

Point-in-Time Inflation Surprise and the Equity Size Spread: A Corrected Null, a Deployment- Governance Failure, and a Frozen Forward Experiment

Arhan Canli

Founder, System Architect, and Quantitative Researcher
Canli Capital / AlphaC Algorithms

VERSION	1.0.0
RELEASED	23 August 2026
REGISTRY KEY	alphavintage_macro_surprise
STATUS	Working paper preprint / not peer reviewed

Research simulation and, where explicitly identified, Alpaca paper evidence. No funded performance, peer review, independent replication, external acceptance, or future return is claimed.

A corrected null, a deployment-governance failure, and a frozen forward experiment

Author: Arhan Canli, Founder and Quantitative Researcher, Canli Capital

System: ALPHAC / AlphaC Algorithms · AlphaVintage sleeve

Version: 1.0, 2026-08-23

Status: public working paper; not peer reviewed

Capital boundary: historical research simulation and a separately labelled Alpaca paper account

Abstract

This paper reports a point-in-time inflation-surprise strategy applied to the dollar-neutral IWM-minus-SPY equity size spread. The fixed specification constructs headline and core CPI surprises from expanding AR(3) forecasts fit within each historical macro-data vintage, enters after the vintage date, clips exposure, and charges 6 basis points one-way per leg. The original active-day evaluation appeared to pass its four preregistered reading rules. Two subsequent integrity findings changed that conclusion: a missing adjacent CPI observation could cause an old change to be reused under a new vintage, and zero-exposure trading sessions were omitted from the portfolio calendar. A single sealed correction, with no retuning and zero new hypotheses, produced net Sharpe **0.2298**, Newey-West t-statistic **1.2673**, and maximum drawdown **-25.1%** over 6,296 sessions. The t-statistic failed the locked 1.5 threshold, so the research verdict is **KILLED**.

AlphaVintage had entered an Alpaca paper account six days before that corrected verdict. It was not removed after the result because the forward composition had already been frozen and because historical portfolio diagnostics showed lower dependence and drawdown despite a slightly negative book-Sharpe contribution. This is not presented as validation. It is a prospective paper-trading experiment whose continuation, correction history, and conflict between standalone alpha and portfolio diversification are disclosed here. The central result is therefore not a profitable anomaly claim. It is an auditable case study in point-in-time data engineering, falsification, and deployment governance.

1. Claim boundary

The corrected historical signal is **not research-admitted**. The Alpaca account is paper-only; its broker-derived record is self-published, short, and not an independent attestation. Nothing in this paper establishes funded performance, a forward Sharpe, an expected live drawdown, causal price underreaction, or future returns. Repository publication is not peer review.

The paper distinguishes three decisions that are easy to conflate:

1. **Did the standalone historical signal pass its locked rule?** No.
2. **Did the sleeve historically diversify the specified four-sleeve book?** Descriptively yes on the measured window, though not prospectively established.

3. Should it remain allocated? No conclusion is drawn here. The current paper allocation is a frozen experiment, not an endorsement.

2. Motivation and prior evidence

Macroeconomic announcements can move equities through discount-rate, risk-premium, and cash-flow channels. Pearce and Roley (1984) emphasize the unexpected component of announcements; Wachter and Zhu (2018) study the concentration of equity premia around scheduled announcements; Modugno and Palazzo (2025) relate aggregate activity and price news to equity-market and firm outcomes. Those papers motivate studying macro news, but none establishes the specific delayed IWM-minus-SPY effect tested here.

The data problem is equally important. Croushore and Stark's real-time-data work shows why the vintage available to an historical decision maker can differ materially from today's revised series. The Philadelphia Fed's Real-Time Data Set for Macroeconomists publishes vintage snapshots for that purpose. CPI is released on a schedule maintained by the U.S. Bureau of Labor Statistics. The strategy deliberately uses historical vintages rather than a modern revised history and delays entry until the first trading close strictly after the recorded vintage date.

The proposed mechanism was that unexpectedly high inflation disadvantages small companies through financing and margin channels, producing a gradual relative rotation from IWM toward SPY. This is a falsifiable economic story, not evidence that the effect exists. The experiment does not isolate financing costs, margins, institutional flows, or announcement-time reactions, so the mechanism remains conjectural.

3. Frozen specification

The implementation is `scripts/probe_cpi_surprise_size.py`. Exactly one configuration was tried:

Element	Locked value
Tradable instruments	IWM and SPY
Placebo market input	QQQ, used only for the QQQ-minus-SPY falsification test
Portfolio	dollar-neutral IWM minus SPY spread
Macro inputs	point-in-time headline CPI (<code>PCPI</code>) and core CPI (<code>PCPIX</code>) vintages
Forecast	expanding ordinary-least-squares AR(3), fit strictly before the new observation
Surprise	forecast residual divided by in-sample residual standard deviation, clipped to +/-3
Composite	simple mean of headline and core standardized surprises
Weight	negative clipped surprise, bounded to [-1, +1]
Entry	close of first trading day strictly after the vintage date
Exit	next vintage date
Trading costs	6 bp one-way on each leg at rebalance
Configurations tried	1

Growth is differenced **within each vintage column**. Differencing a first-release series across vintages can mix seasonal re-estimation and base changes. The corrected implementation additionally requires the expected release month and its immediately preceding month to be finite. It skips an unusable vintage rather than filling, interpolating, carrying forward, or relabelling an older change.

The portfolio return calendar retains every exchange session between first and last exposure. Sessions with no position receive an explicit zero. This is necessary because annualizing only active observations changes the time basis of the estimator.

4. Preregistered decision rule and trial lineage

The candidate was to be added only if all four conditions held:

- standalone net Sharpe exceeded its correlation-adjusted book hurdle;
- Newey-West t-statistic was at least 1.5;
- the apparent portfolio benefit came from mean return rather than variance dilution; and
- a QQQ-minus-SPY placebo remained insignificant.

Failure of any condition meant **KILLED**. The CPI candidate spent one configuration. Five later macro-vintage variants were declared and reported as the same information family, not independent sleeves. The public family lineage therefore contains seven immutable identities. The correction repaired the existing CPI identity and spent **zero** additional return hypotheses.

5. Integrity corrections

5.1 Missing-release fail-open defect

The historical lake contains no October 2025 headline or core CPI level. A later vintage could contain November while the immediately preceding October value remained absent. The old `diff().dropna()` path could then leave an older computable change at the tail and treat it as the new release. The corrected rule requires the two adjacent latest months and otherwise emits no signal.

5.2 Active-day annualization defect

The original curve persisted only days with nonzero exposure. That removed legitimate zero-return sessions from the first-to-last portfolio interval. The corrected curve contains 6,296 portfolio sessions, of which 5,998 are active. The former active-day net Sharpe of **0.3382** is superseded; it is retained solely in the correction record.

5.3 Sealed re-evaluation

The correction protocol was written before revised performance was opened. AR lags, direction, clipping, costs, timing, instruments, scale, and gates were unchanged. The revised result replaced the prior result whether better or worse.

Measurement	Corrected value	Reading
Gross annualized Sharpe	0.44100360008192313	diagnostic
Net annualized Sharpe	0.2298	below any strong standalone claim
Newey-West t-statistic	1.2673	fails locked 1.5 gate
QQQ-SPY placebo Sharpe	-0.0205	placebo remains dead
QQQ-SPY placebo NW t	-0.11700964029171129	placebo remains dead
Maximum drawdown	-25.1%	economically material
Portfolio sessions	6,296	correct calendar basis
Active sessions	5,998	diagnostic only
Verdict	KILLED	binding result

Three of four reading checks remained true; the significance check failed. A near-pass is a fail under a preregistered conjunction. No threshold was lowered to preserve the sleeve.

6. Deployment chronology and governance finding

AlphaVintage first held a position in its dedicated Alpaca paper account on **2026-08-10**. The corrected re-evaluation was sealed on **2026-08-16**. The public state continued displaying the superseded 0.3403 Sharpe and 1.82 t-statistic until **2026-08-19**. That three-day publication lag and the fact that deployment preceded the final corrected verdict are governance failures, not footnotes.

The present system records the sleeve as paper-only, labels the corrected **KILLED** verdict, and requires any composition change to begin a new declared forward epoch. Current broker reconciliation confirms a dedicated active paper account and exact overlapping equity marks, but it does not convert the historical null into alpha evidence.

The process lesson is concrete: deployment eligibility must require a machine-verified sealed outcome, and public claims must be generated from the same artifact rather than copied into prose. The project now treats correction latency and source binding as engineering controls.

7. Portfolio evidence: diversification without standalone validation

On the current 1,061-row common research window, AlphaVintage's correlations to the other sleeves are +0.0393 to crypto carry, -0.0620 to equity momentum, and -0.0444 to managed-futures trend. Its fixed-to-cash marginal contribution changes full-book research Sharpe from 1.7909 without the sleeve to 1.7846 with it, a **-0.0063** delta. Thus the current exact study does not show a positive historical Sharpe contribution.

An older composition comparison on the same row count found that including AlphaVintage reduced average pairwise correlation from +0.0724 to +0.0260, increased the diversification ratio from 1.6425 to 1.9427, and reduced historical maximum drawdown from 3.81% to 2.75%. These are retrospective composition diagnostics. They were known before this paper, are not independent, and do not authorize a reweighting decision.

The distinction matters: a weak-mean component can reduce covariance and drawdown, but variance dilution alone is not evidence of alpha. Conversely, deleting a frozen sleeve after seeing the comparison would select the composition in sample. The forward experiment exists to resolve that tension prospectively.

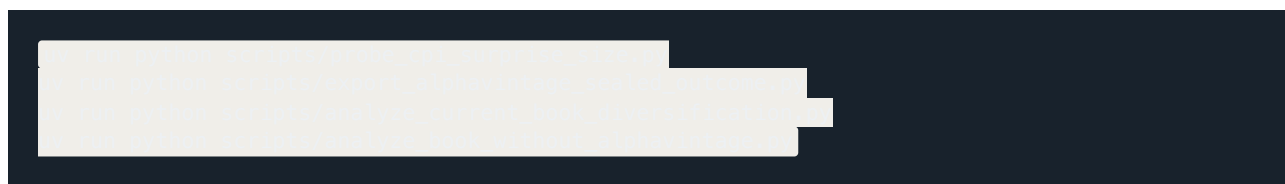
8. Execution, capacity, and unresolved realism

IWM and SPY are liquid exchange-traded funds, and monthly turnover is low by construction. That does not make the backtest executable by default. The historical model uses a fixed 6 bp one-way cost per leg and does not fully model bid-ask variation, short borrow availability and fees, dividends on the short leg, locate failures, market impact, auction behavior, tax, or financing. The historical maximum drawdown is large relative to the weak mean estimate. A publication-grade capacity claim therefore remains **not established**.

The Alpaca paper record can test order generation, broker marks, holdings, fills, and operational continuity. It cannot test borrow economics or market impact with real capital. Funded deployment would require a new capital decision and a separately governed evidence standard.

9. Reproducibility and provenance

Canonical commands:



A PEP 723/`uv --isolated` path reacquired CPI vintages from the Philadelphia Fed and IWM, SPY, and QQQ histories from Yahoo, then recomputed the core signal, returns, costs, significance, and placebo outside the repository data tree. It matched 6,296 portfolio sessions, 5,998 active sessions, both gates, and the `KILLED` verdict. Net-Sharpe and Newey-West-t differences were 0.0000006327801884142836 and 0.000004964325143186343; the largest of six metric differences was 0.00001345836069367723. Yahoo data were not byte-identical (maximum relative price difference 0.0000013794), so the receipt says numerically near-identical core reproduction. Its four-decimal tolerance was selected after the observed drift and is diagnostic, not preregistered. This author-run replay excludes the separate diversification-strategy generation, vendor-row redistribution, and independent human replication.

All three upstream benchmark strategies now have completed author-run replays, with different outcomes. A temporary workspace extracted pinned source commit

`577555f12636e4df81e42a3940184678d0cceb7e`, copied a pre-refresh sealed snapshot of AlphaTrend's 408-file ETF lake plus its instrument and experiment state, and reran the declared 2003-01-01 through 2026-08-24 trend gauntlet. The resulting 5,191-row `mf_live_fwd/equity.parquet` is byte-for-byte identical to the curve consumed by this trial; every timestamp, equity value, and log return matches, and 466 of 467 output files match. Only `walkforward.json` differs: the committed source sees 13 trials, whereas the historical run used the later 228-identity DSR union. The historical DSR remains the published value and is not regraded.

AlphaMax was replayed in a clean workspace from a newly acquired 375-instrument Polygon panel. Fifty-eight earlier lookback sessions were unavailable under the current entitlement, and the fresh-vendor curve did not reproduce the surviving historical curve. Timestamps and configuration match, but the maximum absolute equity-path difference is \$574.69; annualized Sharpe is 0.9071 in the historical artifact and 0.9213 in the replay. Historical and replay DSR values are 0.3406 and 0.3479, so both fail the 0.95 gate. The stored historical artifact remains unchanged. The replay closes execution debt but not exact historical-input or output equivalence.

The historical `prereg_investment` comparison was replayed from the surviving Sharadar archives in a separate temporary workspace. The loader rebuilt 24,085,990 price rows and the exact 6,835-name, 11,359-interval XUSE membership state. The strategy then reproduced all 779 output files byte for byte, including the 5,385-row equity parquet and its 75-trial validation block. The first disclosed attempt failed before strategy execution because the predecessor profile rebuilt only 1,050 XUSE identities. The corrected attempt bound three run-critical files from the first post-run commit, which restored the documented wide-universe state. This was source-lineage adjudication, not performance tuning. `prereg_investment` remains a non-sleeve historical gate input whose original run was not covered by its later preregistration.

These replays establish two exact historical strategy-output equivalences and one completed fresh-vendor divergence. They do not recover every dirty historical source tree, grant vendor-row redistribution rights, validate the non-preregistered comparison prospectively, or constitute independent reproduction.

The sealed corrected curve is `artifacts/probe/cpi_surprise_size/equity.parquet` with SHA-256 `d277c63ddf2bed6e9314aa863dbbf6adf3f4adb55bd89e8166aee4a19aab415f`. The result is `artifacts/probe/cpi_surprise_size/result.json` with SHA-256 `686654b4617efb8322f6fda37f2e71a375423aecc6de03477a0186cf2844e509`. The auditable restatement is `artifacts/engineering/alphavintage_sealed_outcome.json`. Trial identities are in `artifacts/research/macro_economic_trend_family.json` and the union trial manifest.

The public bundle includes a versioned PDF, HTML and LaTeX source, machine-readable paper metadata, environment bindings, a data-rights manifest, checksums, a software bill of materials, RO-Crate, and the portable receipts. Data-rights review, exact AlphaMax historical-input recovery, a complete multi-sleeve end-to-end reproduction, and independent human reproduction remain explicit blockers before the release can be called externally replicated.

10. Limitations

- The surprise is an AR residual, not a survey-based market expectation.
- The vintage stamp is conservative but is not the exact BLS release timestamp for every month.
- The IWM-SPY spread is an imperfect proxy for firm-size exposure and can load on sector, quality, profitability, duration, and financing conditions.
- The Newey-West estimator does not solve all overlap, regime, and specification uncertainty.
- The 1.5 threshold is a preregistered reading rule, not a universal evidentiary standard.
- The family contains related macro tests; a single-configuration label does not erase broader research multiplicity.

- Historical diversification is measured on a post-2022 window and is not crisis-complete.
- The forward record is too short to support a performance or drawdown conclusion.

11. Conclusion

The point-in-time CPI-surprise size-spread hypothesis did not survive its corrected locked test. That result remains published because a research platform is more credible when failed signals, software defects, delayed corrections, and awkward deployment decisions are first-class evidence.

AlphaVintage's low measured dependence makes it a useful frozen forward experiment; it does not make it a validated alpha sleeve. Success for this project means resolving that distinction with time, broker-reconciled evidence, and independent reproduction, not changing the standard after seeing the answer.

References

- Croushore, D., and T. Stark (1999/2001), “A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?”, Federal Reserve Bank of Philadelphia Working Paper 99-21 and *Journal of Econometrics* 105, 111-130. <https://www.philadelphiafed.org/the-economy/macroeconomics/a-real-time-data-set-for-macroeconomists-does-the-data-vintage-matter>
- Modugno, M., and D. Palazzo (2025), “Decoding Equity Market Reactions to Macroeconomic News,” Federal Reserve Finance and Economics Discussion Series 2025-007. <https://doi.org/10.17016/FEDS.2025.007>
- Pearce, D. K., and V. V. Roley (1984), “Stock Prices and Economic News,” NBER Working Paper 1296. <https://doi.org/10.3386/w1296>
- U.S. Bureau of Labor Statistics, “Schedule of Releases for the Consumer Price Index.” https://www.bls.gov/schedule/news_release/cpi.htm
- Wachter, J. A., and Y. Zhu (2018), “The Macroeconomic Announcement Premium,” NBER Working Paper 24432. <https://doi.org/10.3386/w24432>

Authorship and contribution statement

Arhan Canli directed the research, specified the strategy and governance, built the software and evidence system, investigated the corrections, and wrote the paper. AI assisted development and editing; he remains responsible for every claim, decision, and correction.