

Specification Choices and the Interpretation of Predictive Evidence

Jorge Cid Moreno

European Central Bank

8 October 2026

This paper should not be reported as representing the views of the European Central Bank (ECB) or the Eurosystem. The views expressed are those of the author and do not necessarily reflect those of the ECB or the Eurosystem.

Abstract

Two competent researchers analysing the same data can write very different headlines: news improves prediction, or no improvement is detected. We examine how defensible choices produce these findings and what each permits the researcher to claim. Using GDELT news indicators and Eurostat government-bond yields, we compare forecasts of volatility in a euro-area yield differential. Matched linear forecasts yield favourable evidence for news levels, but not after subtracting recent averages. Adding prespecified slower market summaries reduces the twenty-session forecast-error reduction from about 13.5% to 6.9%; no comparison against the expanded reference remains significant after adjustment for multiple comparisons. Controlled simulations examine interpretation when information content is known. A reporting rule can credit uninformative news with market-driven gains. With the stronger added signal and the same expanded reference, 58.0% of datasets yield detection with levels but not with deviations. This paired disagreement shows why a negative result needs a check of what the procedure detects when information is present. We connect matched comparisons and justified alternatives to controls with and without additional information in a claim-comparison protocol. The method helps researchers distinguish a supported conditional gain, dependence on specification choices and an unresolved informational question.

Keywords: specification choices; predictive information; forecast evaluation; research process; model risk; text as data

JEL codes: C12; C18; C52; C53.

SSRN: <https://papers.ssrn.com/abstract=7618698>

Harvard Dataverse DOI: <https://doi.org/10.7910/DVN/J7JWXM>

Non-technical summary

Two researchers study the same news data. One writes: “News improves prediction.” The other writes: “We find no improvement from adding news.” Both can have defensible reasons for their methods and calculate their results correctly. This paper examines how such different headlines arise and how to keep the conclusion within what the evidence supports.

Our main application asks whether news helps forecast fluctuations in a euro-area government-bond yield differential. We combine indicators of news topics and tone from GDELT with interest-rate data from Eurostat, covering March 2015 to August 2026. We compare forecasts using past market movements with forecasts that also use news indicators. The central exercise uses the same estimation method and observations for both, so adding news is the deliberate difference. A companion monthly exercise covers ten countries.

The first contrast concerns what counts as signal. One researcher retains the news indicators as measured; the other subtracts their recent averages to focus on departures from the background. Levels improve longer-horizon forecasts in the later evaluation period, while the transformed inputs provide no detected improvement. Before disagreeing about the finding, the researchers have chosen different parts of the data for their forecasts to use.

The second contrast concerns how much the reference forecast already knows. Adding slower market summaries reduces the twenty-session news gain from approximately 13.5% to 6.9% on identical samples. The remaining gain does not pass the statistical adjustment for the number of comparisons. It becomes smaller and less conclusive; this does not establish zero benefit. Improvements over simpler references likewise cannot establish that news, rather than better use of market data, produced the gain.

Artificial datasets let us examine these interpretations when we know whether the news variables contain additional information. One experiment shows how a reporting rule can credit uninformative news with market-driven gains. Another supplies a persistent signal beyond observed market information. Under the stronger signal, 58.0% of the same datasets yield detection with levels but not with deviations from recent averages, against the same expanded reference. The signal exists, yet one procedure does not detect it. These frequencies describe the experiment, not how often researchers reach misleading conclusions.

Our method asks researchers to state the question, compare procedures on equal terms, justify consequential alternatives and examine controlled examples with and without additional information. They report a supported conditional gain, dependence on methodological choices, or insufficient evidence to decide. A claim of negligible benefit requires a meaningful bound and uncertainty analysis; a nonsignificant result alone is insufficient. The aim is to show how much a conclusion depends on the decisions that made it possible. A useful finding may need a narrower interpretation; an unresolved question can also be an appropriate outcome.

1. Introduction

“News improves prediction” and “We find no improvement from adding news” are very different headlines. Could competent researchers working with the same data, acting in good faith and making defensible methodological choices produce evidence for each? Our application and controlled experiments show concrete ways in which they can. The question is what each analysis permits its author to conclude about the information source.

Our main example asks whether news helps predict how much a euro-area government-bond yield differential will fluctuate over the next five, ten or twenty market sessions. The differential is the broad government-bond composite’s yield minus that of an AAA-rated composite; its volatility measures the size of movements in that gap. We combine Eurostat yield data with sixteen indicators of news topics and tone from GDELT, a database of machine-coded news metadata, over March 2015–August 2026. The forecasting models take these numerical news summaries as inputs. A separate monthly panel of ten countries provides complementary evidence on changes in yield spreads over Germany.

The central comparison is between a forecast using past market movements and the same forecasting method with news indicators added. The market-only forecast is the benchmark: the reference against which an improvement is measured. The two procedures use the same observations, estimation method and updating schedule. We then examine two substantive choices: whether to retain the levels of the news indicators or subtract their recent averages, and whether the market reference includes additional slower movements. Controlled simulations complement these real-data comparisons by creating datasets in which the extra information is known to be absent or present.

Consider two choices made for reasons that do not depend on their results. One researcher retains the measured levels of the news indicators because persistent coverage may contain useful information. Another subtracts a trailing average, a transformation called detrending, to focus on departures from the recent background. Before disagreeing about the finding, they have decided differently which part of the data their forecast should use. Their conditional results can both be correct. The problem arises when either result becomes a general verdict on whether news contains useful information. A p-value crossing a threshold alone does not establish a difference in underlying effects.

The matched comparison gives this contrast empirical content. In the later evaluation period, January 2025–August 2026, news levels improve longer-horizon forecast accuracy, with four of twenty-four comparisons significant after Holm adjustment, which allows for testing multiple comparisons; detrended news does not. The positive interpretation then faces a second question: how much does the market reference already know? Adding market summaries over 60, 120 and 250 sessions, fixed before that calculation, reduces the twenty-session forecast-error reduction from about 13.5% to 6.9% on the same observations. No expanded-reference comparison survives adjustment. The gain is attenuated, while its remaining magnitude is uncertain.

Controlled experiments allow the negative interpretation to be challenged as well, because the information content is known. With the stronger added persistent signal and the same expanded market reference, 58.0% of the same datasets yield detection with levels but not with detrended inputs. This is a paired disagreement: the signal is present in each dataset, yet the two procedures

reach different detection decisions. The positive control therefore asks a practical question about a negative finding: what does this procedure detect when additional information is deliberately supplied? A separate experiment shows how a positive reporting rule can credit uninformative text with genuine market-driven gains.

The broader specification exercise supplies a related illustration. The Superior Predictive Ability (SPA) test asks whether any comparison in a declared set shows an improvement, allowing for the search across that set. In the later daily evaluation, its p-value changes from 0.0002 when the set includes both a historical-average reference and a market-history reference to 0.6934 when only the latter remains. Both the question tested and the comparison with the largest test statistic change. This is neither a paradox nor a defect in the test: improvement somewhere in the broader set does not establish improvement beyond the market-history reference.

The study builds on several established literatures. Specification sensitivity has a long history in econometrics (Leamer, 1983). Researcher degrees of freedom and data-contingent decisions can affect false-positive findings even without deliberate searching (Simmons, Nelson and Simonsohn, 2011; Gelman and Loken, 2013). Multiverse and specification-curve analyses expose variation across reasonable choices, while evidence across research teams shows another source of variation (Steege et al., 2016; Simonsohn, Simmons and Nelson, 2020; Menkveld et al., 2024). Multiple testing is also central to financial discoveries (Harvey, Liu and Zhu, 2016).

Forecast evaluation supplies the relevant statistical framework. Diebold and Mariano (1995) compare predictive accuracy; Giacomini and White (2006) connect evaluation to forecasting methods and estimation uncertainty. White (2000), Hansen (2005) and Romano and Wolf (2005) address searching across alternatives. Hansen and Timmermann (2012) and Rossi and Inoue (2012) show why the estimation/evaluation split and window choices also require attention. These contributions motivate the comparisons and qualifications here; citing them does not make every asymptotic result applicable to our expanding, overlapping forecasts.

Our contribution is to connect these two interpretive vulnerabilities in one empirical and controlled design. The matched and expanded comparisons discipline the attribution of a positive gain; the paired signal experiments show why interpreting a nonrejection also requires evidence about what a procedure can detect. The claim–comparison protocol turns that connection into a practical method: state the informational question, pair procedures, examine justified alternatives, use controls with and without added information, and limit the conclusion to the evidence obtained. It combines established tools around the interpretation of both positive and negative findings. The empirical extensions use already inspected data; Section 4.4 records the timing of their design and execution. Sections 2–3 define the claims and method; Section 4 describes the designs; Section 5 presents the evidence; Section 6 develops implications for research and validation.

2. From a forecast comparison to a discovery claim

2.1. What is being compared?

A forecast comparison needs a target (what is predicted), a horizon (how far ahead), a benchmark (the reference forecast) and a loss function (how errors are scored). A specification combines these

with the inputs, fitting method and evaluation observations. A comparison family is the set of such comparisons considered together. Equation (1) expresses the basic calculation: the reference forecast's error score minus the candidate forecast's error score.

Let s denote a forecasting specification, including its target, horizon h_s , information set, estimation procedure, benchmark, loss function and evaluation sample. At forecast origin t , let $f_{s,t}$ denote the candidate forecast and $b_{s,t}$ the benchmark forecast of $y_{s,t+h_s}$. Define the loss differential as

$$d_{s,t} = L_s(y_{s,t+h_s}, b_{s,t}) - L_s(y_{s,t+h_s}, f_{s,t}). \quad (1)$$

A positive differential favours the candidate. Its sample mean describes the realised improvement over the evaluated observations. Inference about an expected gain additionally depends on the forecasting procedures, temporal dependence and sampling assumptions. Neither a positive mean nor a small p-value identifies which component of the candidate generated that gain.

The benchmark is part of the quantity being evaluated. Improving on a forecast with minimal information and improving on a procedure that uses market history are distinct results. Both comparisons can be informative. Neither the simplicity of the first nor the stronger realised performance of the second is sufficient, by itself, to decide which should define the research question. Their roles should follow the intended claim and information available, rather than whichever reference produces the preferred conclusion after evaluation.

The same discipline applies to the target and loss. Different horizons concern different future outcomes, and different losses give different weight to forecast errors. Changing the sample may change the population to which the result applies. An effect that varies across these comparisons need not be contradictory; it may be conditional on a substantively relevant choice.

Table 1. Claims and the comparisons that support them

Claim	Comparison and evidence required	Scope in this study
Improvement over a reference	Aligned candidate and reference losses, with uncertainty and selection addressed	Evaluated for each named reference, horizon and window
Incremental contribution of text	A comparison isolating text from other changes in inputs and estimation	Procedure-level contribution isolated by pairing; broader informational interpretation challenged by expanded market controls
Performance across specifications	A justified set of alternatives and reported variation in effects and uncertainty	Examined within the documented enumeration
Usefulness for a decision	A decision rule and evaluation of benefits, costs and constraints	Not evaluated by the forecasting losses

Notes: These are distinct claims. A useful result about one comparison need not establish all four. Broadening the conclusion requires evidence for the broader claim.

2.2. Positive and negative overinterpretation

A statistical test starts from a null hypothesis, here that a specified procedure offers no expected predictive gain. Its p-value measures how incompatible the observed test statistic is with that null

under the test’s assumptions; it is not the probability that the null is true. A statistical false positive concerns the truth of this specified null. For a fixed comparison with expected loss gain μ_s , rejecting a true null of no gain is a false rejection. Searching across comparisons, or allowing their definition to depend on outcomes, can change the error probability of the procedure that ultimately reports a discovery. A nominal threshold attached only to the selected comparison need not describe that procedure.

An unsupported attribution is different. A researcher may correctly reject no improvement over a simple benchmark, yet attribute the improvement to text even though the candidate also changes the treatment of market inputs. The statistical finding can be genuine for the stated comparison while the attribution remains unidentified. Similarly, beating a historical-average forecast does not establish improvement beyond a forecast using recent market movements. No assumption that the first finding is false is needed to identify this logical gap.

In the observed data, we do not know the true process generating the outcomes and cannot identify true population nulls. The simulation supplies a known no-additional-information condition and evaluates a stated rule that maps evidence into an attribution claim. Its false-attribution frequency is distinct from test size, the probability of rejecting the equal-predictive-loss null when that null is true: gains due to market dynamics may be real even when crediting text is wrong.

A negative overinterpretation can arise without a false rejection. Failure to reject no predictive gain is not a finding that additional information is absent. The procedure may estimate a useful signal imprecisely, or its representation may fail to preserve the predictive component. Conversely, genuine information need not improve realised out-of-sample loss for an estimated forecasting method. We therefore distinguish known information content, the expected performance of a procedure and its finite-sample declaration.

In positive-control simulations, where the data generator deliberately supplies additional information, a nonrejection is called missed detection only relative to the information known to exist in the generator. It is not labelled a Type II error—a failure to reject a false null—of an equal-predictive-loss test without separately establishing that test’s alternative for the fitted procedure. A broad negative claim about the source requires evidence beyond the reported p-value, just as a broad positive attribution does.

2.3. What a joint rejection establishes

A joint test examines a collection at once. Here the null says that none of its comparisons improves expected forecast accuracy. Rejecting that null supports “at least one improves”, not “all improve” or “news caused the improvement”.

For a declared comparison set S , write the joint null as

$$H_0(S): \mu_s \leq 0 \text{ for every } s \in S. \quad (2)$$

Subject to the test’s assumptions, rejection provides evidence that at least one comparison in the set improves. If every member shares a reference and task, the existential statement concerns that reference and task. If the set mixes references, the conclusion is correspondingly broader: some

included procedure improves on some included reference. It does not imply improvement on every reference or identify the information source responsible.

This distinction survives any valid adjustment for selection. A method can account for searching within S while correctly rejecting because of comparisons against a simple benchmark. Its statistical protection does not transform that rejection into evidence against a more informative benchmark. Restricting S changes both the null and the candidates that can attain the observed maximum. The resulting p-value must be interpreted under that changed question.

A model confidence set (MCS) addresses another question: which members of a declared collection remain statistically unresolved as candidates for best expected performance under a stated loss? Retaining a model does not establish that it uniquely improves on the reference, and retaining two models does not establish their equivalence. Marginal comparisons, joint tests and model confidence sets therefore provide complementary descriptions of evidence. They answer different questions about the collection.

3. Reasonable choices and the risk of overinterpretation

3.1. Decisions that shape the reported result

Choices determine what is measured, how forecasts are constructed, which comparison is evaluated and which result is reported. A plausible variable label can lack measurement support; a motivated horizon can become the headline because it looks most favourable; a joint rejection can be read as identifying a component that changes together with the estimation method. Describing the selected procedure does not reveal what other relevant decisions would have changed.

Table 2. Routes from reasonable choices to unsupported conclusions

Route	Where interpretation can exceed evidence	Appropriate response
Measurement	A plausible variable label is treated as a validated construct	State construction, coverage, timing and validation limits
Benchmark choice	A gain over a simple reference becomes a claim of added information beyond market history	Report references that correspond to each claim and preserve their distinct roles
Transformation	Removing persistent variation yields no improvement and is taken to rule out information in the original source	Evaluate substantively justified representations and state what each preserves
Negative inference	A nonsignificant gain becomes evidence of zero or negligible usefulness	Report uncertainty and assess sensitivity; a negligible-effect claim requires a prespecified meaningful bound
Procedure changes	Gains from a combined change in inputs and estimation are attributed to text alone	Use matched comparisons for attribution or limit the claim to complete procedures
Selection	A favourable horizon or specification is treated as if uniquely fixed beforehand	Document the examined family and address selection within its stated scope

Route	Where interpretation can exceed evidence	Appropriate response
Joint inference	Rejection somewhere in a mixed set is transferred to a narrower benchmark claim	State the joint null and identify the comparisons behind the maximum
Sensitivity analysis	Several tests become either a search for approval or a demand for unanimity	Give each calculation a question and distinguish the principal analysis from sensitivities

Notes: These routes organise distinct evidential problems; they are not cumulative acceptance requirements.

3.2. A proportional evidential standard

The principal question should determine the principal comparison. Sensitivities then examine justified alternatives, rather than accumulate hurdles until nothing survives. Improvement at one horizon need not hold at every horizon, and a loss chosen for one decision need not agree with all other losses. Conversely, a well-motivated alternative that changes the conclusion deserves explanation. The relevance of an alternative cannot be decided by whether it helps significance.

Reasonable choices also differ from errors. An incorrect target construction or an inferential calculation without its required conditions calls for correction or qualification. A legitimate change in benchmark may instead change the question. Treating both as generic robustness failures obscures the distinction. A favourable conditional result should be retained at the level supported by its comparison.

3.3. What inferential adjustment can establish

Examining many comparisons creates more opportunities for a chance finding. Multiple-testing adjustments address that problem within a declared family. Magnitude, direction and uncertainty still precede significance counts. Holm (1979) controls familywise error—the probability of at least one false rejection in the family—when the individual p-values are valid. Benjamini–Hochberg (BH) instead targets the expected proportion of false rejections among rejections, under its dependence conditions (Benjamini and Hochberg, 1995). Romano–Wolf uses the joint dependence structure to identify individual superior comparisons (Romano and Wolf, 2005). These methods address selection within a specified family. They do not identify which input generated a complete procedure's gain. We use Holm and report the other relevant families explicitly rather than introduce another acceptance test.

A nested comparison is one in which a simpler model can be obtained by restricting a larger one. Clark–West adjusts forecast-loss differences for a particular estimation effect in such comparisons and requires a separate interpretation (Clark and West, 2007). The empirical complete procedures alter estimation as well as inputs, so the nesting required for that interpretation is not established. Their adjusted differentials remain auxiliary diagnostics. Even where nesting is valid, an adjusted differential and realised forecast loss are different quantities. The matched comparison therefore reports realised loss differences and root mean squared error (RMSE) ratios directly. RMSE is the square root of the average squared forecast error. Dividing the candidate's RMSE by the reference's gives a ratio below one when the candidate is more accurate; 0.90 means a 10% reduction in that

error measure. Significance thresholds do not replace substantive interpretation (American Statistical Association, 2016).

3.4. Timing and the limits of the evidence

A decision record distinguishes choices made before evaluation from those made after inspecting outcomes. It cannot make a retrospectively chosen procedure prospective or establish the correctness of its assumptions. The observed application does not identify a population false-positive rate. The simulations evaluate specified declaration rules under known information conditions; they neither measure researchers' intentions or behaviour nor provide a comprehensive size-and-power calibration. Their frequencies belong to the stated generators and procedures. Non-rejection is not evidence of equivalence or absence of useful information.

These limits apply throughout the paper. Country-month observations, overlapping forecast horizons and enumerated specifications are not independent replications. Monte Carlo simulation repeats an experiment on independently generated artificial datasets. Only those datasets are replications for the simulated declaration frequencies; trying several models on one dataset does not create several independent experiments. Appendix B states the temporal dependence calculations, and Appendix D distinguishes numerical reproduction from scientific validation.

3.5. A method for linking conclusions to comparisons

The practical task is to move from a plausible headline to a claim that its comparison can support. The following five steps ask what would have to remain true for that interpretation to survive reasonable changes in the analysis.

The protocol begins with an explicit claim and a record of its scope. The researcher names the target, horizon, information available at the forecast origin, relevant market reference, loss and intended interpretation. Improving one implementable forecast is different from showing information beyond an available market history. The record labels choices made before evaluation and decisions made after inspection.

The second step pairs the comparison to that claim. When the question concerns text, the procedures share estimation, training and evaluation rows, timing and updating rules. The researcher also states which available market information the reference actually represents. Pairing removes simultaneous procedural changes; it does not make an incomplete reference informationally sufficient. A limited market expansion justified by the substantive mechanism can test that vulnerability without selecting the strongest realised benchmark after the fact.

The third step fixes a bounded set of alternatives that could change interpretation. Each branch receives a reason unrelated to its realised significance. Here levels and trailing deviations distinguish a persistent state from departures relative to recent background, while the slow market controls examine an omitted-dynamics explanation. Results retain magnitudes and uncertainty across the whole family. No branch is added, dropped or promoted solely because it supplies the desired positive or negative headline.

The fourth step checks the process in both directions. Where feasible, a controlled generator first makes the source redundant conditional on all available market information, then supplies additional

information under a stated alternative. The first control asks whether a reporting rule credits the source with a gain generated elsewhere. The positive control asks what the procedure detects when relevant information is present. Applying the alternative representations to the same simulated datasets makes disagreement directly observable. A redundant proxy can still improve an incomplete benchmark; a supplied signal can still be missed after transformation or estimation. These checks diagnose mechanisms and sensitivity in the chosen scenarios, without identifying the empirical data-generating process.

Finally, the researcher writes the conclusion at the level reached by the evidence. A supported gain names the procedure, reference, representation and period. A materially changing result is reported as specification-dependent, and an unresolved gain remains unresolved. A claim of negligible predictive benefit requires a prespecified economically meaningful margin and an uncertainty bound for the relevant performance quantity; we do not supply such an equivalence test here. Neither universal source informativeness nor universal absence follows from this finite family. The resulting record is a claim, a matched comparison, justified alternatives, controlled diagnostics and an explicit limit on interpretation. This is a reporting and design method, not a mechanically validated acceptance rule.

4. Data, forecasting models and controlled experiments

The daily application measures fluctuations in the yield differential between a broad euro-area government-bond composite and an AAA-rated composite. The text measures relate to established work on news-based financial and economic measurement (Tetlock, 2007; Baker, Bloom and Davis, 2016).

Three tasks form the central argument: comparing complete forecasting procedures, isolating the addition of news through matched and expanded-market comparisons, and testing interpretations in simulations with known information content. A companion monthly panel of spread changes in ten countries supplies complementary evidence and a simple illustration of directional accuracy. For consistency with the tables and replication files, the monthly panel is called Block A and the daily series Block B.

4.1. Data, targets, and information timing

Text variables are constructed from a fixed, archived extraction of GDELT’s Global Knowledge Graph (GKG) 2.0. This database provides machine-generated theme, tone and location information, including processing of translated news (GDELT Project, 2015). The main extraction contains 153,034 aggregate records covering March 2015 to August 2026. These are aggregated cells rather than distinct articles. Country attribution and aggregation differ between the two blocks.

Block A attributes news by subject, using the countries mentioned in location fields. It covers Austria, Belgium, Finland, France, Greece, Ireland, Italy, the Netherlands, Portugal and Spain. Germany supplies the reference yield and is not a panel unit. An article mentioning several countries can contribute to several country cells. Block B instead aggregates news attributed by publishing source across the outlet sample and access categories. Its features describe coverage by those sources rather than the same subject-attributed population used in A. The two blocks therefore describe different news populations.

The sixteen news inputs describe the subject matter and tone of coverage. Nine thematic shares concern sovereign debt, fiscal policy, austerity and adjustment, banking, financial markets, central banks, inflation and prices, elections, and protest or conflict. The remaining inputs are a general-economy theme share, four tone summaries, and mean negative and positive word shares. The tone summaries capture the mean, standard deviation, interquartile range and a quantile-based asymmetry measure. These are numerical features derived from news metadata, not assessments that individual articles are economically informative. Within each aggregation cell, thematic counts are divided by the article count; multiple themes can occur in the same article. Daily cells are aggregated using article-volume weights. The main transformation subtracts a lagged trailing average, over twelve months in A and sixty calendar observations of the daily text series in B. The level variants retain the corresponding unadjusted variables. Article counts and a novelty measure are present in the analytical data but are not selected by the forecasting routines, which use the sixteen columns carrying the transformed-feature suffix.

Measurement checks are part of the design. The study records document verification of individual theme codes, substantial variation in corpus coverage and the withdrawal of an unsupported accessibility interpretation. A nonempty thematic group does not verify each code within it, and normalising article counts does not establish stable source composition. Appendix A reports these checks and distinguishes what the frozen inputs reproduce from what the historical records document.

Yield data come from Eurostat. Block A uses the monthly long-term convergence-criterion yield series, `irt_lt_mcby_m`, and constructs each country's yield minus Germany's yield, expressed in basis points (one basis point is 0.01 percentage point). Let $s_{i,t}$ denote this spread. The modelled target is

$$y_{i,t,h} = \frac{s_{i,t+h} - s_{i,t}}{\sigma_{i,t}}, h \in \{1, 3\} \text{ months.} \quad (3)$$

Standardisation expresses each country's spread change relative to its own recent variability. The scale $\sigma_{i,t}$ is the rolling standard deviation of monthly spread changes over twenty-four months, with at least twelve observations, lagged one month. The alternative basis-point evaluation multiplies the standardised predictions by this known scale; it does not fit a separate model to an unstandardised target. The complete analytical panel contains 1,380 country-month rows, including observations used to construct lags and scales.

Block B uses Eurostat's daily ten-year spot yield curves, `irt_euryld_d`. Its premium p_t is the difference between the euro-area government-bond composite and the AAA composite, expressed in basis points. The volatility proxy combines the squared daily changes over the future horizon; larger values mean larger movements. Its formula is

$$RV_{t,h} = \sqrt{\frac{252}{h} \sum_{k=1}^h (\Delta p_{t+k})^2}, h \in \{5, 10, 20\} \text{ sessions.} \quad (4)$$

The models forecast $\log RV_{t,h}$, the logarithm of this volatility measure. This is an annualised root of future squared daily premium changes, not an intraday realised-variance measure. The analytical daily series contains 2,919 rows. Squared error in these log-volatility forecasts is the primary

descriptive loss. The specification exercise also evaluates quasi-likelihood loss (QLIKE) on the corresponding variance proxy, an alternative scoring rule that weights forecast errors differently. For $u = \log RV - \text{forecast}(\log RV)$, the implemented losses are u^2 and $\exp(2u) - 2u - 1$. Reported ratios are RMSE ratios for squared error and mean-loss ratios for QLIKE; their magnitudes are not pooled into a single effect measure.

Forecasts are constructed with lagged inputs, and model fitting excludes target outcomes that would not yet be available at the forecast date. In the monthly panel, month- t text is aligned with the change from t to $t+h$. In the daily series, text inputs are shifted by one market session before forecasting subsequent premium changes. Training rows are restricted by target availability: monthly origins no later than $t-h$ are eligible in A; the daily implementation uses rows strictly before index $i-h$ at forecast index i . These cut-offs concern the realised targets used for outer estimation. They do not certify historical publication times or database vintages, which are not reconstructed in this study.

The daily feature builder marks an interval in June–July 2025 as missing rather than interpolating it. Evaluation uses complete observations. The number of usable forecasts is consequently smaller than the calendar span, and the resulting gaps require attention when calculating temporal uncertainty.

4.2. Forecasting procedures and benchmarks

A benchmark supplies the reference forecast; a candidate is the procedure evaluated against it. The benchmarks here range from a simple no-change or historical-average rule to linear models using market history. The candidates combine market and news inputs using dimension reduction, penalised regression, decision trees or an average of forecasts. Their short labels (B0, B2, B3 and M1–M4) are identifiers for tables and code, not rankings of quality. In particular, B0 denotes different simple rules for the monthly and daily targets.

Block A compares a zero-change forecast, B0, with a linear market benchmark, B2. B2 includes country indicators, two lags of the standardised spread change, the lagged German yield level and change, lagged own-country spread volatility, and a lagged leave-one-out cross-country factor. These controls define a market-information comparison; they do not establish B2 as the strongest feasible procedure.

For the daily series, the simplest reference (B0) predicts the training-sample mean of log volatility, updated whenever the model is refitted. The market-history reference (B3) is a heterogeneous autoregressive (HAR) regression: a linear forecast using recent movements at daily, weekly and monthly scales. Its inputs are the logarithm of the lagged absolute premium change and the logarithms of its five- and twenty-two-session averages. The use of components over daily, weekly and monthly horizons is motivated by Corsi (2009). The code label HAR-RV refers to this implementation on daily premium changes, not to a benchmark using intraday realised variance. We report B0 and B3 because minimal-information and market-history references answer different questions. B3's stronger realised performance is evidence about this application, not the general rule by which every reference should be selected.

The dimension-reduction model (M1) first uses principal component analysis (PCA) to compress the combined market and news inputs into a few linear summaries, then regresses the target on those

summaries. It includes country indicators in the monthly panel. Component counts are fixed at five in the monthly panel and three in the daily series, subject to dimensional limits.

The penalised-regression model (M2) applies elastic net to the combined inputs. This method penalises coefficient magnitudes to limit complexity and can set some coefficients to zero. The tree-based model (M3), used only in the daily series, is histogram gradient boosting: a sequence of decision trees fitted to improve the forecast. Its maximum depth is three, with 150 iterations and learning rate 0.05. The combination model (M4) equally averages the market-history and candidate forecasts: B2, M1 and M2 in the monthly panel, and B3, M1, M2 and M3 in the daily series.

At each forecast date, the training sample contains observations eligible for fitting; the evaluation observations are those on which forecast errors are assessed. Refitting means estimating the model again as this sample grows. Predictor scaling and PCA are fitted within each outer training sample. Monthly models are refitted at each origin; daily models use a twenty-one-session schedule, with sixty-three-session refits as an alternative. The schedules operate over indexed evaluation origins, skipping unavailable observations. Main results use detrended features and, in A, all ten countries and the standardised target.

For elastic net, an inner validation exercise chooses the penalty settings within the outer training sample. The software implementation, ElasticNetCV, uses three ordinary folds, l1_ratio values of 0.1, 0.5 and 0.9, and twenty penalty values. The observations are partitioned into three subsets, or folds. An integer cv selects ordinary KFold cross-validation for this estimator (scikit-learn, version 1.8.0). Scaling occurs on the outer training sample before inner tuning. The inner validation therefore does not represent a sequence of prospective folds with separate scaling, and overlapping targets are not explicitly purged between them. This limits the interpretation of tuning; the outer target-availability restriction remains separate and is not by itself violated by that inner procedure.

M1–M4 change estimation along with the input set. Their complete-procedure comparisons therefore do not isolate text. The paired linear extension in Section 4.5 holds the estimation procedure fixed and changes the text input set.

4.3. Evaluation windows and specification enumeration

Development ends in December 2020. Evaluation covers January 2021–December 2024, labelled pseudo-out-of-sample (pseudo-OOS), and January 2025–August 2026, labelled holdout. Out-of-sample means that each forecast is evaluated on an observation not used to fit that forecast. Training expands over time; later forecast dates may use earlier observations whose targets have become available. The holdout label identifies the second evaluation window. It does not imply that it remained unseen when all analyses reported here were chosen.

Table 3. Forecasting design and temporal evaluation units

Dimension	Block A	Block B
Analytical structure	Ten-country monthly panel	Daily euro-area aggregate
Target	Standardised spread change; basis-point evaluation variant	Log annualised volatility proxy of daily premium changes
Horizons	1 and 3 months	5, 10 and 20 market sessions
Reference models	B0 zero change; B2 linear market controls and country indicators	B0 mean log volatility; B3 implemented HAR-style regression
Candidate procedures	M1 PCA; M2 elastic net; M4 combination	M1 PCA; M2 elastic net; M3 boosting; M4 combination
Pseudo-OOS evaluation	48 months per horizon	1,010 observed sessions per horizon
Holdout evaluation	19 months at h=1; 17 at h=3	405, 400 and 390 observed sessions at h=5, 10 and 20
Unit of marginal inference	Mean loss difference across countries in each month	Session-level loss difference on the market-date grid

Notes: The monthly inference concerns the fixed panel or the specified panel excluding Greece. Country-month rows are not treated as independent temporal replications. Horizons overlap; neither the daily row counts nor the monthly counts equal a demonstrated number of independent observations. Holdout B spans 418, 413 and 403 market-grid positions, respectively, because some evaluation observations are missing.

The documented enumeration contains 192 comparisons in A: two feature transformations, two country samples, two evaluation windows, two target scales, two references, two horizons and three candidate procedures. There are 96 comparisons within each window. The B enumeration contains 384: two transformations, two refit schedules, two windows, two references, three horizons, four candidates and two losses, giving 192 within each window. These totals count reported comparisons, not independent model fits or independent experiments.

4.4. Inference and the timing of analytical choices

The principal accuracy comparison uses realised loss differences without the Clark–West correction. Consecutive forecast errors can be related: for example, two forecasts of volatility over the next twenty days share many of the same future days. The uncertainty calculation therefore allows for dependence over time rather than treating every forecast as independent. In A, differences are averaged across countries within month before inference, so country-month rows are not treated as independent temporal replications. In B, session-level differences retain their position on the market-date grid. Heteroskedasticity and autocorrelation consistent (HAC) variances allow for unequal variability and correlation over time. The implementation uses Bartlett weights, the bandwidth rule in Appendix B, a finite-sample variance correction and a one-sided normal approximation for individual comparisons. The bandwidth specifies how far apart observations can be in the covariance calculation. Wider bandwidths and the Harvey–Leybourne–Newbold (HLN) finite-sample factor with a t approximation are sensitivities.

Holm adjustment by block and evaluation window is the conservative reference for the enumerated comparisons. BH by design family and alternative family definitions are reported separately. Design

families contain twenty-four comparisons in A and forty-eight in B; block/window families contain ninety-six and 192. Neither separate-window reporting nor these adjustments controls selecting whichever window looks best. BH dependence conditions are not established for this enumeration, and Holm also relies on valid marginal p-values. The inferential qualifications in Section 3.4 apply.

The model confidence set described in Section 2.3 uses a range statistic and sequentially removes the model with the largest adverse studentised pairwise difference; studentisation scales a difference by its estimated uncertainty. Resampling approximates uncertainty by drawing new samples from the observed loss series. Moving-block resampling draws consecutive groups of observations to retain some dependence within each group. The main settings are 5,000 moving-block resamples, three-month blocks in A and twenty-session blocks in B. Alternative block lengths appear in Appendix C. The SPA comparison in Section 5.4 uses the same aligned observations and resamples for its full and restricted sets. Clark–West differentials are retained as auxiliary diagnostics; the nesting needed for an attribution based on them is not established.

The study includes decisions made after results had been inspected, including the specification enumeration and final inferential choices. The inference plan was recorded before executing its calculations, but after forecast results were known. Both the paired extension and the subsequent expanded-market and signal experiments were likewise designed after inspection of the application. Each protocol was fixed locally before its own execution, and all planned results are retained. These records document that sequence, not independent preregistration or a new unseen evaluation. Main settings designate the reference presentation; multiplicity adjustment within each declared family does not account for all earlier analytical discretion.

The replication package regenerates the panels from frozen source inputs, estimates forecasts, computes loss series and inference, and generates the numerical tables and manuscript. Replication follows the declared procedures without adapting them to the reproduced outcomes. Appendix D describes the entry point, checks and remaining limits.

4.5. Isolating the addition of news and testing attribution

The matched empirical comparison asks what changes when news is the only additional input. It fits the market-history HAR regression and the same regression plus the sixteen news variables, both by ordinary least squares (OLS) with an intercept. OLS chooses coefficients to minimise squared errors in the training sample. Predictor standardisation is learned separately inside each training sample. Both members use identical complete rows, target-availability cut-offs, evaluation observations and refit origins. Both fits use singular value decomposition (SVD), a numerical method for solving the least-squares problem. Rank diagnostics, which reveal redundant predictor columns, are retained. Adding text can change fitted market coefficients, but not the estimation method. The question is the performance consequence of supplying those features to this fixed procedure, rather than a structural causal effect of news.

The extension crosses two text transformations, refits every twenty-one or sixty-three indexed origins, three horizons and two evaluation windows. It reports all twenty-four paired comparisons under log squared error, with Holm adjustment across the entire collection, including both windows.

The main display uses twenty-one-session refits and both transformations. The sixty-three-session results and wider-HAC calculations are reported in Appendix C.

The simulation separates information content from observed loss significance. Sixteen artificial news variables follow a stable Gaussian (normally distributed) first-order autoregressive process, AR(1): each value depends on its previous value and a new random disturbance. The autoregressive coefficient is 0.8; each variable has unit stationary variance, with contemporaneous correlation 0.3 between variables. Independent Gaussian market innovations drive a state z according to $z_{t+1} = \rho z_t + \gamma x_{t,1} + \epsilon_{t+1}$. Here ρ measures persistence in the target state and γ determines the contribution of the first news variable. The scenarios are $(\rho, \gamma) = (0.5, 0), (0.9, 0)$ and $(0.9, 0.25)$. Under $\gamma=0$, the entire text history contains no additional information about future z conditional on market history. Under $\gamma=0.25$, the first text variable enters the state transition.

The forecast target is the mean of the next five, ten or twenty state observations. It is a stylised persistent forecasting target, not a reconstruction of the empirical volatility proxy. Market predictors are the current state and its trailing five- and twenty-two-observation means. The augmented OLS adds text levels or deviations from the previous sixty-observation average. After a burn-in of 300 observations, which are generated and discarded to reduce the effect of the initial state, evaluation uses 400 origins beginning at index 800; expanding estimation admits training origins strictly before $t-h$. Refit schedules are twenty-one and sixty-three observations.

Each of 500 replications per scenario produces twelve augmented procedures and twenty-four loss comparisons: twelve against a fitted historical mean and twelve against matched market-only OLS. Four reporting rules declare additional text information on rejection: any nominal comparison in the mixed family; SPA over the mixed family; SPA over the matched family; and Holm over the matched family. The two SPA calculations use the same observations and 999 moving-block draws of length twenty, with common retained-column studentisers and recentering. Monte Carlo p-values use $(r+1)/(B+1)$. Appendix C reports the sampling uncertainty of the declaration frequencies. The complete protocol, seeds and every replication-level record are supplied with the package.

4.6. Challenging the market reference and the treatment of signal

A positive news finding may weaken if the reference forecast already captures more of the market's persistent movements. Conversely, a transformation intended to remove background variation may make persistent information harder for a fitted model to use. The following empirical comparison and second simulation examine these two possibilities.

The empirical follow-up augments the three archived log HAR inputs with the trailing means of log absolute premium changes over 60, 120 and 250 market sessions. The new components average log absolute changes, whereas the original weekly and monthly inputs take the log of average absolute changes. Thus the expansion adds differently constructed slow summaries, rather than varying the averaging horizon alone. Each component requires its full window and is shifted one session. All three enter together, with no choice based on the result. The archived HAR can already approximate long persistence (Corsi, 2009); the extension tests whether these additional summaries change the attribution, not whether the shorter model has literally no memory beyond twenty-two sessions.

Both benchmarks and both text representations are fitted on identical complete rows. Requiring the longest component moves the first eligible training row to 25 February 2016. Re-estimating the short benchmark on those same rows separates the market-input change from that sample restriction. The follow-up crosses both benchmarks with the original two transformations, two refit schedules, three horizons and two evaluation periods. Main Holm adjustment spans all forty-eight comparisons; within-benchmark adjustment over twenty-four is retained as a sensitivity, together with the original and wider calendar-aware HAC calculations. Loss, estimator and target-availability cut-offs are unchanged.

The second simulation separates market information that is already observed from an additional state that can be learned through the artificial news input. It makes the available information set explicit. An observed market state m and an independent state s follow Gaussian AR(1) processes with coefficient 0.98 and unit stationary variance. The target state follows $z_{t+1} = 0.5 z_t + 0.5 m_t + \gamma s_t + \epsilon_{t+1}$, with independent unit-variance Gaussian epsilon. The first text input is $m_t + s_t$; the other fifteen are independent unit-variance AR(1) noise processes with coefficient 0.8. Available market information includes the histories of both z and m . A short HAR uses current z and trailing five- and twenty-two-observation means; the expanded benchmark also observes m_t .

Under $\gamma=0$, text adds only independent s and noise to the observed market history. It therefore has no additional information for future z conditional on that history, even though it can proxy m for the short benchmark. Under $\gamma=0.25$ or 0.50 , text and observed m reveal s_t , which influences the target. This construction avoids inferring a no-information condition from a zero coefficient while allowing text to reveal an otherwise unavailable market state.

Each scenario has 500 independent datasets after a 1,000-observation burn-in, with 800 initial origins and 400 evaluation origins. Targets, horizons, refits, strict training cut-offs, trailing-60 text deviations and expanding OLS follow the earlier experiment. All four benchmark/transformation combinations use the same rows. SPA and Holm are evaluated for each six-comparison path, each benchmark's twelve-comparison family and the full twenty-four-comparison family. The 999 moving-block resamples of length twenty, studentisers and recentering are shared across paths. Parameters and seed were fixed before execution without fitting the generator to the empirical gain. These stylised mean-state targets are not a reconstruction of realised yield-premium volatility. Appendix C reports uncertainty and paired disagreement; Section 3.4 states their inferential scope.

5. Evidence across positive and negative analysis paths

The results answer three questions in sequence. Sections 5.1–5.4 show what comparisons of complete forecasting procedures establish and how the reference changes the finding. Sections 5.5–5.6 isolate the addition of news and examine false attribution in controlled data. Sections 5.7–5.8 challenge the favourable result with slower market controls and the negative interpretation with known signals. Section 5.9 provides the complementary monthly illustration. Throughout, a ratio below one means a lower error measure than the stated reference.

5.1. Accuracy relative to each reference

The daily main specification illustrates the benchmark distinction. Here the simple reference predicts historical average log volatility; the market-history reference uses recent daily, weekly and monthly movements. Tables retain their compact labels B0 and B3, respectively. In holdout, the market-history reference has RMSE ratios of 0.677, 0.673 and 0.694 against the simple reference at five, ten and twenty sessions. Each of the four complete candidate procedures improves on the simple reference at all three horizons. Against the market-history reference, eleven of twelve candidate ratios exceed one; the exception is the elastic-net model at five sessions, with a ratio of 0.998. A historical mean is a limited reference when recent market movements predict the target. The observed advantage of the market-history reference makes that limitation concrete, without making every candidate's improvement automatic.

For this elastic-net comparison, the estimated mean log-squared-loss gain is 0.2867 against the simple reference, with a marginal 95% HAC interval of [0.2049, 0.3685]. Against the market-history reference, it is 0.0011, with an interval of [−0.0095, 0.0117]. These intervals use the stated marginal calculation and do not cover selection of the example. The complete-procedure evidence establishes substantial gains over the simple reference but cannot attribute them to text. The matched extension below directly examines adding text under a fixed linear procedure.

5.2. Unequal consequences of specification choices

Table 4. Point improvements and significance in the specification enumeration

Block	Window	Ref.	N	Point	Nominal	BH design	Holm window
A	Pseudo-OOS	B0	48	1	0	0	0
A	Holdout	B0	48	25	1	0	0
A	Pseudo-OOS	B2	48	36	0	0	0
A	Holdout	B2	48	25	0	0	0
B	Pseudo-OOS	B0	96	67	43	34	8
B	Holdout	B0	96	96	92	91	55
B	Pseudo-OOS	B3	96	42	4	0	0
B	Holdout	B3	96	14	0	0	0

Notes: BH design fixes transformation and country sample in A, and transformation and refit schedule in B, within each window. Holm window includes both references and all comparisons in that block/window. Each row is part of its adjustment family. B includes squared error and QLIKE. Adjustments are applied separately to marginal p-values; effects and intervals remain available at comparison level.

A nominally significant comparison passes the individual 5% test before adjustment for the other comparisons examined. The daily-holdout contrast precedes that multiplicity adjustment: all ninety-six comparisons against the simple reference improve in point estimates and ninety-two are nominally significant; fifty-five remain significant under Holm by block and window. Against the market-history reference, fourteen of ninety-six improve and none reaches 5%; the smallest p-value is 0.131. Adjustment reduces the number of rejections against the simple reference, but the lack of rejections against the market-history reference is already present before adjustment.

Selection matters elsewhere. Four comparisons against the market-history reference in the earlier daily evaluation are nominally significant and none survives BH by design family. The monthly holdout has one nominal rejection against the no-change reference, $p=0.0336$, which also does not survive adjustment. These changes describe the selected comparison's family context, not the truth of its population null.

Choices are not equally consequential. Across both daily windows and losses, 84.9% of comparisons improve on the simple reference versus 29.2% on the market-history reference. The shares are 67.2% for levels and 46.9% for detrended inputs, versus 56.8% and 57.3% for the two refit schedules. These are descriptive proportions in a balanced enumeration, not causal effects or a variance decomposition. A single favourable specification would conceal both the variation and the unequal influence associated with these choices.

5.3. Temporal and inferential sensitivity

Daily point improvements against the market-history reference fall from forty-two of ninety-six in pseudo-OOS to fourteen in holdout; the simple reference counts rise from sixty-seven to ninety-six. The contrast cautions against extrapolating relative performance across periods but does not identify a structural break or its cause. The estimation/evaluation split and window choice are substantive decisions, as emphasised by Hansen and Timmermann (2012) and Rossi and Inoue (2012).

For the simple reference, wider-HAC Holm counts are four and thirty-seven in the two windows, compared with eight and fifty-five under the main calculation. The HLN-factor/t sensitivity gives six and fifty-two. None of these calculations produces Holm rejections against the daily market-history reference or against either reference in the monthly panel. Adjusting separately by reference gives twelve and eighty-five rejections against the simple daily reference, reflecting a smaller family. Appendix C reports the alternatives together; they are not compounded as successive filters.

The daily evaluation includes overlapping targets and gaps. HAC preserves their positions on the market-date grid, whereas MCS and SPA resample consecutive observed rows and compress missing periods. Block-length sensitivity does not remove that limitation. The monthly estimates are especially imprecise, and are reported in Appendix A rather than used to infer absence of predictive content.

5.4. A joint rejection and its changing scope

Table 5. Model confidence sets under the main resampling settings

Window	Block	Horizon	Models retained
Pseudo-OOS	A	1 month	B0, B2, M1, M2, M4
Pseudo-OOS	A	3 months	B0, B2, M1, M2, M4
Pseudo-OOS	B	5 sessions	B3, M2, M3, M4
Pseudo-OOS	B	10 sessions	B3, M2, M3, M4
Pseudo-OOS	B	20 sessions	B0, B3, M2, M3, M4
Holdout	A	1 month	B0, B2, M1, M4
Holdout	A	3 months	B0, B2, M1, M2, M4
Holdout	B	5 sessions	B3, M2, M3
Holdout	B	10 sessions	B3, M2
Holdout	B	20 sessions	B3, M2

Notes: Main blocks are three months in A and twenty observed sessions in B, with 5,000 resamples and seed 20260922. Membership at 90% uses accumulated p-values and $(r+1)/(B+1)$; raw exceedance frequencies are also retained. Membership does not establish equivalence or a universal ranking.

Several procedures remain statistically unresolved in the model confidence sets. The market-history reference is retained in every daily set, including the block-length alternatives. At twenty sessions in holdout, blocks of ten retain the market-history reference alone while blocks of twenty and forty retain the market-history reference and the elastic-net model. In the monthly one-month holdout, the elastic-net model is retained with two-month blocks but not with three or four. Membership describes unresolved relative performance within the collection; it is not equivalence or unique optimality.

The SPA calculation exposes a separate decision: what enters the comparison family. The daily-holdout collection has 192 comparisons across the simple reference and the market-history reference, a maximum studentised statistic of 5.0135 and bootstrap exceedance frequency 0.0002. Restricting it to the ninety-six comparisons against the market-history reference gives 1.0254 and 0.6934. Both use the same 390 common evaluation sessions and 5,000 resamples. The 384-comparison total for the daily exercise includes both evaluation windows; it is not the size of the holdout SPA family.

The full maximum comes from the component-regression model against the simple reference at five sessions with text levels, sixty-three-session refits and squared error. The restricted maximum comes from the elastic-net model against the market-history reference at five sessions with levels, twenty-one-session refits and QLIKE. Retained loss series, studentisers and recentering terms are unchanged. Under the same draws, the restricted bootstrap maximum never exceeds the full maximum; its 95th percentile falls from 2.8386 to 2.7801. The higher restricted p-value follows the changed observed maximum and hypothesis, rather than a wider bootstrap tail confronting the same winner.

A reader could regard the first rejection as confirmation of useful news after allowing for specification search. What it establishes is improvement somewhere in a family that includes the simple reference. It does not establish improvement over the market-history reference or isolate the

contribution of text. A selection-adjusted p-value therefore still needs the membership rule and its substantive rationale. The empirical SPA retains the archived exceedance estimate r/B , whereas simulated SPA and MCS use the finite-resample convention $(r+1)/(B+1)$, where r is the number of exceedances and B the number of resamples. Here 0.0002 is one exceedance in 5,000 draws. Applying the latter convention to these same full and restricted draws gives 0.0004 and 0.6935, respectively, preserving the contrast; the reporting difference is not its cause.

5.5. Adding text under the same forecasting procedure

Table 6. Matched linear forecasts with refits every 21 sessions

Window	Text	h	N	RMSE ratio	Raw p	Holm p
Pseudo-OOS	Detrended	5	1010	1.012	0.8542	1.0000
Pseudo-OOS	Detrended	10	1010	1.022	0.9438	1.0000
Pseudo-OOS	Detrended	20	1010	1.030	0.9576	1.0000
Pseudo-OOS	Levels	5	1010	0.996	0.3712	1.0000
Pseudo-OOS	Levels	10	1010	1.004	0.6082	1.0000
Pseudo-OOS	Levels	20	1010	1.001	0.5339	1.0000
Holdout	Detrended	5	405	0.994	0.3285	1.0000
Holdout	Detrended	10	400	1.006	0.7152	1.0000
Holdout	Detrended	20	390	1.004	0.6137	1.0000
Holdout	Levels	5	405	0.947	0.0130	0.2550
Holdout	Levels	10	400	0.914	0.0024	0.0496
Holdout	Levels	20	390	0.864	0.0013	0.0297

Notes: Ratios compare HAR plus text with market-only HAR under the same OLS procedure and complete observations. Values below one favour adding text. One-sided p-values concern unadjusted log-squared-loss differences under the grid-aware HAC calculation. Holm includes all twenty-four paired comparisons across both windows, transformations, schedules and horizons. Wider-bandwidth results are retained in `matched_summary.csv`. This is a post-inspection analysis.

Here the two headlines become concrete. Retaining news levels supports a favourable longer-horizon holdout finding; using departures from the recent news background provides no detected improvement. The contrast concerns the same source and a matched forecasting procedure, but a different representation of the inputs. Table 6 shows the magnitudes and uncertainty behind each account.

The matched comparison changes the empirical conclusion in a specific way. With detrended inputs and twenty-one-session refits, holdout RMSE ratios are 0.994, 1.006 and 1.004. With levels they are 0.947, 0.914 and 0.864. Thus the level specification improves realised accuracy, particularly at the longer horizons. Its ten- and twenty-session Holm p-values are 0.0496 and 0.0297 after adjustment across all twenty-four comparisons, including both windows.

The sixty-three-session level forecasts show the same pattern: ratios 0.948, 0.915 and 0.862, with Holm p-values 0.0297 and 0.0160 at the longer horizons. All four longer-horizon level comparisons

remain below 5% under the wider-HAC calculation. Neither detrended inputs nor the earlier evaluation window provides a nominal 5% rejection. The two schedules use the same observations and are not independent confirmations. Table C.4 supplies the full alternative-schedule results.

This evidence supports a conditional statement that level features improve holdout forecasts in the specified linear procedure under the reported inference. The timing and selection limits in Section 4.4 apply. Persistence, changing corpus composition and sample-specific associations remain possible explanations for why levels differ from detrended inputs; the exercise does not distinguish them.

The matched result is not inconsistent with the complete-procedure findings. The four candidate procedures process market and text inputs differently from augmented OLS. Their lack of demonstrated improvement over the market-history reference cannot establish that every use of text fails. Equally, favourable augmented OLS performance cannot validate every news-based procedure. Reporting the specific intervention in the input set gives the positive finding its evidential scope.

5.6. Known information content and false attribution

Table 7. Declarations of additional text information in the controlled experiment

Scenario	Nominal mixed	SPA mixed	SPA matched	Holm matched
No signal, rho 0.5	11.2 (56)	1.2 (6)	0.4 (2)	0.2 (1)
No signal, rho 0.9	99.8 (499)	97.2 (486)	0.4 (2)	0.2 (1)
Signal, rho 0.9	100.0 (500)	98.4 (492)	14.8 (74)	9.2 (46)

Notes: Each cell gives the percentage and count of declarations among 500 independent replications. A rule declares additional text information if its specified comparison rejects at 5%. Under $\gamma=0$, declarations falsely attribute information to text; a mixed loss-null rejection need not itself be a statistical Type I error. The signal scenario sets $\gamma=0.25$. Both SPA rules use the same 999 moving-block resamples per replication and $(r+1)/(B+1)$. Monte Carlo uncertainty appears in Table C.5.

When the simulated market state has $\rho=0.5$ and text adds no information, any nominal comparison in the mixed family triggers a text declaration in 11.2% of replications. Mixed SPA reduces that to 1.2%. With $\rho=0.9$ and still no text information, the corresponding frequencies are 99.8% and 97.2%. Search adjustment therefore does not protect the attribution when predictable market dynamics produce the gain that is credited to text. In that persistent setting, the Wilson 95% interval for the mixed-SPA declaration rate is [95.4%, 98.3%].

Matched SPA declares text information in 0.4% and 0.4% of the two no-signal scenarios, while matched Holm gives 0.2% and 0.2%. Each matched-SPA estimate corresponds to two declarations in 500 replications, with a Wilson interval of [0.1%, 1.4%]. These frequencies concern the stated declaration rules under known absence of additional information. They do not establish that the mixed loss-null rejection was statistically false or that the matched tests have uniformly correct size.

The positive-control scenario reveals the other side of the comparison. With $\gamma=0.25$, matched SPA detects the additional information in 14.8% of replications and matched Holm in 9.2%, with Wilson intervals of [12.0%, 18.2%] and [7.0%, 12.1%]. The low detection frequencies rule out treating non-rejection as a reliable diagnosis of no information. Mixed SPA declares in 98.4%, scarcely distinguishing this signal scenario from the persistent no-signal scenario.

The mechanism is an attribution mismatch rather than a newly discovered property of SPA. A mixed comparison asks whether any complete procedure beats any included reference. When the intended claim concerns an added information source, its rejection can answer a broader question. The numerical frequencies quantify that separation in the specified experiment.

5.7. What changes when the market reference includes slower components

Table 8. Holdout text gains against short and expanded market histories

Text	h	Short ratio	Expanded ratio	Short Holm	Expanded Holm
Levels	5	0.943	0.972	0.4235	1.0000
Levels	10	0.910	0.968	0.0793	1.0000
Levels	20	0.865	0.931	0.0360	1.0000
Detrended	5	0.995	0.994	1.0000	1.0000
Detrended	10	1.008	1.011	1.0000	1.0000
Detrended	20	1.008	1.009	1.0000	1.0000

Notes: Refits every 21 sessions. Both benchmarks and transformations use the same complete training and evaluation rows. The expanded model adds lagged means of log absolute market changes over 60, 120 and 250 sessions. Ratios compare each augmented model with its own paired benchmark. Holm covers all 48 follow-up comparisons. Neither ratio is a comparison of the two augmented forecasts against each other.

The added value of news depends also on what the market reference does not yet capture. This comparison asks how much of the favourable level-text result remains when both procedures receive slower summaries of market history.

The common-sample short-benchmark results remain close to the archived matched comparison. For level text and twenty-one-session refits, the holdout RMSE ratios are 0.943, 0.910 and 0.865. Against the expanded benchmark they are 0.972, 0.968 and 0.931. Thus attenuation concerns the estimated magnitude, not just crossing a significance threshold. The sixty-three-session schedule gives a similar pattern. Detrended text does not provide comparable evidence under either reference, and the earlier period supplies no adjusted rejection against the expanded benchmark.

No expanded-benchmark comparison survives Holm across forty-eight comparisons or the within-benchmark family of twenty-four, under either HAC bandwidth. At twenty sessions and refit twenty-one, the raw one-sided p-value is 0.0279, while within-benchmark Holm gives 0.6413. The remaining estimated RMSE reduction is about 6.9%, compared with about 13.5% against the short benchmark. It would therefore be incorrect to describe the gain as zero. Under the common-sample short benchmark, three comparisons survive the main forty-eight-comparison adjustment and four survive within-benchmark adjustment; the original extension's four-of-twenty-four result retains its original definition.

The short and expanded market-only RMSEs at that horizon are 0.4598 and 0.4169; their text-augmented RMSEs are 0.3977 and 0.3883. The expanded market reference closes part of the gap while the augmented forecast also improves. This is consistent with text partly substituting for slower market summaries, but it is not a causal decomposition of the gain and does not identify all

information in market history. The favourable empirical finding is consequently narrower: text levels improve a particular shorter HAR in the later period, while their advantage over the specified expanded reference remains uncertain.

5.8. Known information and disagreement between analysis paths

Table 9. SPA declarations with market-dependent text

Scenario	Short levels	Short detrended	Expanded levels	Expanded detrended
No added information	5.8 (29)	4.8 (24)	0.2 (1)	0.4 (2)
Signal 0.25	48.0 (240)	19.4 (97)	28.6 (143)	4.6 (23)
Signal 0.50	91.2 (456)	32.4 (162)	76.6 (383)	19.2 (96)

Notes: Each cell gives percentage and count among 500 independent datasets. Each route evaluates six comparisons: three horizons and two refit schedules for the indicated benchmark and text transformation. The expanded benchmark observes the omitted market state. All routes share bootstrap draws within a dataset. Separate routes represent alternative analysis paths, not independent confirmations. Joint families across transformations and all 24 comparisons are also reported in the replication files.

The positive control gives “no improvement detected” a direct diagnostic challenge: additional information is present by construction. Evaluating both representations on the same datasets distinguishes the existence of that signal from its detection by each fitted procedure.

When text only summarises observed market information plus independent noise, the level-text SPA path rejects in 5.8% of datasets against the short benchmark and 0.2% against the expanded benchmark. The corresponding detrended rates are 4.8% and 0.4%. The contrast is modest compared with the original mixed-reference experiment and is retained without parameter search. These frequencies describe an attribution rule: a real gain over an incomplete procedure need not imply additional information beyond the available market history.

With the moderate additional signal, level text against the expanded benchmark is detected in 28.6% of datasets, compared with 4.6% for detrended text. With the stronger signal, the rates are 76.6% and 19.2%. The joint event in which levels reject and deviations do not occurs in 25.8% and 58.0% of the same datasets, respectively. Appendix C reports the reverse event and uncertainty through the archived complete summaries.

These are plausible opposing paths rather than two claims that must have identical answers. Retaining levels asks the linear forecaster to use a persistent information state; subtracting recent background focuses it on deviations and can attenuate the usable slow component. The finite feature representation and fitted procedure differ, so the experiment does not claim that a transformation destroys every possible informational recovery from the full text history. A competent analyst choosing the latter path can obtain a negative empirical result even when additional information exists in the generator. Reporting that result as source-wide absence is the unsupported step. Choosing the favourable representation only after comparing its p-value would introduce a separate selection issue, which is why the combined families are retained.

The benchmark expansion and controlled information cases perform complementary checks. The empirical test challenges an attribution using market data actually available in the application. The simulation establishes what can happen when information content is known. Neither determines the true information content of the empirical news measures or guarantees a nominal error rate in other processes.

5.9. Complementary evidence on monthly directional accuracy

The complementary monthly illustration measures the hit rate: the proportion of forecasts that correctly predict whether the spread will rise. For the elastic-net model at three months, the hit rate is 61.0% in pseudo-OOS and 61.8% in holdout: 293 of 480 and 105 of 170 eligible country-month forecasts. Under the same positive-versus-nonpositive convention, always predicting a nonpositive change is correct in 52.3% and 67.6%. Stable accuracy does not establish stable improvement over that simple rule. Appendix A supplies the zero convention, sample limits and the separate within-sample fit illustration.

6. Implications for empirical research and model validation

6.1. Two defensible paths and the conclusions they support

The first analyst's headline is "News improves prediction." The analyst retains levels as a persistent information state and adds them to a familiar short HAR. The matched holdout comparison supports a positive finding. A second analyst's headline is "We find no improvement from adding news." This analyst uses trailing deviations to focus on departures from the recent background. Even under the same short reference, that path supplies no detected gain. Both motivations are intelligible before observing the result; neither needs to be inferred from a preferred outcome.

Each headline nevertheless needs a qualification. For the first analyst, adding slow market summaries reduces the gain and removes adjusted significance against the expanded reference. The defensible statement is that news levels improve the specified shorter HAR in the later period, with weaker evidence beyond the expanded reference. For the second, the controlled signal cases show that non-detection can coexist with additional information in the source. The defensible statement concerns the transformed inputs and fitted procedure; it does not establish that news is uninformative.

These researchers can calculate their results correctly and still tell different stories. Our method makes the dependence of each story on its analytical choices visible. It turns an apparent contradiction about news into conditional findings that can be inspected, while leaving the broader informational question unresolved. The empirical contrasts locate that uncertainty; the controlled cases show why neither a positive nor a negative headline settles it.

The SPA comparison of complete procedures provides a different positive path: correcting search within a family can establish an improvement somewhere in that family while failing to identify the news contribution. Matched and expanded comparisons address attribution; positive controls test how a negative finding should be interpreted when the procedure is confronted with a known signal. Each step resolves a different uncertainty.

6.2. Make consequential decisions inspectable

Alongside the estimate, readers need consequential choices, their reasons and timing, the relevant alternatives examined and what those alternatives change. The benchmark, text transformation, estimation method and comparison-family membership are particularly visible here. Refit frequency has much less association with outcomes in the reported enumeration, so equal reporting space for every possible choice would be unhelpful.

The comparison family should reflect the search relevant to the claim. Omitting searched alternatives can understate selection; combining references can change the null into a broader question. Neither expanding the family indefinitely nor choosing the family with the smallest adjusted p-value solves that problem. A principal analysis and justified sensitivities should be specified, and their implications reported together.

The attenuated level-text estimate warrants a new evaluation with the market reference and representation frozen before outcomes are seen. Reusing the same holdout cannot supply that evidence. Reproducible text counts also do not automatically validate their economic interpretation.

6.3. Validation and reproducibility

These distinctions also matter for model validation. The ECB's guide to internal models asks validation to challenge modelling choices and examine alternatives and out-of-sample performance, particularly for machine learning (European Central Bank, 2025). Model-risk guidance also addresses the relationship between model use, limitations and validation (Board of Governors of the Federal Reserve System, 2026). Our application is not a regulatory-model validation exercise; it illustrates a problem relevant to that task: the evidence used to support a model can concern a narrower comparison than the proposed interpretation or use.

The replication package traces frozen inputs through estimation, paired losses, every simulated declaration and manuscript tables. Reproducing alternatives makes dependence on analytical choices visible. The package does not certify historical data vintages or remove the limitations stated in Sections 3–4.

7. Conclusion

Two researchers can analyse the same data competently and in good faith, yet produce evidence for very different headlines. In this study, “news improves prediction” and “no improvement is detected” can follow from defensible choices about the representation of news and the market reference. The risk lies in extending either conditional finding into a general verdict on the source of information.

The matched empirical comparison supports a favourable longer-horizon holdout result for news levels, whereas deviations from recent averages provide no detected improvement. Adding prespecified slow market summaries reduces the twenty-session forecast-error reduction from about 13.5% to 6.9% on identical samples and removes adjusted significance against the expanded reference. The remaining estimate is positive. The mixed-reference SPA example illustrates a related attribution limit: improvement somewhere in a broader family does not establish the incremental contribution of news.

The controlled experiments sharpen the contrast because information content is known. With the stronger added persistent signal and the same expanded reference, 58.0% of the same datasets yield detection with levels but not with deviations. The source contains additional information, yet its representation changes what the fitted procedure detects. A separate experiment shows how a reporting rule can credit uninformative news with market-driven gains. Together these results motivate controls in both directions: challenge what a positive result is attributed to, and examine what a procedure can detect before interpreting its nonrejection as absence of information. The 58.0% is specific to the reported scenario, not a general frequency of disagreement.

The claim–comparison protocol connects these diagnostics to empirical practice. Researchers state the informational question, pair procedures, justify consequential alternatives, and use controlled cases with and without additional information to assess interpretation. A supported conditional gain, a result dependent on specification choices and an unresolved question are distinct outcomes. The method helps determine which account the evidence supports and what further evidence a broader conclusion would require.

References

- American Statistical Association (2016). American Statistical Association Releases Statement on Statistical Significance and P-Values. 7 March.
<https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf>
- Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593–1636.
<https://doi.org/10.1093/qje/qjw024>
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, 57(1), 289–300.
<https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Board of Governors of the Federal Reserve System (2026). Revised Guidance on Model Risk Management. SR 26-2, 17 April. Supersedes SR 11-7.
<https://www.federalreserve.gov/supervisionreg/srletters/sr2602.htm>
- Clark, T. E., and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.
<https://doi.org/10.1016/j.jeconom.2006.05.023>
- Corsi, F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
<https://doi.org/10.1093/jjfinec/nbp001>
- Diebold, F. X., and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
<https://doi.org/10.1080/07350015.1995.10524599>
- European Central Bank (2025). ECB guide to internal models. July.
https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf
- GDELT Project (2015). GDELT 2.0: Our Global World in Realtime.
<https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>
- Gelman, A., and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Working paper, 14 November.
<https://stat.columbia.edu/~gelman/research/unpublished/forking.pdf>
- Giacomini, R., and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica*, 74(6), 1545–1578.
<https://doi.org/10.1111/j.1468-0262.2006.00718.x>
- Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics*, 23(4), 365–380.
<https://doi.org/10.1198/073500105000000063>
- Hansen, P. R., and Timmermann, A. (2012). Choice of Sample Split in Out-of-Sample Forecast Evaluation. CREATES Research Paper 2012-43, Aarhus University.
<https://pure.au.dk/portal/en/publications/choice-of-sample-split-in-out-of-sample-forecast-evaluation/>

- Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, 79(2), 453–497.
<https://doi.org/10.3982/ECTA5771>
- Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the Cross-Section of Expected Returns. *The Review of Financial Studies*, 29(1), 5–68.
<https://doi.org/10.1093/rfs/hhv059>
- Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.
[https://doi.org/10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4)
- Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
<https://www.jstor.org/stable/4615733>
- Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *The American Economic Review*, 73(1), 31–43.
<https://www.jstor.org/stable/1803924>
- Menkveld, A. J., et al. (2024). Nonstandard Errors. *The Journal of Finance*, 79(3), 2339–2390.
<https://doi.org/10.1111/jofi.13337>
- Newey, W. K., and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708.
<https://doi.org/10.2307/1913610>
- Romano, J. P., and Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica*, 73(4), 1237–1282.
<https://doi.org/10.1111/j.1468-0262.2005.00615.x>
- Rossi, B., and Inoue, A. (2012). Out-of-Sample Forecast Tests Robust to the Choice of Window Size. *Journal of Business & Economic Statistics*, 30(3), 432–453.
<https://doi.org/10.1080/07350015.2012.693850>
- scikit-learn (version 1.8.0). ElasticNetCV documentation. Accessed 5 October 2026; the integer-cv behaviour was also checked in the installed replication environment.
https://scikit-learn.org/1.8/modules/generated/sklearn.linear_model.ElasticNetCV.html
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), 1359–1366.
<https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4, 1208–1214.
<https://doi.org/10.1038/s41562-020-0912-z>
- Steege, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5), 702–712.
<https://doi.org/10.1177/1745691616658637>
- Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3), 1139–1168.
<https://doi.org/10.1111/j.1540-6261.2007.01232.x>

White, H. (2000). A Reality Check for Data Snooping. *Econometrica*, 68(5), 1097–1126.
<https://users.ssc.wisc.edu/~behansen/718/White2000.pdf>

Use of generative AI and AI-assisted technologies

Claude (Anthropic) and ChatGPT (OpenAI) assisted with English-language drafting and restructuring, Python code development and review, data harmonisation, execution of the empirical analysis and controlled simulation, methodological checks, reproducibility checks, and preparation of the manuscript and replication materials. The author directed the research and the substantive choices concerning design, methods and interpretation, and is responsible for the manuscript, code, results and conclusions.

AI assistance is part of the research process and does not constitute independent scientific validation. The timing of analytical decisions and the limitations of the evaluation are described in Section 4.4. Reproduction from the supplied inputs requires no access to AI tools or conversations.

Appendix A. Measurement, procedures and decision records

A.1. Text measurement and coverage

The frozen extraction includes aggregate GDELT cells, theme and tone definitions, source mappings and the Eurostat series used for prices. Subject attribution in A allows an article to contribute to more than one country. Source attribution in B describes the publishing-source sample. The sixteen selected text variables comprise nine theme shares, a general-economy share, four tone summaries and mean negative and positive word shares. Counts and novelty measures are constructed in the data but are not forecasting inputs.

The theme-definition record reports that fourteen of twenty-seven initially proposed codes were not found in its verification exercise; the retained mapping documents checks on GKG observations from 1–7 June 2019. This is evidence about the documented check, not a new exhaustive certification of the coding system. Checking only whether a grouped theme has observations could conceal a missing member code.

Coverage can be checked directly in the frozen extraction. For subject-attributed news across the ten panel countries, mean daily country-article mentions fall by 64.7% between 2016 and 2022. Germany is excluded from this calculation; mentions across countries are not unique articles. This variation cautions against treating the corpus as a fixed measurement population. Shares and detrending address scale and slow variation but do not establish that selection into coverage is stable.

Detrending subtracts lagged trailing averages: twelve months in A and sixty observations of the daily text series in B. The daily inputs are then shifted by one market session. These transformations do not establish that changes in corpus composition have been removed. The missing daily interval from 15 June to 1 July 2025 is not imputed. Historical publication vintages are not reconstructed.

An attempted accessibility interpretation was withdrawn after the study's validation exercise failed to support the predicted discrimination between outlet groups. The associated comparison of open-access and paywalled news was retired, as recorded in amendment H. The decision and its basis are retained in the measurement records. The failure does not establish a different construct, and the accessibility interpretation is not used to support the forecasting conclusions.

A.2. Implemented procedures

The executed feature and model definitions in Section 4 govern replication. M1 uses fixed component counts on the combined predictor matrix. M2 uses ordinary three-fold inner validation and outer-sample scaling. Settings for temporal tuning and component selection retained in the design configuration are not executed by these estimators. This distinction makes the procedures explicit without implying that their unimplemented alternatives have been evaluated.

The monthly basis-point variant rescales the standardised predictions using the available volatility scale; it does not fit a raw-target model. The daily main refit interval is twenty-one sessions and the alternative is sixty-three. Internal identifiers in the code are mapped to the English descriptions in the package dictionary. Labels used for subsets identify their membership and do not confer statistical admissibility on them.

The directional illustration uses the main M2 forecasts at three months in A, with missing realised targets or predictions excluded. Positive means strictly greater than zero; zero counts as nonpositive for both the model and the simple rule. There are ten zero targets in pseudo-OOS and one in holdout. The package records the full confusion counts and eligible months. The always-nonpositive comparator describes the same samples; it is not presented as a prospectively selected strategy.

The OLS illustration uses the already lagged text columns in the daily panel, an intercept and log HAR inputs, with the same complete rows for the base and augmented fits. Each horizon has 1,473 origins between March 2015 and December 2020. The origin-defined window includes five, ten and twenty targets ending after December 2020, respectively. A separate sensitivity requires all targets to be realised by that cutoff and is supplied in `development_fit.csv`. Neither calculation is an unseen forecasting evaluation. The augmented design has twenty columns but rank nineteen, meaning one column is linearly redundant; fitted values and R^2 are reproducible, while individual coefficients are not uniquely identified. No nominal coefficient or F-test claim is drawn from this diagnostic.

A.3. Decision timing

The package retains the study's design document, amendment log, corpus-quality record and evaluation records. Some choices, including specification enumeration and final inference, followed inspection of results. The final inferential plan records its choices before their execution, without claiming prospective status for the study as a whole. The holdout-opening record remains unchanged during deterministic reproduction. These records provide transparency about the documented process; they do not independently certify every historical information set.

A.4. Monthly accuracy and within-sample fit

The monthly main specification is less precise than the daily comparison. At one month, M1, M2 and M4 have slightly higher RMSE than either reference. At three months, all improve on B0 in point estimates, while only M4 improves on B2. Its ratios are 0.953 and 0.992. The corresponding mean standardised-squared-loss gains are 0.1642 and 0.0263, with marginal 95% HAC intervals of $[-0.1725, 0.5009]$ and $[-0.1600, 0.2125]$. Seventeen evaluation months at this horizon leave substantial uncertainty. Table 4 and Tables C.1–C.2 retain the full monthly inference.

R^2 describes the proportion of variation accounted for within the fitted sample. For daily development origins, adding the sixteen text variables to an illustrative OLS with log HAR inputs raises R^2 by 4.69, 4.74 and 4.67 percentage points. The ranks, identical samples and target-realisation sensitivity are specified in Section A.2. The increase describes within-sample fit, not forecast accuracy. The rolling paired exercise in Section 5.5 supplies the separate out-of-sample comparison; its training samples and refit schedule differ from this single development fit.

Appendix B. Inferential calculations and comparison families

B.1. Marginal loss comparisons

The unadjusted loss differential is defined in equation (1), so a positive mean favours the candidate. In A, country losses are averaged within each month before inference; the countries form a fixed panel. In B, the series is placed on the archived market-date grid. Let G be the grid length, n the number of observed differentials, and $\bar{d}$ their mean. For observed dates define $u(t)=d(t)-\bar{d}$, and assign a zero score to missing dates only for the covariance sum. This convention excludes missing-date pairs; it does not impute a forecast loss. The implemented Bartlett HAC variance, following the covariance construction of Newey and West (1987), is

$$\hat{V}(\bar{d}) = \frac{n}{n-1} \frac{1}{n^2} \left(\sum_{t=1}^G u_t^2 + 2 \sum_{j=1}^L \left(1 - \frac{j}{L+1} \right) \sum_{t=j+1}^G u_t u_{t-j} \right). \quad (\text{B.1})$$

The first sum runs over the full grid. The main bandwidth is

$$L = \min \{ G-1, \max \{ h-1, \lfloor 4(n/100)^{2/9} \rfloor \} \}. \quad (\text{B.2})$$

Here Φ denotes the standard normal cumulative distribution function. The one-sided marginal p-value is $1-\Phi(\bar{d}/\sqrt{V})$, and the reported 95% marginal interval is $\bar{d} \pm 1.96\sqrt{V}$. These intervals are not simultaneous or adjusted for selecting examples after inspecting point estimates. The examples in Section 5.1 were selected for illustration after such inspection.

The wider-bandwidth sensitivity replaces $h-1$ by $2h-1$. A separate sensitivity multiplies the main statistic by $c(n,h)$, the finite-sample factor associated with Harvey, Leybourne and Newbold (1997), and uses the upper tail of a t distribution with $n-1$ degrees of freedom:

$$c(n,h) = \sqrt{\frac{n+1-2h+h(h-1)/n}{n}}. \quad (\text{B.3})$$

This factor-HLN/t calculation is an explicitly reported approximation using the same HAC estimate and observed count n . Applying that factor does not establish a finite-sample validity result for these estimated, overlapping forecasts or the incomplete daily grid. The normal calculation, wider bandwidth and factor-HLN/t sensitivity are not successive acceptance criteria.

The adjusted Clark–West differential is also recalculated with the same temporal unit and HAC procedure. It adds the squared difference between candidate and reference predictions to the unadjusted squared-error differential. The resulting p-values are retained as diagnostics because the nesting and estimation conditions required for a structural interpretation have not been established for these procedures.

B.2. Multiplicity families

Table B.1 gives the family definitions. BH and Holm are applied separately to the unadjusted marginal p-values. Missing or non-finite p-values, if present, are retained in the declared family with a value of one for adjustment. The table defines the comparisons being adjusted, not independent tests or a claim of global control over unrecorded analytical choices.

Table B.1. Adjustment families within each evaluation window

Family	Fixed dimensions	A comparisons	B comparisons
BH design	A: feature transformation and country sample; B: transformation and refit schedule	24	48
BH window	Block and window	96	192
Holm window	Block and window	96	192
Holm reference	Block, window and reference	48	96

Notes: Each family is formed separately in pseudo-OOS and holdout. The block/window Holm family is the conservative reference for the enumeration. Neither separate-window reporting nor the listed adjustments provides joint protection for selecting whichever window looks most favourable. Holm depends on valid marginal p-values; the dependence conditions for BH control are not established for the enumerated specifications.

B.3. Model confidence sets and joint diagnostics

The MCS follows Hansen, Lunde and Nason (2011), using the range statistic, the maximum absolute studentised pairwise loss difference. At each step it eliminates the model with the largest adverse studentised pairwise difference, rather than the largest unstandardised mean loss. Stepwise p-values are accumulated monotonically, and membership at 90% uses the Monte Carlo convention $(r+1)/(B+1)$, with $B=5,000$ resamples. The reported sensitivities use block lengths of two, three and four months in A and ten, twenty and forty observed sessions in B. The random seed is 20260922.

SPA comparisons are formed separately by evaluation window. The full and restricted daily-holdout sets in Section 5.4 use the same aligned 390 observations and the same resamples, allowing the role of the changed maximum to be inspected directly. Empirical SPA exceedance frequencies retain r/B ; simulated SPA and MCS use $(r+1)/(B+1)$. Section 5.4 reports both conventions for the empirical headline contrast. No pooled-window SPA result is used as substantive evidence: resampling separate windows independently would require an additional dependence justification. MCS and SPA resample consecutive observed rows and compress missing periods; their block-length sensitivities do not remove that limitation.

Appendix C. Inferential sensitivities

C.1. Alternative adjustments and marginal approximations

Table C.1 reports counts below the 5% threshold for the unadjusted loss comparison. The final two columns apply Holm by block/window to the wider-bandwidth and factor-HLN/t marginal p-values, respectively. They therefore change the marginal approximation while retaining the same multiplicity family. Counts concern the same forecasts and should not be added across columns.

Table C.1. Significant comparisons under the reported alternatives

Block	Window	Ref.	BH design	BH window	Holm window	Holm ref.	Wider HAC	HLN/t
A	Pseudo-OOS	B0	0	0	0	0	0	0
A	Holdout	B0	0	0	0	0	0	0
A	Pseudo-OOS	B2	0	0	0	0	0	0
A	Holdout	B2	0	0	0	0	0	0
B	Pseudo-OOS	B0	34	34	8	12	4	6
B	Holdout	B0	91	91	55	85	37	52
B	Pseudo-OOS	B3	0	0	0	0	0	0
B	Holdout	B3	0	0	0	0	0	0

Notes: BH design and BH window use the main HAC marginal calculation, as do both Holm columns. The wider-HAC and HLN/t columns use Holm-window adjustment. The equality of the two BH counts here is an observed feature of these results, not an equivalence of the family definitions. The number of comparisons per row is 48 in A and 96 in B. Table 4 supplies the corresponding point improvements and nominal counts.

C.2. Complete MCS block-length sensitivities

Tables C.2 and C.3 report all evaluated block lengths for the main forecasting specification under squared error. They make the membership changes visible without requiring agreement across every resampling choice. Horizons are in months for A and market sessions for B.

Table C.2. Monthly model confidence sets at 90%

Window	h	2 months	3 months	4 months
Pseudo-OOS	1	B0, B2, M1, M2, M4	B0, B2, M1, M2, M4	B0, B2, M1, M2, M4
Pseudo-OOS	3	B0, B2, M1, M2, M4	B0, B2, M1, M2, M4	B0, B2, M1, M2, M4
Holdout	1	B0, B2, M1, M2, M4	B0, B2, M1, M4	B0, B2, M1, M4
Holdout	3	B0, B2, M1, M2, M4	B0, B2, M1, M2, M4	B0, B2, M1, M2, M4

Notes: Three-month blocks are the main setting. Five models enter each initial set: B0, B2, M1, M2 and M4. Retained models remain statistically unresolved within the declared collection; membership does not establish equivalence.

Table C.3. Daily model confidence sets at 90%

Window	h	10 sessions	20 sessions	40 sessions
Pseudo-OOS	5	B3, M2, M3, M4	B3, M2, M3, M4	B3, M2, M3, M4
Pseudo-OOS	10	B3, M2, M3, M4	B3, M2, M3, M4	B0, B3, M2, M3, M4
Pseudo-OOS	20	B3, M2, M4	B0, B3, M2, M3, M4	B0, B3, M2, M3, M4
Holdout	5	B3, M2, M3	B3, M2, M3	B3, M2, M3
Holdout	10	B3, M2	B3, M2	B3, M2
Holdout	20	B3	B3, M2	B3, M2

Notes: Twenty-session blocks are the main setting. Six models enter each initial set: B0, B3, M1, M2, M3 and M4. These sets concern the main specification, not selection across the entire specification curve. All calculations use 5,000 resamples and the same declared seed; detailed stepwise and accumulated p-values are archived.

C.3. Paired empirical comparison

Table C.4. Matched linear forecasts with refits every 63 sessions

Window	Text	h	N	RMSE ratio	Raw p	Holm p
Pseudo-OOS	Detrended	5	1010	1.013	0.8601	1.0000
Pseudo-OOS	Detrended	10	1010	1.022	0.9360	1.0000
Pseudo-OOS	Detrended	20	1010	1.030	0.9572	1.0000
Pseudo-OOS	Levels	5	1010	0.990	0.2299	1.0000
Pseudo-OOS	Levels	10	1010	0.996	0.3764	1.0000
Pseudo-OOS	Levels	20	1010	0.996	0.3892	1.0000
Holdout	Detrended	5	405	0.994	0.3420	1.0000
Holdout	Detrended	10	400	1.006	0.7130	1.0000
Holdout	Detrended	20	390	1.004	0.6093	1.0000
Holdout	Levels	5	405	0.948	0.0128	0.2550
Holdout	Levels	10	400	0.915	0.0013	0.0297
Holdout	Levels	20	390	0.862	0.0007	0.0160

Notes: Ratios compare HAR plus text with market-only HAR under the same OLS procedure and complete observations. Values below one favour adding text. One-sided p-values concern unadjusted log-squared-loss differences under the grid-aware HAC calculation. Holm includes all twenty-four paired comparisons across both windows, transformations, schedules and horizons. Wider-bandwidth results are retained in `matched_summary.csv`. This is a post-inspection analysis.

The paired main display is Table 6. Each forecast pair uses the same training and evaluation rows, with the daily cut-off strictly before $i-h$. The package supplies the fit origin, last eligible training index, sample count and design ranks for every fit. The augmented design has twenty columns including the intercept; redundant text columns can reduce rank without preventing fitted-value reproduction. The input matrix is standardised inside each training sample and fitted by SVD. No inference about individual coefficients is drawn.

The Holm family includes both windows, unlike the window-specific families used for the complete-procedure enumeration. Its twenty-four members comprise two transformations, two refit schedules, three horizons and two windows. The wider-HAC Holm p-values for the four significant holdout level comparisons range from 0.0191 to 0.0485. The full comparison-level file preserves means, standard errors, marginal intervals, nominal p-values and both Holm calculations.

C.4. Simulation design and Monte Carlo uncertainty

Table C.5. Monte Carlo uncertainty for the declaration rules

Scenario	Rule	Count	Rate %	MC SE pp	Wilson 95% %
No signal, rho 0.5	Nominal mixed	56	11.2	1.41	[8.7, 14.3]
No signal, rho 0.5	SPA mixed	6	1.2	0.49	[0.6, 2.6]
No signal, rho 0.5	SPA matched	2	0.4	0.28	[0.1, 1.4]
No signal, rho 0.5	Holm matched	1	0.2	0.20	[0.0, 1.1]
No signal, rho 0.9	Nominal mixed	499	99.8	0.20	[98.9, 100.0]
No signal, rho 0.9	SPA mixed	486	97.2	0.74	[95.4, 98.3]
No signal, rho 0.9	SPA matched	2	0.4	0.28	[0.1, 1.4]
No signal, rho 0.9	Holm matched	1	0.2	0.20	[0.0, 1.1]
Signal, rho 0.9	Nominal mixed	500	100.0	0.00	[99.2, 100.0]
Signal, rho 0.9	SPA mixed	492	98.4	0.56	[96.9, 99.2]
Signal, rho 0.9	SPA matched	74	14.8	1.59	[12.0, 18.2]
Signal, rho 0.9	Holm matched	46	9.2	1.29	[7.0, 12.1]

The innovation vector for the sixteen artificial text inputs has covariance 0.3 times the all-ones matrix plus 0.7 times the identity. The AR coefficient is 0.8 and the innovation multiplier is 0.6, giving stationary unit marginal variance. The market innovation has unit variance and is independent of all text innovations. Text enters the next-state equation only through $\gamma x_{t,1}$; when $\gamma=0$ the conditional no-additional-information statement follows from the generator, not from a permutation assumption.

The artificial state starts at zero, followed by the specified burn-in. Current and trailing market inputs and the prior sixty-observation text average use only information available at the forecast origin. All horizons share 400 evaluation origins. Within each replication the two SPA families share block starts, studentisation and recentering for retained columns. The seeds are deterministic and separate the data generator from resampling. No parameters are calibrated to the empirical significance results.

Rates use independent datasets as their denominator, not the twelve procedures or twenty-four comparisons. Monte Carlo standard errors measure the sampling variation from using a finite number of simulated datasets; Wilson confidence intervals summarise that uncertainty and remain informative when the observed rate is zero or one. The scenarios illustrate a mechanism and limited detection under one signal strength. Their interpretation follows Section 3.4.

C.5. Complete follow-up results and paired disagreement

Table C.6. Complete common-sample empirical follow-up

Period	Market	Text	Refit	h	Ratio	Holm 48	Wide 48
Later	Long	Dev	21	5	0.994	1.0000	1.0000
Later	Long	Dev	21	10	1.011	1.0000	1.0000
Later	Long	Dev	21	20	1.009	1.0000	1.0000
Later	Long	Dev	63	5	0.994	1.0000	1.0000
Later	Long	Dev	63	10	1.011	1.0000	1.0000
Later	Long	Dev	63	20	1.011	1.0000	1.0000
Later	Long	Level	21	5	0.972	1.0000	1.0000
Later	Long	Level	21	10	0.968	1.0000	1.0000
Later	Long	Level	21	20	0.931	1.0000	0.9911
Later	Long	Level	63	5	0.973	1.0000	1.0000
Later	Long	Level	63	10	0.969	1.0000	1.0000
Later	Long	Level	63	20	0.928	0.9589	0.8406
Later	Short	Dev	21	5	0.995	1.0000	1.0000
Later	Short	Dev	21	10	1.008	1.0000	1.0000
Later	Short	Dev	21	20	1.008	1.0000	1.0000
Later	Short	Dev	63	5	0.996	1.0000	1.0000
Later	Short	Dev	63	10	1.008	1.0000	1.0000
Later	Short	Dev	63	20	1.008	1.0000	1.0000
Later	Short	Level	21	5	0.943	0.4235	0.4913
Later	Short	Level	21	10	0.910	0.0793	0.0373
Later	Short	Level	21	20	0.865	0.0360	0.0188
Later	Short	Level	63	5	0.944	0.4190	0.4840
Later	Short	Level	63	10	0.911	0.0436	0.0158
Later	Short	Level	63	20	0.863	0.0168	0.0088
Earlier	Long	Dev	21	5	1.005	1.0000	1.0000
Earlier	Long	Dev	21	10	1.015	1.0000	1.0000
Earlier	Long	Dev	21	20	1.022	1.0000	1.0000
Earlier	Long	Dev	63	5	1.006	1.0000	1.0000
Earlier	Long	Dev	63	10	1.014	1.0000	1.0000
Earlier	Long	Dev	63	20	1.018	1.0000	1.0000

Period	Market	Text	Refit	h	Ratio	Holm 48	Wide 48
Earlier	Long	Level	21	5	0.994	1.0000	1.0000
Earlier	Long	Level	21	10	0.999	1.0000	1.0000
Earlier	Long	Level	21	20	1.003	1.0000	1.0000
Earlier	Long	Level	63	5	0.988	1.0000	1.0000
Earlier	Long	Level	63	10	0.990	1.0000	1.0000
Earlier	Long	Level	63	20	1.009	1.0000	1.0000
Earlier	Short	Dev	21	5	1.014	1.0000	1.0000
Earlier	Short	Dev	21	10	1.025	1.0000	1.0000
Earlier	Short	Dev	21	20	1.034	1.0000	1.0000
Earlier	Short	Dev	63	5	1.015	1.0000	1.0000
Earlier	Short	Dev	63	10	1.025	1.0000	1.0000
Earlier	Short	Dev	63	20	1.034	1.0000	1.0000
Earlier	Short	Level	21	5	0.995	1.0000	1.0000
Earlier	Short	Level	21	10	1.000	1.0000	1.0000
Earlier	Short	Level	21	20	0.993	1.0000	1.0000
Earlier	Short	Level	63	5	0.988	1.0000	1.0000
Earlier	Short	Level	63	10	0.989	1.0000	1.0000
Earlier	Short	Level	63	20	0.984	1.0000	1.0000

Notes: Later denotes the reused holdout, earlier the pseudo-OOS period; Long denotes the expanded benchmark. The first common complete training origin is 25 February 2016. There are 1010 evaluation observations per horizon in the earlier period and 405, 400 and 390 in the later period. Raw p-values, marginal intervals, base and augmented RMSE, and Holm within each benchmark's 24 comparisons are retained in `empirical_summary.csv`.

The longer market components are means of log absolute daily premium changes, whereas the original weekly and monthly HAR inputs are logs of mean absolute changes. This distinction is intentional and follows the fixed follow-up design. No zero daily changes occur in the archived data; the undefined first difference is excluded. The first complete 250-session window therefore follows that initial observation. The same complete rows are used for both benchmarks, both transformations and each corresponding refit/horizon pair.

Table C.7. Monte Carlo uncertainty and paired disagreement

Scenario	Route	Count	Rate %	MC SE pp	Wilson 95% %
No added information	Short levels	29	5.8	1.05	[4.1, 8.2]
No added information	Short detrended	24	4.8	0.96	[3.2, 7.0]

Scenario	Route	Count	Rate %	MC SE pp	Wilson 95% %
No added information	Expanded levels	1	0.2	0.20	[0.0, 1.1]
No added information	Expanded detrended	2	0.4	0.28	[0.1, 1.4]
No added information	Expanded level only	1	0.2	0.20	[0.0, 1.1]
Signal 0.25	Short levels	240	48.0	2.23	[43.7, 52.4]
Signal 0.25	Short detrended	97	19.4	1.77	[16.2, 23.1]
Signal 0.25	Expanded levels	143	28.6	2.02	[24.8, 32.7]
Signal 0.25	Expanded detrended	23	4.6	0.94	[3.1, 6.8]
Signal 0.25	Expanded level only	129	25.8	1.96	[22.2, 29.8]
Signal 0.50	Short levels	456	91.2	1.27	[88.4, 93.4]
Signal 0.50	Short detrended	162	32.4	2.09	[28.4, 36.6]
Signal 0.50	Expanded levels	383	76.6	1.89	[72.7, 80.1]
Signal 0.50	Expanded detrended	96	19.2	1.76	[16.0, 22.9]
Signal 0.50	Expanded level only	290	58.0	2.21	[53.6, 62.2]

Notes: Expanded level only means levels reject while detrended text does not, on the same dataset and expanded market benchmark. This is a joint event, not a difference of marginal rejection rates. Intervals quantify Monte Carlo uncertainty under the specified generator, not uncertainty over researcher behaviour. All reverse-disagreement and Holm results are archived.

The simulation's no-information condition refers to all observed market history, including m . Under $\gamma=0$, z evolves without s ; text is $m+s$ plus independent noise, so conditioning on observed m and z leaves no channel by which text predicts future z . Under either positive γ , current s is revealed by subtracting m from the first text input. Its effect on the expected future target is nonzero. The expanded linear benchmark is deliberately simple and observes m directly; it is an informational control, not a claim that the empirical slow averages recover a true state.

All scenarios share generation streams by replication number and all paths share resampling draws. Datasets are independent within each scenario, but results across scenarios and paths are paired. Separate scenario rates therefore cannot be treated as independent estimates when comparing them. Complete summaries retain all SPA and Holm paths, combined families, reverse disagreement and both/neither-reject events. Comparison-level outputs retain forecast-loss gains, RMSE ratios and marginal p-values for every dataset.

Appendix D. Reproducing the study

D.1. Computational entry point

The replication package contains frozen GDELT and Eurostat extraction inputs, panel construction and estimation code, decision records, final inference, reference results, table generators and

manuscript sources. The replication package is identified by Harvard Dataverse DOI 10.7910/DVN/J7JWXM; the abstract page provides links to the paper and replication materials. Reproduction requires no external data retrieval or AI access.

From the extracted package root, create an environment and execute:

```
python3 -m venv .venv
.venv/bin/pip install -r requirements.txt
.venv/bin/python run.py --full
```

The full run reconstructs panels, re-estimates forecasts including the holdout, regenerates specification losses and final inference, recomputes the descriptive illustrations, executes the paired extension, expanded-market follow-up and both controlled experiments, creates the numerical results tables, and verifies outputs against supplied references. The holdout record is preserved; deterministic reproduction is not another independent evaluation. The `--inference` option starts from supplied forecasts and losses and omits model estimation. The `--extension` option executes the paired empirical exercise and all 1,500 simulation replications, assembles their tables and verifies the outputs. The `--diagnostics` option recomputes the descriptive illustrations from existing panels, forecasts and curves, assembles the text and verifies results; it performs no forecast re-estimation.

D.2. Environment and verification

The numerical pipeline was tested on Linux with Python 3.12.14, pinned dependencies, single-threaded numerical libraries and seeds 20260922 for the empirical resampling and 20261005 for the controlled experiment. The dependency file and execution report record the exact environment. Verification compares all rows and columns of the final curves, summary counts, MCS output and six panel, prediction and loss files, as well as the SPA mechanism record and the directional, OLS-fit, choice-summary and coverage diagnostics. Numeric tolerances are relative 1e-9 and absolute 1e-10. Missing patterns and categorical values must agree.

Additional checks link numerical tables and selected displayed claims to their source results. HAC checks compare the complete-series calculation with statsmodels and verify missing-date covariance pairs directly. The distribution checksum manifest identifies file contents; numerical agreement does not imply byte-identical output or validate every statistical assumption.

D.3. Tables and manuscript

Final inference, matched forecasts and simulation outputs reside in results/. The table generator supplies Tables 4–9 and C.1–C.7. Tables 1–2 are conceptual; Table 3 describes the design and checked evaluation counts; Table B.1 defines the checked adjustment families. Comparison-level files preserve effects, marginal intervals and alternative adjustments, and a claims file records selected numerical statements and their source values. The reporting-diagnostics module generates the descriptive illustrations; the extension module produces matched forecasts and simulated declarations. Their displayed values and result tables are inserted automatically during manuscript assembly.

The `--paper` option assembles the authored prose and generated tables, builds Word and converts it to PDF. The README specifies additional Python packages, Pandoc, LibreOffice and fonts. Rendering may vary across environments; the supplied PDF is the reference presentation. Pagination starts

visibly at 3 on the non-technical summary and continues through references, the AI disclosure and lettered appendices.

The package reproduces the study from frozen inputs. It does not reconstruct historical publication vintages or require another paid extraction. Source scripts and documentation are supplied for provenance. The measurement, tuning, selection and inferential limitations discussed in the paper remain even when numerical reproduction passes.

D.4. Reproducing the extension

`analysis/extension_protocol.json` fixes the empirical design, three data-generating scenarios, replication count, seed, declaration rules and resampling settings. Its hash is recorded before execution in `extension_protocol_record.json`. This local chronology is not an independently certified preregistration. The extension was designed after inspection of the application and is identified as such in the manuscript.

`analysis/extension.py` reproduces all paired forecasts and the simulation. The paired base predictions are checked against archived B3 predictions; fit-time target cut-offs and independent least-squares predictions are verified. The simulation uses cumulative sufficient statistics for expanding OLS and exact block sums for bootstrap means. These are checked against direct SVD fitting and explicitly indexed bootstrap samples. The verifier also replays selected independent replications from their seeds, checks every aggregate rate and interval, and compares archived reference outputs. No AI access or external data download is required.

D.5. Reproducing the 8 October follow-up

```
python3 run.py --followup
```

This command runs `analysis/followup.py` for the empirical and simulation exercises, verifies their results and regenerates manuscript tables. The fixed design is `analysis/followup_protocol.json`, with its local pre-execution hash in `followup_protocol_record.json`. Results are in `results/followup_2026-10-08`. Verification reconstructs all 48 empirical summaries, multiplicity adjustments and simulation rates, checks temporal cut-offs, and replays the first dataset of each of the three scenarios against direct SVD fits and explicit bootstrap indexing. The follow-up uses seed 20261008. Its runtime dependency versions are recorded separately from the original execution environment.

The follow-up is also included in `--full`. The manuscript's added Tables 8–9 and C.6–C.7 and numerical tokens are generated by `followup_tables.py`. The prior manuscript template and parts are preserved in `paper/archive_2026-10-05`. A timing-check index was corrected from 250 to 251 to accommodate the undefined initial daily difference; no estimates or design choices were changed by this implementation correction. The protocol and input hashes remained unchanged.