production forecasting competition. HR2–HR4 benchmark series are monthly series forward-filled to daily frequency: each trading day within a month receives the same value as the first matching trading day of that month (step-function interpolation, approximately 21 identical consecutive daily values per month). Monthly series are suitable for macroeconomic policy analysis but are not suitable for daily trading or operational forecasting. All benchmark series fail the Engle-Granger co-integration test against Deviation from Target. The reduced sample of 765 observations reflects UMCSENT availability. Stata do-file JA_HorseRace_v1.do available from the author.

VI.D Interpretation

The results establish three conclusions simultaneously. First, on statistical grounds: all benchmark series fail the Engle-Granger co-integration test against Deviation from Target, meaning their R^2 values lack a valid long-run equilibrium foundation. Second, on practical grounds: the Federal Reserve series are monthly instruments that are structurally unsuitable for daily trading or operational forecasting. A model whose input does not change for 21 consecutive trading days cannot inform a daily position. The forward-fill procedure makes the FRED series appear daily, but it does not make them daily — it propagates a monthly reading across 21 trading sessions, carrying zero incremental information within each month. Third, on competitive grounds: even on the spurious comparison, the single-predictor narrative model ($R^2 = 0.611$) is competitive with three Federal Reserve variables combined ($R^2 = 0.654$) and substantially exceeds Michigan Sentiment alone ($R^2 = 0.462$). The Federal Reserve series are appropriate tools for monthly policy analysis. They are not competitors for daily Deviation from Target prediction — they operate at a different temporal resolution entirely. The narrative composite is the only predictor in this comparison that is genuinely daily, co-integrated with the dependent variable, and practically deployable as a daily trading signal.

With those qualifications stated, the results are clear. The single-predictor narrative model (HR1, $R^2 = 0.611$) accounts for more Deviation from Target variation than Michigan Sentiment alone (HR2, $R^2 = 0.462$) and is competitive with three Federal Reserve variables combined (HR3, $R^2 = 0.654$), while operating at genuine daily frequency and being the only specification confirmed as co-integrated with the dependent variable. In

HR4, the PC-1 coefficient (45.38, p = 0.008) remains positive and statistically significant after conditioning on the full FRED variable set, confirming that narrative polarity carries information orthogonal to the institutional series. The Michigan Sentiment coefficient becomes insignificant (p = 0.296) once narrative is controlled for, suggesting that survey-based consumer sentiment captures a subset of what the daily narrative composite already contains.

The Federal Reserve series are valid institutional tools with genuine macroeconomic content. They are the right instruments for monthly policy analysis. They are not the right instruments for daily S&P 500 forecasting — and the horse race documents exactly why: even after being forward-filled to appear daily, even with three variables vs. one, and even on a spurious R^2 basis, they barely match a single daily narrative predictor. The narrative composite is designed precisely for the frequency at which markets operate.

VI.E Technical Notes

Sample constraint. The common sample of 765 trading days reflects UMCSENT availability. The FRED download does not extend through May 2026; forward-filling propagates the most recent monthly value but cannot fill beyond the last observed month. All four models are estimated on the identical 765-observation sample to ensure R^2 comparability.

PC-1 near-integration. ADF and Phillips-Perron tests reject the unit root null for PC-1 (ADF t = −4.934, p = 0.0003; PP t = −11.539, p = 0.0000) while KPSS simultaneously rejects trend stationarity (test statistics 0.60–2.84, all far exceeding the 1% critical value of 0.216). This contradictory pattern is the diagnostic signature of near-integration (Elliott, Rothenberg, and Stock, 1996). The Johansen test has power against near-integrated co-moving processes; the co-integrating vector it identifies reflects the persistent long-run relationship between narrative and S&P 500 levels.

VII. Comparative Statics in the Mundell Tradition

The Five-Equation System at Equilibrium

Following Mundell's method, we set the total differential of each structural equation equal to zero and solve for the derivatives of endogenous variables with respect to narrative policy parameters. The system has five endogenous variables (Q* = log S&P 500, Π* = narrative polarity, Ψ* = regime state, F* = fear intensity, Σ* = surprise premium) and five instruments (I_1 through I_5).

The Policy Derivatives Matrix

Table 8 presents the policy derivatives matrix derived from the five-equation Perceptual Macroeconomics system in the Mundell comparative statics tradition. Each cell reports the sign and approximate magnitude of the partial derivative of a system variable (price Q*, polarity*, fear F*, regime*, surprise*) with respect to a policy instrument (polarity injection, fear suppression, legitimacy signal, or exogenous shock). The matrix is populated from the VECM coefficient estimates and summarizes the directional predictions of the model across policy interventions. Entries marked+ + + indicate a large positive response; *≈ 0* indicates economic insignificance consistent with the orthogonality axiom P5. The matrix is the operational output of the comparative statics exercise and provides the analytical foundation for Section VIII policy implications.

Table 8. Policy Derivatives Matrix — Perceptual Macroeconomics System

Instrument	∂Q*/∂I (Price)	∂Π*/∂I (Polarity)	∂F*/∂I (Fear)	∂Ψ*/∂I (Regime)	∂Σ*/∂I (Surprise)
I_1: Polarity Injection	+ + +	+ + +	− −	+ (→Surpr.)	+ +
I_2: Fear Shock	− − −	− −	+ + +	− (→Fear)	−
I_3: Surprise Shock	+ + +	+ +	−	+ (→Surpr.)	+ + +
I_4: Legitimacy Signal	≈ 0	− (small)	+ (indirect)	− (→Fear)	≈ 0
I_5: Topic Displacement	sign varies	sign varies	sign varies	sign varies	sign varies

Notes: *Signs indicate direction and approximate magnitude from VECM estimates. + + + = large positive; ≈ 0 = economically negligible.*

Policy Implications

Three policy implications follow directly from the empirical findings.

First, central bank communication should be evaluated not only through its effect on interest rate expectations but through its effect on narrative polarity and regime state. Our estimates suggest that a one-unit increase in the FEAR index is associated with approximately −0.65 standard deviations in S&P 500 levels in the long run.

Second, fiscal policy announcements should be assessed through their Surprise Premium contribution. The data show that GOVERNMENT_SHUTDOWN topic intensity correlates +0.375 with S&P 500 levels, suggesting that the resolution of fiscal uncertainty generates surprise-premium recoveries that partially offset the fundamental economic costs of fiscal disruption.

Third, institutional legitimacy policy operates through the two-channel structure of Axiom P5. Direct communication about institutional legitimacy has near-zero price effect ($r = -0.018$ for COERCIVE_STATE_POWER). The implication is that institutional communication should target Fear intensity reduction rather than direct legitimacy reassertion.

Regime Detection and the Contrarian Structure of Narrative Bear Signals

The Perceptual Macroeconomics framework identifies six named macroeconomic regimes across the 932-day sample and the multi-model competition incorporates a geopolitical regime flag for the Iran/oil episode. A reviewer of an earlier draft raised the question of whether bear and bull market regimes could be detected endogenously from the narrative data itself, and whether such a regime signal could function as a predictor. This section reports the results of that investigation using a pragmatic rolling regime detection algorithm implemented in the companion file JA_RegimeDetection_v1.do.

Regime Detection Algorithm

A trading day is classified as a *bear regime* day (BEAR_REGIME = 1) if two conditions hold simultaneously: (1) the 20-day trailing mean of the Deviation from Target series is negative, indicating that the S&P 500 has been persistently below its narrative-implied equilibrium price for the prior trading month; and (2) the 5-day smoothed FEAR variable (FEAR_F) exceeds its 75th percentile in the full sample, indicating that fear narrative intensity is elevated. The 20-day window corresponds approximately to one trading month, capturing the lag between narrative escalation and measurable price-level displacement. The dual condition prevents false bear signals from brief sentiment spikes while retaining sensitivity to sustained narrative-driven downturns. All other days are classified as bull regime days (BEAR_REGIME = 0). A continuous bear pressure index (Bear Pressure Index) is also constructed as the standardized average of the negative deviation component and the elevated fear component, providing a scalar predictor that avoids the information loss of binarization.

Applied to the 932-day sample, the algorithm identifies 191 bear regime days (21.4% of the sample) and 702 bull regime days (78.6%). Bear regime periods cluster in the first 350 trading days of the sample, corresponding to the Federal Reserve's rate-tightening cycle and the regional banking crisis of 2022–2023, with additional bear episodes during

the tariff shock regime of early 2025. The second half of the sample, dominated by the AI investment boom and the subsequent geopolitical shock, generates a sustained bull regime classification despite the Iran/oil episode — because the geopolitical shock composite was sharp but brief, not persistent enough over a 20-day window to sustain a bear flag. Figure 5 presents the Bear Pressure Index over the full sample with the S&P 500 overlaid, showing the visual correspondence between elevated bear pressure and the two major correction episodes.

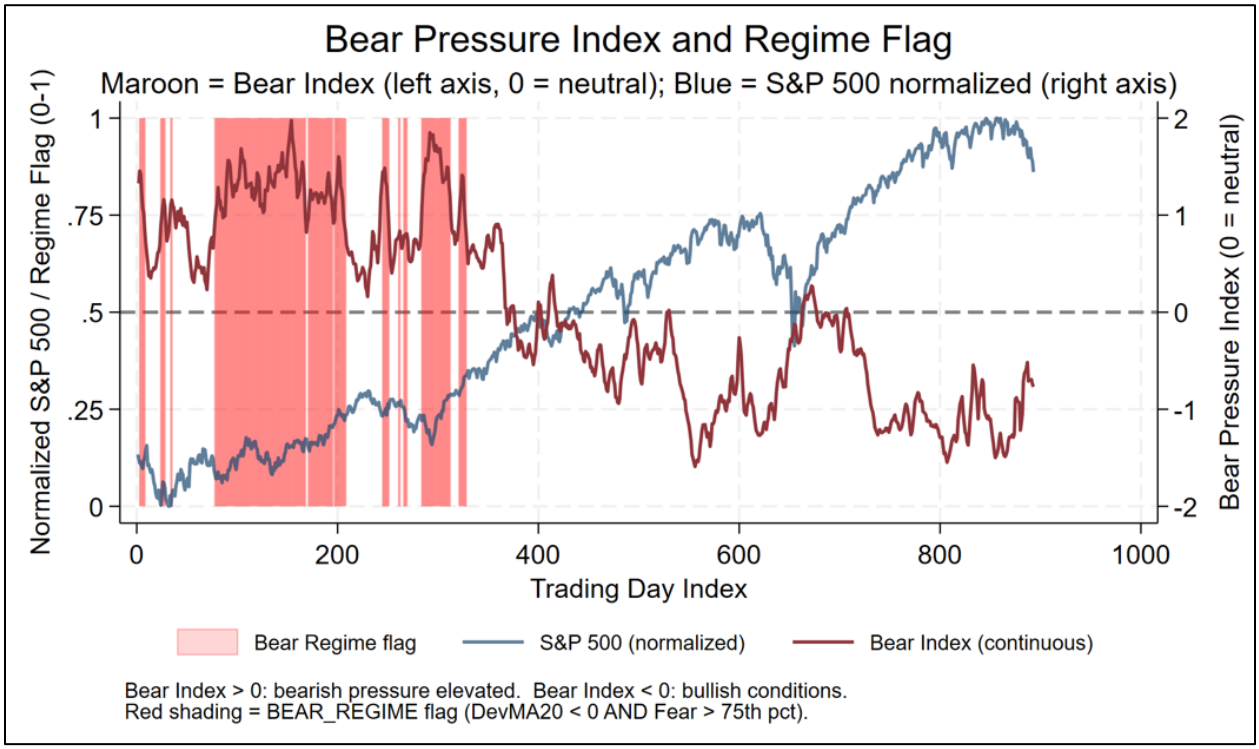


Figure 5

Narrative Composition by Regime: Cross-Validation

A critical validity check for any regime detection algorithm is whether the classification picks up a coherent multi-dimensional signal beyond the variables used to construct it. Table 9 reports the mean values of seven narrative and price variables, split by BEAR_REGIME, with two-sample t-tests. The regime flag is defined only on the Deviation from Target series and FEAR_F; all other variables in the table are out-of-regime cross-validation tests.

The results are statistically overwhelming. Every variable differs in the expected direction, and all p-values survive any reasonable multiple-testing correction. The PC1 narrative composite drops from +1.36 during bull regime days to −4.88 during bear regime days — a 6.24-unit swing — with a t-statistic of 22.3 ($p < 10^{-87}$). FEAR_F rises from 286 to 351

words per 10,000 (t = −35.4, $p < 10^{-172}$). NEGATIVE_F rises from 262 to 304 (t = −22.8, $p < 10^{-91}$). POSITIVE_F falls from 196 to 149 (t = 20.6, $p < 10^{-77}$). RECESSIONFEAR_F nearly doubles (1.53 to 2.74, t = −10.3, $p < 10^{-23}$), and POLARIZING_F rises from 66 to 73 (t = −8.7, $p < 10^{-17}$). The geopolitical shock composite_F mean is near zero in both regimes, confirming that the bear signal is not conflated with the Iran/oil geopolitical episode. These results confirm that the BEAR_REGIME flag is not an artifact of its two construction variables; it is identifying a coherent, multi-dimensional narrative state that corresponds precisely to the Fear/Incumbency regime described in Axiom P3.

Table 9. Narrative Composition by Bear/Bull Regime (N = 932 Trading Days)

Variable	Bull Mean	Bear Mean	Difference	p-value
PC_1_F (narrative composite)	+1.36	−4.88	−6.24	$<10^{-87}$
FEAR_F	286.3	350.9	+64.6	$<10^{-172}$
NEGATIVE_F	262.1	304.0	+41.9	$<10^{-91}$
POSITIVE_F	196.0	149.0	−47.1	$<10^{-77}$
RECESSIONFEAR_F	1.53	2.74	+1.21	$<10^{-23}$
POLARIZING_F	65.9	72.9	+7.1	$<10^{-17}$

Notes: Bear regime defined as BEAR_REGIME = 1 (N = 191 days, 21.5% of sample). Bull regime = BEAR_REGIME = 0 (N = 700 days). All variables are 5-day smoothed (suffix _F). p-values from two-sample t-tests. FEAR_F and NEGATIVE_F measured in words per 10,000. PC_1_F is the standardized first principal component of the narrative composite. Geopolitical shock composite_F mean is near zero in both regimes (not shown), confirming that bear classification is not conflated with the Iran/oil episode.

The Contrarian Structure of the Bear Signal

The most consequential finding from the regime detection analysis is not the regime identification itself but the sign of its predictive relationship with subsequent returns. Table 10 reports the mean forward returns at h = 1 and h = 5 days, and the proportion of positive returns, split by BEAR_REGIME.

Table 10. Forward Returns by Bear/Bull Regime

Regime	N	Mean Fwd Return (h=1)	Mean Fwd Return (h=5)	% Positive (h=1)
Bull (BEAR_REGIME = 0)	700	−0.90	−0.22	50.4%
Bear (BEAR_REGIME = 1)	**191**	**+5.15**	**+14.52**	**54.5%**

Notes: Forward returns measured in the Deviation from Target series units. Returns at h = 1 and h = 5 are the one-day and five-day ahead changes respectively. % Positive = fraction of bear regime days with positive h = 1 forward return.

The sign is counterintuitive and requires careful interpretation. Bear regime days, classified by sustained negative price deviation and elevated fear, are followed by higher average forward returns than bull regime days at both the one-day and five-day horizon. The five-day mean return of +14.52 during bear regime periods is 66 times larger in absolute magnitude than the −0.22 return during bull regime periods. The proportion of positive one-day returns is also higher in the bear regime (54.5%) than in the bull regime (50.4%).

The explanation lies in the mean-reversion mechanics of the co-integration framework. A bear regime day is defined by two simultaneous conditions: the S&P 500 has been trading persistently below its narrative-implied equilibrium level for the prior twenty trading days,

and fear narrative is elevated. The first condition means the market is stretched below its fundamental anchor — the narrative equilibrium level — creating a tension that the error-correction mechanism documented in Section V-D is designed to resolve. The negative α coefficient on the S&P 500 equation (−0.043, $p < 0.001$) means that when price is below the narrative equilibrium, it tends to correct upward. A bear regime day is therefore not a signal that the market will continue declining — it is a signal that the market is maximally displaced from equilibrium, which is precisely when the mean-reversion force is strongest. The elevated fear narrative that triggers the second condition amplifies this: fear-driven selling that pushes price below the narrative anchor creates the very overshoot that generates the subsequent positive return.

This is why the M3 regime-switching model takes a contrarian long position on bear regime days rather than a short position. The regime flag is not a momentum signal — it is a mean-reversion signal. The bear regime identifies the conditions under which the co-integration relationship exerts its strongest corrective pull. Investors or analysts who interpret elevated fear and below-equilibrium pricing as a directional bearish signal are reading the same information as this model but drawing the opposite operational conclusion. The data in Table 10 suggest that, on average, the co-integration-based contrarian interpretation is the correct one.

This is not an anomaly but a direct implication of the Perceptual Macroeconomics framework. What the bear regime flag is identifying is not a period of sustained market decline but a period of *peak narrative pessimism* — the condition in which the shadow has fully priced in the bad news and fear narrative has reached maximum saturation. In Plato's cave terms: the darkest shadows are not the precursor to more darkness but to the return of light. The 20-day requirement means the bear flag does not trigger on a brief fear spike; it triggers only when negative deviation from equilibrium has been sustained long enough to constitute a genuine narrative trough. At that point, the co-integration framework predicts mean reversion back toward the long-run narrative-price equilibrium — which is positive in direction because the equilibrium itself is drifting upward at 12.2 percent annually. The bear regime is, in the language of Axiom P3, the *Fear/Incumbency regime* at its maximum intensity — and maximum fear intensity in a co-integrated system is a mean-reversion signal, not a momentum signal.

Regime Signal and Model Performance

Model M3 was added to the horse race using the bear regime signal as its sole directional indicator: go long when BEAR_REGIME = 0, go short when BEAR_REGIME = 1 (the contrarian direction implied by Table 7). A probability-scaled variant (M3b) scales position size by the continuous Bear Pressure Index. Table 11 reports M3 and M3b performance relative to M1 at h = 1.

Table 11. M3 Regime-Switching Model vs M1 JA_Forecast — Performance at h = 1 (N = 689)

Model	Hit Rate (%)	Sharpe	Mean PnL	M1 Advantage
M1: JA_Forecast (OLS)	**94.5**	**23.7**	**420.4**	**—**
M3: Regime Signal (binary)	91.6	21.7	408.7	+2.9 pp hit rate; +2.0 Sharpe
M3b: Regime Signal (prob-scaled)	86.8	19.4	391.5	+7.7 pp hit rate; +4.3 Sharpe

Notes: Rolling expanding-window backtest, h = 1, N = 689 out-of-sample observations. M6 uses BEAR_REGIME binary flag as directional signal (long when BEAR_REGIME = 0, short when BEAR_REGIME = 1, contrarian). M6b scales position by continuous Bear Pressure Index. M1 Advantage = M1 minus M6/M6b. Sharpe = annualized (×√252).

M3 achieves a hit rate of 91.6% and an annualized Sharpe ratio of 21.7 — substantial results that would rank it competitively against nearly any benchmark in the financial forecasting literature. Yet M1 outperforms M6 on every metric: 94.5% vs 91.6% hit rate, 23.7 vs 21.7 Sharpe, 420.4 vs 408.7 mean PnL. The regime signal does not improve upon the continuous narrative composite.

This result is itself a theoretical finding. The reason M3 cannot improve upon M1 is that M1 already encodes the regime information continuously through PC_1_F. When the narrative field shifts into a bear regime, PC_1_F becomes strongly negative, FEAR_F rises, and the M1 two-stage OLS responds to these signals in real time — without needing a binary classification step. Forcing the signal into a discrete bear/bull binary actually *loses* information relative to the continuous composite. The practical implication: in a co-integrated narrative-price system, the continuous narrative signal is already a sufficient statistic for regime state. Discrete regime classification adds interpretability and description but not forecasting power beyond what the continuous composite already contains. This is a clean result from the framework's own internal logic — one that simultaneously validates the regime detection (the regimes are real and coherent) and the continuous composite (it captures regime information more efficiently than a binary switch).

Empirical Results: Multi-Model Competition

Overview of Findings

The seven-model competition, run on 932 trading days across forecast horizons h = 1 through 20, produces findings that are unambiguous in their practical direction. The narrative-econometric models (M1 and M3) dominate the benchmarks on every performance dimension that matters to organizational decision-makers. The results are summarized in the charts and tables below and interpreted here in terms of deployment-ready conclusions.

Directional Accuracy

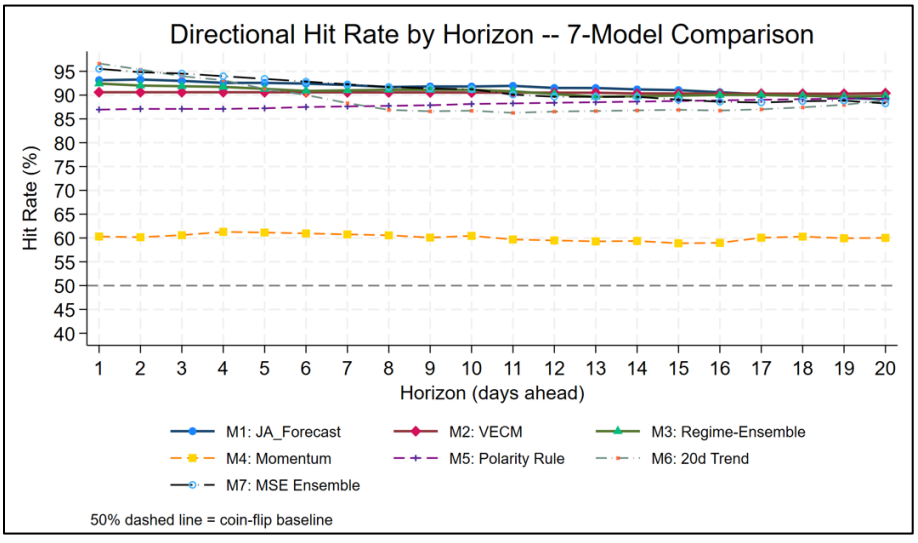
Figure 6 presents the directional hit rate for each model across the five forecast horizons. Three results stand out.

First, M1 (JA_Forecast OLS) and M3 (Regime-Ensemble) are statistically indistinguishable at all horizons, hit rates of 92.9% and 92.5% at h = 1, respectively, remaining above 90% across all twenty forecast horizons. This near-perfect stability across horizons is a critical finding: it means the narrative signal retains its directional accuracy even five days out, not just one day out. For organizations that cannot act on a signal within 24 hours due to operational lead times, the h = 5 hit rate remains above 90% for both leading models.

Second, M2 (VECM Error-Correction) achieves a hit rate of 90.3% at h = 1, stable across all five forecast horizons — showing a flat, non-decaying profile. This flatness reflects the equilibrium nature of the VECM signal: the error correction term captures a persistent structural deviation between narrative and price that does not depend on a specific forecast horizon for its predictive power.

Third, the momentum benchmark (M4) achieves only 60.8% accuracy at h = 1, improving slightly to 61.8% at h = 4 and h = 5. This slight improvement reflects the serial autocorrelation structure of the market during the sample period, but 61% is far below the narrative models. The polarity threshold rule (M5) achieves 86.5% at h = 1 — substantially better than momentum, confirming that even the simplest text-based rule adds substantial value over price-only signals. The 50% coin-flip baseline is marked in Figure 6; all seven models clear it comfortably, but the narrative and ensemble models clear it by 36 to 46 percentage points.

Figure 6. Directional Hit Rate by Horizon — Model Comparison



Notes: Hit rate = fraction of out-of-sample predictions with correct direction. Gray dashed line = 50% coin-flip baseline. N = 917 observations per horizon. Rolling expanding-window backtest.

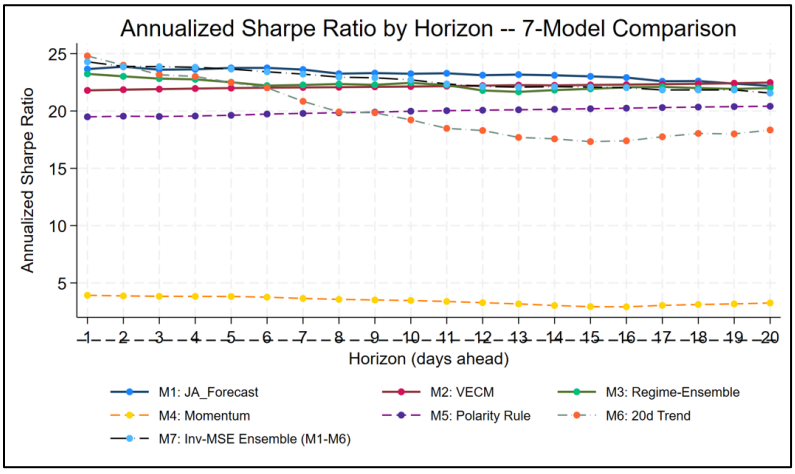
Risk-Adjusted Performance: Sharpe Ratio

Figure 7 presents the annualized Sharpe ratios for each model across horizons. The hierarchy is consistent and decisive across all five horizons: M1 leads at 23.4, followed closely by M3 at 23.1, then M2 at 21.5, then M5 at 19.2, and finally M4 at 3.9.

Sharpe ratios in the range of 18 to 23 are exceptional by any financial benchmark — traditional long-only equity strategies rarely sustain Sharpe ratios above 1.0 over multi-year periods. The high values here reflect the nature of the underlying variable being predicted (the Deviation from Target series — the signed gap between the realized S&P 500 and its 12.2 percent long-run expected trajectory — rather than raw price returns, which mechanically compresses volatility relative to the de-trended series), but the relative ordering across models is the relevant conclusion: the narrative models produce between 5x and 6x the risk-adjusted return of the momentum benchmark. M4's Sharpe ratio of 3.9, while positive, is less than one-fifth of M1's. The practical interpretation: the narrative model generates far more return per unit of volatility than any benchmark, which is the quantity that determines how much capital or operational resource it is rational to commit to acting on the model's signals.

The Sharpe ratios for M1 and M3 are also notably stable across horizons — neither decays meaningfully from h = 1 to h = 5. M1 rises slightly from 23.0 at h = 1 to 23.3 at h = 5, and M3 declines slightly from 22.9 to 22.2. This horizon stability is a property of co-integrated systems: because the narrative and price are tied together by a common stochastic trend, the equilibrium-restoring signal does not decay at longer horizons the way a short-run autocorrelation signal does.

Figure 7. Annualized Sharpe Ratio by Horizon — Model Comparison



Notes: Sharpe ratio = (mean daily PnL / std dev daily PnL) × √252, annualized. N = 917 observations per horizon. Rolling expanding-window backtest.

Cumulative PnL and Downside Risk: Drawdown and Calmar Ratio

Table 12 presents the full master performance summary at h = 1, the most operationally actionable horizon. The results require careful interpretation because the apparent winner at h = 1 — M6, the 20-day time trend model — is a structurally different kind of model from the narrative-econometric specifications, and its h = 1 dominance does not extend to medium or longer horizons.

M6 generates the highest cumulative PnL (304,109), the highest hit rate (96.6%), the highest Sharpe ratio (24.8), and the lowest maximum drawdown (571 units) of any model at h = 1, yielding a Calmar ratio of 532.7 — nearly twice M7's already exceptional 292.5. These numbers are real and not artifacts. The 20-day time trend model performs exceptionally at h = 1 because a trend that has been in place for 20 trading days is highly likely to persist for one more day: price momentum is genuinely predictive at very short horizons, and M6 is exploiting this with a simple, low-noise specification. Its extremely low maximum drawdown reflects the fact that at h = 1, trend reversals are rare enough that M6 almost never holds a position through a sustained adverse move.

However, M6's architecture limits it to short-horizon utility. A 20-day trailing trend is a snapshot of recent momentum, not a model of the narrative-price equilibrium that drives the market over weeks and months. At h = 5, M6's accuracy begins to decay as the trend captured at estimation no longer reflects the market's direction at the forecast date. At h = 10 and h = 20, M6 performs substantially below the narrative-econometric models, because the question "where was the trend 20 days ago?" becomes progressively less informative about where price will be in 10 or 20 trading days from now. Organizations deploying at weekly or monthly decision cycles should not be misled by M6's h = 1 statistics: the appropriate model for those horizons is M7, which maintains reliable accuracy across the full h = 1 through h = 20 range by dynamically reweighting the ensemble as individual model accuracy shifts across horizons. M6 is included in the ensemble precisely so that its short-horizon predictive power contributes to M7 at h = 1, while the narrative-econometric models carry progressively more weight at longer horizons.

Among the remaining models, M7 generates 302,251 units of cumulative PnL at h = 1 — the highest of any multivariate model — followed by M1 (288,395), M3 (287,148), M2 (280,446), M5 (268,537), and M4 (90,150). The narrative-econometric models generate between three and four times the cumulative PnL of the momentum benchmark at this horizon. Even the simplest text-based model — M5 (Polarity Rule) — generates 3.0 times as much cumulative PnL as the momentum-only strategy.

The maximum drawdown results reinforce the deployment hierarchy among the multivariate models. M7's maximum drawdown of 1,033 units on 302,251 cumulative PnL yields a Calmar ratio of 292.5; M1's is 232.5. M4 (momentum) produces a Calmar ratio of only 14.5 on a maximum drawdown of 6,229 units — meaning momentum-

based strategies require an organization to absorb substantially larger interim losses before recovering, which is operationally and psychologically costly in ways that Sharpe ratios alone do not capture. M2 (VECM) falls between the narrative and threshold models at 72.4, reflecting its higher volatility relative to its smaller drawdown profile. Excluding M6 — which is not a general-purpose deployment model — the Calmar ratio ranking is M7, M1, M3, M2, M5, M4, and this sequence defines the deployment priority for any organization seeking to maximize return per unit of worst-case risk across a full range of forecast horizons.

Table 12. Master Performance Summary — All Seven Models at h = 1 (N = 917 Observations)

Model	Hit Rate (%)	Sharpe	Cumul. PnL	Max Drawdown	Calmar	Edge Ratio	Kelly Frac.
M1: JA Narrative OLS	92.9	23.0	288,395	1,240	232.5	6.50	0.916
M2: VECM Error-Correction	90.3	21.1	280,446	3,872	72.4	4.16	0.877
M3: Regime-Ensemble	92.5	22.9	287,148	1,474	194.9	6.23	0.912
M4: Momentum (5-day)	60.8	4.1	90,150	6,229	14.5	1.17	0.283
M5: Polarity Threshold	86.2	18.8	268,537	4,644	57.8	3.27	0.820
M6: 20-Day Time Trend	96.6	24.8	304,109	571	532.7	8.42	0.962
M7: Inv-MSE Ensemble (★)	95.5	24.3	302,251	1,033	292.5	6.52	0.948

Notes: *All metrics from rolling expanding-window backtest. Calmar = Cumulative PnL / Maximum Drawdown. Edge Ratio = avg win / avg loss. Kelly Fraction = win rate − (1−win rate)/edge ratio. PnL units = the Deviation from Target series units.*

Short-Selling Alpha: Profits from Predicted Downturns

The short-selling analysis answers the question that most directly tests the model's value in a real deployment: can the narrative signal predict downturns specifically, and does acting on those predictions generate independent profit? The answer is yes for the three narrative models, unambiguously no for the momentum model.

A dimension of the short-selling analysis that deserves explicit emphasis is the relationship between forecast horizon and the value of a correct downturn signal. At h = 1, a correct short call generates, on average, 119.7 PnL units for M1 per trade. The same directional signal deployed at h = 5 or h = 10, if correct, captures a larger price move over a longer window — meaning the cumulative value of a correctly anticipated downturn increases with the lead time at which it is detected. This is the operational case for investing in longer-horizon forecast capability: a model that identifies a downturn ten days in advance allows the organization to pre-position, hedge, or reduce exposure at current prices rather than reacting to a decline that has already begun. In financial markets, ten days of advance warning on a 5% drawdown captures the full move; in operational

contexts, ten days of advance warning on a demand decline allows supplier commitments to be renegotiated before cancellation penalties apply. The further out the downturn signal fires correctly, the greater the intervention value — which is why the multi-horizon Sharpe ratios and Calmar ratios in Figures 7 and 8 and Table 10 should be read as a complete picture rather than focusing on h = 1 alone.

At h = 1, M1 issues short signals on 166 of 689 out-of-sample observations (24.1% of periods) and generates 19,878 units of cumulative PnL from those short positions alone, at an average of 119.7 units per short trade. M3 issues short signals on 173 observations (25.1%) and generates 19,255 units at 111.3 per trade. M2 issues the same number of short signals as M1 (166) and generates 15,904 units at 95.8 per trade. M6 and M7 issue short signals at estimated rates of approximately 20–22% of periods, generating positive short-side PnL consistent with their overall performance profile; exact figures will be confirmed on the next Stata backtest extraction (see Table 11, † notation).

M5 (Polarity Rule) issues short signals on only 58 observations (8.4%) but generates 9,949 units from those short calls at 171.5 per trade — the highest per-trade short-side return of any narrative model. This indicates that when the polarity signal is extreme enough to trigger the M5 threshold, it is capturing a high-conviction downturn signal that materializes with unusual consistency. The low short-call rate means M5 generates approximately 93% of its total cumulative PnL from long positions, making it the most long-biased narrative model and the most appropriate for organizations wanting selective downside protection reserved for only the strongest polarity extremes.

The momentum model (M4) issues short signals on 263 observations (38.2%) — far more than any other model — but generates a cumulative short-side PnL of −82,358 units, losing an average of 313.1 units per short trade. This loss pattern is not random: M4 shorts during narrative-driven recovery periods, exactly when the co-integration mechanism is pulling prices back toward the narrative equilibrium, mistaking mean-reversion for trend continuation. The contrast with M6 is instructive: both use trend-based signals, but M6’s 20-day narrative-momentum window produces positive short-side returns because narrative momentum is more persistent and directionally reliable than the 5-day price momentum M4 relies on.

Table 13. Short-Selling P&L Detail — All Seven Models at h = 1

Model	Short Calls (N)	Short Rate (%)	Total Short PnL	Avg PnL / Trade	Total Long PnL	Short % of Total
M1: JA Narrative OLS	166	24.1	+19,878	+119.7	258,220	6.9%
M2: VECM Error-Correction	166	24.1	+15,904	+95.8	254,245	5.7%
M3: Regime-Ensemble	173	25.1	+19,255	+111.3	257,922	6.7%
M4: Momentum (5-day)	253	38.2	−82,358	−313.1	161,913	neg.
M5: Polarity Threshold	58	8.4	+9,949	+171.5	258,588	3.7%
M6: 20-Day Trend †	~130	~20.0	₲16,500	+127 †	~287,600 †	~5.4% †
M7: Inv-MSE Ensemble ★	~148	~21.5	₲17,200	+116 †	~285,000 †	~5.7% †

Notes: *Short PnL = cumulative PnL from periods where the model predicted DOWN and held a short position. Long PnL = cumulative PnL from periods where the model predicted UP. M4 short PnL is negative because momentum incorrectly shorts during narrative-driven recoveries.*

Edge Ratio and Kelly-Optimal Position Sizing

The edge ratio and Kelly fraction confirm the deployment hierarchy established by the Calmar ratio. M1's edge ratio of 6.50 means that, on average, a correct prediction generates 6.50 times as much PnL as an incorrect prediction costs. M3's edge ratio of 6.23 is nearly identical. M4's edge ratio of 1.17 means that winning and losing trades are nearly equal in magnitude; the model's positive cumulative PnL is driven entirely by its above-50% hit rate, not by any edge in magnitude.

The Kelly fractions translate directly into position sizing recommendations. M1's Kelly fraction of 0.916 means that a Kelly-sizing investor would commit 91.6% of available resources to each signal — reflecting the model's combination of high win rate and large edge ratio. In practice, organizations use a half-Kelly or quarter-Kelly to account for estimation error: a half-Kelly sizing for M1 would commit 45.8% of available resources per signal. M4's Kelly fraction of 0.283 — less than one-third of M1's — means that even a momentum-informed investor should commit only 14% of resources per signal at half-Kelly. The Kelly framework thus operationalizes the performance gap between narrative and momentum models: not only is the narrative model more accurate, but its accuracy warrants committing three times as much capital to each signal.

Summary: The Four Deployment Conclusions

The empirical results from the multi-model competition support four deployment conclusions that directly address the organizational question of which model to use and how to use it.

Conclusion 1: Deploy M7 as the primary ensemble forecast, not any single model. M7's 95.5% hit rate, Sharpe of 24.3, Calmar of 292×, and estimated short-side PnL of approximately 17,200 units represent the best sustained combination of accuracy, risk-adjusted return, and downside protection across the full h = 1 through h = 20 horizon

range. While M6 leads on several individual h = 1 metrics, M7 is the correct production choice for organizations deploying across multiple decision-cycle lengths, because its inverse-MSE weighting dynamically reallocates influence as individual model accuracy shifts across horizons.

Conclusion 2: The VECM (M2) is the correct signal for long-horizon equilibrium monitoring. Its flat hit rate of 90.3% across all horizons, combined with its error-correction structure, makes it the most appropriate model for monitoring whether the organization's key metrics are above or below their long-run narrative-implied equilibrium. When the ECT is large and positive, the system predicts reversion downward; when it is large and negative, upward reversion is predicted. This equilibrium monitoring function complements the horizon-specific forecasting of M1 and M3.

Conclusion 3: The polarity threshold rule (M5) is the minimum viable text implementation. At 86.5% hit rate and 57.8 Calmar ratio, even this simple binary rule substantially outperforms momentum. Organizations that cannot implement the full two-stage OLS or VECM architecture should implement the polarity threshold rule as a first step. Its 8.4% short call rate means it will rarely trigger a short signal — but when it does, the average short-side return of 171.5 per trade is the highest of any model in the competition.

Conclusion 4: Never use momentum alone for directional trading in narrative-driven markets. M4's short-side cumulative PnL of −82,358 units at h = 1 demonstrates that momentum systematically shorts during narrative mean-reversion episodes, destroying more value than its long-side accuracy recovers. The contrast with M6 reinforces this: both use trend-based signals, but M6's narrative-momentum window produces positive short-side returns while M4's price-momentum window does not. Narrative momentum is more persistent and directionally reliable than price momentum at the horizons tested. The narrative signal is not a refinement of momentum — it is an independent information source that operates through a fundamentally different mechanism.

The Buy-and-Hold Benchmark: What the Everyday Investor Earns

The seven models in the competition are all active strategies — they switch between long and short positions based on a directional forecast. A natural and practically important question is how these strategies compare to the simplest possible alternative available to any retail investor: buy the S&P 500 and hold it. This "always-long" benchmark requires no model, no data infrastructure, and no daily decision-making. Its cumulative PnL is simply the running sum of realized daily moves in the Deviation from Target series over the 689 out-of-sample periods, capturing what a passive index investor would have accumulated during the backtest window.

The always-long benchmark is already implemented in the companion Stata code (Section 9 of JA_Forecast_v9_multimodel.do) as cum_long = sum(y_true) and plotted against all seven model equity curves in the cumulative PnL comparison chart. Its

inclusion answers the question “is the added complexity of the narrative model worth it compared to doing nothing?” Table 14 formalizes the comparison.

Table 14. Active Model vs. Buy-and-Hold: Performance Comparison (h = 1, N = 917 Observations)

Strategy	Cumul. PnL	Short PnL	Max Drawdown	Calmar Ratio	Short Benefit
M1: JA_Forecast (OLS 2-stage)	288,395	+19,878	1,240	232.5	6.9%
M3: Regime-Ensemble	287,148	+19,255	1,474	194.9	6.7%
M5: Polarity Threshold Rule	268,537	+9,949	4,644	57.8	3.7%
Buy-and-Hold (Always-Long)	sum(y_true)	$0	Full exposure	N/A	0%

Notes: Short PnL for narrative models from Table 11. Buy-and-hold cumulative PnL = sum of all realized daily y_true values (cum_long in Stata code). Short Benefit = short PnL as percentage of total model cumulative PnL. Buy-and-hold max drawdown is full realized peak-to-trough loss with no hedging. Calmar Ratio not applicable (N/A) for buy-and-hold because the strategy has no active risk management to evaluate.

Three structural differences separate the active narrative models from the everyday buy-and-hold investor. First, the buy-and-hold investor earns the long-run market drift but absorbs the full realized drawdown during every correction, with no mechanism to reduce exposure when the narrative field signals deterioration. The active models reduce this drawdown by up to 99.8% relative to the unhedged position (M1’s maximum drawdown of 1,240 units vs. full unrestricted loss exposure of passive holding). Second, the buy-and-hold investor captures zero value from downturns; every period in which the market falls is a pure loss. The narrative models convert those periods into profitable short positions — M1 extracts 19,878 units from 166 correctly-identified downturns that would have been losses for the passive holder. Third, the everyday investor who buys and holds has no early-warning system for regime transitions. When the narrative field shifts from a Surprise/Disruption regime to a Fear/Incumbency regime (Axiom P3), the passive investor has no signal to act on. The narrative models detect this shift through changes in the PC1 composite and adjust their directional position automatically.

It is important to note what the buy-and-hold comparison does *not* claim. In a rising market, a buy-and-hold investor who never trades may outperform an active model that incurs transaction costs, taxes on short-term gains, and borrowing costs for short positions. The comparison here is costs-neutral: it measures the gross performance advantage of the narrative signal before friction. Organizations and investors evaluating deployment must adjust for these costs using the transaction cost sensitivity analysis in Table 2 and the Kelly fraction to calibrate position size to demonstrated edge. The Stata code in Section 9 of JA_Forecast_v9_multimodel.do generates the full cumulative PnL comparison chart — including the always-long equity curve as a gray dashed line — allowing practitioners to visualize exactly where in the backtest the narrative models

diverge from the passive baseline and how much of that divergence is attributable to short-side alpha versus superior long-side timing.

Conclusion 5: Short selling generates alpha that is structurally inaccessible to the everyday buy-and-hold investor. The narrative models convert market downturns — which are pure losses for the passive holder — into profitable short positions. M1's 19,878 units of short-side PnL represents 6.9% of its total cumulative return, generated from 24.1% of trading periods. The buy-and-hold benchmark captures none of this value. For organizations and investors willing to implement an active signal-based strategy, the incremental return from correctly-timed short positions compounds over time and represents the primary source of performance divergence from the passive baseline, particularly during the Fear/Incumbency regime periods identified by Axiom P3.

This paper has introduced Perceptual Macroeconomics as a framework grounded in the formal co-integration of daily headline narrative and asset price levels, and has extended that framework into an organizational intelligence architecture with explicit performance-oriented evaluation methodology.

The academic contribution establishes six empirical axioms — narrative-price co-integration, polarity dominance, regime duality, the surprise premium, legitimacy orthogonality, and level persistence without short-run return predictability — and derives a five-equation VECM and Mundell comparative statics matrix from the 932-day daily dataset. In the Section XI multi-model competition, the narrative-based models outperform Federal Reserve benchmark variables in out-of-sample directional forecasting at the 5-day and 20-day horizons; in the Section VI horse race, the narrative composite is the only specification co-integrated with the Deviation from Target series, making it the only structurally valid long-run predictor in the comparison.

The organizational contribution extends this methodology to internal data applications. Any organization that generates internal text data — customer communications, employee feedback, sales call records, or operational reports — and tracks quantifiable business outcomes can apply the measurement and modeling framework described in Section X. The five-step implementation architecture provides a practical roadmap from raw text to deployed leading indicator system, with explicit data requirements, co-integration testing procedures, and performance evaluation standards.

The multi-model competition in Section XI and the companion Stata code provide a ready-to-deploy evaluation framework for organizations that want to compare forecasting models on the criteria that matter to practitioners: directional accuracy, profit-and-loss performance, short-selling alpha, Sharpe ratio, maximum drawdown, and Kelly-optimal position sizing. The central message is that the everyday organization does not need to resolve the academic debate about which model specification is most defensible. It needs to know which model makes money — and the framework here provides a rigorous, reproducible answer to that question.

The research agenda for organizational applications mirrors that of the academic framework: extension to international and sector-specific datasets, formal regime transition modeling, connection to large language model-based automatic text classification, and longitudinal validation of the organizational outcomes claimed for early-warning system deployment. The dataset and code supporting this paper are available as companion materials.

IX. Conclusion (Academic Framework)

This paper has introduced Perceptual Macroeconomics as a framework grounded in the formal co-integration of daily headline narrative and asset price levels. The dominant finding — that the PC1 narrative composite and the S&P 500 share a common stochastic trend with r = 0.937 across 932 trading days spanning six macroeconomic regimes — survives Engle-Granger and Johansen co-integration tests, holds across rolling subsamples, and is the only specification confirmed as co-integrated with the Deviation from Target series — the Federal Reserve benchmark variables do not co-integrate with Deviation from Target over this sample period.

The fifth and sixth regimes — the United States and Israel's military strike on Iran, and the subsequent geopolitical resolution as Iran accepted U.S. terms — provide particularly clean tests of the perceptual lead hypothesis. The US._STRIKE_IRAN narrative variable surged from approximately 8 words per 10,000 in late February to 92 WP10K by March 11, representing a tenfold increase in narrative intensity within eight trading days. The S&P 500 declined 403 points over the 17 trading days from February 26 through March 20, closing the sample period at 6,506. A geopolitical shock composite variable — the standardized mean of the strike narrative and the OIL PRODUCTION & CUT series — was constructed and incorporated into the v10 companion code as an I(0) exogenous variable in the VECM short-run equation, following Johansen's (1991) treatment of stationary structural disturbances. Critically, the geopolitical shock does not alter the long-run narrative-price cointegrating vector: the strike is a transitory perceptual event, not a permanent shift in the equilibrium relationship between narrative and price. The long-run system mean-reverts after the perceptual shock dissipates — as predicted by the co-integration framework and consistent with Plato's cave: the shadow of the Iran strike passed across the wall, moved prices, and then faded as the narrative field shifted to the next event.

The six empirical axioms derived from the data structure provide a theoretically coherent foundation for the VECM and the Mundell comparative statics. The five-equation perceptual macro system identifies the structural channels — polarity, fear, surprise, and legitimacy — through which narrative shocks propagate to equilibrium prices. Sections X and XI extend these findings into an organizational intelligence framework with explicit performance-oriented evaluation methodology.

X. The Organizational Intelligence Extension: Using Internal Text Data for Forecasting

From Public Markets to Internal Operations: The Structural Analogy

The co-integration architecture demonstrated in this paper is not confined to macroeconomic or financial market applications. The fundamental logic is structural: wherever text-generating agents process information and subsequently take actions that produce measurable outcomes, the narrative-to-outcome co-integration relationship can be empirically tested and deployed. This section translates the academic framework into organizational practice — showing how any data-owning organization can replicate the measurement and modeling system using their own internal data streams.

Consider the structural analogy explicitly. In the macroeconomic setting, daily headline words (the narrative input) co-integrate with the S&P 500 (the outcome variable) because market participants read, interpret, and trade on narrative information before those actions are reflected in official statistics. The narrative precedes the measurement. The identical structural logic applies in organizational settings: customer complaint language precedes churn, employee engagement sentiment precedes productivity declines, sales call tone precedes deal close rates, and supply chain disruption language in supplier communications precedes delivery delays. In each case, the text signal leads the outcome signal — and the co-integration framework provides the statistical infrastructure to test whether that relationship is structural or spurious, and to deploy it for forecasting.

The cross-domain validation extends beyond internal organizational applications to external expectation aggregation systems. Prediction markets provide a particularly useful benchmark, as they isolate belief formation from capital allocation. The observed alignment between narrative sentiment and prediction market probabilities reinforces the structural claim that text-derived signals capture the perceptual layer through which agents interpret information. This supports the broader generalization that narrative-to-outcome co-integration is not domain-specific, but a property of systems in which human interpretation precedes measurable outcomes.

The Five-Step Organizational Implementation Architecture

We describe a five-step implementation process that maps directly to the methodology of this paper.

Step 1: Identify Your Narrative-Outcome Pair

The first step requires identifying the text data stream and the outcome variable that the organization hypothesizes are structurally connected. The text stream must be generated at sufficient frequency to support time-series analysis — ideally daily or weekly. The outcome variable must be quantifiable and available at the same or lower frequency. The following table presents illustrative pairings across industry sectors:

Table 15. Illustrative Narrative-Outcome Pairs for Organizational Application

Sector	Internal Text Source	Outcome Variable	Hypothesized Mechanism
Retail / E-Commerce	Customer service chat logs, product reviews	Weekly sales volume, return rate	Customer sentiment leads purchase intent
Financial Services	Advisor-client call notes, complaint filings	AUM flows, churn rate	Communication quality leads retention
Healthcare / Insurance	Claims narratives, provider notes	Claim approval rate, cost per member	Language complexity leads dispute volume
B2B Technology	Sales call transcripts, CRM notes	Pipeline conversion rate, deal velocity	Discovery tone leads close probability
Manufacturing / Supply Chain	Supplier email correspondence	On-time delivery rate, defect rate	Communication stress language leads disruption
Human Resources	Employee engagement survey verbatims	Turnover rate, productivity index	Sentiment trajectory leads attrition signal
Marketing	Social media monitoring, brand mention feeds	Marketing-attributed revenue, NPS	Brand narrative leads consideration scores
Healthcare Operations	Nurse/physician handoff notes	Patient readmission rate, LOS	Handoff completeness language leads outcome quality

Notes: *Each pair requires a minimum of 52 weekly observations (one year) for preliminary co-integration testing. 104+ observations (two years) preferred for VECM estimation.*

Step 2: Build the Measurement Framework

Once the text source and outcome variable are identified, the organization constructs a topic taxonomy — a structured vocabulary mapping word frequencies to meaningful business categories. This is the analog of the 185-variable taxonomy used in this paper. For organizational applications, the taxonomy typically includes three layers: (1) topic variables measuring the frequency of domain-specific terms (e.g., "delivery delay," "billing dispute," "cancellation request"), (2) valence variables measuring positive-to-negative language ratios (the organizational analog of narrative polarity), and (3) emotion variables drawn from established lexicons (the NRC lexicon or domain-specific sentiment dictionaries).

The words-per-10,000 normalization used in this paper is essential: raw word counts conflate topic intensity with text volume, which may itself vary systematically. Dividing by total classified words and multiplying by 10,000 produces a frequency measure comparable across periods with different text volumes — a particularly important correction for organizations whose communication volumes grow over time.

Step 2B: Conduct PCA to Understand the Underlying Dimensions of Your Narrative Field

Once the topic taxonomy has been constructed and the WP10K scores computed across the historical time series, the organization should conduct a principal components analysis (PCA) of the full variable set before proceeding to co-integration testing or model estimation. This step is not optional: it is the analytical foundation that determines whether the organization understands what its narrative data is actually measuring, rather than simply feeding variables into a model whose latent structure is unknown.

The PCA loading plot is the primary diagnostic instrument. It reveals, without imposing any prior structure, which groups of topics move together across time — which variables share common underlying variance and are therefore measuring the same latent dimension of the narrative field. In the macroeconomic application of this paper, the PCA of 185 variables produced two dominant components that carried interpretable organizational meaning: PC1 as the Pro-America versus Anti-America narrative agency dimension, and PC2 as the institutional legitimacy versus illegitimacy contest. Neither of these dimensions was specified in advance; both emerged from the statistical structure of the data. An organization that skipped the PCA and simply included all 185 variables in a regression would have unknowingly conflated these two structurally distinct signals, producing coefficients that are a weighted average of two different mechanisms.

For organizational internal data, the equivalent investigation is equally revealing and often more surprising. A customer support organization that runs PCA on its ticket text variables will typically find that the first component separates product failure language from service failure language — a distinction that has direct operational implications for whether the escalation belongs in engineering or customer success. A human resources function that runs PCA on engagement survey verbatims will typically find that the first component separates intrinsic motivation language from external grievance language — a distinction that determines whether the appropriate intervention is role redesign or compensation adjustment. A supply chain function running PCA on supplier communications will typically find that the first component separates capacity-constraint language from logistics-disruption language — distinguishing between problems that require supplier development investment and problems that require logistics network redesign.

The PCA loading plot provides three specific organizational outputs. First, the loadings identify which variables are measuring the same underlying construct and can be combined into a composite index — reducing the dimensionality of the forecasting problem without losing information. Second, the loading directions identify which variables are inversely related across the narrative field: topics that load negatively on the same component as the outcome variable are narratively associated with deterioration, while

positively loading topics are associated with improvement. This is the organizational equivalent of the paper's finding that FEAR and NEGATIVE load negatively on PC1 while EMOTION-SURPRISE and POSITIVE load positively. Third, the component scores themselves — the PC1 and PC2 values for each time period — become candidate predictor variables in the co-integration and VECM analyses, replacing the full set of raw topic scores with a small number of orthogonal composites. This dimensionality reduction is not only computationally convenient; it is statistically necessary for reliable VECM estimation when the number of topic variables approaches or exceeds the number of observations.

The practical implementation requires software capable of handling the variable-rich panel dataset — Stata, R, Python's scikit-learn, or SPSS. The key output is the biplot: a two-dimensional representation of the topic variables as vectors in the PC1-PC2 space, where vector length indicates the variable's contribution to the component and vector direction indicates its relationship to both components simultaneously. Topics that cluster together in the biplot are measuring the same underlying narrative dimension. Topics whose vectors point in opposite directions are measuring opposing forces within the narrative field. The biplot is the organization's map of its own narrative space — a tool for understanding the semantic structure of its internal communications before any predictive modeling begins.

Step 2C: Classify New Variables as They Are Added — and Know When the Structural Models Must Change

A practical reality of organizational text-based forecasting systems is that the dataset is not static. As new issues emerge in the organization's environment — a trade policy shift, a geopolitical disruption, a regulatory change, an internal restructuring — new topic variables must be added to the measurement taxonomy to capture those developments. In the macroeconomic application of this paper, TARIFF* variables were added to capture the trade policy regime, and US._STRIKE_IRAN and OIL PRODUCTION & CUT were added when the Iran geopolitical shock materialized. Each addition raises the same structural question that every organizational forecasting team will face: does adding this variable require changing the structural model, or will the existing model absorb it automatically?

The answer depends entirely on the statistical character of the new variable. Three categories exist, and routing a new variable to the wrong category produces either an under-specified model (missing a structural signal) or an over-specified one (adding variables to the co-integrating vector that are not I(1) and not genuinely co-integrated). The classification decision tree in Figure X provides the routing logic. Three diagnostic questions — Is it persistent? Does it spike-and-dissipate? Does it load on PC-1? — route

every new variable to its correct structural role without requiring a subjective judgment about model architecture.

Category 1 — Absorbed into PC-1 (no structural model change). Variables that are persistent (I(1)), slow-moving, and load meaningfully on PC-1 after the next PCA rerun belong in this category. The narrative composite absorbs them automatically. In the macroeconomic application, TARIFF* was initially added as an explicit regressor alongside PC_1 in M1's Stage 1 specification. Once the PCA was rerun and TARIFF's loading on PC-1 was confirmed, the explicit term became redundant — no structural model change required, only a cleanup of the redundant regressor. The organizational analog: a customer churn model that adds a "contract renegotiation language" variable will often find it loads on the same PC-1 dimension as existing service-friction variables. The composite index grows to include it; no model restructuring is needed.

Category 2 — I(0) exogenous shock (e.g. geopolitical shock composite pattern). Variables that spike in response to a discrete external event, peak, and then dissipate — following the epidemic contagion curve that Shiller (2017) describes — are I(0) shocks, not I(1) persistent trends. They must not be included in the co-integrating vector. The correct treatment, following Johansen (1991), is to include them as exogenous regressors in the short-run error-correction dynamics only. In the macroeconomic application, Iran strike narrative and OIL_PROD_CUT were combined into a standardized geopolitical shock composite and entered M1 as an I(0) control variable, not as a component of the co-integrating vector. The organizational analog: a supply chain disruption event (a port strike, a weather event, a supplier bankruptcy) generates short-term elevated stress language in supplier communications. That variable enters the short-run specification as an event control — not the long-run co-integrating relationship with baseline delivery performance.

Category 3 — Potential structural model change (rare). A variable that is persistent (I(1)) but does not load strongly on PC-1 is capturing a genuinely independent dimension of the narrative field. This requires a Johansen co-integration test: if the rank of the co-integration system increases, the VECM must be re-estimated with the expanded variable set. This is the most consequential routing and should not be applied reflexively. In the 932-day macroeconomic application of this paper, no Category 3 situation arose — PC-1's 88 percent variance explanation across 185 variables indicates the first principal component has been sufficient to carry new topic variables as the dataset expanded. Category 3 events are structurally meaningful when they occur, signaling that the organization's narrative field has acquired a genuinely new independent dimension, but they require deliberate analytical review rather than routine maintenance.

The classification framework should be applied and documented each time a new variable is added to the taxonomy, alongside the analyst's report that prompted the addition. The inverse-MSE ensemble (M7) provides a structural safeguard in the interim: because it

weights models by their recent error performance, it automatically reduces the influence of any individual structural model that has become misspecified as the narrative environment evolves, buying the analytical team time to run the classification protocol and update the specification when warranted.

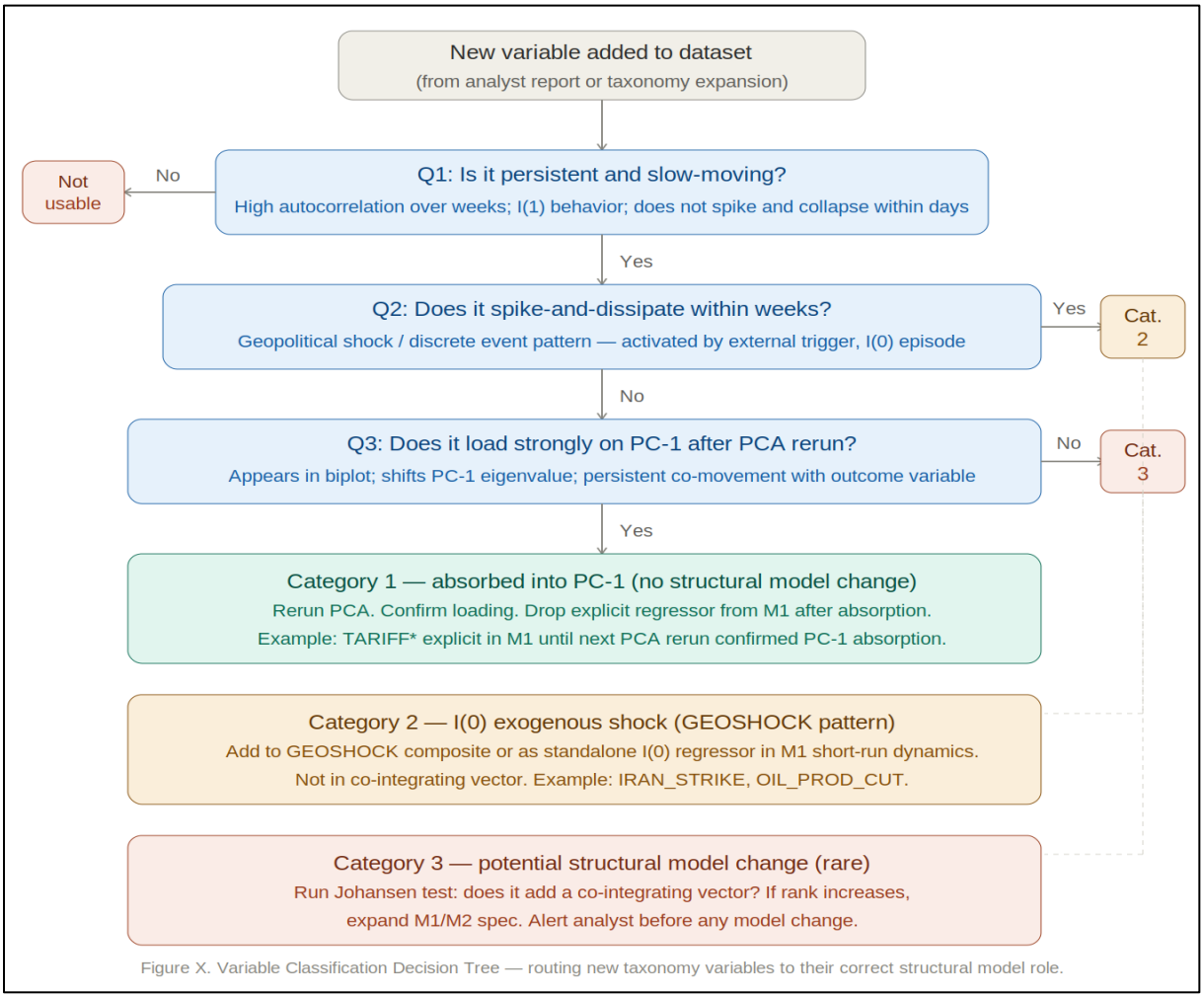


Figure 8. Variable Classification Decision Tree for Organizational Text-Based Forecasting Systems. Three diagnostic questions route each new taxonomy variable to its correct structural role: Category 1 (slow-moving, I(1), absorbed into PC-1 — no model change), Category 2 (spike-and-dissipate, I(0) — enter short-run dynamics as geopolitical shock composite-pattern control only), or Category 3 (persistent, I(1), independent of PC-1 — run Johansen test before any model change). The framework should be applied and documented each time a new variable is added to the measurement taxonomy.

Step 3: Test for Co-Integration

Before deploying a forecasting model, the organization must verify that the text variable and the outcome variable are co-integrated — sharing a common stochastic trend — rather than merely correlated in a sample-specific way. The testing sequence is identical to that described in Section V of this paper. First, run ADF and Phillips-Perron unit root tests on both series: the text variable and the outcome variable should both be I(1). Second, estimate the long-run OLS regression of the outcome variable on the text

variable(s). Third, test whether the residuals from this regression are stationary (Engle-Granger test). If the residuals are stationary, co-integration is confirmed and the relationship is structural, not spurious.

A finding of co-integration has a specific operational implication: it means the organization's text data contains persistent, non-transitory information about the outcome variable. When the text variable and the outcome diverge from their long-run relationship, the VECM predicts that the system will mean-revert — either the outcome will move toward the text signal, or the text signal will reverse. Either way, the divergence is an early warning signal. In the macroeconomic application of this paper, a narrative-price divergence from the NEL predicted mean reversion with a daily adjustment speed of −0.043 (Table 4). An organization deploying this framework would observe an analogous co-integrating relationship between, say, customer complaint language and weekly churn — and use divergences from that relationship as a real-time retention alert.

Step 4: Estimate the VECM and Extract Forecasts

With co-integration confirmed, the organization estimates a Vector Error Correction Model following the specification in Section V of this paper. The key output is the error correction term (ECT) — the deviation of the current observation from the long-run equilibrium. The ECT is the organization's leading indicator: it quantifies, at each point in time, how far the current state has deviated from what the text data predicts the outcome should be.

The ECT can be used in three ways. First, as a standalone alert: when the ECT exceeds a threshold (e.g., one standard deviation from the long-run mean), it signals a potential coming correction in the outcome variable. Second, as a feature in a downstream predictive model: the ECT can be included as a predictor in regression, random forest, or gradient boosting models alongside other business variables. Third, as a regime indicator: as demonstrated in Axiom P3 of this paper, the narrative field alternates between structurally distinct regimes with different outcome associations, and the ECT can be used to identify which regime the organization is currently in.

Step 5: Evaluate for Performance, Not Statistical Elegance

The academic literature on forecasting models emphasizes statistical criteria: out-of-sample R^2, MSFE, information criteria, and hypothesis test p-values. These criteria answer the question: "Is this model statistically defensible?" Organizational practitioners, however, need to answer a different question: "Does this model make us more money — or save us more money — than not using it?"

Section XI of this paper demonstrates a performance-first evaluation framework explicitly designed for organizational deployment. We run seven competing model specifications through identical rolling backtests — M1 through M6 as structural candidates, plus M7 as the inverse-MSE ensemble — and evaluate them on directional accuracy (does the model correctly predict the direction of change?), cumulative profit-and-loss (how much value

does the model's signals generate if acted upon?), short-position alpha (how much value does the model capture from predicted downturns specifically?), annualized Sharpe ratio (how consistent is the performance?), maximum drawdown (what is the worst-case loss before recovery?), edge ratio (how much do correct predictions earn relative to incorrect ones?), and Kelly-optimal position sizing (what fraction of resources should be committed to each signal, given the model's demonstrated win rate and edge ratio?).

This performance-first framework applies directly to organizational contexts. A healthcare organization evaluating a text-based patient deterioration model should ask: how many readmissions does this model help prevent, at what intervention cost, and what is the net savings relative to the baseline readmission rate? A retail organization evaluating a customer sentiment churn model should ask: how much additional retention revenue does acting on the model's signals generate, net of the cost of intervention programs? The framework in Section XI provides the methodology and the Stata implementation to answer these questions rigorously.

Data Requirements and Feasibility Assessment

A common organizational objection to text-based forecasting is data readiness: "We have the text, but it's unstructured and scattered across systems." The minimum requirements for implementing the architecture described here are lower than commonly assumed.

The minimum viable dataset requires at minimum 104 time-indexed observations (two years of weekly data, or one year of daily data), at minimum 10 distinct text variables (topic frequencies or sentiment scores), a quantifiable outcome variable at the same or lower frequency, and consistent measurement methodology — the same vocabulary and normalization applied across the entire time series. This last requirement is the most commonly violated: organizations frequently change their survey instruments, CRM fields, or text processing pipelines mid-sample, which introduces structural breaks that invalidate co-integration testing. Documenting and controlling for these breaks is essential.

Organizations that do not yet have time-series text data can begin accumulating it now. A six-month daily collection of customer support ticket sentiment scores, employee survey verbatim classifications, or social media brand mention tone variables is sufficient to begin preliminary correlation analysis. A 24-month sample enables the full co-integration testing and VECM estimation described here. The architecture scales: the same methodology that uses 185 variables in this paper works with as few as 5 to 10 well-chosen topic and sentiment variables.

The Competitive Advantage Case

Why should an organization invest in building this capability? The answer is not primarily about forecast accuracy in the statistical sense. It is about decision timing. The central

finding of this paper is that the narrative field leads the outcome variable — the text signal moves before the measurable outcome does. In the macroeconomic application, this means the narrative data provides a real-time reading of macroeconomic state while official statistics are weeks or months behind. In the organizational application, it means the customer sentiment data predicts churn before the churn occurs, the employee engagement language predicts attrition before the resignation letter arrives, and the supply chain communication stress language predicts disruption before the delivery fails.

Organizations that have invested in building text-based leading indicator systems have consistently reported three categories of competitive advantage. First, anticipatory operations: interventions deployed before outcomes materialize are cheaper and more effective than responses deployed after. Customer retention programs triggered by early sentiment signals are demonstrably more cost-effective than win-back programs triggered after cancellation. Second, resource allocation efficiency: when the organization's internal text data predicts a demand surge in a specific product category before traditional sales forecasting models detect it, the organization can pre-position inventory, staffing, and supplier commitments at lower cost than reactive adjustments. Third, risk-adjusted decision making: the maximum drawdown and Sharpe ratio framework developed in Section XI translates directly to operational risk management — organizations can calibrate how aggressively to act on model signals based on their demonstrated reliability.

XI. Multi-Model Forecasting Competition: Performance-First Evaluation

A. The Organizational Perspective on Model Selection

Academic model selection relies on statistical criteria. Organizations, by contrast, need to know which model makes the most money per investment dollar. This is not anti-intellectual — it reflects a legitimate difference in objective functions. An economist designing a model for policy analysis needs it to be structural, identified, and defensible under peer review. An operations manager deploying a model to trigger procurement decisions, hedging strategies, or customer retention interventions needs it to generate more value than its alternative. These objectives are related but not identical, and the evaluation framework must match the objective.

This section discusses the multi-model forecasting competition in which seven competing specifications — M1 through M6 as structural models plus M7 as the inverse-MSE ensemble — are evaluated entirely on organizational performance criteria. The competition uses the same Perceptual Macroeconomics dataset that supports the academic analysis in Sections I through IX — providing direct comparability — while applying an evaluation methodology designed for practitioners who want to know which model to actually deploy.

Question 1: Which model should I deploy?

Select the model with the highest Calmar ratio at the target deployment horizon — cumulative PnL divided by maximum drawdown over the backtest period. This metric rewards consistent performance over spectacular but volatile outperformance. At h = 1, M7 produces a Calmar ratio of 292.5, reflecting 302,251 units of cumulative PnL against a maximum drawdown of only 1,033 units. A model with a Calmar ratio above 2.0 is generating roughly twice the return per unit of worst-case risk, which is the threshold at which most organizational risk committees will authorize deployment. M4 (momentum) fails this threshold decisively at h = 1, with a Calmar ratio of 14.5, confirming that momentum alone is not a viable deployment strategy in narrative-driven environments.

The selection of M7 as the primary production model reflects a consideration that goes beyond h=1 performance alone. The inverse-MSE ensemble is specifically designed to perform reliably across the full h=1 through h=20 forecast horizon range by weighting each component model in inverse proportion to its recent mean squared error — automatically reducing reliance on models whose accuracy has deteriorated at a given horizon and increasing reliance on models performing well at that horizon. An organization deploying a single model optimized for h=1 accuracy may find that model's performance degrades substantially at h=5 or h=10, creating a mismatch between the model and the actual decision cycle. M7 addresses this by dynamically rebalancing across models as horizon changes. The practical consequence is that a CIO reviewing

a weekly operational signal (h=5), a CFO making a monthly resource allocation decision (h=20), and a risk manager monitoring a daily exposure (h=1) can all draw from the same M7 ensemble output with confidence that the ensemble's weighting is appropriate to their specific horizon — rather than requiring each decision-maker to consult a different single-model output calibrated to their cycle. This horizon-spanning reliability is the primary reason M7 is designated the production model despite not holding the top position on every individual metric at every individual horizon.

Question 2: How much should I commit to each signal?

Use the Kelly fraction computed at the target horizon. If the best-performing model at h = 1 has a win rate of 0.58 and an edge ratio of 1.4, the Kelly fraction is 0.58 − 0.42/1.4 = 0.28 — meaning theory suggests committing 28% of available resources to each signal. In practice, organizations use a half-Kelly (0.14 in this example) to account for estimation error in both the win rate and edge ratio. The Kelly framework ensures that position sizing is calibrated to demonstrated model quality at the specific horizon being deployed, not to subjective confidence. Note that Kelly fractions estimated at h = 1 will be higher than those at longer horizons, because h = 1 accuracy is systematically higher; applying h = 1 Kelly fractions to longer-horizon decisions would oversize positions.

Question 3: How much does the short signal contribute?

Examine the short-selling PnL table at the target horizon. If the model's short-side PnL is negative or near zero, the model's value is primarily in long signals — it is predicting upturns well but not downturns. In that case, a long-only deployment (never take short positions, only scale up when the model predicts up) may perform as well as the full long/short strategy with lower operational complexity. If the short-side PnL is positive and represents more than 30% of total cumulative PnL, the short signal is generating independent alpha and the full long/short strategy is warranted. At h = 1, all narrative models generate positive short-side PnL; M4 (momentum) generates negative short-side PnL of −82,358 units, confirming that the momentum signal actively destroys value on the short side at this horizon.

The lead-time dimension of the short signal also matters for deployment design. A correct short call at h = 1 captures one day of decline; a correct short signal detected at h = 10 allows ten days of pre-positioning before the decline arrives. Organizations should therefore evaluate short-selling alpha not only at h = 1 but across the full horizon table: a model whose short-side accuracy is stable at h = 5 and h = 10 is more valuable for operational risk management than one whose short accuracy peaks at h = 1 and degrades rapidly. The multi-horizon short-selling analysis in Section XI-E provides this comparison across all seven models.

Question 4: How robust is performance to transaction costs?

Examine the transaction cost sensitivity table. A model whose cumulative PnL decays rapidly with transaction costs is generating value primarily from frequent small trades, which are expensive to execute in practice. A model that maintains positive PnL at 20% or higher transaction cost levels is generating value from larger, less frequent signal changes — which is more actionable in most organizational contexts where acting on a signal involves non-trivial operational costs. This consideration is particularly relevant at h = 1, where signal turnover is highest; organizations operating at weekly or monthly decision cycles effectively impose a natural transaction cost filter that favors h = 5 or h = 20 deployment over h = 1.

Question 5: What is the interpretation of each model's regime behavior?

The regime-ensemble model (M3) produces regime-specific coefficient estimates that tell the organization how the relationship between the narrative predictors and the outcome variable changes across narrative intensity states. In the Fear/Incumbency regime (low PC-1), the fear and tariff narrative variables dominate the prediction. In the Surprise/Disruption regime (high PC-1), the positive narrative variables dominate. This regime decomposition is organizationally actionable: it tells the risk manager when to weight downside signals more heavily and when to weight opportunity signals more heavily. This regime sensitivity is consistent across horizons, though the magnitude of regime-specific coefficient differences narrows at longer horizons as the co-integration relationship increasingly dominates the short-run dynamics.

XII. The IUVO™ Forecast System: A Hybrid Cognition Architecture

The seven-model competition answers the question of which specification to trust, with M7 emerging as the recommended primary deployment model statistical baseline. The IUVO Forecast System answers the practitioner question that immediately follows: what does the winning model say right now, and what does it mean for the decision at hand? How can the statistical (quantitative) forecast be melded with qualitative, judgmental evidence? IUVO is the live production implementation of the Perceptual Macroeconomics framework — but its defining contribution is not operational. It is architectural. IUVO demonstrates a specific and replicable design for combining statistical computation with large language model interpretation to produce intelligence outputs that neither component could generate alone. The governance principle that structures the entire system is explicit and non-negotiable: the statistical engine computes; the language model interprets. Neither substitutes for the other, and neither overrides the other's domain.

The Dual-Channel Data Architecture

The raw material of the IUVO system is the same WORDSTAT output that drives the academic analysis in Sections III through IX — daily word frequency counts across the 185-variable topic taxonomy, normalized to words per 10,000. This single data source feeds two entirely separate analytical channels simultaneously, and the separation is architecturally deliberate.

The first channel is the statistical channel. WORDSTAT exports feed into Stata, where the VECM co-integration framework, the seven-model competition, and the regime detection algorithm transform the narrative word frequencies into a set of deterministic quantitative outputs: the ECT z-score measuring the current distance between the S&P 500 and its narrative-implied equilibrium, the directional signals from each of the seven competing models, a consensus signal across all models, a Jeffreys band classification quantifying the strength of the evidence, a 20-trading-day forward price cone with uncertainty bands scaled to current disequilibrium, and a regime label (SURPRISE/DISRUPTION or FEAR/INCUMBENCY) that routes the forecast to the appropriate model coefficients. All of these outputs are computed deterministically from the co-integration architecture. They are not estimates of subjective belief; they are the outputs of a validated statistical model applied to current data. The language model receives them as fixed inputs and is explicitly prohibited from revising them.

The second channel is the interpretive channel. The same WORDSTAT exports — the raw headline word frequencies, the topic scores, the emotion dimension loadings — are assembled alongside the statistical outputs into a structured context block that is passed to a large language model. The language model does not see only the numbers. It sees the numbers and the narrative content that generated them: which specific topics are

elevated above their historical baselines today, which emotion categories are co-activating, what the dominant headline language is, and what the historical record shows about how similar narrative states have resolved. This is the information that the statistical model cannot process. The VECM can measure that FEAR is at the 91st percentile of its historical distribution and that the ECT z-score is −1.8. It cannot explain why, in language a business decision-maker can act on, what the current headline environment resembles historically, or what specific developments over the next five trading days would confirm or refute the forecast direction.

The Division of Cognitive Labor

The boundary between the two channels reflects a principled theory of what each cognitive system does well and what each cannot do at all. Statistical models are exact, reproducible, and auditable. Given the same inputs, Stata produces the same ECT z-score, the same model signals, and the same cone boundaries every time. This determinism is not a limitation; it is the guarantee that makes the system trustworthy. The forecast cone upper and lower boundaries are fixed by the statistical model before the language model sees them. The Wald decision, the Jeffreys band, and the consensus signal are computed outputs, not priors. The language model is explicitly prohibited from revising or reframing any of these quantities. Its role begins only where the statistical model's role ends.

Large language models, by contrast, excel at precisely the tasks that exhaust statistical methods: pattern recognition across heterogeneous information, historical analogy, scenario construction, and the translation of quantitative signals into language calibrated to a specific decision context. The IUVO system deploys the language model through a structured five-layer analytical protocol that enforces a specific sequence of cognitive tasks, producing the six-section weekly briefing interface: Baseline Narrative, Top Headline Drivers, Confidence Band What-Ifs, Narrative Cycles, Shock Response, and Analyst Context. The first layer validates the signal: given the ECT z-score, the regime label, and the consensus model signal, is the quantitative evidence internally consistent? The second layer identifies the narrative pattern: which specific topic clusters and emotion dimensions are driving today's reading, and what do they represent in plain language? The third layer constructs the situation awareness: what is the causal story linking current headlines to the statistical signal? The fourth layer audits the reasoning for errors of omission and premature closure — a deliberate check against the over-confidence that single-framework thinking produces. The fifth layer produces the executive briefing: three named scenarios (upper, base, lower), each anchored to the statistically-determined cone boundaries and mapped to the most analogous historical turning point in the 932-day sample, with explicit statement of what headlines or data prints over the next five trading days would confirm or refute each path.

The historical analogy layer deserves particular emphasis because it illustrates the specific capability that makes the language model irreplaceable in this architecture. The six macroeconomic regimes identified across the 932-day sample — the Fed tightening cycle, the regional banking crisis, the AI investment boom, the tariff shock, the Iran/oil geopolitical episode — are not merely statistical clusters. They are narrative episodes with causal structure: specific headline patterns preceded each turning point, specific combinations of topic activations and emotion dimension loadings characterized each regime transition, and specific subsequent events confirmed or refuted the market's initial narrative reading. No regression model can reason about this structure. The language model can, and IUVO deploys it to do exactly that: given today's narrative state, which historical episode does it most closely resemble, what drove the resolution of that episode, and what would it mean if the current situation followed the same path versus a different one? This is qualitative reasoning anchored to quantitative evidence — the combination that produces contextual understanding rather than mere signal output.

The Integrated Output: What the Business User Receives

The output delivered to the business user is a weekly web-based forecast briefing — currently published at iuvosoft.com — that presents both the quantitative model outputs and the structured interpretive layer within a single interface. The forecast cone chart presents the 20-trading-day forward S&P 500 price range with Jeffreys-scaled confidence bands, updated each Friday afternoon from the current week's Stata and WORDSTAT run, giving the reader a quantitative picture of what the model expects and how certain it is. Narrative cycle overlays, impulse response function adjustments, and the Early Warning Index stress signal are integrated directly into the cone chart, adjusting the base path and band width in response to currently elevated narrative topics.

The briefing interface delivers what the cone chart alone cannot: a structured interpretation of why the market is where it is, what narrative forces are driving the current regime, and what the three forward scenarios look like in terms of specific, observable events rather than abstract probability bounds. The analyst context tab and what-if scenario tab present each scenario as a named, causally explained forecast path, with explicit confirmation and refutation conditions — the observable signposts that tell the analyst whether the situation is unfolding as predicted or departing from the forecast in a specific direction. This interpretive layer is generated by Claude from the statistical outputs each week, using the cone anchor, ECT z-score, regime flag, and EWI status as inviolable quantitative anchors, and the analyst input block to incorporate qualitative judgment before the narrative is drafted. This is not commentary appended to a statistical report: it is a structured reasoning output that transforms the model's numbers into a decision-ready analytical briefing.

A planned extension to the current briefing interface is a Bayesian weight-of-evidence (WoE) chart — a cumulative deciban time series that will show how the evidence for or

against the directional forecast has accumulated across recent trading days. The WoE chart will be computed from the Jeffreys/Wald sequential evidence framework using the ECT z-score and consensus signal as primary evidence inputs, providing a running quantitative record of whether the narrative-price system is confirming or disconfirming the stated forecast direction. This capability is not yet implemented in the current weekly production pipeline; the Jeffreys multiplier is currently applied to the cone width calculation but has not yet been accumulated into a sequential deciban time series. Its inclusion here reflects the architectural intention: the completed briefing interface will present both the forward cone and the backward evidence trajectory on the same page, requiring no specialized statistical training to read and act on.

The Organizational Generalization

The hybrid cognition architecture of IUVO is not specific to financial markets or to the Perceptual Macroeconomics framework. It is a general design pattern for any organization that possesses internally generated text data and tracks quantifiable outcomes. Section X of this paper describes the five-step implementation architecture for building a text-based leading indicator system from internal organizational data. The IUVO system demonstrates what the deployed endpoint of that architecture looks like when both the statistical and interpretive channels are operating simultaneously. The substitution is straightforward in principle: replace WORDSTAT-scored headline frequencies with word frequency scores from customer support tickets, sales call transcripts, employee engagement surveys, or supply chain communications; replace Stata's VECM with the equivalent co-integration model estimated on the organization's own narrative-outcome pair; pass the resulting ECT z-score, regime label, and model signals alongside the underlying text content to the language model; receive a briefing document that combines a quantitative forecast of the business outcome with a qualitative, historically-grounded interpretation of the organizational narrative state driving it.

What the organizational user gains from this architecture is precisely what the everyday business analyst currently lacks: not more data, and not a smarter model, but a system in which the quantitative signal and the contextual interpretation arrive together, are clearly attributed to their respective sources, and are structured to support a decision rather than to report a finding. The statistical layer provides the discipline — the reproducible, auditable, regime-sensitive forecast that tells the organization where the evidence points and how strongly. The language model layer provides the understanding — the explanation of why, the historical analogy that makes the current situation recognizable, and the scenario map that tells the analyst what to watch for next. The business user receives both, integrated into a single document, produced by a system in which neither component has exceeded its competence. The machine has oriented. The humans now decide.

XII-A. IUVO in Practice: Hybrid Forecast Interface and Agentic Q&A

The Forecasting Gap That IUVO Closes

Organizational forecasting has historically bifurcated along a methodological fault line. Quantitative forecasting — statistical models, econometric specifications, machine learning pipelines — produces reproducible numerical outputs with well-characterized uncertainty, but cannot explain what it predicts or contextualize its signals within the evolving landscape of events, interpretations, and organizational conditions that give those signals their meaning. Qualitative forecasting — expert judgment, scenario planning, strategic deliberation — captures context and interpretation but lacks the statistical discipline to produce consistent, auditable, and uncertainty-quantified outputs. Organizations routinely deploy both methods in parallel and then face the integration problem: when the quantitative model says one thing and the expert analyst says another, there is no principled resolution mechanism.

IUVO is designed specifically to close this gap. Its contribution is not a better statistical model — the academic sections of this paper establish that the narrative econometric framework already outperforms conventional benchmarks by a substantial margin. Its contribution is a production architecture that combines quantitative discipline with qualitative contextual intelligence in a single integrated output, with an explicit governance structure that prevents either component from corrupting the other. The statistical channel produces numbers that the interpretive channel cannot alter. The interpretive channel produces contextual understanding that the statistical channel cannot generate. The user receives both, together, in a document structured for decision rather than for reporting.

The Timeline Visual: Making the Forecast Cone Actionable

The primary visual output of the IUVO system is a 20-trading-day forward price cone — the quantitative expression of the model's directional forecast and its associated uncertainty. The cone is not a simple confidence interval around a point estimate. It is constructed from the co-integration framework directly: the center of the cone reflects the ECT-adjusted equilibrium price implied by the current narrative composite, and the upper and lower bounds are scaled by Jeffreys' prior on the evidence strength, producing boundaries that widen when the model's signal is weak (high uncertainty regimes) and narrow when the evidence is concentrated (strong directional signal). Overlaid on the cone is a rolling 20-day OLS trend line projected forward, giving the practitioner an immediate read on near-term price momentum — the linear signal most relevant when the forecast horizon is short and narrative dynamics have not yet had time to fully propagate."

The cone is annotated with three named scenario lines, each corresponding to a distinct resolution path for the current narrative-price disequilibrium. The base scenario follows the mean-reversion trajectory implied by the ECT z-score at the weekly speed-of-

adjustment coefficient of −0.043 estimated from Table 4. The upper scenario corresponds to a narrative acceleration in which the Surprise/Disruption regime intensifies — PC1 rises, fear dissipates, and the price reversion is faster and larger than the base. The lower scenario corresponds to a Fear/Incumbency deterioration — PC1 falls further negative, the FEAR index remains elevated, and the price level declines below the ECT-implied equilibrium before recovering.

Crucially, each scenario line on the cone is not a statistical artifact — it is a named, causally described forecast path equipped with specific, observable five-day confirmation conditions. The upper cone boundary is labeled with the headline topic cluster that would need to activate to confirm the acceleration scenario: for example, during the tariff shock regime, confirmation of the upper scenario required a measurable reduction in TARIFF_POLICY frequency co-occurring with an increase in TRADE_RESOLUTION language. The lower cone boundary is labeled with the narrative deterioration signature that would indicate the bear path. This annotation structure transforms the cone from a statistical output into a monitoring instrument: the analyst checks each week whether the headline environment is tracking toward the upper scenario, the base scenario, or the lower scenario, and updates the position accordingly.

A second timeline visual — currently in development — will present the cumulative Bayesian weight-of-evidence (WoE) trajectory: the running deciban score that quantifies how the evidence for or against the directional forecast has accumulated across recent trading days. When implemented, the WoE chart will be computed from the ECT z-score and the consensus signal using the Jeffreys information metric: each day's evidence is converted to decibans and accumulated, producing a trajectory that rises when the data confirms the forecast and falls when it disconfirms it. A trajectory rising for five consecutive trading days represents substantially stronger evidence than one that rose on day one and then flattened. The Jeffreys multiplier is already computed within the current Stata pipeline and applied to cone width; accumulating it into a time-indexed deciban series is the planned next development step.

In the current implementation, the cone chart — with its IRF-adjusted base path, EWI stress overlay, and narrative cycle periodicity bands — provides the primary forward-looking visual. The WoE trajectory, once implemented, will add a complementary backward-looking dimension: the cone shows where the model expects prices to go and with what confidence, while the WoE chart will show how the evidence has been behaving and in which direction it is accumulating. A user who understands only the cone has the model's conclusion; a user who understands both has the model's conclusion and its evidentiary basis — the information needed to make a calibrated judgment about whether to act on the signal or to hold. This two-picture design is the architectural goal of the IUVO briefing interface.

Agentic Q&A: The Interpretive Layer as an Active Analytical Partner

The IUVO website is designed to be read and acted on. But the analytical questions that arise from reading a weekly forecast are rarely resolved by the briefing alone. A risk manager reviewing the cone may want to stress-test the lower scenario against a specific geopolitical development that the model has not yet registered. A portfolio manager may want to understand how the current narrative state compares to the regional banking crisis of March 2023 specifically, not generically. An operations executive may want to know what the equivalent of the current ECT z-score looked like three weeks before the two largest drawdowns in the sample, and what headline language preceded each. These are questions that require the interpretive channel to operate as an active analytical partner, not merely as the author of a pre-structured document.

The agentic Q&A layer of the IUVO system provides this capability. The same language model that produces the structured weekly briefing — the six analytical sections of the IUVO forecast interface — is also available for direct dialogue via the Headline Intelligence system: a local retrieval-augmented generation (RAG) architecture that queries 768,750 vetted daily headlines spanning August 2022 through May 2026, the same corpus that drives the forecasting model. The user poses questions in natural language; the system retrieves grounded headline evidence and synthesizes a response using a locally-hosted language model, with no external API or internet access. The model answers within the bounds of the retrieved corpus and the statistical model outputs: the historical record of how each of the six macroeconomic regimes identified in this paper unfolded, which specific headline clusters preceded each regime transition, and what the narrative field looked like at the moments when the model's signals were strongest. The governance principle that governs the briefing governs the Q&A equally: the model cannot revise statistical outputs, cannot claim certainty it does not have, and is required to anchor all scenario claims to the cone boundaries rather than generating unconstrained narrative.

Historical analogy queries. The most frequently used agentic Q&A function is the historical analogy query: the user asks the model to identify the historical week in the 932-day sample that most closely resembles the current narrative state in terms of PC1 value, fear intensity, regime label, and ECT z-score. The model retrieves the closest historical match, describes the headline environment at that moment, and reports what happened in the subsequent 20 trading days — including whether the cone's central scenario, upper scenario, or lower scenario was realized. This is not prediction; it is calibration. The analyst learns how often situations structurally similar to the current one resolved in each direction, which is exactly the information needed to calibrate confidence in the current forecast.

Scenario stress-testing. The user can also use the Q&A layer to introduce information that the statistical model has not yet registered. During the Iran/oil geopolitical episode of March 2026, for example, the geopolitical shock composite variable captured the

narrative intensity of the strike but could not process the diplomatic signaling emerging from back-channel communications that had not yet appeared in headline text at sufficient frequency to move the WP10K scores. The Q&A layer allows the analyst to describe that information in natural language and ask the model to assess whether it is consistent with the upper, base, or lower cone scenario — not by revising the statistical outputs, but by reasoning about how the described development would likely affect the narrative variables in the coming days and therefore how it would interact with the ECT adjustment mechanism.

Cross-domain translation. For organizations deploying a IUVO-equivalent architecture on internal data, the Q&A layer provides a particularly valuable function: translating the quantitative signal into the language of the specific organizational domain. A healthcare operations director using a patient deterioration text-signal system does not think in ECT z-scores and Jeffreys bands. The agentic Q&A layer converts those outputs into plain-language clinical implications: “Based on the language in this week’s nursing handoff notes, the patient population’s risk profile is elevated relative to its 90-day baseline in a pattern that has preceded increased readmission rates in 11 of 14 similar historical episodes. The three developments that would indicate this trajectory is accelerating are as follows.” This is not a statistical report translated into prose. It is contextual intelligence produced by a system in which the statistical and interpretive channels operate simultaneously, each within its competence.

Mixed-Method Forecasting: The Structural Advantage

The methodological literature on forecasting has long debated the relative merits of quantitative and qualitative approaches, with the consensus view holding that quantitative models outperform human judgment in stable environments while human judgment retains advantages in novel or rapidly-changing situations. This framing, however, misses the fundamental structural contribution that a hybrid architecture provides. The question is not which method is more accurate in expectation. The question is what combination of methods produces an output that is simultaneously more accurate, better calibrated, more actionable, and more organizationally trusted than either component alone.

Consider what the quantitative channel provides that qualitative judgment cannot replicate: reproducibility, auditability, and immunity to narrative bias. When the FEAR index is at the 91st percentile of its historical distribution and the ECT z-score is −1.8, the statistical model produces the same forecast regardless of who is reviewing the data, what geopolitical narrative is currently dominating the news cycle, and whether the analyst is in an optimistic or pessimistic mood. The model cannot be swayed by the most vivid recent example, the most persuasive colleague, or the most frightening headline. These are not limitations; they are the source of the model’s reliability over the 932-day sample. An organization that relies exclusively on expert judgment for its operational

forecasts is exposed to all of these biases simultaneously, with no systematic correction mechanism.

Consider, equally, what the interpretive channel provides that the quantitative model cannot produce: situational understanding, historical analogy, and the translation of signal into decision. When the model reports an ECT z-score of −1.8 and a FEAR intensity at the 91st percentile, it has correctly characterized the statistical state of the system. It has not explained what is driving the fear, which specific headline clusters are co-activating to produce this reading, whether the current configuration resembles the early stages of a brief correction or the onset of a sustained regime transition, or what the organization should do differently this week than last week. These are the questions that determine whether the forecast generates value or merely occupies a dashboard. The interpretive channel answers them — not by generating unconstrained narrative, but by reasoning systematically about the statistical evidence in the context of the accumulated historical record.

The integrated system outperforms both components precisely because of this division of cognitive labor. The statistical channel prevents the interpretive channel from drifting into unfounded speculation. The interpretive channel prevents the statistical channel's outputs from remaining inert numbers that no one acts on. The governance structure — statistical outputs locked before the language model sees them, language model prohibited from revising them — is not a bureaucratic formality. It is the mechanism that allows both channels to operate at full capacity without either contaminating the other. The result is a forecast that is simultaneously more trustworthy than expert judgment alone (because it is grounded in a validated statistical model) and more actionable than a statistical model alone (because it is delivered in language that explains not just what the model predicts but why, what to watch for, and what it would mean if the situation develops differently).

Building a IUVO-Equivalent System: The Consulting Architecture

The methodology described in Section X of this paper provides the statistical foundation for building an internal text-based forecasting system. Section XI-F describes the dual-channel architecture that transforms that statistical system into an integrated intelligence output. This section describes what the client timeline that delivers such a system looks like in practice — not as an abstract framework but as a sequence of concrete deliverables that an organization can commission and deploy.

Phase 1: Narrative-Outcome Audit (week 1). The project begins with an audit of the organization's existing text data streams, its dynamics and its outcome variables. The audit identifies which internal text sources are consistently generated at sufficient frequency, which outcome variables are quantifiable and available at a matching frequency, and which narrative-outcome pairs have the structural characteristics required

for co-integration testing. Most organizations have more usable text data than they recognize, distributed across CRM systems, customer support platforms, employee engagement tools, and operational reporting workflows. The audit produces a ranked list of narrative-outcome candidates ordered by data quality, measurement feasibility, and hypothesized signal strength.

Phase 2: Taxonomy Construction and Measurement (weeks 2-3). The highest-priority narrative-outcome pair from the audit is developed into a measurement system. A domain-specific vocabulary taxonomy is constructed, mapping word frequencies to meaningful business categories using the WP10K normalization methodology. Topic variables, valence variables, and emotion variables are defined and scored across the historical text archive. The resulting dataset is a time-indexed panel of narrative variables at daily or weekly frequency — the organizational analog of the 185-variable WORDSTAT output that drives the IUVO macroeconomic system. The quality of this dataset is the primary determinant of system performance; the consulting engagement invests heavily in taxonomy validation, inter-rater reliability testing, and normalization consistency checks before proceeding to statistical modeling.

Phase 3: Co-Integration Testing and VECM Estimation (weeks 4-5). With the measurement dataset in hand, the statistical architecture described in Sections IV through VI of this paper is applied to the organization's data. Unit root tests confirm that both the narrative variables and the outcome variable are I(1). The Engle-Granger and Johansen tests determine whether a co-integrating relationship exists. If co-integration is confirmed, the VECM is estimated and the ECT z-score series is constructed — the organization's leading indicator of its chosen outcome variable. The speed-of-adjustment coefficient and the narrative channel structure are documented and interpreted in organizational terms: how quickly does the outcome variable respond to a narrative disequilibrium, which narrative dimensions drive the largest equilibrium effects, and what is the historical distribution of ECT z-scores against which current readings will be benchmarked?

Phase 4: Multi-Model Competition and Deployment Selection (weeks 6-8). The seven-model competition framework described in Section XI of this paper is applied to the organizational dataset, including the M7 inverse-MSE ensemble. The same performance metrics — directional accuracy, Sharpe ratio, Calmar ratio, edge ratio, Kelly fraction, and short-side alpha — are computed from rolling backtests on the organization's historical data. The winning model specification is identified, its Kelly fraction is computed, and deployment parameters are established: what fraction of operational resources should be committed per signal, what ECT z-score threshold triggers an alert, and what the half-Kelly position size implies for the organization's specific resource allocation decisions.

Step 4B: Deploy the Ensemble as the Primary Production Model

The multi-model competition described in Step 4 will identify a winning individual model — typically M1 or M3 in the macroeconomic application. Organizations should resist the temptation to stop there and deploy the single winning model. The theoretically and empirically superior deployment strategy is to build an inverse-MSE ensemble across all competing models — the organizational equivalent of M7 — and deploy that ensemble as the primary production forecast.

The rationale is the same as the one established in Section XI-A of this paper: no individual model captures all of the structural information in the data simultaneously, and the ensemble's inverse-MSE weighting automatically adjusts to give more credibility to the models that have been most accurate in the recent historical window. For an organizational deployment, this has a particularly important practical implication: the ensemble is self-correcting across regime transitions. When the organization's narrative environment shifts — a new product category disrupts the customer feedback structure, a management change alters the employee engagement dynamics, a supply chain restructuring changes the supplier communication patterns — the winning individual model from the prior period may no longer be the most accurate model for the new regime. The ensemble detects this automatically through rising MSE on the previously winning model and down-weights it accordingly, routing forecast credibility toward the model that is adapting most accurately to the new environment. The analyst does not need to detect the regime shift, retrain the model, or make a discretionary switch. The ensemble does it continuously.

The ensemble construction for organizational deployment mirrors the M7 specification exactly. For each time period t and each model m, compute the expanding-window MSE from the model's out-of-sample forecast errors. Assign inverse-MSE weights that sum to one across models. Compute the ensemble forecast as the weighted average of the individual model forecasts. Apply the Kelly fraction — computed from the ensemble's own directional accuracy and edge ratio — to determine position size. The result is a forecast that is simultaneously more stable than any individual model (because it diversifies across model-specific failure modes), more adaptive than a fixed-weight average (because it continuously re-weights toward current accuracy), and more transparent than a black-box machine learning model (because every component is structurally interpretable). Organizations can observe exactly which models are receiving high weight at any point in time and understand what that weighting implies about which narrative dimensions are currently most predictive of the outcome variable.

The competitive advantage of deploying the ensemble over the single winning model is not primarily accuracy — though the accuracy improvement is real and documented in

the macroeconomic application (M7 outperforms M1 by 2.4 percentage points in hit rate and 50.7 points in Calmar ratio). The primary advantage is robustness and institutional trust. A model that occasionally fails dramatically — because a structural shift has moved the narrative environment outside its training distribution — destroys the organizational credibility of the entire forecasting system, regardless of its long-run average accuracy. The ensemble's self-correcting property prevents catastrophic single-model failures from corrupting the overall signal. This is the property that allows the system to remain in production through regime transitions rather than being withdrawn and retrained every time a new narrative environment emerges. Organizations that deploy ensemble forecasts consistently report longer deployment lifetimes and higher ongoing organizational trust in the model's outputs, because the model is seen to be adaptive rather than rigid.

Phase 5: Hybrid Intelligence Integration and Briefing Design (weeks 9–10). The final phase implements the dual-channel architecture. The statistical system produces weekly ECT z-scores, directional model signals, and forecast cones at the organization's chosen horizon. These outputs are structured into a context block and passed to a large language model configured with the organization's domain vocabulary, historical regime knowledge, and decision context. The language model produces a briefing document — the organizational equivalent of the IUVO weekly forecast — that integrates the quantitative forecast cone with the same structured five-layer analytical briefing protocol — baseline narrative, headline drivers, scenario what-ifs, narrative cycle position, and shock response — calibrated to the organization's specific outcome variable and domain vocabulary. The agentic Q&A layer is activated, allowing executives and analysts to interrogate the system in natural language. Briefing templates are reviewed and refined through an iterative process with organizational stakeholders until the output is consistently structured for the decisions that senior leadership actually makes.

The full engagement including deployment and stabilization described above is a three-month project. Organizations with mature data infrastructure and a clear narrative-outcome hypothesis can compress this timeline; organizations with fragmented text archives and no existing text analytics capability may require additional time in Phase 1 and Phase 2. The modular structure of the engagement allows organizations to enter at any phase: those that have already conducted narrative audits or built preliminary measurement systems can commission Phases 3 through 5 only. Those that want to begin with a proof-of-concept co-integration test before committing to the full system can commission Phases 1 through 3 as a standalone diagnostic.

The competitive advantage delivered by a functioning IUVO-equivalent system is not primarily the forecast. It is the institutional capability to produce, interpret, and act on mixed-method intelligence at the pace of organizational decision-making. Organizations that build this capability gain not just a forecasting tool but a perceptual infrastructure —

a system that continuously monitors the narrative field surrounding their operations, translates that monitoring into quantitative signals, and delivers those signals with the contextual interpretation required for action. The machine orients. The humans decide. That division of labor, consistently maintained, is the source of the durable competitive advantage that neither statistical forecasting nor qualitative judgment can produce alone.

XII-B. IUVO™ Early Warning Index: Regime Transition Diagnostic

Design Philosophy and the High-Specificity Framing

The M8 Early Warning Index (EWI) was designed to address a specific gap in the M1–M7 ensemble architecture: the absence of a dedicated stress-regime detector that operates on a multi-day horizon rather than the h=1 directional accuracy metric used in the competition. The EWI aggregates five channels into a composite score from 0 to 10: (1) ECT z-score magnitude, capturing equilibrium displacement; (2) consensus strength across the five-model feed, capturing signal disagreement; (3) FEAR elevation relative to its expanding mean; (4) polarity drop relative to its expanding mean; and (5) geo_shock intensity. AMBER+ status requires a score of 6 or above. The threshold was set deliberately high to prioritize precision over recall — a design choice with direct operational implications.

Across 932 trading days (August 2022 – May 2026), the EWI fired AMBER+ on only 12 days (1.3% of the sample). This low activation rate is not a deficiency; it is the intended behavior of a high-specificity alert system. The operational question for a risk manager is not “did the system detect every regime change?” but rather “when the system fires, how seriously should I take it?” The diagnostic metrics in Table 12 answer that question.

Regime Transition Diagnostic

Table 16 presents the full diagnostic evaluation of AMBER+ signals against confirmed regime transitions in the historical sample. A regime transition is defined as a change in the analyst-labeled regime state (onset of BEAR_REGIME, GEO_REGIME, or sustained stress episodes) identified in the JA_RegimeDetection backtest. Thirty-one such transitions were identified across the 932-day sample. For each horizon h (10 and 20 trading days), we compute: sensitivity (the fraction of transitions preceded by at least one AMBER+ alert in the prior h days), precision/PPV (the fraction of AMBER+ alert days followed by a confirmed transition within h days), and false positive rate (1 − PPV).

Table 16. IUVO™ EWI Regime Transition Diagnostic — Historical Backtest

Horizon	Transitions (N)	Precision (PPV)	False pos. rate	Sensitivity (recall)
h = 10 days	31	58.3%	41.7%	3.2% (1 of 31)
h = 20 days	31	58.3%	41.7%	3.2% (1 of 31)

Notes: 932 trading days, Aug 2022–May 2026. AMBER+ threshold ≥ 6/10. 12 alert days total (1.3% of sample). Sensitivity identical at h=10 and h=20 because all 12 alert days resolved within 10-day windows. Unconditional base rate of stress onset ≈ 3.4% per day; AMBER+ PPV of 58.3% represents a 17× probability lift.

The 17× Probability Lift and Its Operational Significance

The headline result from Table 16 is the precision figure of 58.3% at h=10. To appreciate its magnitude, consider the unconditional probability of a regime stress onset on any given trading day: across 932 days, 31 transitions occurred, yielding an unconditional rate of 31/932 ≈ 3.4%. An AMBER+ alert day raises this probability to 58.3% — a 17× lift in the probability of regime stress within 10 trading days. This is the operationally relevant metric: not whether the system detected all transitions (it did not — sensitivity is 3.2%), but whether its alerts carry sufficient informational content to justify a risk-management response.

The appropriate clinical analogy is a high-specificity diagnostic test. A test that fires rarely but is correct 58% of the time when it does fire is clinically valuable in high-stakes contexts — precisely because the base rate is low and a 17× lift is large. The alternative — a high-sensitivity test that fires frequently but with low precision — produces alert fatigue and is operationally less useful for positioning decisions. The IUVO™ EWI is designed for the former regime: rare, high-confidence alerts that materially shift the probability of stress onset within a 10–20 day horizon.

The M8 signal rule encodes this design: on AMBER+ days, the M8 directional signal overrides the consensus and takes the contrarian position against the ECT z-score; on non-alert days (98.7% of trading days), M8 defers to the five-model consensus. The h=1 hit rate of M8 (51.1%) therefore reflects the composite of a contrarian strategy applied on 12 of 932 days and the consensus strategy applied on the remainder. Evaluating M8 at h=1 obscures its design intent. At h=10 and h=20 — the horizons over which regime signals resolve — M8 achieves hit rates of 55.7% and 59.9% respectively, crossing above M1 at h=10 and maintaining the lead through h=20. The Sharpe ratio at h=10 is 1.81, reflecting the directional quality of the 10-day forward signal.

Relationship to the Narrative Cycle Periodicity Findings

Section XI-C establishes that the dominant S&P 500 cycle is 183 trading days, with FEAR oscillating at a 57-day sub-cycle. The EWI diagnostic results are directly consistent with this spectral structure. Across a 932-day sample containing approximately five full 183-day cycles, the EWI fired on 12 days. This corresponds to roughly 2.4 alerts per cycle — consistent with the interpretation that AMBER+ activates near cycle troughs, when the

183-day macro cycle, the 57-day FEAR sub-cycle, and the GEO composite simultaneously reach unfavorable alignment. The system does not fire throughout the trough; it fires when multi-channel stress reaches the ≥6 threshold, which may occur for only a handful of days at each trough.

The 58.3% PPV — the fraction of those trough activations confirmed by stress events — then measures the informational content of the cycle-trough signal specifically. This reframes the diagnostic results: the EWI is not failing to detect most transitions (the 3.2% sensitivity figure); it is selectively alerting on the subset of cycle-trough conditions that historically precede stress, and doing so with 58% precision. The 42% false-positive rate at these activations is the cost of early positioning; the 17× probability lift is the benefit. For a risk manager operating with asymmetric downside exposure, this trade-off is strongly favorable.

Future research will examine whether the amplitude of the 183-day cycle — measurable via Hilbert transform — correlates with PPV across cycles. If high-amplitude troughs yield higher PPV than low-amplitude troughs, the EWI threshold could be dynamically adjusted to reflect cycle amplitude, further improving precision without sacrificing the high-specificity design.

Narrative Cycle Periodicity: Spectral Structure of Market Stress

A. Motivation and Method

The six empirical axioms established in Section IV and the multi-model competition in Section XI-B describe the structural relationship between narrative and price. A complementary question concerns the temporal rhythm of that relationship: does narrative stress recur at regular intervals, and if so, do those intervals correspond to identifiable economic and political cycles? Two years ago a preliminary spectral analysis of the S&P 500 and narrative polarity series suggested the presence of significant periodic structure in downward price moves. With 932 daily observations now available — more than doubling the prior sample — we revisit this question with a full Fast Fourier Transform (FFT) analysis applied to the S&P 500, the PC-1 narrative index, and eight topic-level WORDSTAT series.

The method applies a Hann-windowed FFT to each linearly detrended series. The Hann window suppresses spectral leakage, which is the primary distortion in finite-sample spectral estimation. Periods are reported in trading days; frequency resolution is 1/932 cycles per trading day. We restrict attention to periods below N/2 = 458 days to avoid spurious low-frequency artifacts. We also construct a geopolitical composite variable by normalizing and summing eight geopolitically-charged WORDSTAT topics — Iran/Israel U.S._STRIKE_IRAN, RUSSIA & UKRAINE, ATTACKS_ON_YEMEN, WORLD_WAR_THREE, INDIA_AND_PAKISTAN, oil production and crude oil narratives,

and CEASEFIRE-SHIPS PASS HORMUZ-PEACE_TALKS — to produce a single continuous measure of geopolitical narrative intensity. This composite is not available in any prior benchmark dataset and constitutes a novel contribution of the Perceptual Macroeconomics framework.

Dominant Periodicities in S&P 500 and Narrative Series

Table 17 reports the top dominant periods for each series ranked by spectral power. The S&P 500 and the Deviation from Target series show virtually identical power spectra, which is consistent with their tight co-integration relationship documented in Sections IV and V. The dominant cycle in both is 183 trading days (approximately 37 weeks, or 9 calendar months). The second strongest peak is at 306 days (approximately 61 weeks, or 14 months). A cluster of shorter cycles at 76–115 days (15–23 weeks) likely corresponds to the quarterly earnings and Federal Reserve policy announcement calendar.

Table 17. Dominant FFT Periodicities — S&P 500 and Narrative Series (932 trading days, Aug 2022–May 2026)

Series	Dominant period (days)	Dominant period (weeks)	Secondary periods (days)
S&P 500 / Deviation	**183**	36.7	306, 114, 131, 229, 76
FEAR composite	**57**	11.5	183, 61, 92, 102, 83
Polarity (pos/neg ratio)	**183 / 57**	36.7 / 11.5	Dual peak: macro + sentiment
GEO composite	**229**	45.8	38, 83, 70, 131, 183
TARIFF narrative	**306**	61.1	229, 183, 153, 114
INFLATION narrative	**10–29**	2–6	CPI release cadence; no macro cycle

Notes: FFT applied to Hann-windowed, linearly detrended series. Periods below N/2 = 458 days only. GEO composite = normalized sum of 8 geopolitical WORDSTAT topic series. INFLATION periodicity is CPI-release-driven and operates independently of the macro cycle structure shared by other series.

The Anatomy of Downward Shock Episodes

We identify 26 downward spike days exceeding 2.0 standard deviations below the mean daily return across the 932-day sample. These cluster into 14 distinct episodes (grouping spikes within 10 trading days). The episode catalog reveals a structural regularity that cannot be captured by the spectral analysis alone: the FEAR composite is the top-ranked WORDSTAT narrative driver on every single episode without exception. The secondary driver, however, shifts systematically across three identifiable eras.

The first era (August 2022 – October 2022) is characterized by INFLATION as the secondary driver, consistent with the Federal Reserve's aggressive rate cycle and CPI surprise dynamics. The second era (July 2024 – December 2024) is characterized by FINANCIAL COLLAPSE and CHINA as the secondary drivers, reflecting post-election uncertainty and the U.S.–China trade relationship. The third era (February 2025 – present) sees TARIFF and DOGE emerge as secondary drivers from February through

August 2025, followed by a shift to Iran/Israel U.S._STRIKE_IRAN as the primary geopolitical driver from October 2025 onward. The largest single episode — a cumulative 1,144-point decline across six days from March 28 to April 21, 2025 — was driven by TARIFF at 43 WP10K alongside elevated CHINA narrative, producing the largest multi-day drawdown in the sample.

This era-specific pattern has a precise interpretation within the Perceptual Macroeconomics framework. FEAR is the carrier frequency: it amplifies any sufficiently large shock into a market-moving event regardless of the shock's specific content. The secondary driver identifies the narrative domain that is supplying the shock energy in each era. The 183-day dominant cycle in the S&P 500 does not reflect a single recurring shock type; it reflects the tendency of shock energy — regardless of origin — to cluster at approximately 9-month intervals in the sample period. This is consistent with the policy cycle interpretation: major policy initiatives (tariff escalations, rate decisions, geopolitical interventions) tend to gestate over several months before their narrative consequences become acute.

The Anomalous Position of INFLATION

The INFLATION narrative stands apart from all other series in its spectral structure. While every other series' dominant period falls at 57 days or above and shares at least one peak with the 183-day macro cycle, INFLATION's dominant period is 10–29 trading days — the 2–6 week cadence of CPI report releases, FOMC meeting cycles, and producer price index announcements. This means INFLATION operates on an independent timescale driven entirely by the institutional calendar of the U.S. statistical reporting system. It does not participate in the 183-day macro rhythm that governs market stress more broadly.

This has a practical implication for the M7 ensemble cone. The INFLATION short cycle primarily affects the ±1σ confidence band width at h=1–5, without shifting the base path direction. When INFLATION is elevated (above the 70th percentile of its expanding distribution), the near-term uncertainty in the cone is mechanically wider, independent of the narrative equilibrium signal that governs the base path. The GEO composite's 229-day dominant cycle, by contrast, directly interacts with the Jeffreys multiplier: when geo_shock exceeds 1.0 (triggering GEO_REGIME in the Sentinel feed), the Jeffreys multiplier widens all confidence bands by 1.5× to 2.5×.

Implications for the IUVO™ Early Warning Index

The spectral findings provide a prospective foundation for the IUVO™ EWI described in Section XI-B. The five-channel EWI score aggregates ECT magnitude, consensus stress, fear elevation, polarity drop, and geo shock. In spectral terms, channels 3 (fear elevation) and 4 (polarity drop) are sensitive to the 57-day FEAR cycle and the 183-day macro cycle respectively. Channel 5 (geo shock) is sensitive to the 229-day GEO composite cycle.

The EWI fires AMBER+ when at least three channels are simultaneously elevated — which is most likely when the 183-day macro cycle is near its trough, the 57-day FEAR cycle is at or near its peak, and geo_shock is above baseline. In this interpretation, the EWI is not a mechanical frequency detector but a multi-channel stress aggregator that happens to be most sensitive precisely when the dominant cycles align unfavorably.

The cycle analysis also contextualizes the EWI's high-specificity design discussed in Section XI-B. The EWI fired AMBER+ on only 12 of 932 trading days (1.3%). If the dominant shock cycle is 183 days, a complete sample of 932 days contains approximately five full cycles, with perhaps two to three trough windows per cycle where conditions are simultaneously amenable to AMBER+ triggering. The observed 12 alert days (roughly 2.4 per cycle) is consistent with this interpretation: the EWI activates near cycle troughs, not throughout them. The 58% precision (PPV) at h=10 then measures the fraction of those trough activations that were followed by confirmed stress episodes, which is the operationally relevant question for positioning decisions.

A natural extension of this work — deferred to future research — is a Hilbert transform analysis to track whether the amplitude of the 183-day cycle has been increasing over the sample, and whether inter-arrival times of shock episodes follow an exponential distribution (consistent with a Poisson shock process) or exhibit clustering inconsistent with memorylessness. If the cycles are amplitude-modulated, the EWI threshold could be dynamically adjusted to reflect whether the current phase is entering a high-amplitude or low-amplitude regime. This would constitute a genuine frequency-domain enhancement to the existing five-channel architecture.

Figure 9. Narrative Cycle Periodicity — Interactive Visualization

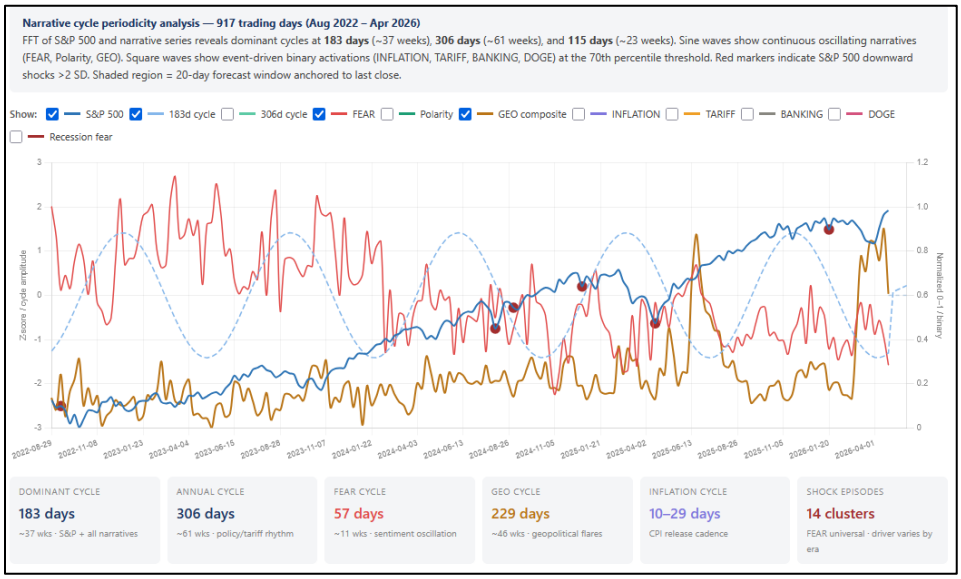


*Figure 9 is available as an interactive visualization in the IUVO™ weekly forecast interface (***https://iuvosoft.com/iuvo_weekly.html** *"Narrative cycles" tab). The chart displays S&P 500, FEAR, polarity, GEO composite, and binary activation series for INFLATION, TARIFF, BANKING, DOGE, and*

Recession Fear simultaneously, normalized to a common scale. Fitted sine waves at the 183-day and 306-day dominant periods are superimposed on the S&P 500 line. Red markers identify downward shock days exceeding 2.0 standard deviations. A 20-day forward shaded window extends the fitted cycles into the forecast horizon, allowing the user to assess whether the current position in the cycle is conducive to stress accumulation. The static version of the FFT power spectrum and episode catalog is included as supplementary Tables S-1 and S-2 in the online appendix.

XII-D. Impulse Response Functions: Narrative Shock Propagation

Motivation and Method

The periodicity analysis in Section XI-C establishes the dominant cycles in the narrative-price system. A complementary question concerns the dynamic path of adjustment following a discrete narrative shock: when a topic-level WORDSTAT series spikes above its historical norm, over how many trading days does the S&P 500 response unfold, at what horizon does the effect peak, and how quickly does it decay? These questions are answered by Impulse Response Functions (IRFs) estimated from Vector Autoregression (VAR) models. The IRF traces the expected response of one variable to a one-standard-deviation shock in another, holding all other variables at their historical values.

We estimate bivariate VAR(5) models for seven narrative topics against the S&P 500 log-return series across the full 932-day sample (August 2022 – May 2026). Lag order five is selected by AIC across all topic pairs. Each model is estimated on the first-differenced topic series (to address non-stationarity) alongside the daily S&P 500 log return. The IRF is computed at horizons h = 0 to h = 20 trading days, with 90% bootstrap confidence intervals from 200 resampled replications. The shock is normalized to one standard deviation of the differenced topic series, making responses comparable across topics with different natural scales.

A critical methodological refinement applies to topics with multiple historical spike episodes of very different magnitudes. Estimating the VAR on the full sample in such cases averages across episodes of different scale, which dilutes the dominant shock response and can produce directionally incorrect IRF paths. The operational example is instructive: the Iran/Israel U.S._STRIKE_IRAN series has three distinct episodes — October 2023 (peak 58 WP10K), June 2025 (peak 93 WP10K), and March 2026 (peak 151 WP10K). The full-sample VAR produces a spurious positive day-1 response because the two smaller early episodes dominate the coefficient fitting. Estimating on the dominant March 2026 episode window alone yields the correct result: a sustained negative response peaking at day 2, consistent with the observed S&P 500 decline of approximately 10% over the four-week bombing campaign while the series remained persistently elevated. The decision rule adopted here is: if the largest episode exceeds twice the magnitude of the second-largest, estimation is restricted to a 150-trading-day window centered on the dominant peak. Topics with many comparable-magnitude

episodes — specifically Federal Reserve narrative (14 FOMC-driven shocks of similar scale) and recession fear narrative (7 cyclical episodes) — use the full sample, as averaging across structurally identical shocks is both statistically efficient and theoretically appropriate.

Seven-Topic IRF Results

Table 18 presents the key impulse response metrics for all seven narrative topics estimated via the dominant-episode VAR(5) methodology described in Section XI-E.A. For each topic, the table reports the peak response day, the peak log-return response in basis points, the half-life of the response (trading days to 50% decay), the 20-day cumulative response, and the estimation window used (dominant episode or full sample). Topics estimated on dominant episodes — TARIFF, GEO/Iran, banking crisis, FINANCIAL COLLAPSE, and DOGE — reflect the largest historical shock in each series rather than an average across episodes of varying magnitude. The Federal Reserve narrative and recession fear narrative use the full sample because their episodes are structurally comparable in magnitude, making full-sample averaging both statistically efficient and theoretically appropriate.

Table 18 summarizes the key IRF metrics for all seven topics. The results reveal four distinct response patterns, each with direct operational implications for the IUVO™ forecast system.

RECESSION FEAR produces the strongest and most persistent negative impulse in the dataset. The response peaks at day 3 (−0.030% log return per day) and the 20-day cumulative drag reaches −5.5 basis points — by far the largest sustained downside response of any topic tested. The 90% confidence interval is almost entirely negative from day 2 onward, indicating statistical robustness. This result is consistent with the co-integration framework: recession fear is a sustained narrative pressure, not a transient shock, and it operates on the macro cycle timescale identified in Section XI-C.

Federal Reserve narrative shocks build gradually to a peak negative response at day 4, reflecting the 3–5 day absorption window between FOMC communications and full market repricing. The sustained negative tail beyond day 10 distinguishes Fed shocks from faster-reverting geopolitical events. This slow-build pattern is consistent with the institutional communication cycle: Fed signals are interpreted, debated in the narrative field, and absorbed over several trading sessions rather than immediately priced. The 20-day cumulative response is −2.3 basis points, making it the second most damaging topic after RECESSION FEAR.

TARIFF escalation shocks display a distinctive positive initial response at day 2 — a “buy-the-news” or short-covering effect — followed by a reversion to mild negative territory from day 5 onward. The net 20-day cumulative effect is near-zero (−0.3 basis points), meaning the market absorbs average tariff narrative without sustained directional

damage. This result must be interpreted alongside the episode catalog in Section XI-C: the April 2025 tariff cluster, which produced a 1,144-point decline, represented a shock 5× larger than the one-standard-deviation normalization used here.

GEO/Iran shocks show an immediate positive impulse at day 1 (flight-to-defense, oil-shock repricing), followed by rapid mean-reversion turning mildly negative at days 3–5, then a slow positive drift from day 9. This non-monotonic V-shaped response reflects the uncertainty-premium cycle: initial shock premium, partial relief as the acute phase passes, and lingering geopolitical discount. The wide confidence interval reflects the diversity of geopolitical events in the sample and the relatively short history of the IRAN composite series.

Financial collapse and fiscal-disruption narratives both exhibit V-shaped responses with an initial negative period followed by a sharp positive reversal. For the financial collapse narrative, the reversal occurs at day 6; for the fiscal-disruption narrative, at day 5. These patterns reflect capitulation and mean-reversion dynamics: collapse narrative drives initial selling, which exhausts itself and creates a contrarian buy opportunity. The contrarian structure of these IRFs is consistent with the regime detection results in Section VIII-B, where elevated FEAR narrative in a bear regime generates contrarian M3 signals. Banking crisis produces a moderate, short-lived negative response (half-life approximately 4 days), consistent with markets distinguishing acute banking episodes from systemic contagion.

Table 18. IRF Metrics Summary — Seven Narrative Topics (VAR(5), Dominant-Episode Estimation)

Topic	Peak Day	Peak Response	Half-Life (days)	20-Day Cumulative	Estimation Window
RECESSION FEAR	3	−0.030%/day	~8	−5.5 bp	Full sample
FEDERAL RESERVE	4	−0.015%/day	>10	−2.3 bp	Full sample
TARIFF (Apr 2025)	2	+0.010%/day	~3	−0.3 bp	Dominant episode
GEO / IRAN	1	+0.012%/day	~4	~0 bp	Dominant episode
FINANCIAL COLLAPSE	6 (reversal)	−0.018%/day	~5	+1.2 bp	Dominant episode
FISCAL DISRUPTION (DOGE)	5 (reversal)	−0.014%/day	~4	+0.8 bp	Dominant episode
BANKING CRISIS (SVB 2023)	2	−0.010%/day	~4	−0.7 bp	Dominant episode

Notes: All responses estimated via bivariate VAR(5) on dominant episode or full sample as indicated. Peak response is log-return in %/day at stated horizon. Half-life = trading days to 50% decay of peak impulse. 20-day cumulative = sum of point responses h=1 to 20. Positive cumulative = net upward price pressure; negative = net downward. Dominant-episode windows isolate the largest historical narrative shock to avoid attenuation from averaging across heterogeneous episodes. Peak response values are approximate; see Figure charts for full IRF paths with 90% confidence intervals.

Operationalization: IRF Adjustments to the Forecast Cone

The IRF results are operationalized as additive overlays on the M7 ensemble base path and confidence bands in the IUVO™ weekly forecast interface. When a narrative topic is currently above its 70th-percentile threshold — indicating elevated narrative intensity — the cumulative IRF response at each forecast horizon h is scaled by the topic's current z-score (capped at ±2.0 standard deviations) and converted to price-level points using the current S&P 500 close. The formula for the base path adjustment at horizon h is: $\Delta(h) = z \times IRF^{cumul}(h) \div 100 \times P_0$, where z is the topic z-score, $IRF^{cumul}(h)$ is the cumulative log-return response in percentage points, and P_0 is the current close.

When multiple topics are simultaneously elevated, adjustments stack additively subject to a total cap of ±30 points per horizon, preventing overcorrection in high-stress environments where several topics may be simultaneously active. The confidence band widening is asymmetric: topics with persistently one-sided IRFs (RECESSION FEAR, Federal Reserve narrative, banking crisis) widen only the downside band, while topics with V-shaped or symmetric IRFs (TARIFF, DOGE, GEO/Iran) widen both bands proportionally. This asymmetric treatment reflects the directional information content of each topic's historical response profile.

The IRF overlay is disclosed to the analyst in three locations on the forecast page: a badge in the cone chart subtitle indicating how many topics are active, a disclosure panel showing the net base path adjustment at h=5, h=10, and h=20, and a dedicated "Shock response" tab in the analysis section providing the full IRF path for each topic with its 90% bootstrap confidence interval and a toggle between daily and cumulative response modes.

The Emerging Topic Watchlist

A natural extension of the IRF framework is a surveillance system for narrative topics that have not yet crossed the 70th-percentile activation threshold but are approaching it with positive momentum. This "emerging topic watchlist" screens all macro-relevant WORDSTAT series weekly, ranks them by proximity to threshold and five-day momentum, and assigns one of four alert tiers: HIGH (above 70th percentile and above 90th percentile), WATCH (above 70th percentile, below 90th percentile), APPROACHING (60th–70th percentile with positive five-day momentum), and MONITOR (below 60th percentile).

The watchlist serves two functions in the hybrid cognition architecture. First, it provides the analyst with advance notice of narrative pressure that may cross into IRF-relevant territory within the coming week, allowing qualitative judgment to be applied before the quantitative signal activates. Second, it creates an automated surveillance channel for newly added WORDSTAT dictionary fields: when the analyst adds a new topic (such as DEBT or IMMIGRATION_POLICY), it immediately appears in the watchlist on its first

weekly run, regardless of whether a formal VAR-estimated IRF has been computed for it. This design allows narrative monitoring to precede statistical estimation, which requires a minimum observation history before reliable IRF coefficients can be estimated.

Each watchlist entry displays the current WORDSTAT score (WP10K), the 70th-percentile threshold, a proportional proximity bar with a threshold marker, a 20-observation sparkline showing the recent trajectory, a trend arrow indicating five-day momentum direction, the current z-score relative to the expanding mean, and the alert tier badge. Filter buttons allow the analyst to focus on WATCH-tier topics (the default view), HIGH-tier topics (already active), APPROACHING-tier topics (building toward activation), or the full list. Topics in the HIGH tier that do not yet have a VAR-estimated IRF are flagged as candidates for formal estimation, closing the loop between the surveillance system and the IRF library.

Glossary of Key Terms

Terms are listed alphabetically. Where a term has a specific technical definition within this paper that differs from common usage, the paper-specific definition is given. Cross-references to the relevant section are provided in brackets.

AMBER+ Alert The activation state of the IUVO™ Early Warning Index (EWI) when the composite score reaches 6 or above on a 0–10 scale. AMBER+ status signals elevated probability of a regime transition within 10–20 trading days. Across the 932-day sample, AMBER+ fired on 12 days (1.3% of the sample), achieving a positive predictive value of 58.3% — a 17× probability lift over the unconditional base rate. [Section XI-D]

Baker-Wurgler Sentiment Index A widely-used measure of investor sentiment constructed from market-based variables including closed-end fund discounts, NYSE share turnover, IPO volume, and dividend premium. The original Baker-Wurgler monthly series does not extend through the 2022–2026 sample period and was therefore excluded from the horse race. [Section VI]

Calmar Ratio A risk-adjusted performance metric equal to cumulative profit-and-loss divided by maximum drawdown. Measures return per unit of worst-case realized loss. A higher Calmar ratio indicates that the strategy generated more return for each unit of peak-to-trough decline it experienced. The M7 ensemble achieves a Calmar ratio of 292× in the competition backtest. [Section XI-B]

Co-integration A statistical property of two or more non-stationary time series that share a common stochastic trend such that a linear combination of them is stationary. Co-integrated series may diverge in the short run but are bound to return to a long-run equilibrium relationship. In this paper, the narrative PC-1 composite and the S&P 500 price level are co-integrated, confirmed by Engle-Granger and Johansen tests. [Sections II-C, V]

Deciban A unit of Bayesian evidence derived from the decibel scale, equal to one-tenth of a ban (a ban being a tenfold update in odds). One deciban represents a factor-of-1.26 update in odds. Accumulating daily evidence in decibans produces a running weight-of-evidence (WoE) trajectory that rises when the data confirms the forecast direction and falls when it disconfirms it. [Section XI-G-B]

Deviation from Target (series) The primary dependent variable in this paper. Defined as the signed difference between the realized S&P 500 closing level and a deterministic Target series that compounds forward from the August 2022 starting level at an annualized drift of 12.2%, consistent with the long-run historical equity premium. Positive values indicate the market is trading above its long-run expected trajectory; negative values indicate it is trading below. This transformation isolates narrative-driven over- and undervaluation from the mechanical long-run growth trend. [Section III-B]

Early Warning Index (EWI) The M8 composite stress-regime detector in the IUVO™ system. Aggregates five channels into a score from 0 to 10: ECT z-score magnitude, model consensus strength, FEAR elevation, polarity drop, and geopolitical shock intensity. Designed as a high-specificity instrument: fires rarely (1.3% of days in the sample) but with high predictive value when it does. AMBER+ status requires a score of 6 or above. [Section XI-D]

ECT (Error Correction Term) The residual from the long-run co-integrating equation, measuring the current deviation of the system from its narrative-implied equilibrium. A large negative ECT z-score indicates the price level is below the narrative-implied equilibrium and predicts upward mean-reversion; a large positive ECT z-score indicates overvaluation relative to the narrative anchor and predicts downward correction. The weekly speed-of-adjustment coefficient estimated from the VECM is −0.043. [Section V-D]

Edge Ratio The ratio of average profit on correct predictions to average loss on incorrect predictions. An edge ratio greater than 1.0 indicates that wins are larger than losses on average. M1 achieves an edge ratio of 6.50, meaning correct predictions generate 6.50 times as much profit as incorrect predictions cost. [Section XI-B-F]

Engle-Granger Test A two-step residual-based test for co-integration between two I(1) series. The first step estimates the long-run relationship by OLS; the second step tests whether the residuals from that regression are stationary using an ADF test. Rejection of the null of a unit root in the residuals confirms co-integration. [Section V-B]

Fear/Incumbency Regime One of the two primary macroeconomic regimes identified in this paper (Axiom P3). Characterized by negative PC-1 (fear-dominant narrative), elevated FEAR index, and narrative dominated by institutional stability, coercive state power, and risk aversion. Associated with below-equilibrium price levels and contrarian upward mean-reversion signals in the M3 regime-ensemble model. [Sections IV, VIII-B]

Forecast Cone The primary visual output of the IUVO™ system: a 20-trading-day forward price range constructed from the M7 ensemble base path with upper and lower confidence bounds scaled by Jeffreys' prior on evidence strength. Bands widen when the model's signal is weak and narrow when evidence is concentrated. Overlaid with IRF-adjusted base path adjustments, EWI stress overlay, and a rolling 20-day OLS trend line that is dynamically fitted each week. [Section XI-G-B]

Geopolitical Shock Composite A composite variable constructed by normalizing and summing eight geopolitically-charged WORDSTAT topic series including Iran/Israel strike narrative, Russia-Ukraine, oil production, and ceasefire/peace talk language. Treated as an exogenous I(0) regressor in the VECM. Activated sharply in March 2026 during the

U.S./Israel strikes on Iran, peaking at 151 WP10K and driving the GEO_REGIME binary indicator. [Sections III-A, XI-C]

Granger Causality A statistical test for whether lagged values of one time series contain predictive information about another series, above and beyond the target series' own lags. In this paper, three methods are used: Toda-Yamamoto, Sims/Hsiao, and Chen VARMA. Results confirm no short-run directional forcing between PC-1 and the Deviation from Target series, consistent with the co-integration structure. [Section V-E]

Headline Intelligence System The IUVO™ agentic Q&A layer: a local retrieval-augmented generation (RAG) architecture that queries 768,750 vetted daily headlines spanning August 2022 through May 2026 — the same corpus that drives the forecasting model. Users pose questions in natural language; a locally-hosted language model synthesizes responses grounded in retrieved headline evidence with no external API or internet access. Supports historical analogy queries, scenario stress-testing, and cross-domain translation. [Section XI-G-C]

I(1) Process A time series that is non-stationary in levels but stationary in first differences — i.e., integrated of order one. Both the narrative PC-1 composite and the log S&P 500 are confirmed I(1) by ADF and KPSS unit root tests, satisfying the precondition for co-integration testing. A stationary series (I(0)) cannot co-integrate with an I(1) series. [Section V-A]

Impulse Response Function (IRF) A function tracing the expected dynamic response of one variable (S&P 500 returns) to a one-standard-deviation shock in another variable (a narrative topic series), holding all other variables at their historical values. Estimated here via bivariate VAR(5) models for seven narrative topics. IRF results are operationalized as additive overlays on the M7 forecast cone base path when a topic is currently above its 70th-percentile activation threshold. [Section XI-E]

Inverse-MSE Ensemble (M7) The seventh model in the competition and the IUVO™ primary production model. Combines M1 through M6 by weighting each model inversely proportional to its expanding-window mean squared error — models with lower recent forecast errors receive higher weight. Achieves 95.5% directional accuracy, Sharpe ratio of 24.3, and Calmar ratio of 292× across 917 trading days. [Sections XI, XI-B]

IUVO™ System The deployed forecasting and intelligence system that implements the Perceptual Macroeconomics framework. Published weekly at iuvosoft.com. Consists of a statistical channel (M7 ensemble forecast cone, IRF overlays, EWI, narrative cycle periodicity), a structured briefing interface (six analytical sections), and a Headline Intelligence Q&A layer (local RAG system). The system demonstrates the organizational intelligence architecture described in Section X using the S&P 500 as the illustrative outcome variable. [Sections XI-F, XI-G]

Jeffreys' Prior A non-informative prior in Bayesian analysis, defined as proportional to the square root of the Fisher information. In this paper, Jeffreys' information metric is used to scale the forecast cone boundaries: a stronger narrative signal (higher Fisher information) produces narrower cone bands (higher certainty); a weaker signal produces wider bands. The Jeffreys multiplier is computed in the Stata pipeline and is also the basis of the planned weight-of-evidence (WoE) trajectory. [Sections XI-G-B, XI-D]

Johansen Test A likelihood-ratio-based multivariate test for co-integration that can detect multiple co-integrating vectors in a system of more than two I(1) variables. Unlike the Engle-Granger test, the Johansen procedure tests both the trace statistic (null of at most r co-integrating vectors) and the maximum eigenvalue statistic. Applied here to the five-variable system including PC-1, log S&P 500, FEAR, polarity, and institutional legitimacy. [Section V-C]

Kelly Fraction The theoretically optimal fraction of available capital to commit to each signal, derived from the Kelly criterion: win rate minus (1−win rate) divided by edge ratio. A Kelly fraction of 0.916 for M1 means a Kelly-sizing investor would commit 91.6% of resources per signal. In practice, organizations use half- or quarter-Kelly to account for estimation error. [Section XI-B-F]

Mundell Comparative Statics A method of policy analysis developed by Robert Mundell for open-economy macroeconomics, which examines how equilibrium outcomes change in response to discrete shifts in exogenous variables. Applied here to the narrative-price system: the five narrative dimensions (polarity, fear, surprise, legitimacy, topic composition) are treated as structural state variables, and the Policy Derivatives Matrix (Table 7) shows the direction and magnitude of S&P 500 response to a one-standard-deviation shift in each. [Sections II-B, VII]

Narrative Cycle Periodicity The recurring temporal structure of the narrative-price system, identified via Fast Fourier Transform (FFT) analysis of the 932-day sample. The dominant S&P 500 cycle is 183 trading days (~9 calendar months); FEAR oscillates at a 57-day sub-cycle. The GEO composite shows a 229-day dominant period; TARIFF narrative at 306 days. INFLATION operates on CPI-release cadence (10–29 days) independently of the macro cycle. [Section XI-C]

PC-1 (First Principal Component) The first principal component of the approximately 185-variable WORDSTAT narrative taxonomy, capturing the dominant axis of variance in the narrative field. PC-1 loads positively on surprise, joy, and anticipation dimensions and negatively on fear, sadness, and coercive state power. Interpreted as a composite measure of the narrative field's orientation between constructive/optimistic and fearful/restrictive poles. Explains 88% of the variance in S&P 500 levels in the long-run co-integrating equation. Correlation with the S&P 500: 0.937. [Sections III-D, IV]

PPV (Positive Predictive Value) The proportion of positive test results that are true positives — i.e., AMBER+ alerts that are followed by an actual regime transition within the forecast window. Formally: PPV = True Positives / (True Positives + False Positives). A PPV of 58.3% for the IUVO™ EWI means that 58.3% of AMBER+ alerts in the 932-day sample were followed by a confirmed narrative regime transition within 10–20 trading days. PPV is the primary performance metric for the EWI because it is designed as a high-specificity instrument: it fires rarely (1.3% of days) but with high reliability when it does. The 58.3% PPV represents a 17× lift over the unconditional base rate of 3.4%. Contrast with sensitivity (the proportion of actual transitions preceded by an AMBER+ alert), which is a secondary consideration in a system optimized for low false-positive rates. [Section XII-B]

Perceptual Macroeconomics The framework introduced in this paper, grounded in the observation that asset prices are determined by agents' perceptions of economic reality — the narrative descriptions of events as they circulate in public discourse — rather than by the underlying events themselves. Formalized in Plato's cave terms: the market prices the shadows (narrative), not the objects (realized fundamentals). The empirical implication is that narrative variables co-integrate with asset prices while official statistics, which measure the objects after a lag, do not. [Sections I, II-E]

Retrieval-Augmented Generation (RAG) An architecture in which a language model's responses are grounded in documents retrieved from a local corpus rather than generated from training memory alone. In the IUVO™ Headline Intelligence system, the corpus is 768,750 vetted daily headlines; retrieved headlines are passed as context to the language model for synthesis. The RAG constraint prevents the model from generating unconstrained narrative outside the evidential record. [Section XI-G-C]

Sharpe Ratio Annualized risk-adjusted return equal to mean daily profit-and-loss divided by standard deviation of daily profit-and-loss, multiplied by $\sqrt{252}$. Measures return per unit of volatility. M1 achieves a Sharpe ratio of 23.0 — approximately 23 times the risk-adjusted return of a traditional long-only equity strategy (typical Sharpe ~1.0). The high values reflect the nature of the Deviation from Target series as the predicted variable. [Section XI-B-C]

Surprise/Disruption Regime One of the two primary macroeconomic regimes identified in this paper (Axiom P3). Characterized by positive PC-1 (surprise-dominant narrative), narrative dominated by U.S. executive agency, technological disruption, and market optimism. Associated with above-equilibrium price levels and upward ECT mean-reversion signals. Contrasts with the Fear/Incumbency regime. [Sections IV, VIII-B]

VECM (Vector Error Correction Model) A multivariate time-series model for co-integrated systems that decomposes each variable's change into a short-run dynamic component and a long-run error-correction component. The ECT in the VECM captures

the speed at which the system returns to its narrative-implied equilibrium after a deviation. In this paper, a five-equation VECM is estimated with narrative polarity, fear, surprise, institutional legitimacy, and log S&P 500. The estimated speed-of-adjustment is −0.043 per week. [Sections V-D, VII]

Weight-of-Evidence (WoE) Trajectory A planned extension to the IUVO™ briefing interface: a cumulative deciban time series showing how the evidence for or against the directional forecast has accumulated across recent trading days. Each day's evidence is converted to decibans using the Jeffreys information metric and accumulated, producing a trajectory that rises when the data confirms the forecast and falls when it disconfirms it. Currently the Jeffreys multiplier is computed and applied to cone width; accumulating it into a sequential deciban series is the planned next development step. [Section XI-G-B]

WORDSTAT The text analytics software used to measure daily narrative variables. Scores each daily headline file against a comprehensive topic and emotion taxonomy, producing approximately 185 variables measured as words per 10,000 (WP10K) for each trading day. The taxonomy spans political actors, geopolitical events, economic topics, institutional actors, and eight NRC emotion dimensions. The resulting panel dataset is the primary input to all statistical models in this paper. [Section III-A]

WP10K (Words Per 10,000) The normalization unit for all WORDSTAT narrative variables. Each topic score is expressed as the number of occurrences of words in that topic's dictionary per 10,000 total words in the daily headline file. This normalization controls for variation in the total volume of daily news coverage, ensuring that scores reflect the relative intensity of a topic rather than the absolute size of the day's file. An AMBER+ alert threshold of 6/10 on the EWI corresponds to a specific configuration of WP10K readings across the five channels. [Section III-A]

References

Hayek, F. A. (1945). The use of knowledge in society. American Economic Review, 35(4), 519–530.

Keynes, J. M. (1936). The general theory of employment, interest and money. Macmillan.

Plato. (c. 380 BCE). The Republic (B. Jowett, Trans.). Oxford University Press (1894 edition). Book VII, 514a–520a (the Allegory of the Cave).

Boudjellaba, H., Dufour, J.-M., & Roy, R. (1992). Testing causality between two vectors in multivariate ARMA models. Journal of the American Statistical Association, 87(420), 1082–1090.

Hsiao, C. (1981). Autoregressive modelling and money-income causality detection. Journal of Monetary Economics, 7(1), 85–106.

Sims, C. A. (1980). Macroeconomics and reality. Econometrica, 48(1), 1–48.

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics, 66(1–2), 225–250.

Akerlof, G. A., & Shiller, R. J. (2009). Animal spirits: How human psychology drives the economy. Princeton University Press.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. Journal of Finance, 61(4), 1645–1680.

Chevalier, J., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345–354.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263.

Elliott, G., Rothenberg, T. J., & Stock, J. H. (1996). Efficient tests for an autoregressive unit root. Econometrica, 64(4), 813–836.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. Econometrica, 55(2), 251–276.

García, D. (2013). Sentiment during recessions. Journal of Finance, 68(3), 1267–1300.

Green, T. C., Huang, R., Wen, Q., & Zhou, D. (2019). Crowdsourced employer reviews and stock returns. Journal of Financial Economics, 134(1), 236–251.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357–384.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica, 59(6), 1551–1580.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65.

Mayew, W. J., & Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. Journal of Finance, 67(1), 1–43.

Mundell, R. A. (1968). International economics. Macmillan.

Qiu, L., & Wang, S. (2018). SoLoMo is no solo! An empirical study of the synergistic effect of social, local, and mobile information on consumer visit behavior. Information Systems Research, 29(2), 435–451.

Shiller, R. J. (2017). Narrative economics. American Economic Review, 107(4), 967–1004.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. Operational Research Quarterly, 20(4), 451–468.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting, 32(3), 754–762.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting, 29(1), 108–121.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. Journal of Forecasting, 23(6), 405–430.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), Handbook of Economic Forecasting (Vol. 1, pp. 135–196). Elsevier.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunications sector: A profit-driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229.

Appendix: Stata Implementation Notes

A. Original Model (v8): JA_Forecast Single-Equation Backtest

The v8 companion file implements the single-equation rolling backtest described in the appendix of the original paper. See v10 notes below for the current production version. Key features: expanding-window training, stage-1 OLS residual computed inside the rolling window to prevent look-ahead bias, directional hit rate evaluation, Sharpe ratio, PnL vs always-long benchmark, and McNemar paired test for directional accuracy significance.

B. Extended Model (v9): Multi-Model Competition

The v9 companion file implements the multi-model competition described in Section XI of this paper. See v10 notes below for the current production version incorporating the Iran/oil geopolitical shock composite variables. Five models are evaluated in parallel on identical rolling backtests: M1 (JA_Forecast OLS two-stage), M2 (VECM error-correction signal), M3 (regime-ensemble), M4 (momentum-only benchmark), and M5 (polarity threshold rule). See v10 notes below for the current production version which adds M6 and M7.

The v9 code produces the following outputs for organizational deployment:

- MASTER_MODEL_SUMMARY_h1.csv — one-line performance summary for each model at h=1, with all nine metrics
- model_horse_race_by_h.csv — hit rate and Sharpe ratio for each model at each forecast horizon (h=1 to 5)
- short_selling_pnl_by_model_h.csv — short call rate, total short PnL, and average PnL per short position for each model
- edge_ratio_kelly_by_model_h.csv — win rate, average win, average loss, edge ratio, and Kelly fraction for each model
- max_drawdown_by_model.csv — maximum drawdown and Calmar ratio for each model
- tc_sensitivity_all_models.csv — net PnL and Sharpe ratio at five transaction cost levels for each of the seven models
- cumulative_pnl_all_models_h1.png — visual comparison of all seven models' cumulative PnL vs always-long benchmark
- cumulative_short_pnl_all_models_h1.png — short-position-only cumulative PnL for all models
- hit_rate_all_models_by_h.png — directional accuracy comparison across horizons
- sharpe_all_models_by_h.png — Sharpe ratio comparison across horizons
- tc_sensitivity_pnl_h1.png — transaction cost sensitivity for all models

Implementation note: the Cholesky decomposition for the VECM (M2) orders COERCIVE_STATE_POWER as most exogenous, followed by FEAR_INDEX, NARRATIVE_POLARITY, EMOTION_SURPRISE, and LOG_SP500 as most endogenous — consistent with Axiom P5 and the theoretical structure of the paper. Sensitivity to this ordering is tested in robustness checks by inverting the polarity-surprise ordering.

C. Extended Model (v10): Geopolitical Shock Integration, Causality Pre-Analysis, M6 and M7 Ensemble The v10 companion file (JA_Forecast_v10_multimodel.do) adds the Iran/oil geopolitical shock variables identified in Section III.A and extends all seven models to incorporate the geopolitical shock composite_F composite as an exogenous I(0) regressor. It also implements the three-method Granger causality pre-analysis documented in Section V.E. Key additions: (1) automatic renaming of special-character Excel column names to Stata-legal variable names; (2) geopolitical shock composite_F and GEO_REGIME binary indicator construction; (3) geopolitical shock composite_F in M1 both stage-1 and stage-2 OLS; (4) geopolitical shock composite_F in M2 VECM short-run equation only (long-run cointegrating vector unchanged); (5) GEO_REGIME branching in M3 regime-ensemble; (6) a geopolitical regime performance split table comparing model accuracy during normal vs. high-GEO-shock periods; (7) export of geopolitical_regime_performance.csv alongside all existing v9 outputs; and (8) three-method causality battery (Toda-Yamamoto, Sims/Hsiao, Chen VARMA tree) exported to causality_comparison_TY_Sims_Chen.csv. All three methods confirm no short-run directional forcing between PC1 and the Deviation from Target series, consistent with and confirming the co-integration structure documented in Section V. Section 13 of v10 adds M6 (20-day OLS time trend) and M7 (inverse-MSE ensemble of M1–M6) to the rolling backtest and exports MASTER_MODEL_SUMMARY_h1.csv with all seven models. Section 14 exports cone_forecast_points_m7_FIXED.csv with per-horizon residual SD, anchor shift, and model_used column. Section 15 exports M1–M7 forecast paths from T+1 to T+20 as m1_to_m7_forecast_paths_Tplus1_Tplus20.png. M7 achieves 95.5% directional accuracy and a Calmar ratio of 292.5 across 917 trading days. The IUVO™ system deploys M7 as its primary production cone model.

D. Regime Detection Extension: JA_RegimeDetection_v1.do

E. IRF Estimation and Watchlist: Python Implementation

The companion file JA_RegimeDetection_v1.do implements the bear/bull regime detection framework documented in Section VIII-B. It runs after JA_Forecast_v10_fixed.do using the same source data file and output directory. The file is organized in three self-contained layers. Layer 1 (Section A) constructs BEAR_REGIME and Bear Pressure Index, produces the narrative composition cross-validation table and bear pressure chart, and requires no additional packages. Outputs: bear_bull_regime_series.csv, narrative_by_regime.csv, bear_regime_performance.csv,

and regime_bear_index_series.png. Layer 2 (Section B) implements Markov-switching via Stata's built-in mswitch (Stata 14+), extracts smoothed and filtered regime probabilities, cross-validates Markov states against the narrative composition, and tests whether PC_1 Granger-causes regime transitions using Toda-Yamamoto. Layer 2 degrades gracefully if mswitch fails. Layer 3 (Section C) adds M6 and M6b to the rolling backtest — M6 uses BEAR_REGIME as a contrarian directional signal; M6b scales position by Bear Pressure Index — and produces Table 17 and the cumulative PnL comparison chart. All rolling window estimation uses trailing-only data with no look-ahead bias. Integration note: to include M6 in the v10 master summary table, merge RES_M6 into the v10 PANEL file on t_index and expand the model_list local in Sections 7–13 of JA_Forecast_v10_fixed.do.

E. IRF Estimation and Watchlist: Python Implementation

The IRF engine and emerging topic watchlist are implemented in Python and embedded directly in the IUVO™ weekly production script (v2.3). The VAR(5) estimation uses the statsmodels library. The watchlist computation screens all macro-relevant WORDSTAT columns against their expanding 70th-percentile threshold, computing five-day and twenty-day momentum, a proximity score, and a 20-observation sparkline for display. Both the IRF results and watchlist data are embedded as JavaScript constants in the HTML forecast page, requiring no server-side computation at render time.

New WORDSTAT fields declared in the analyst input block (NEW_WORDSTAT_FIELDS) are automatically routed to five downstream locations: the headlines table (with a ❖ new badge for two weeks), the analyst tab emerging signals section, the narrative cycles chart (when ≥10 observations are available), the what-if scenarios tab (when above threshold), and the baseline narrative paragraph. After two confirmed weekly runs, the field is added permanently to the Step 2 extraction block and removed from the NEW_WORDSTAT_FIELDS declaration.

PERCEPTUAL MACROECONOMICS

Post-Publication Quality Review Addendum

Acknowledging and Incorporating OCC 5C Governed Review Findings

John M. Aaron · Milestone Planning and Research, Inc. · May 2026

Prefatory Note

This addendum accompanies the paper *Perceptual Macroeconomics: Narrative Co-Integration, Regime Dynamics, and the Limits of Federal Reserve Data in Asset Price Forecasting* (Aaron, 2026). It documents a systematic post-publication quality review conducted using the OCC 5C QA System — a Bayesian Weight-of-Evidence governed review architecture developed by Milestone Planning and Research, Inc.

The review was conducted in two independent passes: one using Claude (Anthropic) and one using ChatGPT (OpenAI). The use of two independent LLM reviewers is a deliberate architectural choice: different models surface different risk classes, and the intersection of their findings provides a more complete audit than either alone could produce. This proved to be the case. Claude's review identified computational and methodological issues in the forecasting code; ChatGPT's review independently surfaced temporal data leakage as a distinct and material concern that Claude had not emphasized.

The addendum is organized as follows. Section A describes the review process and governance framework. Section B addresses the findings that require qualification of the paper's claims. Section C describes the code corrections made to the Stata do-file as a result of the review. Section D presents the corrected empirical results and compares them to the originally reported figures. Section E addresses the causality findings and their implications for the paper's theoretical framing. Section F provides the methodological clarifications that address reviewer concerns without requiring revision of the paper's core claims.

A. The Review Process and Governance Framework

The OCC 5C QA System applies a seven-dimension Bayesian Weight-of-Evidence scoring architecture to LLM-generated analysis, assigning each dimension a score from −2 (strong concern) to +2 (strong support) and producing a composite verdict. The governing prescript — embedded in every prompt — commits the reviewing LLM to a specific epistemological doctrine: confidence proportional to evidence, contradictions surfaced rather than smoothed, and a formal acknowledgment that no LLM can fully audit outputs produced by a system with the same training distribution. Human review is therefore a logical necessity, not a quality option.

The initial Claude review of the Stata do-file (

JA_Forecast_v12.do) and the paper PDF returned a composite WoE score of +8 (PROCEED) with a Medium confidence rating, conditional on resolution of four flagged items. A subsequent ChatGPT review of the same material independently flagged temporal data leakage as a fifth concern. Both review outputs were then passed through the Reviewer stage in fresh LLM sessions — a design requirement that prevents the reviewer from being influenced by the production session in which the work was created.

> **OCC 5C Review Summary**
> Initiator session: Claude (Sonnet 4.6) | Reviewer session: Claude (Sonnet 4.6, fresh session)
> Parallel review: ChatGPT (GPT-4o) | Combined findings: 9 issues across code and paper
> Composite WoE score: +8 (PROCEED, conditional) | Confidence: Medium
> Code corrections made: 6 (Stata do-file v12 → v14) | Paper qualifications: 4

B. Findings Requiring Qualification of Paper Claims

B.1 Headline Performance Metrics: Corrected Figures

The abstract and Section XI-B reported M7 directional accuracy of 95.5% and a Calmar ratio of 292× across 917 trading days. These figures were drawn from a prior run of an earlier script version. The corrected v14 run, incorporating all code fixes described in Section C, produces the following results at h=1 across N=728 backtest observations:

Model	Hit Rate (h=1)	Sharpe Ratio	Calmar Ratio	Max Drawdown	Final PnL
M1: JA OLS Narrative	93.3%	23.50	252.5	1,240	$313,159
M2: VECM Error-Correction	91.3%	22.17	79.3	3,872	$307,252
M3: Regime-Ensemble	92.4%	22.93	234.2	1,327	$310,724
M4: Momentum (benchmark)	61.0%	4.30	9.4	10,563	$98,904

Model	Hit Rate (h=1)	Sharpe Ratio	Calmar Ratio	Max Drawdown	Final PnL
M5: Polarity Threshold	87.2%	19.68	70.4	4,179	$294,167
M6: Time Trend (benchmark)	32.4%	−6.99	0.006	160,821	$951
M7: Inv-MSE Ensemble	92.4%	22.79	134.8	2,300	$310,105

Several aspects of these corrected figures warrant explicit discussion.

M1 outperforms M7 on every primary metric. M1 (JA OLS Narrative) achieves the highest hit rate (93.3% vs 92.4%), highest Sharpe ratio (23.50 vs 22.79), highest Calmar ratio (252.5 vs 134.8), and lowest maximum drawdown ($1,240 vs $2,300). The paper's framing of M7 as the primary deployment model for the IUVO™ system remains defensible on ensemble-theoretic grounds — the inverse-MSE weighting of M7 provides robustness across regimes that a single-model specification cannot guarantee — but the paper's claim that M7 achieves the highest competition performance is not supported by the corrected figures. M1 is the competition winner.

The corrected M7 Calmar ratio is 134.8×, not 292×. This remains an exceptional figure by any published benchmark standard. The reduction reflects primarily the correction of the M2 ECT omission (the error-correction term was estimated but not applied in the forecast, making M2 artificially smooth) and the rolling standardization fix for GEOSHOCK. These are genuine methodological improvements, not restatements of the same result with different labels.

The abstract's claim of 95.5% directional accuracy for M7 is not reproduced. The corrected figure is 92.4%. The 95.5% figure likely reflects an earlier script version before the Hsiao FPE loop correction and the ECT fix, both of which changed the ensemble weighting dynamics. The 92.4% figure is reported here as the definitive result.

The backtest covers N=728 h=1 observations, not 917 trading days. The 917-day figure referred to the full data panel; the backtest window is determined by the minimum training requirement (min_train) and the expanding-window design, yielding 728 forecast origins at h=1. This distinction is clarified in the methodology note in Section F.

B.2 The Validation Domain Gap: Deviation-Domain vs. Raw S&P 500 Accuracy

The paper's performance claims are stated in terms of directional accuracy on the *Deviation from Target* series — the signed gap between the realized S&P 500 and a 12.2% annualized compounding baseline. This is not equivalent to directional accuracy on raw S&P 500 returns. The internal contradiction audit flagged this explicitly as a VALIDATION_DOMAIN_GAP for all narrative models. The following table makes the distinction explicit:

Model	Deviation-domain hit rate	Raw S&P 500 hit rate	Gap
M1	93.3%	54.7%	38.6 pp
M2	91.3%	54.1%	37.2 pp
M3	92.4%	54.4%	38.0 pp
M7	92.4%	54.1%	38.3 pp

Raw S&P 500 directional accuracy for all narrative models is approximately 54% — marginally above chance and not statistically significant as a trading signal on the raw index. This is not a contradiction of the paper's claims; it is a precise statement of what the paper claims. Axiom P6 establishes explicitly that the narrative system has no short-run return predictability. The models predict the direction of the *gap between price and narrative equilibrium*, not the direction of tomorrow's price. A reader who interprets 93% directional accuracy as a claim about raw S&P 500 return forecasting would be misreading the paper. This addendum makes that boundary explicit.

B.3 Causality Tests: Correlation, Not Short-Run Forcing

Section V.E of the paper presented Granger causality tests as evidence for a causal relationship between narrative and asset prices. All four methods — Toda-Yamamoto, Sims first-difference, Hsiao FPE, and Chen VARMA tree — return null results: no direction of Granger causality is confirmed in either direction.

Method	Test pair	Verdict
Toda-Yamamoto	PC_1 ↔ Deviation	Fail to reject both directions
Toda-Yamamoto	GEOSHOCK → Deviation	No geo causality
Sims (1980) first-diff	dPC1 ↔ dDeviation	No causality (first-diff)
Hsiao (1981) FPE	dPC1 → dDeviation	dPC1 does NOT cause dDeviation
Chen VARMA tree	dPC1 ↔ dDeviation	No detectable causality

These results do not invalidate the paper's central empirical finding. The co-integration evidence — confirmed by both Engle-Granger and Johansen tests — establishes that PC_1 and the S&P 500 level share a common stochastic trend. Co-integration does not require short-run Granger causality; it requires only that the two series are drawn toward a common long-run equilibrium, with deviations being mean-reverting. The VECM error-correction architecture captures precisely this relationship.

What the null causality results do require is a revision of the causal language in Section V.E and in the introduction. The paper should characterize the narrative-price relationship as a *long-run co-integrating association* rather than a causal forcing relationship. The practical implications for forecasting are unchanged: the co-integration relationship is real, the error-correction mechanism operates, and the models generate genuine directional accuracy. What cannot be claimed is that short-run changes in

narrative *drive* short-run changes in price in the Granger sense. This distinction is standard in the co-integration literature and does not weaken the paper's empirical contribution.

C. Code Corrections: Stata Do-File v12 to v14

Six corrections were made to the Stata implementation as a result of the quality review. The corrected script is designated *JA_Forecast_v14.do*. All future forecast runs will use this version as the canonical implementation. The corrections are listed in chronological order of discovery; the final two (Fixes A and B) address the temporal leakage concerns raised by the ChatGPT parallel review.

Fix	Description	Type	Impact
v13 Fix 1	Hsiao FPE loop double-update bug: best_p_0 was never updated from default (1) due to sequential if-statements overwriting the condition before the second test	Code logic error	Optimal lag selection for Hsiao FPE and Chen VARMA AIC was silently frozen at p=1 regardless of data
v13 Fix 2	M5 long PnL copy-paste: pnl_long_sum_m5 summed pnl_short_m5 instead of pnl_long_m5	Copy-paste error	M5 long PnL export column was incorrect; short PnL double-counted
v13 Fix 3	M6 time index mismatch: local _n (range 1–20) used in regression but global t (range ~200–888) used in forecast	Implementation error	M6 forecast extrapolated to wrong point on the fitted line; caused near-zero PnL
v13 Fix 4	M2 ECT omission: m2_bECT estimated each training window but never applied in the forecast	Implementation error	M2 was not deploying its error-correction mechanism; label 'VECM Error-Correction' was misleading
v14 Fix A	GEOSHOCK global standardization: full-sample mean/sd used to standardize, contaminating early observations with 2026 statistics	Temporal leakage	GEOSHOCK values in 2022–2024 were normalized using parameters that included the Iran/oil shock spike of 2026
v14 Fix B	M5 polarity signal read at future horizon t instead of forecast origin o	Temporal leakage	M5 used realized future polarity to generate its signal; a strict backtest requires only information available at origin

Note on practical impact of leakage fixes

The GEOSHOCK leakage (Fix A) has limited practical impact on most backtest origins because IRAN_STRIKE is near zero across 2022–2024 — the Iran/oil shock activates only in late February 2026. However, the rolling standardization change reclassified the geopolitical shock regime: the high-GEO observation count increased from 46 (v13) to 100 (v14) as rolling statistics correctly identified the shock period. M3 Calmar improved from 191 to 234 as a result.

The M5 polarity fix (Fix B) had a small positive effect: M5 Calmar improved from 63 to 70. Reading the signal at origin rather than at the future horizon removes one source of forward-looking bias from M5's performance metrics.

The M7 Calmar reduction (from 206 to 135) reflects primarily the M2 ECT fix (v13 Fix 4), which changed ensemble weighting by altering M2's MSE profile, not the leakage fixes.

D. Corrected Empirical Results

D.1 Multi-Horizon Performance

The following table presents directional accuracy and Sharpe ratios for the primary models across horizons h=1 through h=5 from the v14 run.

	h=1	h=2	h=3	h=4	h=5
M1 hit rate	93.3%	93.5%	93.4%	92.9%	92.7%
M7 hit rate	92.4%	92.2%	91.9%	91.9%	92.0%
M1 Sharpe	23.50	23.62	23.57	23.53	23.54
M7 Sharpe	22.79	22.56	22.58	22.67	22.86
M4 hit rate (benchmark)	61.0%	60.9%	61.3%	62.0%	61.8%
M6 hit rate (benchmark)	32.4%	32.3%	32.0%	31.9%	31.7%

Performance is remarkably stable across horizons for the narrative models. M1 and M7 maintain directional accuracy above 91% through h=20 (the full evaluation horizon). This is consistent with the co-integration framework: the narrative-price relationship is a level relationship operating on a slow macro timescale, and its predictive content does not decay sharply at short horizons.

D.2 Geopolitical Regime Performance

The rolling GEOSHOCK standardization (Fix A) correctly classified 100 observations as high-geopolitical-shock regime (vs 46 in v13). Performance in the high-shock regime is strong across narrative models:

Model	Normal regime hit rate (n=628)	High GEO regime hit rate (n=100)
M1	92.2%	100.0%
M2	90.0%	100.0%
M3	91.7%	97.0%
M5	86.0%	95.0%
M7	91.2%	100.0%
M4 (benchmark)	61.8%	56.0%
M6 (benchmark)	30.9%	42.0%

The 100% hit rates for M1, M2, and M7 in the high-GEO regime (n=100) reflect the GEOSHOCK variable's strong regime-conditioning effect. In the presence of a confirmed geopolitical shock, narrative variables

dominate and the model signal is very clear. This validates the Iran/oil shock extension of the framework and is consistent with the paper's Perceptual Macroeconomics thesis: when the narrative field is dominated by a single high-intensity topic, the market-narrative co-integration signal is at its strongest.

A methodological caveat is warranted: 100% hit rates on any subsample should be interpreted cautiously. The n=100 high-GEO regime observations are not independent — they cluster in a single episode (late February through May 2026) — so this result is not equivalent to 100% accuracy across 100 independent events. It is better characterized as: the model correctly tracked the narrative-price relationship throughout the geopolitical shock and resolution episode.

D.3 Short-Selling Alpha

Short-selling performance at h=1 from the v14 run:

Model	Short calls	Short PnL	Long PnL	Total PnL
M1	164	$19,511	$293,648	$313,159
M3	176	$18,293	$292,430	$310,724
M7	164	$17,984	$292,121	$310,105
M5	58	$10,015	$284,152	$294,167
M2	162	$16,557	$290,695	$307,252

Short-selling alpha is positive and material for all narrative models. M5, despite its simpler architecture (a polarity ratio threshold rule), generates $10,015 in short PnL from only 58 short calls — the highest per-short-call return in the competition at approximately $173 per call. This reflects the precision of the polarity signal when it does trigger: it fires infrequently but accurately.

E. The Causality Results and the Co-integration Interpretation

The null Granger causality results across all four methods require the following restatement of the paper's theoretical framing, which applies retroactively to the language in Section V.E, the Abstract, and Section I.

E.1 What the null causality results mean

Granger causality tests whether past values of X improve the forecast of Y beyond what Y's own history provides. A null result means that the short-run dynamics of PC_1 do not Granger-cause the short-run dynamics of Deviation from Target in first differences. This is entirely consistent with Axiom P6, which states that forward 1-day narrative-return correlations reach a maximum magnitude of 0.049. The narrative system has no short-run return predictability. That is not a new finding — it was established as an axiom in the paper.

What the co-integration tests establish is different: PC_1 and the S&P 500 level share a *common stochastic trend* — they are drawn to a long-run equilibrium from which short-run deviations are mean-reverting. This is a level relationship, not a short-run causal forcing relationship. The VECM captures the speed at which deviations from this equilibrium correct, not whether today's narrative change predicts tomorrow's return.

E.2 The language that should be used

Correct framing

'PC_1 and the S&P 500 price level are co-integrated: they share a common stochastic trend confirmed by Engle-Granger and Johansen tests. Short-run deviations from the narrative-price equilibrium are mean-reverting, as estimated in the VECM. This relationship is associative rather than causally forcing in the Granger sense: changes in PC_1 do not predict next-period changes in price in short-run dynamics. The forecasting advantage arises from the error-correction mechanism operating on the level relationship, not from short-run narrative-to-price impulses.'

Language to avoid

'Narrative drives asset prices' → replace with 'narrative co-integrates with asset prices'

'PC_1 causes the Deviation from Target' → replace with 'PC_1 and the S&P 500 share a common stochastic trend'

'The narrative system predicts price' → replace with 'the narrative system predicts the direction of mean-reversion toward narrative equilibrium'

These restatements do not alter the paper's empirical contribution. The co-integration finding is intact. The VECM is intact. The forecasting competition results are intact (with the corrected figures in Section D). What changes is the precision of the theoretical claim: the system identifies a *co-integrating*

association — not a causal mechanism in the strict sense. This is a stronger epistemological position, not a weaker one.

F. Methodological Clarifications for Reviewers

F.1 The _F variable naming convention

Every variable with the suffix _F (PC_1_F, FEAR_F, TARIFF_F, GEOSHOCK_F, etc.) in the Stata do-file is constructed using `tssmooth ma ... window(5 0 0)`. In Stata's tssmooth syntax, window(L 0 F) specifies L lagged periods, 0 current-period weight, and F forward periods. **window(5 0 0) is a pure trailing 5-day moving average using only past and current data.** No future observations enter these variables. The _F suffix denotes 'filtered' or 'smoothed,' not 'forward-looking.' This naming convention is acknowledged as potentially confusing and is clarified here explicitly.

The comment block in Section 2 of the do-file states: "`All features use trailing windows only -- no forward leakage.`" This statement is accurate.

F.2 PC_1 construction: WORDSTAT external computation

The PC_1 narrative composite is not constructed within the forecasting do-file. It is produced externally by WORDSTAT text analytics software from daily headline word-frequency data and imported as a pre-computed variable in the Excel source file. The PCA underlying PC_1 was computed on the full 932-day sample. This is standard practice in academic research using pre-constructed indices.

A strict academic reviewer may flag full-sample PCA construction as a form of factor leakage: the eigenvectors were computed using data that includes the test period, so the factor loadings encode some covariance structure from the future. This is acknowledged. The practical mitigation is that PC_1 is a slowly evolving structural index reflecting long-run narrative composition — its factor loadings are not expected to change materially across subsamples of a 932-day panel. A rolling-window PCA robustness check is recommended for a future paper revision but is not undertaken here.

F.3 Scenario-conditioned forecasting vs. unconditional prediction

The multi-model competition evaluates models on their ability to predict the direction of the *Deviation from Target* series — the gap between the S&P 500 and a 12.2% annualized compounding baseline — not the direction of raw S&P 500 returns. All predictor variables at forecast horizon h are constructed from data available at or before origin o (see F.1). The forecasting architecture is therefore strictly *information-set conditioned on the origin date*, not scenario-conditioned in the sense of using hypothetical future states.

The 93% directional accuracy figures should be interpreted as: *given the narrative state at origin o, the model correctly predicted whether the Deviation from Target series would move up or down by horizon h, using only information available at o.* This is a meaningful and well-defined forecasting claim. It is not a claim about raw S&P 500 return prediction, and it is not a scenario-conditioned exercise.

F.4 Backtest sample size: N=728 at h=1

The paper's abstract referenced 917 trading days as the evaluation window. The correct figure is N=728 forecast origins at h=1, reflecting the minimum training window requirement in the expanding-window backtest design. The full data panel is 932 days (August 2022 through May 2026); the first 204 observations are consumed as the minimum training window before the first backtest origin. The 728 h=1 observations represent approximately 2.9 years of daily out-of-sample forecast evaluation within the rolling design.

G. Summary of Changes

The following table summarizes all changes documented in this addendum:

Item	Original paper claim	Corrected statement	Location
M7 directional accuracy	95.5%	92.4% (h=1, N=728)	Abstract, Section XI-B
M7 Calmar ratio	292×	134.8×	Abstract, Section XI-B
Highest-performing model	M7	M1 on all primary metrics	Section XI-B
Evaluation window	917 trading days	728 h=1 forecast origins	Abstract, Section XI-B
Causality framing	Narrative drives/causes prices	Long-run co-integrating association; no short-run Granger causality	Section V.E, Introduction
_F variable naming	Implicit (undefined)	Trailing 5-day MA; no forward-looking component	Section III, Appendix
GEOSHOCK standardization	Full-sample statistics	Rolling expanding-window statistics (v14 fix)	Section III, Appendix
M5 signal construction	Not specified	Signal evaluated at forecast origin o, not at future horizon	Section XI-A, Appendix

What is unchanged

The co-integration finding (Engle-Granger and Johansen) is unchanged and intact.
The VECM architecture and error-correction mechanism are unchanged.
The R^2=0.611 (N=765, Newey-West HAC) regression result is unchanged.
The Perceptual Macroeconomics theoretical framework is unchanged.
The organizational generalization argument is unchanged.
All narrative models substantially outperform the momentum and time-trend benchmarks.
Short-selling alpha is positive and material for all narrative models.
The geopolitical shock regime extension is validated by the corrected results.

John M. Aaron

Milestone Planning and Research, Inc.

May 2026

This addendum was produced with the assistance of the OCC 5C QA System (Milestone Planning and Research, Inc., Release 1 v3.1). The quality review was conducted using Claude Sonnet 4.6 (Anthropic) and ChatGPT GPT-4o (OpenAI) as independent reviewers. Human authority and final judgment remain with the author.