

When information breaks the historical pattern: preliminary evidence from a deployed Bitcoin forecaster

Kibaek Kim, Kiok Kim, Danielle Ahn

8 September 2026

Abstract. Historical-pattern forecasts may fail when information moves prices beyond historical support. We audit next-session forecasts from a deployed Bitcoin system. Across 1,973 BTC/USD and 1,125 BITO origins, directional accuracy was 48.8% and 50.6%, with near-nominal interval coverage. On return-defined large-move sessions, accuracy fell to 36.7% and 37.5%, and all nominal-80% intervals missed. Headline-count spikes increased standardized error for both instruments, but directional degradation appeared only for BTC and was imprecise. These exploratory results do not measure investor behavior; they motivate research into which news changes behavior and how responses propagate into prices.

Keywords: Bitcoin; forecast evaluation; information shocks; investor attention; historical analogs.

JEL: G14, G17, C53.

1. Introduction

Price-pattern forecasters assume that a market state resembling a historical state has an informative historical continuation. This premise can be useful when the relevant drivers evolve gradually. It is more fragile when new information changes beliefs, constraints, or attention faster than those changes can enter a price-based state vector. A forecast may therefore look adequate on average while failing conditionally on the observations that matter most.

This paper reports an exploratory audit of one deployed historical-analog forecasting system for BTC/USD and the ProShares Bitcoin Strategy ETF (BITO). It asks three narrow questions:

1. Is its next-session direction or nominal 80% interval informative on average?
2. Does forecast performance change on days with unusually high headline counts?
3. What happens when the realized move is far outside the recent volatility scale?

The third question is diagnostic rather than predictive: a large-move day is defined using the realized return and cannot be known at forecast time. The headline-count condition is observable during the target day, but it is also not a causal news classification. The contribution is consequently a preliminary boundary finding, not a universal theorem about news or human behavior.

Prior research documents weak and data-sensitive performance for technical rules (Brock et al., 1992; Sullivan et al., 1999), strong relationships between information, attention, and market activity (Chan, 2003; Da et al., 2011; Tetlock, 2007), and distinctive cryptocurrency risk (Liu & Tsyvinski, 2021; Urquhart, 2016). The present audit connects these literatures by showing how the distribution

from a live pattern-based product changes conditionally around information-intensive and extreme-return observations.

2. Forecasts, data, and design

2.1 Audited system

The authors operate the audited service. The production system is described by its operator as a seven-session historical-analog model: it maps recent price and indicator states to similar historical states and reports a probability of an increase, a median, and P10/P90 paths. This study treats that engine as a black box.

The public repository does not contain the proprietary forecast-generation code. The 2026 directional signals were frozen from the deployed service, while the multi-horizon forecast distributions were reconstructed with the production engine in a point-in-time harness. The public materials therefore permit reanalysis of the archived forecast-score rows, but they do not permit an independent third party to regenerate the forecasts from raw prices. This boundary is material and should not be confused with full model reproducibility.

We use 1,973 BTC/USD origins from 5 April 2021 through 29 August 2026 and 1,125 BITO origins from 4 March 2022 through 27 August 2026. Prices are licensed Alpaca daily bars. Headline counts use the licensed Benzinga feed available through Alpaca from February 2024. Headline text is not used in the public analysis artifact and is not redistributed; only counts and cryptographic fingerprints are retained.

2.2 Scores and conditions

The analysis is limited to the next session, avoiding the repeated-target overlap created by scoring seven horizons together. Directional accuracy is

$$H_t = \mathbf{1} \{ \mathbf{1}(\hat{p}_t \geq 0.5) = \mathbf{1}(R_{t+1} > 0) \},$$

where $\hat{p}_t$ is the forecast probability of an increase and R_{t+1} is the realized next-session return. Interval coverage is one when the return lies between the forecast P10 and P90. Standardized absolute surprise is

$$|s_t| = \left| \frac{R_{t+1} - \hat{q}_{50,t}}{(\hat{q}_{90,t} - \hat{q}_{10,t})/2.5631} \right|.$$

The denominator maps the P10-P90 width to a normal-distribution scale. It is a reporting normalization, not an assumption that returns are normal.

A **headline-count spike** has a target-day count at least twice its trailing 30-day median, with one added to numerator and denominator to stabilize low counts. The count identifies information intensity, not content, novelty, credibility, or investor response. A **large realized move** has

$$|z_{t+1}| = \frac{|r_{t+1}|}{\hat{\sigma}_t^{(20)}} \geq 2.5,$$

where the volatility estimate uses only information through the forecast origin. Because the event label uses r_{t+1} , large-move results describe conditional failure after the fact. In particular, zero coverage is partly mechanical when a 2.5-volatility move is compared with an interval approximately 1.3 volatilities wide.

Point estimates and 95% intervals use a circular moving-block bootstrap with seven-observation blocks and 4,000 deterministic draws. Conditional differences resample the complete ordered sequence with the event indicator attached, rather than separately resampling filtered event and comparison arrays. The analysis and contrasts are exploratory and were not preregistered.

3. Results

3.1 Average performance is close to a direction coin

Across all origins, BTC/USD next-session directional accuracy is 0.488 (95% block interval 0.467 to 0.510), and BITO accuracy is 0.506 (0.479 to 0.532). Nominal 80% interval coverage is 0.784 (0.764 to 0.802) for BTC and 0.777 (0.752 to 0.801) for BITO. Thus the predictive intervals are approximately calibrated marginally, while their medians provide little directional separation.

Table 1. Next-session performance by condition. Intervals in the all-origin rows are seven-observation moving-block intervals. Conditional-difference intervals are reported in the text. Large realized moves are outcome-defined diagnostics.

instrument	condition	n	directional accuracy	P10-P90 coverage	mean absolute surprise
BTC/USD	all origins	1,973	0.488	0.784	0.852
BTC/USD	quiet-news days	847	0.499	0.783	0.844
BTC/USD	headline-count spikes	95	0.389	0.684	1.148
BTC/USD	large realized moves	98	0.367	0.000	3.593
BITO	all origins	1,125	0.506	0.777	0.889
BITO	quiet-news days	612	0.500	0.776	0.868
BITO	headline-count spikes	34	0.529	0.676	1.270
BITO	large realized moves	48	0.375	0.000	3.682

3.2 Headline intensity is associated with larger forecast errors, but not uniformly with wrong direction

On BTC headline-count spike days, accuracy is 0.389 versus 0.499 on quiet-news days, a difference of -0.110 (95% block interval -0.216 to 0.003). Coverage falls by 0.099 (-0.205 to 0.006), while absolute surprise rises by 0.304 (0.036 to 0.605).

BITO does not show the same directional pattern: accuracy is 0.529 on 34 spike days and 0.500 on 612 quiet-news days, a difference of 0.029 (-0.138 to 0.229). Its coverage difference is -0.100 (-0.273 to 0.074), while absolute surprise increases by 0.402 (0.040 to 0.822).

The common signal is therefore wider forecast error on high-count days, not a replicated directional reversal. Headline volume alone is too coarse to identify which information matters. It mixes anticipated and unexpected releases, favorable and unfavorable content, repeated reporting, and stories with different credibility or audience reach.

Forecast performance changes when information intensity or realized moves are large

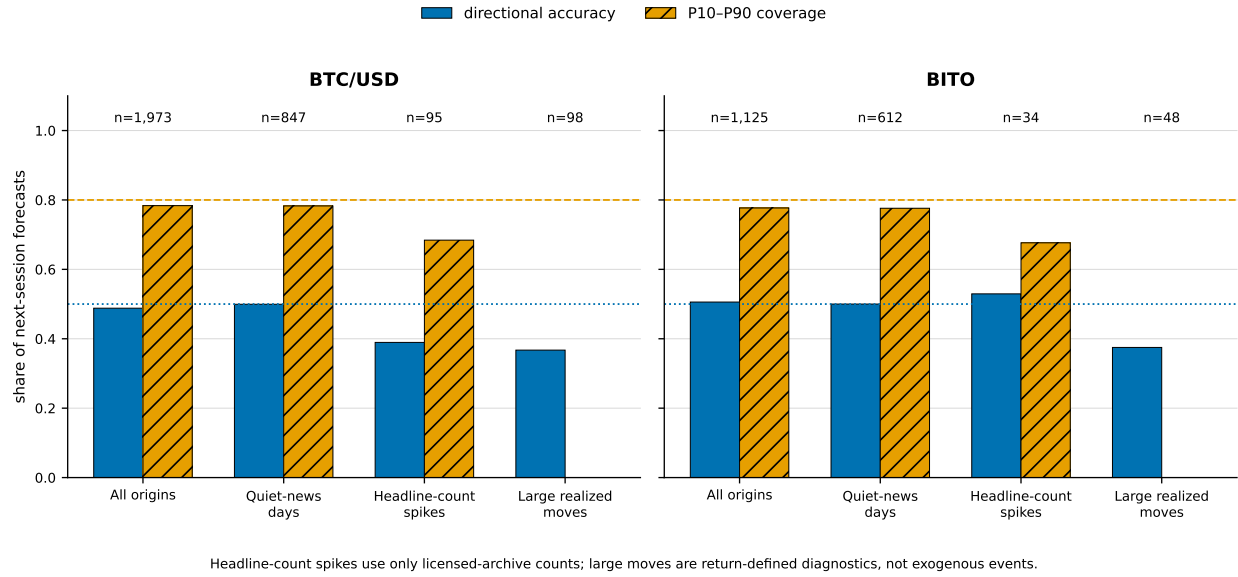


Figure 1. Directional accuracy and nominal-80% interval coverage for all origins, quiet-news days, headline-count spikes, and return-defined large-move sessions. The dotted 0.5 and dashed 0.8 lines are directional and coverage references.

3.3 Large realized moves expose conditional distribution failure

On 98 BTC large-move sessions, forecasts issued one session earlier have directional accuracy of 0.367, compared with 0.494 otherwise. The difference is -0.127 (-0.220 to -0.030). On 48 BITO large-move sessions, accuracy is 0.375 versus 0.512, a difference of -0.137 (-0.276 to 0.002). Every P10-P90 interval misses in both large-move samples.

The interval result should not be read as an independent test: the event threshold is based on the same realized return used to score coverage. The directional result is less mechanical but remains conditional on an outcome-defined subset. Together, the results locate where this specific historical-pattern distribution fails; they do not establish that every such move is news-caused or that all price-pattern models share the same failure rate.

Figure 3 shows the same failure case by case for the twelve largest BTC/USD declines with a forecast on record. The forecast issued at the prior close carried a mean probability of an increase of 0.49, and the deployed consensus was Sell in two of the twelve. The realized next-session close fell below the forecast P10 in all twelve panels; by the seventh session the realized path was back inside the band in five. Because the panels were selected on the realized decline, they illustrate the conditional failure described above rather than adding an independent test.

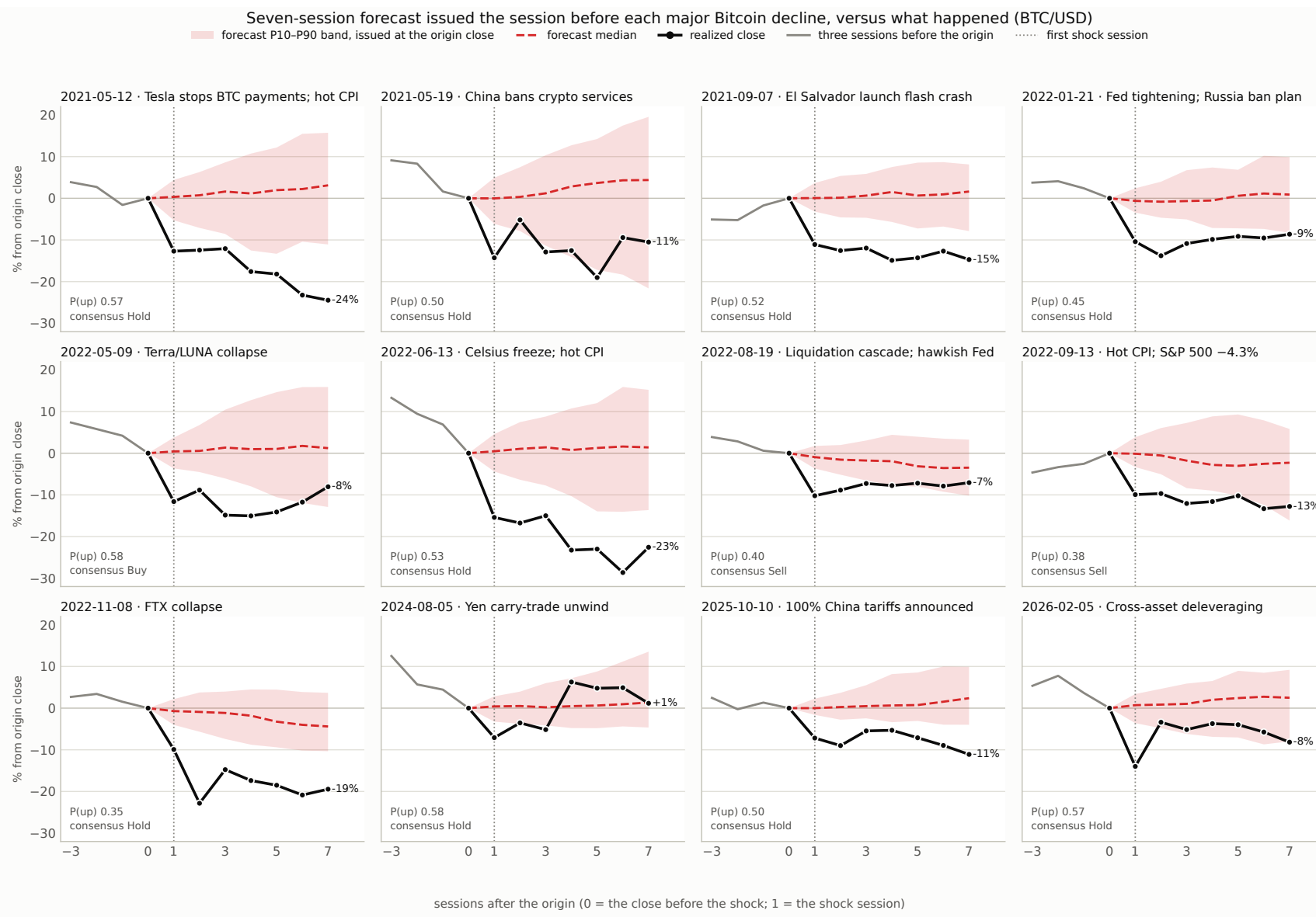


Figure 3. The seven-session forecast issued at the close before each of twelve large Bitcoin declines (P10–P90 band and median) against the realized path, in % from the origin close. The twelve sessions were selected ex post by realized decline size: the ten largest single-session BTC/USD declines with a forecast on record, plus the two largest liquidation events of the headline-archive period. Panel labels are descriptive context taken from contemporaneous reporting, not a causal news classification. P(up) and consensus are the deployed signal at the origin.

3.4 A motivating information event

On 18 August 2026, before a reported White House crypto-policy meeting and public push for the CLARITY Act, the BTC model assigned a next-session median return of -0.05% and a P90 of $+1.15\%$. BTC rose 7.15% the next session and 21.10% over three sessions; the three-session P90 had been $+2.61\%$. The deployed consensus was Sell/confirmed at the origin.

For BITO, the next-session median was $+0.32\%$ and P90 was $+2.06\%$; realized returns were $+5.96\%$ in one session and $+19.24\%$ in three sessions, versus a three-session P90 of $+3.82\%$. BITO’s direction was correct, but its distribution did not represent the magnitude. Public reporting contemporaneously tied the rally to the policy event (CNBC, 2026; Ray, 2026).

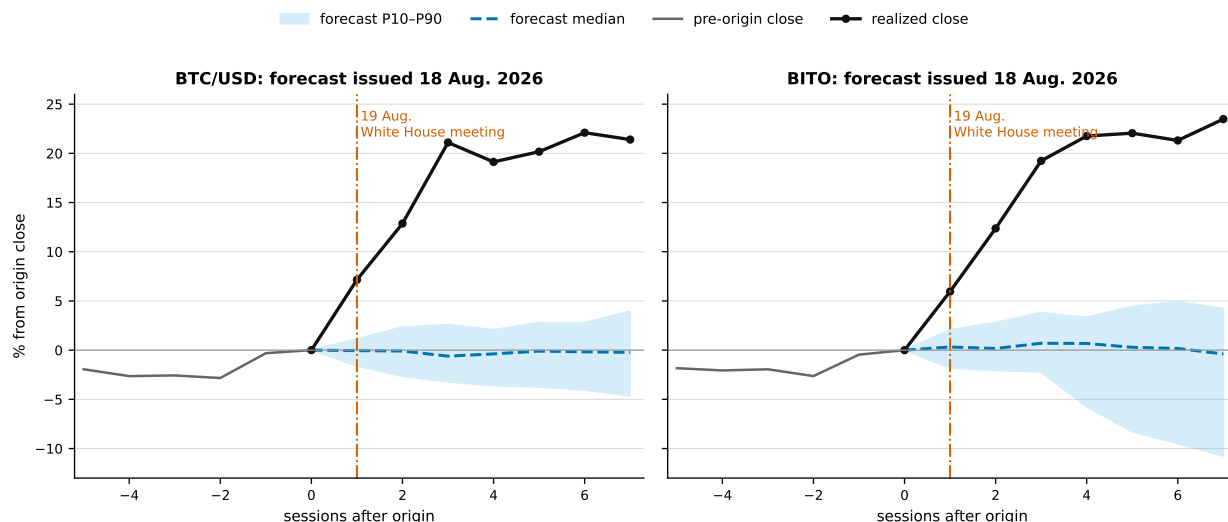


Figure 2. Point-in-time forecasts issued at the 18 August 2026 close and subsequent realized paths. Values are percentages from the origin close. The event line marks 19 August.

This example illustrates the paper’s limited claim: new information can move the realized path outside the historical range represented by a deployed pattern forecaster. A single event cannot identify a general behavioral mechanism.

4. Interpretation and research agenda

The results separate three ideas that are often conflated.

First, marginal calibration does not guarantee conditional reliability. Approximately 78% coverage over all origins coexists with extreme conditional misses. A user who sees only average coverage may not know when the interval is least trustworthy.

Second, “news” is not one variable. The headline-count proxy is associated with larger standardized errors in both instruments, but directional degradation is visible only for BTC and is estimated imprecisely. A useful gate requires content, novelty, timing, credibility, expectedness, and likely audience—not count alone.

Third, this audit does not observe human behavior. Price continuation, model latency, or a large forecast error cannot by itself distinguish attention, belief updating, herding, forced liquidation, in-

stitutional rebalancing, or market-microstructure constraints. Those mechanisms require additional data and research designs.

We propose four next steps:

1. **Build a news taxonomy.** Distinguish scheduled macro announcements, regulation, exchange or protocol failures, liquidation cascades, institutional flows, and repeated commentary. Label expectedness, sign, novelty, and source credibility before examining returns.
2. **Measure responses directly.** Combine timestamped news with order flow, liquidation records, search and social attention, investor-position data, or preregistered experiments. Separate retail from institutional responses where possible.
3. **Estimate propagation.** Model the sequence from information arrival to attention, trading, liquidity, price impact, and any subsequent drift or reversal. Use event time rather than a daily count when timestamps permit.
4. **Validate a reliability gate prospectively.** Freeze the taxonomy and gate before new events, then evaluate whether withholding or widening a forecast improves proper scores out of sample. The gate should be tested independently of the model it protects.

The immediate engineering implication is not to “train harder” on the same price history. For a historical-analog system, adding more calm analogs may leave the information state absent from the feature space. The research question is which observable information and behavioral-response variables improve conditional reliability without introducing look-ahead or narrative labels created after the move.

5. Limitations

This is an exploratory audit of one proprietary product and two related Bitcoin exposures, not a representative sample of commercial systems. The forecast-generation engine cannot be independently inspected or rerun from the public repository. Distributional forecasts before the live archive were reconstructed with production code, while the public 2026 directional signals were frozen. The findings may therefore combine properties of the deployed model with properties of the reconstruction process.

The headline measure uses licensed archive counts, not semantic content, readership, or novelty. Target-day counts may respond to the price move itself, so the association is not causal. Large-move conditioning is explicitly ex post and mechanically selects interval misses. BTC and BITO observations are economically dependent. The seven-observation block length is a pragmatic sensitivity choice rather than an estimated dependence horizon. No order-flow, position, survey, or experimental data identify investor behavior.

Finally, the study was organized after the August 2026 case drew attention to forecast failure. The systematic rules were fixed before this narrowed rerun, but the analysis was not preregistered. Results should be treated as preliminary hypotheses for prospective evaluation.

6. Conclusion

One deployed historical-pattern forecaster is marginally calibrated in interval width yet fails conditionally when realized Bitcoin moves are large. High headline counts are also associated with larger standardized forecast errors, although their directional relationship differs between BTC and

BITO. These findings are evidence of a reliability boundary, not evidence for a specific psychological mechanism.

The next study should identify which information changes investor behavior, measure that response directly, and test prospectively whether those variables can warn when a price-pattern forecast should be withheld or widened.

Data, code, and conflicts

The rights-safe aggregate used for this paper is `research/output/arxiv_prefinding.json`, generated by `research/arxiv_prefinding.py` from archived forecast-score rows. The public repository is <https://github.com/hyper07/finance-signal>. The proprietary forecast-generation engine is not included, so the archived scores can be checked and reanalysed but not regenerated independently. Raw price bars and licensed headline text are not redistributed; only headline counts and SHA-256 fingerprints are public. The authors operate the audited forecasting service. This paper is research, not investment advice.

References

- Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance*, 47(5), 1731-1764.
- Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, 70(2), 223-260.
- CNBC. (2026, August 20). Bitcoin surges 12% in two days as Trump, crypto executives lead final push for Clarity Act. <https://www.cnbc.com/2026/08/20/bitcoin-surges-as-trump-crypto-execs-lead-final-push-for-clarity-act.html>
- Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, 66(5), 1461-1499.
- Liu, Y., & Tsyvinski, A. (2021). Risks and returns of cryptocurrency. *Review of Financial Studies*, 34(6), 2689-2727.
- Ray, S. (2026, August 20). Bitcoin soars above \$70,000 after Trump calls for passage of Clarity Act at White House crypto event. *Forbes*. <https://www.forbes.com/sites/siladitya-ray/2026/08/20/bitcoin-soars-above-70000-after-trump-calls-for-passage-of-clarity-act-at-white-house-crypto-event/>
- Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5), 1647-1691.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139-1168.
- Urquhart, A. (2016). The inefficiency of Bitcoin. *Economics Letters*, 148, 80-82.