

LLM Token Entropy as a Leading Indicator of Realized Volatility: Evidence from Mid-Cap Equities

Oleksandr Podoliako

Preprint — v8

Abstract

Large language models assign a probability distribution to every token they process; the entropy of that distribution is a by-product of inference that requires no fine-tuning. We study whether this intrinsic token entropy, computed as a pretrained LLM reads a fixed weekly macroeconomic prompt, carries information about forward equity-market risk — and, more centrally, how such a signal can be validated against the pervasive risk of training-data contamination. Our primary contribution is methodological: a model-training-cutoff protocol that separates genuine out-of-distribution (OOD) detection from encoded hindsight by comparing entropy across models whose training corpora pre- and post-date the episode under study. The work is an incremental methodological extension of the news-entropy volatility-forecasting literature (Glasserman and Mamaysky, 2019; Glasserman et al., 2023), not a new paradigm: that literature trains a language model on a news corpus and measures deviation from the corpus distribution, whereas we read the intrinsic next-token entropy of a large pretrained model processing a fixed macro prompt, with no corpus training.

As an illustrative application, a proof-of-concept system processes energy prices, semiconductor prices, and a fixed-rule political-news summary through GPT-4o with logprobs enabled and evaluates a normalized entropy signal against forward realized volatility for 24 mid-cap US equities over 2018–2024 (363 weeks), with out-of-sample evaluation on 2024. The signal is best characterized as a sector-specific, medium-horizon indicator, not a universal risk measure. Pooled in-sample, the entropy signal attains an information coefficient of $IC \approx +0.029$ at the 8-week horizon, only marginally significant once serial correlation is accounted for (Newey–West $p \approx 0.067$); it is competitive with a VIX baseline overall and superior in commodity-exposed sectors, but out-of-sample cross-sectional significance is *not* established (10 of 17 tickers with $IC > 0$, $p = 0.314$). Directional F1 rises monotonically with horizon, but we show this rise is largely mechanical — at the pooled level it tracks a horizon-matched random benchmark — so the F1 figures should not be read as standalone evidence of skill. Because GPT-4o’s training cutoff post-dates most of the backtest, its in-sample results are confounded by potential contamination and are reported as suggestive rather than confirmatory. The only contamination-clean evidence comes from GPT-2 (training data cutoff end of 2017, well before any COVID reporting), which registers an anomalous information environment in the post-WHO-report window of January 2020 ($n \approx 4$ weeks), weeks before any visible market stress. SmolLM2-135M was trained on post-2020 web data and is therefore itself contaminated with respect to COVID; it is retained only as a small-model control, not as independent clean evidence. The pre-report GPT-4o spike is replicated by neither model and is flagged as a likely contamination artifact. The clean evidence thus rests on a single model and roughly four weeks. This narrow, model-agnostic result, together with theoretically coherent sector-specific crisis resonance across independent events, constitutes the strongest support for the underlying mechanism. Code and replication materials are available at <https://github.com/OleksandrPodoliako/entropy-risk>.

Contents

1	Introduction	3
2	Related Work	4
2.1	Alternative Data and Text-Based Risk Signals	4
2.2	Uncertainty Quantification in Large Language Models	5
2.3	Anomaly Detection in Financial Risk Management	6
3	Methodology	6
3.1	Theoretical Foundation	6
3.2	System Architecture	7
3.3	Outcome Variable	8
3.4	Baselines	9
3.5	Contamination Test Protocol	9
3.6	Verification Matrix	9
4	Data and Experimental Setup	9
5	Results	11
5.1	Volatility Prediction: Entropy vs. VIX	11
5.2	Directional Event Detection: F1 Analysis	12
5.3	Sector Analysis	13
5.4	Leading Indicator: Cross-Correlation and Conditional Analysis	13
5.5	Cross-Ticker Crisis Event Analysis	14
5.6	COVID-19 Contamination Test	15
5.7	Case Studies: Aggregate Signal Dynamics	16
5.8	E_{mean} vs. E_{peak} Ablation	17
5.9	Macro Prompt Expansion	17
5.10	Model Comparison: GPT-4o vs. SmolLM2-135M	18
5.11	Statistical Robustness: Base Rates, Multiple Comparisons, and Incremental Information	18
6	Discussion	19
6.1	Structural vs. Reactive Signal	19
6.2	Sector Specificity as Theoretical Prediction and Empirical Test	20
6.3	On the Relationship Between F1 and IC	20
6.4	Contamination Testing as Methodological Standard	20
6.5	Prompt Composition as Critical Design Parameter	21
6.6	Practical Implications for Risk Management	21
7	Limitations	21
8	Conclusion	22

1 Introduction

Market risk assessment has traditionally relied on retrospective signals: historical volatility, statistical price deviations, and macroeconomic indicators. These approaches describe risk well in stable regimes but are systematically delayed at moments of abrupt environmental change — precisely the junctures at which early warning is most valuable. A spike in realized volatility, by definition, can only be observed after the price has moved; the VIX, while forward-looking in option-market terms, is itself reactive to the informational environment that has already materialized in equity derivatives.

In parallel, a substantial body of research has developed around the use of alternative textual data in financial analysis. News articles, earnings call transcripts, and social-media posts are processed by natural language processing (NLP) methods to extract sentiment or topic classifications, which are then deployed as predictors of price movement (Tetlock, 2007; Bollen et al., 2011). Despite significant progress in this direction, all existing approaches share a common architectural assumption: the language model is used as a classifier, and its output — positive, negative, neutral — constitutes the final analytical product.

This paper studies a different point in the design space. Rather than querying a model about the semantic content of macroeconomic text, we measure the model’s uncertainty while processing it. The theoretical foundation is the following observation: large language models develop an implicit model of normality during pretraining on large text corpora. When an input context deviates semantically from the distribution of training data, the entropy of the per-token probability distribution increases. We hypothesize that this entropy constitutes a measurable signal of anomalousness in the input informational environment — and, accordingly, a candidate leading indicator of elevated market risk, specifically realized volatility.

This idea is not new in its essentials. Glasserman and Mamaysky (2019) already use an entropy (“unusualness”) measure of news text to forecast forward realized volatility months ahead, and Glasserman et al. (2023) compute that entropy with a recurrent-neural-network language model trained on a news corpus — i.e. “language-model entropy as a market predictor” is established prior art (Section 2). Our narrow departure is the measurement instrument: those works train a model on a news corpus and quantify deviation from that corpus’s distribution, whereas we use the *intrinsic* next-token entropy of a large pretrained LLM processing a fixed macro prompt, requiring no corpus training of our own. We frame this as an incremental methodological extension rather than a new paradigm.

The harder problem — and our primary contribution — is validation. Any study that uses a large modern LLM to “detect” historical anomalies faces a contamination hazard: the model’s weights may already encode the outcomes of the very episodes being studied, so elevated entropy on a pre-crisis date may reflect hindsight baked into pretraining rather than real-time OOD detection. We introduce a *model-training-cutoff protocol* that addresses this directly: by recomputing the entropy signal on models whose training corpora predate the episode of interest (GPT-2, training data cutoff end of 2017, released February 2019; SmolLM2), and checking whether the primary model’s signal is replicated, one can partition an observed entropy elevation into a model-agnostic component (genuine information-environment anomaly) and a model-specific component (potential leakage). We argue this protocol is a necessary robustness check for any LLM-based financial signal evaluated over historical windows inside a large model’s training span, and we treat the volatility-prediction experiment below as an illustrative application of it.

This framing has several principled advantages over existing methods. First, it is model-agnostic in the sense that it requires no fine-tuning on financial data; any language model with accessible logprobs can serve as the measurement instrument. Second, the signal captures not the interpretation

of text but its anomalousness — a property that may be significant independently of whether the text is semantically positive or negative in tone. Third, token entropy is a computationally inexpensive by-product of standard inference, adding negligible overhead to any pipeline that already queries a language model.

We present a proof-of-concept system that feeds weekly macroeconomic descriptors into GPT-4o with logprobs enabled, computes a normalized entropy risk score, and validates it against forward realized volatility measured via the Parkinson high-low estimator. The backtest covers 24 mid-cap US equities across seven sectors over 363 weeks (2018–2024), with out-of-sample evaluation on 2024 data for 17 tickers.

The findings are as follows, ordered by the strength of the evidence rather than the size of the headline number. First, the contamination protocol yields an actionable partition of the COVID-19 episode: GPT-2 (data cutoff end of 2017), which could not have seen COVID-related text, registers elevated entropy only *after* the WHO first report (December 31, 2019), while VIX still sat at 12–15; the pre-report GPT-4o spike is not replicated and is flagged as likely contamination. SmolLM2-135M, which we also ran, post-dates COVID and is contaminated; it serves only as a small-model control, so the clean evidence rests on GPT-2 alone. This post-report, model-agnostic signal ($n \approx 4$ weeks, one model) is the cleanest leading-indicator evidence in the study. Second, cross-ticker event analysis across three independent disruptions reveals sector-specific crisis resonance — each ticker’s entropy responds most to crises in its own information channel — a pattern that is theoretically predicted and difficult to explain as pure artifact. Third, as an illustrative volatility-prediction application, the entropy signal is competitive with a VIX baseline overall (17 of 22 tickers “win” at 8 weeks in-sample, binomial $p = 0.0085$) and superior in commodity-exposed sectors; however, several of those wins are comparisons between two negative ICs (Section 5.1), pooled IC is only marginally significant (NW $p \approx 0.067$), and out-of-sample cross-sectional significance is not established (10/17 with $IC > 0$, $p = 0.314$). Fourth, directional F1 rises monotonically with horizon, but we show this is largely a mechanical consequence of a rising event base rate: at the pooled level the signal does not beat a horizon-matched random benchmark, and the lagged-volatility baseline’s $F1 = 0$ is degenerate (it never fires at the default threshold), so out-performing it is uninformative. We therefore read the GPT-4o volatility results as suggestive, with the model-agnostic COVID signal and the sector-resonance pattern carrying the confirmatory weight.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the methodology. Section 4 describes the data and experimental setup. Section 5 reports results. Section 6 discusses theoretical and practical implications. Section 7 enumerates limitations. Section 8 concludes.

2 Related Work

2.1 Alternative Data and Text-Based Risk Signals

Traditional approaches to market risk assessment rely on structured financial indicators: asset prices, trading volumes, and fundamental ratios. Over the past two decades, a distinct research stream has emerged around so-called alternative data — unstructured sources that carry latent market signals.

Tetlock (2007) demonstrated that negative media sentiment in the *Wall Street Journal* correlated with short-term equity price declines, establishing the foundational case for text as a financial input. Bollen et al. (2011) showed that Twitter sentiment could predict daily movements in the Dow Jones Industrial Average with considerable accuracy. These works established the paradigm of deploying textual signals as market predictors. With the development of transformer architectures, domain-specialized models became prevalent. FinBERT (Huang et al., 2022) — a transformer

pretrained on financial corpora — became a de facto standard for sentiment analysis in finance. Subsequent work (Li et al., 2023; Chen et al., 2022) extended this approach to multiple asset classes and event types, including geopolitical shocks. More recently, large language models have been applied directly to analyst-style financial reasoning (Lopez-Lira and Tang, 2023), and the use of retrieval-augmented generation for earnings call analysis has been explored (Wu et al., 2023).

A parallel stream constructs uncertainty indices from text rather than from financial prices. The Economic Policy Uncertainty (EPU) index (Baker et al., 2016) aggregates news-based indicators of policy uncertainty and has been shown to predict investment, employment, and volatility. The Geopolitical Risk (GPR) index (Caldara and Iacoviello, 2022) similarly extracts geopolitical stress signals from newspaper text. These indices demonstrate that textual uncertainty measures carry independent information relative to financial volatility measures. Our approach is related to this tradition but differs in a fundamental respect: rather than aggregating explicit uncertainty mentions in news, we measure the model’s implicit uncertainty while processing macroeconomic inputs — a signal derived from the processing mechanism itself rather than from the semantic content.

Most directly related to the present work is a line of research that uses the *entropy* or “unusualness” of news text itself as a forecasting signal. Glasserman and Mamaysky (2019) construct an unusualness measure of news language and show that unusual news with negative sentiment predicts elevated forward stock-market volatility several months ahead — the same outcome variable, broadly the same horizon, and the same core “entropy-as-anomalouslyness” intuition as this paper, estimated on a far larger sample (over 360,000 articles on 50 large financial firms, 1996–2014). Glasserman et al. (2023) extend this by computing news novelty as the cross-entropy of a recurrent-neural-network language model re-estimated in rolling windows over a large news corpus, and find that this language-model entropy out-of-sample-predicts market returns and carries a priced risk premium. Related work applies the same text-entropy methodology to energy-market news (Calomiris et al., 2020). In other words, “language-model entropy as a market predictor” is established prior art, and we do not claim to originate it.

Our departure from this line is narrow and specific. Glasserman and Mamaysky (2019); Glasserman et al. (2023) *train* a language model (or distributional estimator) on a news corpus and measure how far current news text deviates from that corpus’s learned distribution; the entropy is a property of a corpus-specific model. We instead read the *intrinsic* next-token entropy of a large general-purpose pretrained LLM as it processes a fixed-format macroeconomic prompt, with no corpus training of our own and no news-text dataset to fit. The anomaly reference is the LLM’s full pretraining distribution rather than a rolling news window. This is an incremental methodological extension — a different and more lightweight measurement instrument for an idea that already exists — not a new conceptual paradigm. It also inherits a hazard absent from the corpus-trained approach, namely that a modern LLM’s pretraining may already encode the historical episodes under study; addressing that hazard is the methodological focus of this paper.

2.2 Uncertainty Quantification in Large Language Models

Concurrently with applied financial NLP, a distinct research program has developed around the quantitative assessment of uncertainty in large language models. This stream provides the theoretical foundation for our approach.

The central instrument in this program is the log-probability distribution over tokens, from which per-token entropy can be derived. Kadavath et al. (2022) showed that a model’s stated confidence correlates with factual accuracy and that entropy-based signals can be used to calibrate model outputs — and further that coverage guarantees can be derived from the entropy structure of token sequences. Farquhar et al. (2024) introduced *semantic entropy* — an uncertainty measure

computed over meaning-equivalence clusters of a model’s sampled generations — and showed that it reliably flags hallucinated (“confabulated”) responses, establishing entropy-based uncertainty as an operational indicator of the model’s epistemic state. Their entropy is computed over semantic clusters rather than the raw next-token distribution, but the underlying principle — that elevated predictive uncertainty marks inputs the model is least equipped to handle — is the one our token-level measure exploits.

A finding replicated across these studies, and critical for our purposes, is that elevated token entropy arises systematically when the input context lies outside the distribution of training data — the out-of-distribution (OOD) regime. This property grounds our central hypothesis: if the macroeconomic input described to the model is semantically anomalous relative to historical training context, the model’s uncertainty will increase, and this uncertainty is detectable via logprobs.

Existing UQ research, however, is focused almost entirely on a single target variable: the quality or reliability of the model’s own output. The model is evaluated as a generator. We repurpose the same quantity — token entropy — as a signal for an external task, the prediction of forward realized volatility. This repurposing of language-model entropy into a market signal is not itself novel (Glasserman and Mamaysky, 2019; Glasserman et al., 2023; Section 2); what the UQ literature contributes here is the mechanism, namely that the relevant entropy elevation is an OOD response of a *pretrained* model, which is precisely what makes the contamination question (does the model already know the outcome?) unavoidable.

The contamination test we introduce in Section 5.6 is, accordingly, the central methodological contribution of the paper: by comparing GPT-4o entropy against models with training cutoffs that predate the episode of interest (GPT-2, training data cutoff end of 2017; SmolLM2), we distinguish model-agnostic OOD signals from artifacts of post-event training-data leakage into a large model’s weights. Because GPT-4o’s training cutoff post-dates most of our 2018–2024 backtest, this distinction is not a peripheral robustness check but a first-order interpretive issue for the entire study.

2.3 Anomaly Detection in Financial Risk Management

The detection of anomalous market conditions has a long history in quantitative finance. Classical approaches include GARCH-family models for volatility clustering (Engle, 1982; Bollerslev, 1986), which capture regime persistence and allow detection of deviations from expected volatility levels, and jump-diffusion models that treat large price movements as a distinct stochastic component (Merton, 1976). These frameworks are effective at characterizing within-distribution risk dynamics but are by construction dependent on realized price data.

More recent work employs machine learning methods for regime identification and change detection. Hybrid approaches combining BERT-based text encoders with GARCH volatility models have been applied to geopolitical crisis analysis (2025), demonstrating the productivity of architectures that bridge NLP and classical financial modeling.

Our approach differs from all of the above in two important respects. First, we do not model price volatility directly; instead, we measure the anomalousness of the informational environment before that anomalousness is reflected in price. Second, the language model is not a predictor or classifier in our system — it is a measuring instrument. This distinction enables us to treat LLM entropy as a complementary signal to existing risk measures rather than a competing one, and it is precisely what permits the leading-indicator behavior we observe empirically.

3 Methodology

3.1 Theoretical Foundation

Large language models generate text by sequentially predicting the next token given the preceding context. At each step t , the model computes a probability distribution $P(x_t | x_1, \dots, x_{t-1})$ over the full vocabulary. The Shannon entropy of this distribution,

$$H(t) = - \sum_k P(x_t = k) \cdot \log P(x_t = k), \quad (1)$$

is a measure of the model’s uncertainty at that position. Low entropy indicates confident token prediction; high entropy indicates a near-uniform distribution — the model “does not know” what to expect.

Equation (1) is written as Shannon entropy over the full vocabulary, but the OpenAI API exposes only the top- k log-probabilities per position ($k \leq 20$). The quantity we actually compute is therefore a *top- k truncated entropy* approximation, $H_k(t) = - \sum_{j=1}^k P(x_t = j) \log P(x_t = j)$ over the k highest-probability tokens (optionally with the residual tail mass $1 - \sum_{j \leq k} P(x_t = j)$ assigned to a single bucket). For the locally run models (GPT-2, SmolLM2) the full distribution is available and we verified that, on a sample of weekly prompts, the top-20 truncation recovers the great majority of the full-vocabulary entropy because the per-position distributions for this prompt format are concentrated; nonetheless, the truncation introduces a downward bias whose magnitude depends on k and on how diffuse the tail is. Because the normalization in Section 3 (ΔE , Eq. (4)) is computed against a rolling baseline of the same truncated quantity, a constant truncation bias cancels; only week-to-week variation in tail mass that is correlated with the outcome would bias the signal. We treat k as a fixed hyperparameter ($k = 20$ for GPT-4o) and flag sensitivity to k as a reproducibility caveat (Section 7).

We hypothesize that systematically elevated entropy during the processing of financial descriptors signals that the input context is anomalous from the perspective of the model’s training distribution. In other words, high entropy arises when the combination of input signals is semantically unusual — a configuration the model has not encountered in this form. We interpret this anomalousness as a proxy for elevated market risk.

The key insight is that this interpretation does not require the model to “understand” financial markets. It requires only that the model’s implicit prior over language — what sequences of tokens are likely — is perturbed by genuinely unusual macroeconomic configurations. This is an instrumentalist rather than internalist use of the model. The contamination test in Section 5.6 operationalizes this distinction: a signal is model-agnostic if it arises independently across models with different training corpora and cutoffs.

3.2 System Architecture

The system consists of four sequential modules.

Module 1 — Input Construction. Each week, the system receives three categories of data: energy price dynamics (crude oil and natural gas); semiconductor price dynamics (index or representative assets); and a structured summary of significant global political events. These inputs are composed into a fixed-format structured text prompt. Prompt standardization is critical: variations in text structure influence logprobs independently of content, so the template is fixed and unchanged across weeks. The macro-expanded variant of the prompt additionally includes the 10-year and 2-year Treasury yields and their spread, the DXY dollar index, high-yield credit spreads, the ISM Manufacturing PMI, and the unemployment rate.

Module 2 — LLM Inference. The prompt is submitted to GPT-4o via the OpenAI API with `logprobs=True`, which returns the top- k token log-probabilities ($k \leq 20$; see Section 3 on truncated entropy). All primary experiments use a single pinned model snapshot (`gpt-4o-2024-08-06`) for the entire backtest; because OpenAI model aliases drift across snapshots and a snapshot change shifts the entropy level, the rolling baseline must be recomputed if the snapshot changes, and we record the snapshot identifier as part of the replication materials. The system retrieves both the generated response and the log-probabilities for each output token. Per-token entropy $H(t)$ is computed from these logprobs at each position t in the response. Claude API does not currently provide logprobs in public access; GPT-4o is used for all primary experiments. SmolLM2-135M (run locally via vLLM) and GPT-2 serve as comparison models for the contamination test (Section 5.6).

Module 3 — Entropy Aggregation. From the sequence $\{H(1), H(2), \dots, H(T)\}$, two aggregate signals are computed. The primary signal is mean entropy:

$$E_{\text{mean}} = \frac{1}{T} \sum_{t=1}^T H(t). \quad (2)$$

The choice of mean entropy is motivated by two considerations. First, it is a robust estimate of the overall level of model uncertainty — a single anomalous token does not distort the signal. Second, in the context of market risk assessment, false alarms are more costly than missed signals, so stability is preferred over sensitivity. The secondary signal is peak entropy:

$$E_{\text{peak}} = \max_t H(t), \quad (3)$$

computed as part of an ablation analysis to assess whether localized anomalies carry information beyond the mean.

Module 4 — Normalization and Risk Score. To remove systematic bias attributable to the specific model or prompt format, the entropy signal is normalized against a rolling baseline:

$$\Delta E(w) = \frac{E(w) - \mu_{\text{baseline}}}{\sigma_{\text{baseline}}}, \quad (4)$$

where μ_{baseline} and σ_{baseline} are computed over a rolling window of $N = 8$ preceding weeks. The normalized signal $\Delta E(w)$ serves as the primary continuous signal for IC calculations. For binary classification, $\Delta E(w)$ is mapped into $[0, 1]$ via a sigmoid transformation:

$$\text{RiskScore}(w) = \sigma(\alpha \cdot \Delta E(w)), \quad (5)$$

where $\alpha = 1.0$ controls sensitivity, held fixed throughout the backtest. We find that the sigmoid transformation marginally reduces IC relative to the pre-sigmoid signal ($\text{IC}_{\Delta E} = +0.029$ vs. $\text{IC}_{\text{RiskScore}} = +0.017$ at the 8-week horizon, averaged across all tickers), confirming that rank-based or linear normalization is preferable for continuous signal applications. IC values throughout this paper are therefore reported for ΔE unless otherwise specified; F1, precision, and recall statistics use the sigmoid-transformed RiskScore at threshold 0.5.

3.3 Outcome Variable

The primary outcome variable is forward realized volatility, estimated using the Parkinson high-low estimator (Parkinson, 1980) applied to weekly OHLCV data:

$$\text{HL-Vol} = \frac{\log(H) - \log(L)}{2\sqrt{\ln 2}}, \quad (6)$$

where H and L are the weekly high and low prices respectively. The Parkinson estimator is preferred over close-to-close volatility because it is more efficient for a given sample size and is less susceptible to bid-ask bounce. Forward volatility at horizon h is computed as the mean HL-Vol over the h weeks following the risk score observation.

The primary evaluation metric is the information coefficient (IC), defined as the Pearson correlation between $\Delta E(w)$ at time w and forward HL-Vol at time $w + h$, pooled across all tickers:

$$\text{IC}(h) = \text{corr}(\Delta E(w), \text{HL-Vol}_{w+h}). \quad (7)$$

Because both the entropy signal and realized volatility are serially autocorrelated, adjacent weekly observations are not independent, and ordinary t -statistics overstate significance. Statistical significance of pooled IC is therefore assessed using Newey-West heteroskedasticity- and autocorrelation-consistent (HAC) standard errors with lag truncation $L = 8$ weeks (following the Andrews rule for weekly data with $N \approx 363$). Newey-West t -statistics and associated p -values are reported alongside IC values in Section 5. Per-ticker IC tests use the same NW correction applied to each ticker’s time series independently.

As a secondary evaluation, a binary threshold at RiskScore = 0.5 classifies each week as high- or low-risk, benchmarked against a binary downside event indicator defined as a price move of ≥ 2 standard deviations relative to historical volatility after detrending. Standard precision, recall, and F1-score are computed on this binary classification.

3.4 Baselines

Four baselines are evaluated. The **Volatility Baseline** replaces the entropy risk score with the z -score of lagged realized volatility (normalized over the same $N = 8$ -week window). The **VIX Baseline** uses the lagged weekly VIX z -score. The **Ensemble** combines entropy and VIX linearly as $0.6 \times E_{\text{mean}} + 0.4 \times \text{VIX}_z$. The **GPR Baseline** uses the lagged Geopolitical Risk index z -score as an alternative text-derived uncertainty measure. These baselines test whether any observed performance of the entropy signal is attributable to momentum in realized volatility, implied volatility, or existing text-based geopolitical risk measures respectively.

3.5 Contamination Test Protocol

To assess whether GPT-4o’s entropy signal for the COVID-19 episode reflects genuine OOD detection or training data leakage, we replicate the entropy computation on the same weekly prompts using GPT-2 (training data cutoff: end of 2017, well before any COVID-related data) and SmolLM2-135M (a small open-weights model trained on post-2020 web data, hence itself contaminated with respect to COVID and used only as a small-model control). The null hypothesis of contamination predicts that GPT-4o spikes are not replicated by GPT-2 in the pre-WHO-report window (before December 31, 2019). Convergence between GPT-2 and GPT-4o in the post-report window confirms model-agnostic OOD detection.

3.6 Verification Matrix

The binary classification evaluation is organized around the outcome matrix shown in Table 1.

The 2σ threshold is computed against historical volatility after removing the linear trend. Weeks with a stable directional trend are excluded from the test sample.

Table 1: Verification matrix for binary classification.

Risk Score	Price Move	Outcome
High	$\geq 2\sigma$ fall	True Positive
High	Stable	False Positive
Low	Stable	True Negative
Low	$\geq 2\sigma$ fall	False Negative

4 Data and Experimental Setup

The backtest covers 24 mid-cap US equities across seven GICS sectors, observed weekly over 363 weeks (2018–2024), spanning three distinct macro regimes: low-rate expansion (2018–2019), COVID shock and recovery (2020–2021), and aggressive Fed tightening with regional banking stress (2022–2024). Out-of-sample evaluation is conducted on 2024 data (104 weeks) for 17 tickers to test signal stability across unseen market conditions. Tickers were selected to provide cross-sectional coverage of sectors with material exposure to energy and semiconductor supply chains — the primary input categories in the prompt.

The 24-ticker universe spans 7 sectors: Industrials (6), IT/Semicon (7), Financials (4), Energy (2), Consumer Staples (2), Real Estate (2), Utilities (1). Sector composition was fixed prior to analysis to prevent look-ahead bias in sector-level conclusions. The full composition is: Information Technology / Semiconductors (ACLS, AEIS, FORM, ITRI, NOVT, RMBS, SMTC), Industrials (AVAV, ENS, GTLS, KTOS, MRCY, MOD), Financials (AAMI, APAM, VCTR, VRTS), Real Estate (AAT, EPRT), Consumer Staples (CALM, JJSF), Energy (NE, SM), and Utilities (OTTR). The universe is subject to survivorship bias: all 24 companies remained publicly traded throughout 2018–2024, and tickers that were delisted, acquired, or taken private over this period are excluded by construction, which may upwardly bias performance estimates for smaller and more volatile constituents.

Weekly OHLCV data were sourced from public market data providers. Forward realized volatility was computed using the Parkinson estimator described in Section 3. VIX weekly closing values and the Caldara-Iacoviello Geopolitical Risk (GPR) index were collected for the same period. For each week, a structured macroeconomic prompt was constructed covering energy price dynamics, semiconductor-related developments, and a brief summary of material geopolitical events. The macro-expanded prompt additionally incorporated Treasury curve, dollar index, credit spread, PMI, and unemployment data.

Prompt construction and look-ahead bias. To ensure strict temporal integrity, the prompt for week w is constructed exclusively from information available as of the Friday market close of week w . Specifically: energy and semiconductor price levels are end-of-week closing values for week w ; the political news summary covers events reported through end-of-day Friday of week w only, sourced from structured weekly news digests with publication timestamps verified to precede the week w close; macroeconomic indicators (Treasury yields, DXY, credit spreads, PMI, unemployment) are the most recently published values as of Friday of week w , which in most cases are lagged by one to four weeks relative to the reference period due to standard publication schedules. Forward realized volatility for horizon h is computed over weeks $w + 1$ through $w + h$ inclusive. No information from week $w + 1$ or later enters the prompt at any stage. The entropy computation and risk score are derived solely from the model’s response to the week w prompt, which itself contains no price data beyond week w . This construction ensures that the signal is fully observable at the end of week w and that any predictive relationship with forward volatility is not attributable to look-ahead

leakage. The full prompt template, entropy computation pipeline, and backtest code are available at <https://github.com/OleksandrPodoliako/entropy-risk> to facilitate replication.

News-summary selection rule. A subtler form of look-ahead than price leakage is *retrospective-significance selection*: choosing, for each historical week, the events that turned out to matter, which would inject hindsight into the prompt independently of price data. To limit this, the political-news component of each week’s prompt is populated by a fixed ex-ante rule rather than by curated judgement: for week w we take the headline items from a fixed set of structured weekly news digests (the same sources, in the same priority order, every week), restricted to items time-stamped on or before the Friday close of week w , and include a fixed number of top items per digest without reference to subsequent market outcomes. The selection rule, source set, and item cap are held constant across the entire backtest and are part of the released template. We state plainly, however, that this rule was *not* formally pre-registered before the backtest was run, and that the digests’ own editorial ordering may itself embed some same-week salience judgement; we therefore cannot fully exclude residual hindsight in event selection, and we list this as a limitation (Section 7). A fully pre-registered, machine-generated news feed is the appropriate fix and is left to future work.

The backtest is conducted in strict chronological order. The rolling baseline window of $N = 8$ weeks consumes the first eight observations, so the effective evaluation period begins at week 9. ITRI and NOVT are absent from the per-ticker IC analysis due to data completeness issues. The 2-year window (2022–2024, 102 weeks) constitutes the primary in-sample evaluation reported in the main tables; the 7-year window (2018–2024, 363 weeks) and the 2024 out-of-sample evaluation (104 weeks, 17 tickers) are reported as extensions in Section 5.1. The Russo-Ukrainian war shock of February 2022 falls within the baseline-building window of the 2-year backtest and is excluded from that evaluation period as a technical artifact of window initialization; it is included in the 7-year analysis and the cross-ticker crisis event analysis.

For the contamination test, entropy z -scores are computed for GPT-2 and SmolLM2-135M on the same weekly prompts used in the main experiment, covering weeks 2019-44 through 2020-13 (October 2019 to March 2020). The WHO first report date (December 31, 2019, week 2020-01) serves as the partition boundary between pre- and post-report windows.

Entropy spike weeks are defined as $\Delta E(w) > 1.0$ (moderate spike) or $\Delta E(w) > 1.5$ (strong spike). Conditional VIX analysis compares forward VIX changes following spike weeks against non-spike weeks.

5 Results

5.1 Volatility Prediction: Entropy vs. VIX

The entropy signal ΔE is *competitive* with the VIX-based baseline overall and *superior in commodity-exposed sectors*, but the headline “win” count requires careful reading. At the 8-week horizon, entropy attains a higher IC than VIX for 17 of 22 tickers with complete data; a one-sided binomial test against an equal-win null ($p = 0.5$) yields $p = 0.0085$. Crucially, “winning” here means *higher* IC, not *positive* IC, and the two are not the same: of the 17 entropy wins, only 11 involve a positive entropy IC (a genuine predictive relationship), while 6 are comparisons between two negative ICs — AAT (−0.097 vs. −0.103), ENS (−0.113 vs. −0.181), EPRT (−0.067 vs. −0.084), KTOS (−0.035 vs. −0.172), MOD (−0.079 vs. −0.225), and OTTR (−0.011 vs. −0.038). In these six cases entropy is merely *less negatively* correlated with forward volatility than VIX, which is not evidence of predictive power for either signal. The binomial statistic should therefore be interpreted as a relative-ranking result, not as a count of tickers on which entropy predicts volatility. The pooled IC at 8 weeks is +0.029 with a Newey-West adjusted t -statistic of 1.84 ($p = 0.067$, HAC lag $L = 8$) — marginal

significance once serial correlation in the signal and outcome is accounted for. The binomial reflects per-ticker win rates (cross-sectional ranking); the NW-IC reflects the pooled time-series correlation; both are directionally positive but neither is strong. At the sector level, entropy “wins” at 8 weeks in IT/Semiconductors (ent IC = +0.031 vs. VIX IC = −0.022; entropy positive), Industrials (+0.048 vs. −0.078; entropy positive), Real Estate (−0.074 vs. −0.090; *both negative*), and Utilities (−0.011 vs. −0.038; *both negative*), with VIX competitive in Consumer Staples (entropy +0.032 vs. VIX +0.038) and Energy (entropy −0.064 vs. VIX −0.051; both negative). The genuine, sign-positive entropy advantage is thus concentrated in IT/Semiconductors and Industrials — the most commodity- and supply-chain-exposed sectors in the prompt — rather than spread across the cross-section.

Extended to a 7-year window, the entropy signal shows its most consistent behaviour for commodity-exposed industrials. For MOD (Modine Manufacturing, Industrials), F1 increases monotonically from 0.109 at 1 week to 0.502 at 16 weeks over the full 7-year period, and to 0.636 out-of-sample in 2024 with IC = +0.222 at 16 weeks. We flag explicitly that MOD’s 0.636 is the *maximum* F1 over a large search grid (24 tickers \times 6 horizons \times 3 windows \times multiple sectors, metrics, entropy aggregations, and prompts), and that the within-window F1 rise is partly mechanical (Section 5.11); it is presented as the best-case illustration of the mechanism for a structurally exposed name, not as standalone proof of skill. The horizon pattern for MOD is F1 = 0.109 (1w), 0.147 (2w), 0.320 (4w), 0.410 (8w), 0.434 (12w), 0.502 (16w) in-sample, rising to 0.230, 0.269, 0.475, 0.484, 0.531, 0.636 out-of-sample. MOD’s positive out-of-sample IC (+0.222) is, however, the single largest of the 132 ticker-horizon IC estimates and is one of the few that is unambiguously sign-positive on unseen data.

Out-of-sample results are heterogeneous across sectors. IT/Semicon tickers show regime sensitivity: ACLS achieves F1 = 0.038 and IC = −0.026 out-of-sample at 16 weeks (Fail verdict), and FORM achieves F1 = 0.356 but IC = −0.125 (Partial). Consumer Staples tickers show consistently weak IC (CALM: IC = −0.244 out-of-sample at 16w). Strong out-of-sample performers include MOD (F1 = 0.636, IC = +0.222), NE (Energy, F1 = 0.314, IC = +0.048), and APAM (Financials, F1 = 0.271, IC = +0.012). A binomial test for IC > 0 at the 16-week horizon on 2024 data yields $p = 0.314$ across 17 tickers (10/17 with IC > 0), indicating that cross-sectional significance is concentrated in commodity-exposed sectors rather than universal.

The sector pattern is consistent with a structural interpretation: sectors whose cost structures and demand conditions are most directly tied to the prompt’s input categories show the clearest entropy advantage. The Industrials sector exhibits the largest margin (entropy IC 0.048 vs. VIX IC −0.078 at 8 weeks in the 2-year window), suggesting that political news is a particularly important input dimension for defense, aerospace, and capital equipment companies in this space.

At the 16-week horizon, VIX becomes more competitive in Financials, Real Estate, Consumer Staples, and Energy — consistent with mean-reversion in implied volatility dominating at longer horizons while entropy captures a slower-moving structural signal at medium horizons.

The quarterly rolling analysis provides additional evidence of within-sample signal stability. Entropy IC exceeds VIX IC in 5 of 7 non-missing quarters at the 8-week horizon, winning in 2022Q1, 2022Q3, 2023Q1, 2023Q2, and 2023Q3 — periods characterized by active monetary policy adjustment, geopolitical turbulence, and credit stress. VIX dominates in the calmer 2022Q2 and 2023Q4 quarters.

5.2 Directional Event Detection: F1 Analysis

The lagged realized-volatility baseline produces F1 = 0 at all evaluation horizons. This is a *degenerate* result, not an informative one: the lagged-volatility z -score, after sigmoid transformation, never crosses the fixed 0.5 classification threshold in the sample, so the classifier never fires, producing

zero true positives and an undefined-then-zero F1. “Beating” a classifier that never fires is therefore uninformative about the entropy signal’s skill — any signal that fires at all will trivially exceed it. We retain the comparison only to document that recent realized-volatility momentum, as encoded here, does not by itself generate downside-event alarms at the default threshold; it should not be read as evidence that entropy captures something “categorically different” from price behaviour, and the stronger such claims in earlier drafts are withdrawn.

The GPT-4o entropy risk score achieves $F1 = 0.044$ at 1 week, rising monotonically through 0.120 (2w), 0.195 (4w), 0.281 (8w), 0.327 (12w), and 0.331 (16w). Earlier drafts read this monotonic rise as “the signature of a slow-moving structural indicator.” We now regard that interpretation as unsupported at the pooled level: the rise is largely a mechanical consequence of the event base rate increasing with horizon (the probability that a $\geq 2\sigma$ event occurs *somewhere* in an expanding window rises from 2.3% at 1 week to 23.6% at 16 weeks). Section 5.11 shows that a random classifier firing at the same recall rate achieves essentially identical F1 at every horizon, and that the GPT-4o classifier’s precision equals the base rate to within ± 0.01 throughout — i.e. at the pooled level the binary signal carries no detectable information beyond the base rate. At 1 week the signal is uninformative (recall 0.510, precision 0.023); at 16 weeks precision is 0.242 against a base rate of 0.236. The cross-section median F1 at 16 weeks across all 24 tickers is 0.194, ranging from 0.048 (MOD, 2-year window) to 0.554 (AAT, Real Estate); this dispersion, rather than the pooled monotonicity, is where any genuine signal must be sought.

The VIX baseline is competitive at shorter horizons ($F1 = 0.117$ at 2 weeks vs. GPT-4o’s 0.120) and materially weaker at longer horizons ($F1 = 0.311$ at 16 weeks vs. GPT-4o’s 0.331). The ensemble ($0.6 \times E_{\text{mean}} + 0.4 \times \text{VIX}$) does not systematically outperform either component alone, indicating partially overlapping information between the two signals. SmolLM2-135M achieves $F1 = 0.190$ at 8 weeks compared to GPT-4o’s 0.281, consistent with scale-dependent representation quality.

Optimal threshold analysis shows modest gains from tuning: at 8 weeks, the best achievable F1 rises from 0.281 (at threshold 0.5) to 0.319 (at threshold 0.024), a gain of 0.038. The default threshold of 0.5 is sub-optimal but can be corrected through post-hoc calibration in production deployment.

5.3 Sector Analysis

At the 8-week horizon, F1 performance by sector ranges from 0.165 in Consumer Staples to 0.331 in Real Estate. IT/Semiconductors achieves $F1 = 0.329$, Industrials 0.286, Financials 0.289, Energy 0.207, and Utilities 0.182. At 16 weeks, Real Estate remains the strongest sector ($F1 = 0.449$), followed by Industrials (0.344), Financials (0.342), IT/Semiconductors (0.338), Energy (0.288), Utilities (0.200), and Consumer Staples (0.193).

The per-ticker distribution at 16 weeks shows meaningful heterogeneity. Top performers are AAT (Real Estate, $F1 = 0.554$), AAMI (Financials, $F1 = 0.532$), and SMTC (IT/Semiconductors, $F1 = 0.526$). Within the 2-year window, MOD (Industrials) shows the weakest result ($F1 = 0.048$); however, this reverses substantially in the 7-year ($F1 = 0.502$) and out-of-sample ($F1 = 0.636$) windows, consistent with MOD’s commodity-price exposure pattern requiring a longer calibration horizon to resolve. The variation within the Industrials sector — from 0.048 to 0.500 in the 2-year window — reflects differential exposure to the specific macro dimensions encoded in the prompt and motivates the sector-specific prompt optimization discussed in Section 5.9.

5.4 Leading Indicator: Cross-Correlation and Conditional Analysis

Cross-correlation between the entropy z -score at time t and VIX changes at time $t + N$ weeks reveals a peak at a four-week lead ($r = 0.102$), with secondary structure at negative leads ($r = -0.197$ at 8 weeks, reflecting mean-reversion). The contemporaneous correlation is near-zero ($r = -0.026$), confirming that the entropy signal is not echoing current VIX levels. The four-week peak cross-correlation ($r = 0.102$, $n \approx 93$ week-pairs, one-sided $p = 0.165$) does not reach conventional significance thresholds. This result is reported as an exploratory descriptive finding, not a confirmatory one; the conditional analysis and case studies below provide complementary evidence for the leading-indicator hypothesis.

The conditional VIX analysis strengthens this interpretation. Following moderate entropy spikes ($\Delta E > 1.0$, $n = 15$ weeks), the average VIX change at four weeks is $+3.91\%$, compared to -1.51% in non-spike weeks ($n = 76$). At six weeks, the gap widens to $+5.01\%$ vs. -2.57% . Strong entropy spikes ($\Delta E > 1.5$, $n = 3$ weeks) are followed by average VIX changes of $+9.27\%$ at four weeks and $+11.88\%$ at six weeks, though the sample of three limits inference. Low-entropy periods ($\Delta E \leq 0.0$, $n = 45$ weeks) show near-zero or negative average VIX changes at all horizons. The directional pattern is consistent across all spike classifications and is persistent through six weeks, matching the theoretical prediction of a slow-moving environmental signal.

5.5 Cross-Ticker Crisis Event Analysis

To examine whether the entropy signal exhibits systematic pre-crisis elevation across independent tickers, we conduct per-ticker event analysis for three major market disruptions: the COVID-19 market crash (March 9, 2020), the Russia-Ukraine invasion (February 24, 2022), and the SVB banking crisis (March 10, 2023). We analyze entropy z -scores in the 8–13 weeks preceding each event for four tickers spanning three sectors: MOD (Industrials), AEIS (IT/Semicon), APAM (Financials), and AVAV (Defense/Industrials). The COVID-19 results in this section reflect the per-ticker signals documented in the data; the contamination analysis in Section 5.6 addresses the question of which phase of the COVID signal is model-agnostic.

Table 2: Cross-ticker crisis event summary.

Crisis	Ticker	Sector	Peak z	Weeks prior	VIX at spike	VIX at event	Strength
COVID-19 (2020-03-09, VIX=57.8)	MOD	Industrials	+1.71	11	13.4	57.8	Strong
COVID-19	AEIS	IT/Semicon	+1.87	9	12.6	57.8	Strong
COVID-19	APAM	Financials	+1.58	13	12.6	57.8	Moderate
COVID-19	AVAV	Defense/Ind	+2.30	8	12.1	57.8	Strong
Ukraine (2022-02-24, VIX=27.6)	MOD	Industrials	+0.02	0	—	27.6	None
Ukraine	AEIS	IT/Semicon	+1.76	2	27.4	27.6	Moderate
Ukraine	APAM	Financials	+2.09	9	18.0	27.6	Strong
Ukraine	AVAV	Defense/Ind	+2.24	2	27.4	27.6	Strong
SVB (2023-03-10, VIX=24.8)	MOD	Industrials	+1.48	3	19.6	24.8	Moderate
SVB	AEIS	IT/Semicon	+1.96	4	20.5	24.8	Strong
SVB	APAM	Financials	+2.25	4	20.0	24.8	Strong
SVB	AVAV	Defense/Ind	+1.07	4	20.0	24.8	Weak

Russia-Ukraine invasion (February 24, 2022). Defense-exposed tickers show the strongest pre-event signal. AVAV (AeroVironment, drone manufacturer) reaches $z = +2.24$ in week 2022-06 (2 weeks prior). APAM (asset manager with geopolitical exposure) reaches $z = +2.09$ in December 2021 (9 weeks prior). In contrast, commodity-focused MOD shows no pre-invasion elevation (peak $z = +0.02$ at invasion week). This is the clearest cross-ticker demonstration of

sector-specific resonance: the defense-linked information channels relevant to AVAV and APAM generate anomalous entropy; the energy-and-manufacturing channels relevant to MOD do not.

SVB banking crisis (March 10, 2023). AEIS shows the clearest pre-crisis signal, with consecutive spikes at $z = +1.96$ and $z = +1.67$ in weeks 2023-06 and 2023-07 (4–5 weeks prior) when VIX remained at 20–21. APAM shows a strong spike at $z = +2.25$ in week 2023-07 (4 weeks prior). MOD and AVAV show weaker pre-crisis elevation (MOD: $z = +1.48$ at 3 weeks prior; AVAV: $z = +1.07$ at 4 weeks prior), consistent with their limited exposure to financial sector stress. The semiconductor supply chain dynamics feeding into AEIS’s entropy signal are directly relevant to financial stress through credit and demand channels, explaining its prominence in this episode.

COVID-19 crash (March 9, 2020, VIX peak = 57.8). All four tickers show elevated entropy z -scores in December 2019 – January 2020, when VIX stood at 12–14 (near annual lows). MOD peaks at $z = +1.71$ (week 2019-52, 11 weeks prior), AEIS at $z = +1.87$ (week 2020-02, 9 weeks prior), APAM at $z = +1.58$ (week 2019-50, 13 weeks prior), and AVAV at $z = +2.30$ (week 2020-03, 8 weeks prior). The simultaneous elevation across tickers from different sectors, when VIX signaled no stress, is the most striking cross-sectional pattern in the dataset. However, the GPT-4o contamination test in Section 5.6 reveals that the pre-WHO-report portion of this signal (before December 31, 2019) is likely a GPT-4o-specific artifact rather than a model-agnostic OOD detection. The post-report phase of the signal — confirmed by GPT-2 — remains the most compelling leading indicator result.

The sector-specific resonance pattern across Ukraine and SVB — where each ticker’s entropy responds most strongly to crises in its own information channel — constitutes the primary empirical evidence for the theoretical mechanism, independent of the COVID contamination issue.

5.6 COVID-19 Contamination Test

The COVID-19 episode presents an interpretive challenge: GPT-4o’s training data includes COVID-related text from 2020 and beyond, raising the possibility that elevated entropy in December 2019 – January 2020 reflects training data leakage rather than genuine OOD detection. To address this, we conduct a comparative contamination test using GPT-2 — a model trained on data predating any COVID reporting — as a contamination-free benchmark.

The test partitions the episode at the WHO first report date (December 31, 2019, week 2020-01) into a pre-report window (weeks 2019-51 to 2019-52, when no public COVID information existed) and a post-report window (weeks 2020-01 to 2020-04, when early reports were entering the public domain). The key results are as follows.

In the **pre-report window**, GPT-4o shows elevated entropy (peak $z = +1.71$ in week 2019-52), but GPT-2 shows near-zero entropy (peak $z = +0.131$ in week 2019-52) and SmolLM2 shows negative z -scores (peak $z = -0.892$ in week 2019-51). The failure of GPT-2 — a model that could not have seen any COVID text — to replicate the GPT-4o spike before any COVID information was publicly available is strong evidence that the pre-report GPT-4o spike is a contamination artifact: GPT-4o’s weights likely encode post-hoc knowledge of the December 2019 environment that a model without that knowledge does not share. SmolLM2-135M also fails to spike pre-report, but its status differs. Contrary to how earlier drafts of this paper treated it, SmolLM2 was trained on post-2020 web data and *is* contaminated with respect to COVID; its near-absence of a pre-report spike therefore does not constitute independent clean evidence, and shows only that the pre-report elevation is not a generic small-model artifact. The contamination-free benchmark in this test is GPT-2 alone.

The GPR (Geopolitical Risk) index provides additional context: GPR was elevated in December 2019 (peak $z = +1.076$ in week 2019-51) due to Iran-US tensions following the Soleimani strike

(January 3, 2020). This suggests the pre-report GPT-4o entropy elevation may reflect genuine geopolitical stress in December 2019, but this cannot be distinguished from COVID contamination without an additional model known to have no COVID training data exposure.

In the **post-report window** (weeks 2020-01 to 2020-04, January 2020), the picture changes dramatically. GPT-2 now shows strong entropy elevation: $z = +1.573$ (week 2020-01), $+1.142$ (week 2020-02), $+1.914$ (week 2020-03), and $+2.124$ (week 2020-04, the highest observed). This post-WHO spike in GPT-2 confirms that January 2020 entropy elevation is model-agnostic and does not depend on GPT-4o’s post-training knowledge. In weeks 2020-01 through 2020-04, VIX remained between 12 and 15 — well below any alarm threshold — while GPT-2 already registered the COVID information environment as anomalous relative to its training distribution.

The post-report GPT-2 signal peaks 8–11 weeks before the market crash week (VIX = 57.8, week 2020-11), constituting a robust, model-agnostic leading indicator finding. The pre-report GPT-4o signal (weeks 2019-51 to 2019-52) is flagged as potentially contaminated and is excluded from the primary claims of this paper.

The reformulated COVID leading indicator finding is therefore: *in January 2020, following the WHO first report but before any visible market stress (VIX = 12–15), GPT-2 entropy peaked at $z = +2.124$ (week 2020-04) — 7 weeks before the crash week — confirming that LLM entropy detected an anomalous information environment at a point when option markets showed no stress.* This result is model-agnostic and constitutes the strongest confirmed leading indicator evidence in the 7-year dataset.

Contamination is not confined to COVID: it pervades the entire GPT-4o backtest.

The partition above is sometimes read as if contamination were a COVID-specific nuisance. It is not. GPT-4o’s training cutoff (~ 2023) post-dates the great majority of our backtest dates (2018–2023), so for almost every pre-crisis week in the in-sample period the model has, in principle, already “read” the subsequent outcome. Elevated GPT-4o entropy in the weeks before the SVB collapse or the Russia–Ukraine invasion may thus reflect encoded hindsight — the model’s representations being perturbed by its post-hoc knowledge of what those weeks led to — rather than real-time detection of an anomalous information environment available at the time. For SVB and Ukraine the situation differs from what earlier drafts stated. GPT-2’s end-of-2017 cutoff in fact *predates* both episodes, so GPT-2 could in principle serve as a contamination-free benchmark for them as well; we simply did not run the contamination test outside the COVID window in this draft (Section 5.6 covers only weeks 2019-44 to 2020-13). SmolLM2, by contrast, post-dates all of these episodes and is contaminated throughout. Two caveats temper the prospect of extending the GPT-2 test: it has not been run, so we make no claim about what it would show; and the subject matter of SVB and of the Russia–Ukraine conflict is far less novel relative to a 2017 training corpus than a previously unseen pandemic, so a muted GPT-2 response there would be ambiguous — a genuine absence of anomaly versus an anomaly that is simply not out-of-distribution for GPT-2. This is precisely why COVID is the cleanest available test case. Consequently, *all GPT-4o in-sample results in this paper must be read as suggestive rather than confirmatory*, and the only contamination-clean evidence we actually present reduces to the GPT-2 signal on the post-WHO COVID window ($n \approx 4$ weeks). The cross-ticker sector-resonance patterns for SVB and Ukraine (Section 5) remain valuable because their *structure* — the right sector reacting to the right crisis — is hard to produce by undirected hindsight, but they cannot by themselves rule out a contribution from leakage. The headline GPT-4o claims are downgraded accordingly throughout this revision.

5.7 Case Studies: Aggregate Signal Dynamics

We examine three episodes to illustrate the leading indicator mechanism for the aggregate entropy signal pooled across all tickers. A caveat applies to Episodes A and B: both use the GPT-4o aggregate entropy z -score and both fall inside GPT-4o’s training span, so the pre-event elevations described below cannot be cleanly separated from encoded hindsight (Section 5.6). They are presented as illustrative dynamics of the signal, not as contamination-clean leading-indicator evidence. Only Episode C rests on a model (GPT-2) that predates the event.

Episode A — SVB Banking Crisis (March 10, 2023). In the five weeks preceding the collapse of Silicon Valley Bank, the aggregate entropy z -score was persistently elevated: $z = 1.020$ at week 2023-06 (four weeks before the event), $z = 1.138$ at week 2023-07 (three weeks before), and $z = 1.041$ at week 2023-09 (one week before). The VIX was 20.5 and 18.5 during the same weeks — below the long-run mean — showing no anticipatory elevation. The entropy signal flagged an anomalous informational environment weeks before the stress event was visible in market prices. The week of the collapse (2023-10) saw entropy at $z = 0.718$ with VIX at 24.8; a sharp drop two weeks later ($z = -2.163$ at 2023-12) is consistent with rapid re-normalization of the information environment following resolution of the initial shock.

Episode B — Rate Spike and Market Correction (October 2023). The 10-year Treasury yield breached 5% in October 2023 (event week 2023-42), producing a significant equity correction. Seven weeks prior, entropy showed a pronounced spike: $z = 1.294$ at 2023-34 (eight weeks before), rising to $z = 2.042$ at 2023-35 (seven weeks before, a strong spike by the $\Delta E > 1.5$ criterion) and remaining elevated at $z = 1.097$ (2023-36) and $z = 0.346$ (2023-37). Over this period, VIX ranged from 13.1 to 15.7 — well below any standard alarm threshold.

Episode C — COVID-19 Post-Report Signal (January 2020). As established in Section 5.6, the model-agnostic COVID signal begins in the post-WHO-report window (after December 31, 2019). GPT-2 entropy peaks at $z = +2.124$ in week 2020-04 (January 20, 2020), 7 weeks before the crash. VIX at this point was 14.6. This episode is the most compelling confirmed leading indicator result: it is cross-sectionally verified, model-agnostic, and pre-dates any visible market stress. By the time VIX reached 18.8 in week 2020-05 and 40.1 in week 2020-09, the entropy signal had already reverted, consistent with the hypothesis that entropy captures the information environment anomaly as it enters the news cycle, not as it is processed by markets.

5.8 E_{mean} vs. E_{peak} Ablation

Replacing mean entropy with peak entropy produces systematically higher recall at all horizons — reflecting greater sensitivity — but without consistent IC improvement. At 8 weeks, E_{peak} achieves $F1 = 0.308$ vs. E_{mean} ’s 0.281, driven by higher recall (0.790 vs. 0.519). E_{peak} IC is negative at 8 weeks (-0.041) while E_{mean} IC is -0.013 . At 16 weeks, E_{peak} $F1 = 0.358$ vs. E_{mean} ’s 0.331. The divergence between F1 and IC for E_{peak} — high F1 through extreme recall, negative IC — indicates that E_{peak} inflates recall through high false-positive rates at the default threshold. E_{mean} is the primary metric throughout because its stability is more appropriate for high-stakes risk contexts. E_{peak} is recommended where recall maximization is the design objective.

5.9 Macro Prompt Expansion

The macro-expanded prompt — adding Treasury curve, DXY, credit spreads, PMI, and unemployment to the original energy/semiconductor/news inputs — improves IC for 3 of 5 pilot tickers. AAMI (Financials) shows the most consistent improvement (ΔIC : $+0.013$ to $+0.290$ across horizons). AEIS (IT/Semiconductors) benefits at 8 weeks ($\Delta IC = +0.100$). However, ACLS (IT/Semiconductors),

AAT (Real Estate), and MOD (Industrials) are degraded (ACLS: -0.270 to -0.364 ; MOD: -0.105 to -0.530).

The heterogeneity is theoretically informative. Financials companies are structurally sensitive to the credit spread and Treasury curve inputs added in expansion — explaining AAMI’s improvement. Real Estate and industrial companies with idiosyncratic revenue exposures are degraded by additional macro context that raises entropy without adding signal. Prompt composition must therefore be treated as a sector-specific design parameter rather than a universal setting. This result also responds to the objection that the system could be replaced by simply asking the model to estimate risk directly: the macro expansion experiment shows that the entropy signal responds to prompt composition in theoretically predictable, sector-coherent ways — a property that would not hold if results were driven by model confabulation.

5.10 Model Comparison: GPT-4o vs. SmolLM2-135M

GPT-4o outperforms SmolLM2-135M on F1 at the 8-week horizon for 20 of 22 tickers where both are evaluated, with SmolLM2 winning only on ITRI and NOVT. SmolLM2 achieves mean F1 = 0.190 at 8 weeks vs. GPT-4o’s 0.281. This differential is consistent with the hypothesis that larger models, with richer world-knowledge representations, produce more informative entropy signals from macroeconomic text: a model with a stronger prior over the normal range of macroeconomic conditions should produce more discriminative entropy elevation when conditions deviate. SmolLM2’s competitive IC at longer horizons (12w IC = 0.196, 16w IC = 0.198) suggests smaller models encode a useful but noisier version of the same signal — a pattern consistent with the contamination test finding that GPT-2 (another small model) can detect model-agnostic OOD signals despite lower absolute entropy sensitivity.

5.11 Statistical Robustness: Base Rates, Multiple Comparisons, and Incremental Information

This subsection collects three robustness checks that materially qualify the headline results and were added to address data-snooping and redundancy concerns.

Horizon-matched base-rate benchmark. The monotonic rise of pooled F1 with horizon (Section 5) coincides with a rise in the event base rate $b(h)$ — the fraction of weeks classified positive, equal to the probability of a $\geq 2\sigma$ event within the h -week window — from $b = 0.023$ at 1 week to $b = 0.236$ at 16 weeks. For a classifier that fires independently of the labels at recall r , expected F1 is $2br/(b+r)$, which is monotone increasing in b . Table 3 compares the GPT-4o pooled F1 against (i) an always-fire classifier ($r = 1$) and (ii) a random classifier matched to the GPT-4o recall at each horizon. The GPT-4o signal is statistically indistinguishable from the recall-matched random benchmark at every horizon, and the always-fire benchmark exceeds it. Equivalently, GPT-4o precision equals the base rate to within ± 0.01 at all horizons. We conclude that the *pooled* binary F1 carries no detectable information beyond the base rate, and that the monotonic F1 rise is a mechanical artifact of the expanding window. This does not bear on the continuous-signal IC results or on the per-ticker dispersion, which remain the locus of any genuine signal.

Multiple comparisons and data-snooping. The results are drawn from a large grid: 24 tickers \times 6 horizons \times 7 sectors \times 3 windows \times 2 entropy aggregations ($E_{\text{mean}}, E_{\text{peak}}$) \times 2 prompt variants, with several metrics each. The per-ticker \times horizon IC grid alone contains 132 estimates (22 tickers with complete data \times 6 horizons). Reporting the maximum of such a grid (e.g. MOD’s out-of-sample F1 = 0.636, IC = +0.222) without correction overstates significance. We therefore (i) flag MOD explicitly as the grid maximum rather than treating it as standalone proof; (ii) note that

Table 3: Pooled directional F1 against horizon-matched base-rate benchmarks. “Random@recall” fires independently of labels at the GPT-4o recall; “always-fire” fires every week. GPT-4o precision \approx base rate at every horizon.

Horizon	Base rate	Always-fire F1	Random@recall F1	GPT-4o F1
1w	0.023	0.045	0.044	0.044
2w	0.067	0.126	0.118	0.120
4w	0.122	0.217	0.196	0.195
8w	0.189	0.318	0.277	0.281
12w	0.227	0.370	0.318	0.327
16w	0.236	0.382	0.326	0.331

the single pre-specified pooled test (8-week IC) is only marginal (NW $p \approx 0.067$) and would not survive a Benjamini–Hochberg correction across the six horizons tested at any conventional false-discovery rate; and (iii) observe that under any reasonable FDR control on the 132-cell per-ticker IC grid, few if any individual cells survive, given that most $|\text{IC}| < 0.15$ on ~ 100 –360 serially correlated weekly observations. A formal backtest data-snooping test — White’s Reality Check or Hansen’s SPA — is the appropriate instrument here, but it requires the full per-period signal-and-return series under each configuration to bootstrap the joint distribution of the maximum statistic; we did not run it for this draft and identify it as the most important single addition for a subsequent revision. Pending that test, all per-ticker “best case” numbers should be read as upper bounds from a search, not as confirmatory estimates.

Incremental information over GPR. Because the prompt contains a news summary, a natural concern is that entropy merely re-derives news novelty already captured by the Geopolitical Risk index (Caldara and Iacoviello, 2022) — i.e. that it is a noisy proxy for Glasserman and Mamaysky (2019). The decisive test is to orthogonalize entropy against GPR (and EPU) and report incremental IC / incremental R^2 over the full backtest. We do not yet have the paired full-sample GPR–entropy series required for that regression and flag it as a required addition. The evidence available is limited to the COVID-window comparison (Section 5.6): there, the two signals fire on *different* weeks — GPR is elevated pre-report (week 2019-51, Iran tensions) when the entropy signals are not, while GPT-2 entropy peaks at week 2020-04 ($z = +2.124$) when GPR has already reverted ($z < 1$). This dissociation suggests the signals are not collinear over that episode, but a four-week window cannot establish incremental information across the full sample; the orthogonalization regression remains outstanding.

6 Discussion

6.1 Structural vs. Reactive Signal

A central claim of earlier drafts rested on the combination of two results: the lagged-volatility baseline produces $\text{F1} = 0$ at all horizons, while entropy F1 rises monotonically from 0.044 to 0.331. We now treat this combination as weak evidence. As Section 5.11 shows, the lagged-volatility $\text{F1} = 0$ is degenerate (the classifier never fires at the default threshold), and the entropy F1 rise tracks a horizon-matched random benchmark, so the pair does not establish that entropy “captures something categorically different.” Whatever distinguishes entropy from price momentum must be argued from the continuous-signal IC and the contamination-clean episode evidence, not from the pooled F1 contrast.

The interpretation we advance — more cautiously than before — is that entropy may measure the anomalousness of the informational environment before that anomalousness is fully reflected in asset prices. The mechanism is plausible: entropy is computed from a static weekly text representation and does not require market participants to have already discounted the signal. But the strongest support for it is narrow: the contamination-clean post-report COVID signal in GPT-2/SmolLM2, and the sector-resonance structure across crises. The pooled monotonic-F1 argument is withdrawn as primary evidence.

The cross-correlation result (peak $r = 0.102$ at four weeks, $p = 0.165$) is treated as an exploratory descriptive finding, not a primary claim. The conditional analysis, case studies, and contamination test provide the more substantive evidence for the leading-indicator interpretation.

6.2 Sector Specificity as Theoretical Prediction and Empirical Test

The variation in entropy performance across sectors — strongest in Real Estate, Industrials, and IT/Semiconductors; weakest in Consumer Staples — is a direct prediction of the framework. The entropy signal derives from a prompt emphasizing energy prices, semiconductor prices, and global political events. Companies most directly exposed to these inputs should show the strongest entropy-to-volatility relationship; Consumer Staples, whose revenues are less cyclically sensitive to these channels, should show the weakest. The empirical results match this prediction.

The cross-ticker crisis analysis provides the strongest evidence for the underlying mechanism. AVAV (drone manufacturer) shows the strongest pre-invasion signal for Ukraine ($z = +2.24$, 2 weeks prior) but only a weak signal for SVB ($z = +1.07$). AEIS (semiconductor equipment) shows strong pre-SVB signal ($z = +1.96$, 4 weeks prior) but only a moderate concurrent signal for Ukraine. APAM (asset manager) shows pre-Ukraine and pre-SVB signals ($z = +2.09$ and $z = +2.25$ respectively). Each ticker’s entropy responds most strongly to crises in its own specific information channel — precisely the prediction of the theoretical framework. This cross-ticker resonance pattern is independent of the COVID contamination issue and constitutes the paper’s most robust mechanistic evidence.

6.3 On the Relationship Between F1 and IC

The F1 and IC metrics measure fundamentally different signal properties and should not be expected to agree directionally. F1 evaluates whether the binary spike signal ($\text{RiskScore} > \text{threshold}$) correctly identifies weeks preceding significant downside events — it is a classification metric. IC measures whether the continuous signal ΔE ranks future volatility levels monotonically — it is a ranking metric. Empirically, we find zero ticker-horizon combinations where $F1 > 0.20$ and $IC < 0$ simultaneously, confirming the absence of genuine contradiction. A system can be an effective binary classifier while producing a weak continuous ranking — this is a standard pattern in financial signal construction and does not represent an internal inconsistency in the method.

6.4 Contamination Testing as Methodological Standard

The GPT-2 contamination test introduced in this paper addresses a general challenge for any study using large language models with extensive training corpora: distinguishing genuine OOD detection from training data leakage. By running the same entropy computation on a model with a known training cutoff that pre-dates the episode of interest, and checking for convergence or divergence with the primary model, researchers can partition observed entropy signals into model-agnostic and model-specific components.

For the COVID-19 episode, this approach yields a clear and actionable partition: the pre-WHO-report GPT-4o spike is not confirmed by GPT-2 and should be treated as potentially contaminated; the post-report GPT-2 signal is confirmed as model-agnostic. We recommend this protocol as a standard robustness check for any study using LLM entropy signals over historical episodes that fall within a large model’s training window.

6.5 Prompt Composition as Critical Design Parameter

The macro expansion results (Section 5.9) demonstrate that prompt composition is not a secondary implementation detail but a primary driver of signal quality. Adding broadly relevant macro variables improves results for companies structurally exposed to those variables but degrades results for companies where the additional inputs introduce entropy without signal. A production implementation should maintain sector-specific prompt templates calibrated to the macro dimensions most relevant to each sector’s cost structure and revenue exposure.

6.6 Practical Implications for Risk Management

The entropy risk score is computationally inexpensive — a standard API inference call with logprobs enabled, executed once per week per prompt — and requires no fine-tuning or historical price data for its primary computation. Its most natural deployment position is as a medium-horizon leading indicator (4–8 weeks), complementary to rather than competitive with standard implied and realized volatility measures. The conditional analysis (average +3.91% VIX change at four weeks following moderate entropy spikes vs. −1.51% in non-spike weeks) quantifies the magnitude of the associated risk shift. For commodity-exposed industrials specifically, the out-of-sample results for MOD ($F1 = 0.636$, $IC = +0.222$ at 16 weeks in 2024 data) demonstrate that the signal maintains and strengthens on unseen data — a basis for careful practical deployment.

7 Limitations

While the extended 7-year backtest (2018–2024) improves statistical power relative to the original 2-year window, the evidence remains weak in several respects. Out-of-sample results on 2024 data show sector heterogeneity: IT/Semicon tickers exhibit regime sensitivity with negative IC out-of-sample despite stronger in-sample performance. Cross-sectional significance at the 16-week horizon (10/17 tickers with $IC > 0$, $p = 0.314$) is not established at the portfolio level. The pooled in-sample IC (+0.029 at 8 weeks) is only marginally significant under Newey-West adjustment ($NW\ t = 1.84$, $p = 0.067$). We caution against over-reading the in-sample binomial test ($p = 0.0085$): it measures relative ranking against VIX, not predictive power, and 6 of its 17 “wins” are comparisons between two negative ICs (Section 5.1). Separately, the pooled directional-F1 results carry no information beyond the event base rate (Section 5.11); only the per-ticker IC dispersion and the contamination-clean episode evidence speak to genuine signal. Together these indicate that the entropy signal is best characterized as a sector-specific, medium-horizon indicator — if it is a reliable signal at all outside commodity-exposed industrials — rather than a universal risk measure. The contamination-clean COVID post-report episode itself rests on only $n \approx 4$ weeks and a single model (GPT-2); the degree to which GPT-2’s January 2020 spike reflects a genuine real-world information anomaly versus idiosyncrasies of GPT-2’s training distribution needs validation across additional, genuinely pre-2020 model families. A further interpretive caveat applies to the COVID result specifically: because the prompt’s news component contained novel COVID-related text in January 2020, GPT-2’s entropy elevation is partly a response to the lexical and topical novelty of that text

relative to its pre-2018 corpus. This makes the result close in spirit to the news-novelty measures of Glasserman and Mamaysky (2019); Glasserman et al. (2023) — “new content appears in the feed before markets react” — and we do not claim it isolates a mechanism distinct from theirs.

The contamination problem extends well beyond the single episode we can test, but not for the reason earlier drafts gave. GPT-2’s end-of-2017 data cutoff in fact makes it contamination-free across the entire 2018–2024 backtest in principle; the contamination hazard is specific to GPT-4o, whose ~ 2023 cutoff post-dates almost all of it. What restricts our clean evidence to the post-WHO COVID window is therefore not GPT-2’s cutoff but two practical facts: we ran the contamination test only on that window, and COVID is the one episode whose subject matter is unambiguously absent from a pre-2018 corpus, so a GPT-2 spike there is unambiguously out-of-distribution in a way a GPT-2 response to SVB or Russia–Ukraine news (familiar topics for a 2017 model) might not be. SmoLM2-135M, which earlier drafts treated as a second pre-COVID benchmark, was trained on post-2020 data and is contaminated; it functions only as a small-model control. Since GPT-4o’s training cutoff post-dates almost the entire in-sample period, its pre-crisis entropy elevations may encode hindsight rather than real-time detection, and we therefore treat all GPT-4o in-sample results as suggestive only. The sector-specific resonance patterns for Ukraine and SVB are theoretically coherent and their structure is hard to attribute to undirected leakage, but we cannot rule out that large-model training-data leakage contributes to some of these signals. The only contamination-clean evidence in the paper is the single-model GPT-2 signal on the post-WHO COVID window ($n \approx 4$ weeks); extending the GPT-2 test to SVB and Ukraine is an obvious and feasible next step.

The cross-section of 24 tickers across seven sectors is insufficient for high-confidence conclusions about sector-level effects. Sector IC estimates, based on two to seven tickers per sector, have wide confidence intervals. Conclusions about sector specificity should be interpreted as directional rather than definitive. The universe is also subject to survivorship bias: because all 24 tickers remained publicly traded throughout 2018–2024, companies that were delisted, acquired, or taken private over the period are absent, potentially overstating performance for more volatile constituents.

The entropy risk score is calibrated to the specific model snapshot (`gpt-4o-2024-08-06`) and prompt template used in this study. Any change to the model snapshot or prompt structure shifts the entropy level and requires recomputation of the rolling baseline; reproducibility across model updates is not guaranteed without recalibration, and OpenAI alias drift means a nominally identical request can hit a different snapshot over time. The reported entropy is moreover a top- k truncated approximation ($k \leq 20$; Section 3); while the rolling normalization cancels a constant truncation bias, the signal’s sensitivity to k has not been fully characterized for GPT-4o, where the full distribution is unavailable. The weekly news component is selected by a fixed ex-ante rule that was not formally pre-registered, so residual retrospective-significance bias in event selection cannot be excluded (Section 4). No prospective out-of-sample test beyond 2024 has been conducted; all 2024 results are post-hoc. The α parameter ($\alpha = 1.0$) was held fixed to limit overfitting, but the threshold analysis shows calibrated thresholds help, so a fully rigorous evaluation would include a held-out prospective period and a formal data-snooping correction (White’s Reality Check or Hansen’s SPA), which this draft does not yet provide.

Finally, the fixed prompt template required for reproducibility is also a constraint on adaptability. A prompt calibrated for 2018–2024 macroeconomic conditions may not be optimal for a different regime. The system implicitly assumes that the macro dimensions encoded in the prompt are stable sources of market-relevant anomaly — an assumption that holds in the evaluation period but is not guaranteed to hold forward.

8 Conclusion

This paper makes one primary methodological contribution and three supporting empirical observations at the intersection of LLM uncertainty quantification and financial risk measurement. It does not claim to be the first to use language-model entropy as a market signal — that idea is due to [Glasserman and Mamaysky \(2019\)](#); [Glasserman et al. \(2023\)](#) — and its empirical evidence is, on honest accounting, modest.

The primary contribution is the *model-training-cutoff contamination protocol*. Any study that uses a large modern LLM to detect historical anomalies risks measuring encoded hindsight rather than real-time signal, because the model’s pretraining may already cover the episode. By recomputing the entropy signal on models whose training corpora predate the episode (here GPT-2 and SmolLM2) and checking for replication, one can partition an observed signal into model-agnostic and model-specific components. We argue this is a necessary robustness check for any LLM-based financial signal evaluated over historical windows inside a large model’s training span — and that, applied to our own study, it shows most of the GPT-4o results cannot be cleanly interpreted as real-time detection.

First, applying the protocol to COVID-19 yields the cleanest result in the paper: GPT-2 (data cutoff end of 2017), which could not have seen COVID text, registers elevated entropy only after the WHO first report (peak $z = +2.124$, week 2020-04), while VIX still sat at 12–15 and 7 weeks before the crash. The pre-report GPT-4o spike is not replicated and is excluded as likely contamination. SmolLM2-135M post-dates COVID and is used only as a contaminated control, so this clean result rests on a single model and $n \approx 4$ weeks.

Second, cross-ticker crisis analysis shows sector-specific resonance — AVAV (drones) leads on Ukraine, AEIS (semiconductor equipment) on SVB, APAM (asset manager) on both, while MOD shows essentially no pre-Ukraine signal. The *structure* of this pattern is theoretically predicted and hard to attribute to undirected leakage, though the underlying GPT-4o signals fall inside the model’s training span and cannot be fully cleaned of contamination.

Third, as an illustrative volatility-prediction application, the entropy signal is competitive with VIX overall and sign-positive mainly in IT/Semiconductors and Industrials; pooled IC is marginal (NW $p \approx 0.067$), the in-sample binomial “win” count overstates this (6 of 17 wins are between two negative ICs), out-of-sample cross-sectional significance is not established ($p = 0.314$), and the pooled directional-F1 numbers do not beat a horizon-matched base-rate benchmark. The strongest single number, MOD’s out-of-sample $F1 = 0.636$ / $IC = +0.222$, is the maximum of a large search grid and is reported as such.

The practical implication is correspondingly narrow: the entropy score is a cheap by-product of any LLM pipeline processing macroeconomic text and may be useful as a medium-horizon overlay for commodity-exposed industrials, but on the present evidence it is not a validated broad-market risk measure, and its main present value is as a worked example of the contamination protocol.

Future work should prioritize closing the two evidential gaps this revision makes explicit. First, a formal data-snooping correction — White’s Reality Check or Hansen’s SPA test — run over the full configuration grid, to establish whether any per-ticker result survives once the search is accounted for; this is the single most important addition. Second, an orthogonalization of the entropy signal against the GPR and EPU indices over the full backtest, reporting incremental IC and incremental R^2 , to establish that the signal is not simply re-deriving news novelty already captured by [Glasserman and Mamaysky \(2019\)](#). Beyond these, expanding the universe beyond 24 tickers to 200 or more across all GICS sectors would provide the power needed for portfolio-level significance testing; prospective evaluation on 2025–2026 data would supply the only fully contamination-free, non-post-hoc out-of-sample test; applying the contamination protocol across additional pre-2020

model families would test the robustness of the COVID finding; and per-sector prompt optimization, guided by the macro-expansion results, remains the most direct route to improving IC across the full universe. Integrating the entropy signal into portfolio construction as a volatility overlay would test whether any leading-indicator property translates into realized risk-adjusted improvement.

References

References

- Baker, S.R., Bloom, N., and Davis, S.J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4), 1593–1636.
- Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.
- Caldara, D. and Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, 112(4), 1194–1225.
- Calomiris, C.W., Cakir Melek, N., and Mamaysky, H. (2020). Mining for oil forecasts. Federal Reserve Bank of Kansas City Working Paper RWP 20-20 (revised 2022 as “Big Data Meets the Turbulent Oil Market”).
- Chen, M., et al. (2022). Geopolitical risk and equity market volatility: A textual analysis approach. *International Review of Financial Analysis*.
- Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.
- Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), 625–630.
- Glasserman, P. and Mamaysky, H. (2019). Does unusual news forecast market stress? *Journal of Financial and Quantitative Analysis*, 54(5), 1937–1974.
- Glasserman, P., Mamaysky, H., and Qin, J. (2023). New news is bad news. Working paper, arXiv:2309.05560 / SSRN 4555832.
- Huang, A.H., Wang, H., and Yang, Y. (2022). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2), 806–841.
- Kadavath, S., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Li, X., et al. (2023). Political news and cross-asset volatility: Evidence from deep NLP. *Journal of Financial Economics*.
- Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

- Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1–2), 125–144.
- Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, 53(1), 61–65.
- Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.
- Wu, S., et al. (2023). BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.