



Working papers available at alexvidovich.com/papers

Alex Vidovich · Working Papers

research homepage: <https://alexvidovich.com>

**Working
Paper**

October 2026
Not investment
advice

A Thermometer, Not an Analyst: Lawful Uses of a Text Judge in a Live Trading Lab

Alex Vidovich^{a,*}, Zivana Zerjal^b

^a Independent researcher · alexvidovich.com

^b Independent researcher · zivanazerjal.com

* Corresponding author.

E-mail addresses: support@alexvidovich.com (A. Vidovich), support@zivanazerjal.com (Z. Zerjal)

ARTICLE INFO

Article history:

Working paper, 9 October 2026

Prepared for SSRN

JEL classification:

G14

G17

C45

C52

C81

C88

Keywords:

large language models

text classification

news analytics

pre-registration

look-ahead bias

research infrastructure

trading systems

ABSTRACT

A text model can sit next to a live trading system the way a thermometer sits next to a patient: it reads, it never prescribes. We follow one such model into a lab running an intraday strategy and a second, overnight strategy on one account. It is a small commercial judge, pinned to one version, that answers only typed questions about headlines. We describe six tests for a text signal: read a plain fact correctly against a real label, beat a simple count of its own headlines, show no memory of famous past events, give the same answer twice, always see the decision-time date, and check each forward verdict rule against random flags. They were assembled over three days, most of them after a failure, and we present them as the checklist we would now apply up front. We compare the judge with a general chatbot on blind rows, list what it may touch (stamps written after the close, counted forward) and what it may never touch (any number in the trading system), and report the plumbing failures that corrupted its readings. Its readings so far are small: one fact read well, one small effect below its bar, several that vanished; the fact was read under one setup, and its margin over a simple headline count is borderline. None of the six tests is new on its own; what is new is using them together on a text instrument beside a live system, and reporting every study it was used in, failures included. This is a paper about checking data and the instrument, not about profit: we make no claim about either strategy's returns.

Figures come from the authors' dated research log and its ledger files, which are not public. Some figures are rounded or withheld to protect the live system. The paper describes how a text model was tested and fenced; it does not describe the trading rules it sits beside.

1. Why a text judge, and why not a chatbot

The door was shut for a good reason. Only a small, dumb, repeatable instrument could open it.

The question that started it. News had been discussed in our lab and then parked. Earlier work had already ruled it out as an input to any trading decision. Every idea of that kind we had tried, ten in all (calendar events and cross-market moves among them), had failed

its own pre-registered test. One narrower use stayed lawful (in this paper "lawful" means allowed by the lab's own written rules; it is not a legal term): a mechanical flag stamped at decision time and measured forward, with its bars written before any judge existed.

The look-ahead trap that kept the door shut. A model that has read the internet may know how 2025 and 2026 turned out. Score old headlines with it and the future can leak into a backtest. Nothing in the score tells you it happened. Studies that score headlines with large language models guard against this by using headlines from after the model's training cutoff (e.g., Lopez-Lira and Tang, 2023). So at the time only a mechanical flag was allowed in any backtest: a headline is present, how many, from which source tier. To lift that rule for one model, we first had to show that the model did not remember the past (Section 2).

Thermometer versus analyst. A thermometer gives the same kind of number every time. It has no opinion about the patient. An analyst designs the question and reads the measurements. In our lab the judge is the thermometer. The humans are the analyst, with a general assistant model as a drafting and analysis aid. The properties we claim for the judge are measured, not asserted:

- *Typed answers only.* The judge returns a probability for a yes/no question, or a probability over a fixed set of choices. It never returns free text. The first run used about 450 tokens per call over 150 rows with no errors. That figure was logged without the input/output split; it is probably lower because 45% of those rows had no headline at all. The second run used about 1,290 input tokens per call at a median latency of 283 ms. Speed and cost matter here only for scale: the judge runs after the close, so its latency never touches a trade. At the vendor's published early-access price of \$42 per billion input tokens (TypeSafe AI, 2026), a call of the hearing's size costs about a two-hundredth of a cent; we did not audit our own bill against it. We did not price the chatbot on the same work. Typed answers are also easy to audit: every stored reading is a number or one word from a fixed list.
- *Repeatable.* On identical input the event probability moves by about ± 0.05 on confident answers and up to 0.07 on mid-range ones, and does not depend on headline order. On empty input it answers about 0.03. A 50-row panel re-asked at the pinned version agreed on the tone choice 96% of the time, with a mean absolute change in event probability of 0.011. A fresh 30-row panel, re-fetched two days after the stamps with unchanged headline counts on all 30,

repeated every tone choice (30 of 30) and moved the event probability by 0.008 on average, across days and within one session alike; the largest move, 0.07, came on a mid-range answer (0.51), and confident answers moved by at most 0.05.

- *Reads the date.* The same scheduled-event headline scored 0.97 with the real decision date and 0.4 with a dummy date; a second case 0.63 against 0.04. Two cases are a demonstration, not a measurement; a replication should score a panel of dated and undated headlines. Even so, the decision-time date now always travels with the headlines (Gate E).
- *Not easily fooled by headline shape.* Among nine synthetic probes, a plain event line scored 0.84, its negation 0.14, the same event for another company 0.04, a stale-dated version 0.12 and empty noise 0.02. This is descriptive, not a validation.

Whether the judge can be talked out of an answer is a design property, not a finding. It has a typed interface, no conversation and no access to outcomes. We have no measurement of persuadability and claim none.

Comparison with a general chatbot. We put a general assistant model (Claude, Anthropic) on the same 83 rows: every calibration row with at least one headline, 39 of them labelled positive. Both models were scored against the proxy label that Section 2 shows to be wrong on many rows, so this comparison cannot rank them; it shows how far they agree. The comparison model was the same family that helped draft this paper (Claude Fable 5.1, see "Use of AI tools"); shuffling the rows and hiding dates and labels limits, but does not remove, that circularity. It judged blind, from headline and symbol only. The chatbot scored an area under the ROC curve of 0.806 against the judge's 0.761. The two sat on the same side of 0.5 on 75 of 83 rows, and their probabilities correlated at $r = 0.90$. On the rows the label flagged their mean probabilities were 0.66 and 0.65; on the rest 0.19 and 0.27. On this sample the judge behaved like a slightly looser version of the same judgement; we did not measure the chatbot's cost or calibration on these rows. Real disagreements were four: three where the judge over-read routine coverage, one the other way on a bare "earnings scheduled" line. On 16 rows both disagreed with the label; the real calendar later confirmed the proxy's errors in aggregate (Section 2). So the case for the small model is not accuracy. The small model was probed for memorisation and pinned to a version. The chatbot was not: its training may well carry earnings calendars, and we did not put it through the same probe. It was used once, as a comparison, not as an instrument we meant to keep; a replication that wants a chatbot as its judge should put it through the same gates.

One caveat on what "no memory" means. With headlines removed, the judge still answered the scope question from what the symbol is: a broad-index fund came back "market-wide", a company "stock-specific", on event and control dates alike. It knows what a ticker is. It does not carry what happened to it on a date. Section 2 is worded accordingly.

Related work. Headline scores from large language models have been shown to correlate with later returns (Lopez-Lira and Tang, 2023), and the look-ahead risk of such studies has been measured directly by removing company names from the headlines (Glasserman and Lin, 2023). The case for pre-registration and for correcting for the number of tests is made across financial economics (Harvey, Liu and Zhu, 2016; Harvey, 2017); backtest overfitting through parameter search is quantified by Bailey et al. (2014); and the hidden cost of data plumbing in machine-learning systems is described by Sculley et al. (2015). Large language models are also used as annotators of text in social science (Gilardi, Alizadeh and Kubli, 2023) and as judges of other models' output, with known position, verbosity and self-enhancement biases (Zheng et al., 2023); data validation as an engineering discipline is described by Breck et al. (2019). None of our six gates is new on its own. What this paper adds is the combination, applied to a text instrument that sits beside a live system; the full record of every study the instrument was used in, failures included; and the plumbing faults found on the way.

Terms. A *hearing* is a validation run on rows nobody had looked at. A *stamp* is a scheduled after-close annotation job. The *board* is the registry of forward-tracked items, and a *look* is a pre-set count at which an item's verdict is read. *Seen* rows had been looked at before a test; *unseen* rows had not.

2. Admission gates for any text signal

Six gates. They were assembled over three days, most of them after a failure; the stamps ran in shadow from the first passed gate on. We present them as the checklist we would now apply up front. The count baseline on the hearing rows was computed afterwards, with no margin fixed in advance, which is one reason we call its margin borderline. The gates were written in working sessions between the first author and the AI assistant named under "Use of AI tools", each recorded in the lab's dated log the day it was adopted; later AI reviews of the numbers and code corrected some of them, Gate F's among them.

Gate A — Fact-check against a real label, before any outcome study

A text signal must first show it can read a *fact* from headlines. Only then does anyone ask whether the fact predicts a return. Our fact was plain: "a scheduled company event is reported or imminent for this name this morning."

First try, against a proxy label. The sample was stratified from every candidate name-morning of 2025–26: all 50 mornings a mechanical, price-derived proxy flagged as event mornings plus 100 unflagged mornings at random, with an 18-hour headline window before the decision and a pre-registered bar of $AUC \geq 0.75$. The judge scored 0.737 and failed by 0.013. On 150 rows that miss is well inside sampling noise (the standard error of an AUC near 0.74 is about 0.045), so the first test could not have told 0.737 from a true 0.75 either; the bar still stood. Nothing was wired, and the bar did not move after the data.

Two causes were found. First, 67 of the 150 mornings had no headline at all in the window. The judge rightly answered about 0.03 on nothing, and on the 83 mornings with headlines it scored 0.761. Over all 150, flagged mornings averaged 0.51 (48% at or above 0.8), unflagged 0.13 (3%). Second, the label itself was wrong. The proxy had flagged one market-wide volatility week as company events on four names, flagged two mornings of one name that had no event, and missed real earnings mornings that the judge caught.

Build the right yardstick. We parsed a real earnings calendar from the news feed's own daily schedule article: 452 trading days from January 2025 to October 2026, a non-empty list of reporting companies on 378 of them, several hundred reporting-days across our universe. Judged against it across every candidate name-morning of the era, the proxy label scored 0.903 overall, but its flag caught only about half of the true event mornings and raised about as many false flags. A yardstick that misses half the events cannot grade a reader of events.

One fresh hearing on unopened rows. This was pre-registered before the calendar was opened: every calendar-positive among the rows no one had looked at (23) plus 23 random negatives, a 36-hour window, rows with no headline reported and excluded (7, none positive), 39 scored. Three things changed at once: the label, the window (doubled, because the 18-hour one had left almost half the rows with nothing to read) and the zero-headline rows, now dropped. The gain cannot be credited to the label alone. The judge scored 0.981 against the proxy's 0.918 on the same rows; mean probability 0.91 on event mornings and 0.31 on the rest;

one miss at 0.25 with five headlines, one confident false call at 0.84. Thirty-nine rows (23 event mornings) is a small sample, so this is a sanity check that the judge can read the fact, not a precise measurement: a bootstrap of the same rows puts the 95% interval at 0.935 to 1.00, against 0.815 to 0.986 for the proxy; the two intervals overlap, which is why we call it a sanity check.

Gate B — Beat the dumbest baseline by a margin

The dumbest baseline for "is there an event" is to count the headlines. On the 39 hearing rows the count alone scores 0.889 (event mornings averaged 15.4 headlines, the rest 4.5); the judge's 0.981 clears it by 0.092. The margin is positive but not wide, and it is borderline at conventional levels. On the same bootstrap the count's interval is 0.765 to 0.978 and the margin's two-sided 95% interval runs from 0.000 to 0.213, so it touches zero; read two-sided, that is about $p = 0.05$, right at the conventional line. Read one-sided, the margin is zero or less in 2.5% of resamples, which clears the 5% line but not by much. Thirty-nine rows cannot settle whether the judge adds much to a count of its own headlines on this fact; a larger hearing can. On the first run's 150 rows, with the proxy label, the pair is 0.729 against 0.737: there the label, not the judge, was the problem. An earlier note in our log quoted the count baseline as 0.85 without naming the rows. No file reproduces it, and we retire it in favour of 0.889.

The house example of *losing* to this baseline came later. A pre-open "news regime" reading asks whether the morning is driven by macro news, by company results, or quiet. It was scored on every session of the two-year era: 70% macro-driven, 30% earnings-driven, none quiet. By the letter it passed its pre-registered bar on the intraday strategy's own outcomes, and it beat a calendar-based baseline. Then we built the baseline the rule actually asks for, from the judge's *own input*: the headline count. Split into equal thirds by that count (110–214, 214–249 and 249–315 headlines), the strategy's historical fills reproduced the gap. Within the low-count tercile the macro and earnings readings were indistinguishable. The gap's t-statistic was 1.64 and it was not spread evenly across fills. The reading became a counter and the follow-up sizing study was withdrawn. The lesson: the baseline must be built from the judge's own input, not from a convenient neighbour.

Gate C — No-memorisation probe

Does the judge remember famous mornings? We handed it six widely reported event mornings from 2022–24 (company results, a bank failure, an IT outage) with the headlines removed and the real date kept. Every one came back at an event probability of 0.03 with tone

"none", identical to four quiet control mornings. A refresh on four further public mornings from 2023–24 (a regional-bank panic, a global risk-off day, a software company's results, a cloud company's leadership change) and three quiet controls, each asked twice, gave 0.03 and "none" on all fourteen answers. The consequence is narrow. The probe shows only that the judge does not recall these events without their text. It does not show that real, dated headlines carry no date-linked bias, so old-headline backtests are cleaner with *this* judge at *this* version than with one that recalls, not proven free of look-ahead. The look-ahead worry of Section 1 still applies to any judge not probed this way. The probe is small and tests one thing, answers with the headlines removed. It does not test a shifted date or a quiet date inside a famous crisis; a replication should add both. A complementary design removes company names from the headlines instead (Glasserman and Lin, 2023).

Gate D — Repeatability, and a drift panel

Repeat noise is about ± 0.05 on confident answers and up to 0.07 on mid-range ones (Section 1); nothing smaller can be seen. The model version is pinned in code. A fixed panel of 50 stamped name-mornings is re-asked weekly, after the close and outside the live process. Tone agreement below 0.90 is declared drift. The line is a judgement call, not a derivation: it sits below the 0.96 agreement the same version shows when simply re-asked, so ordinary noise does not trip it; a drop below the bar is flagged to the operator and moves the since-date of every board row that depends on the judge. The panel's baseline is the 0.96 above. A vendor can retire a pinned version; the panel is how we would learn.

Gate E — The date and window travel with every headline

Because the judge reads dates (Section 1), the real decision-time date is passed on every call. The window is part of the question too: a changed window is a new question with a new since-date (Section 4). Section 4 shows the cost: the same rows read under two window clocks gave two different answers.

Gate F — The forward judge is tested before it ships

Every board row carries a judge written before its data: a rule that says what count and what gap mean integrate (send the item to a full study as a candidate for use), kill (close it) or hold (keep counting). We learned to test the judge itself. One row's judge (a price-only row, same machinery) was run against random flag sets (a standard permutation test; Phipson and Smyth, 2010) and said "integrate" 49.8% of the time over 20,000 draws. The fixed judge requires the flagged mean at or below the 5th

percentile of random same-size flag sets, at least 12 flagged observations and a lower standard deviation. A unit test pins that no more than 3 of 20 random sets integrate. The single-look false-integrate rate is now 5%. Power at the first fixed look (60 observations) is about 11% for an effect of the backtest's size, and nightly re-evaluation after it pushed the false-integrate rate above 13%. So that row is read once at 60 and then only at 120 and 180; the text rows have pre-set looks too (Section 3). Until October 9 the code re-judged them every evening once a count was reached, which repeats the flaw this gate found; since then it reads each look once, on the data up to the evening the count was first reached. No text row had reached its first count when this was fixed. At that power a "hold" at the first look is the expected outcome even for a real effect, not evidence against it; the later looks are the real ones. The first look is kept because most judges also say when a wrong-signed result means "kill", so an item that is clearly going the wrong way can be closed early. Only that price row's judge carries the permutation null so far. The five text-judge rows still use plain gap thresholds and have not yet been run against random flags; until they are, their verdicts carry the flaw this gate found.

3. Where it may act, and where it never acts

The judge may write stamps after the close. It may never write a number into the robot.

Allowed: stamps after the close, counted forward. Four shadow stamps run after the close, outside the live trading process. A fifth job is operations, not research. The five jobs are:

1. the intraday strategy's chosen name: scheduled event (probability), tone, scope and news driver, over the 18 hours before the decision;
2. the overnight strategy's held names: the dominant theme of each name's recent coverage and, as a separate call so the theme answers stay untouched, each name's tone over the prior 72 hours;
3. the pre-open market: the news regime and a risk-off reading over the 12 hours before the open;
4. the names the intraday strategy's own checks turned down, with the same four questions as stamp 1;
5. a corporate-action watchdog over a wider pool of names (delistings, ticker changes, deals, splits, halts) over a 24-hour window, flagged to the operator at probability 0.6.

Each stamp has a board row. Our forward board lists every tracked item. Each row has a judge written before the data, a tag saying whether the item is deployed, shadow-only or a verdict gate, a since-date

that moves when the question or its input changes, and a fixed review count. All five judge rows are shadow-only. Their review counts are 40 picks (catalyst), 40 picks (regime), 60 nights (theme), 300 name-nights (tone) and 40 cases (refused). The counts were set by what the flow can deliver within a few months, not by a power calculation; we have no power figure for any judge row, and none for the price row's later looks either. They are the earliest honest look, not a guarantee of seeing an effect. We have not computed the text rows' power; with counts this small we expect it to be low, so a null at 40 picks is not evidence against an effect of the size seen in backtests. At its count each row returns one of its written verdicts. An inconclusive result says "hold" to twice the count (80 picks, 120 nights, 600 name-nights); the refused-decision row holds to 80 only when a bucket is too thin, and is otherwise re-heard or closed. Since October 9 the code enforces this: each look is read once, on the data up to the evening its count was first reached, and a second look is read only after a "hold". Before that, the second looks existed only in the verdict text and the rows would have been re-judged every evening; none had reached its first count. At the last run before this draft the board held 30 items plus one receipt row, and the judge rows stood at 3 of 40 (none flagged), 2 of 40, 0 of 60, 0 of 300 and 2 of 40, all "too early". A changed question set is a new item and does not inherit the old count.

Post-mortem tags, labelled as seen. Back-stamping the intraday strategy's live fills to date is context, not evidence. Fewer than half of them had no headline in the window, and about a quarter carried an event probability of 0.5 or more. The best-documented winners and losers alike carried the same kind of supportive company news, so the headline could not tell them apart. A related lens was tried and buried: whether the judge's scope reading predicts which mornings the system's own checks turn down. On 46 mornings it does not.

Never: numbers, thresholds, sizing, edge validation, risk measurement, parameter optimisation. The judge judges text, not numbers. A judge inside a parameter search launders overfitting: it supplies a plausible story for whichever cell the search prefers, and parameter search is already the main route to backtest overfitting (Bailey et al., 2014). No threshold in the trading system was ever set, moved or confirmed by it. A hypothetical shows the danger. Search ten values of one threshold and keep the best: by chance alone one will look good. Ask a text model why that value works and it will find a plausible headline story for it, and the story makes a lucky cell look caused. The search, not the story, decides how far the result can be trusted. The same logic sets the stakes of the judge's own errors: in shadow

use a false call (Section 2 shows a confident one at 0.84) costs a mislabeled stamp; read by live code it would cost a trade. That is why its outputs are only counted forward.

Never read by live code. The judge's stamps run after the close, outside the live trading process, and no live code reads any of their fields. Its credential is held outside the research and trading code, is never read by live code, and is validated before use. How this is checked, honestly: by the architecture (the stamps are written by a separate job after the close, to files the trading process does not open), by a code search that finds no reference to the judge in any live module, and by three code reviews in October 2026. It is not yet enforced by an automated test; a test that fails if any live module imports the judge is the obvious next guard.

4. Plumbing lessons

None of these is about the model. All are about the pipe that feeds it, and each cost a verdict or nearly did. That matches the wider experience that most of the risk in a machine-learning system sits in its data plumbing and glue code, not in the model (Sculley et al., 2015).

Page the feed. The news API returns 50 headlines per call with a next-page token. Our market-wide windows were never paged. So every one of the "six-hour" market windows in one study held exactly 50 headlines: the newest 30 to 90 minutes, two-thirds of them analyst price-target notes. A real window, fetched properly, held 330 headlines over 7 pages before filtering. The study's "nothing" was withdrawn and re-run. Per-name windows were barely touched: none of the refused-decision rows hit the cap in one study, 2.8% did in another and 4.9% in a third. The fix pages up to $8 \times 50 = 400$ headlines and drops analyst, call and transcript boilerplate by pattern. It still stops silently on a mid-page error and still caps at 400; we note that and leave it. After paging and filtering, nightly windows held 103 to 379 headlines (median 232) and pre-open windows 44 to 317 (median 232, by coincidence the same). One call failed with a server-overloaded error. That night was scored as missing, not as quiet.

Window depth changes the categories. On the truncated feed the market-story reading called 167 nights "quiet"; on the paged feed, 6. Every other category moved too (the central-bank story 167 to 218, inflation data 29 to 86, bonds 16 to 5). On the full pre-open window the regime reading never once called a morning quiet in any session of the era. The first forward regime rows show it in miniature. The unpaged first morning (50 headlines) read earnings 0.45, quiet 0.32, macro 0.23; the next two, paged (237 and 232 headlines), read

macro 0.97 and 0.87. A category is a function of the window.

The clock window is part of the question. The overnight stamps' 72-hour window ended at the wrong clock, earlier than the overnight strategy's own decision. The documentation said "the last 72 hours" before that decision. That was safe (no look-ahead) but it was not the documented input. The window was moved, and the two rows' since-dates moved with it; both had zero forward counts, so nothing was lost. On the same seeded 1,000 unseen name-nights, the old window gave a supportive-minus-none gap of +19.1 basis points (508 and 362 rows) and the honest window +7.0 (529 and 333), with the tone choice agreeing between windows on 85% of rows. Same judge, same rows, same question: the answer depended on the clock.

When news arrives. A sector-wide negative story at probability 0.7 or above appeared in the window before the overnight decision on only about 8% of nights, most in a single sector. A held name sat in a stressed sector on about half of those nights. The sector story that prompted the study was printed only after the position had already been decided. A rule cannot use what is not yet printed.

Firing rate. A trigger that fires on more than a quarter of nights is a size change, not a signal. The quarter is a rule of thumb we wrote down, not a derived number. The reasoning: a flag that is on most of the time mostly changes how much is held on average, and its effect cannot be told apart from simply holding less. The "rates up" reading fired on 39% of nights and the combined macro flag on 41%. Both failed the rule as triggers; it was written down before either run's results came in. The first was dropped, and the second survives only as a size-down insurance idea (Section 5).

Pin the version, re-ask weekly. This is Gate D in practice. The version string the vendor returns is stored with every answer and compared with the pin weekly, on the drift panel; a changed string moves every since-date. We do not know whether the vendor would return an error or substitute silently if the pinned version were retired; checking the returned version on every call would be stronger.

A stamp test must call the stamp. A stray code comment swallowed an assignment in the overnight stamp. Every held name would have raised an error that night. The after-close run would have reported the failure without flagging it, and the board would have shown "no wire yet" indefinitely. The existing test only searched the source text for the function name. The fix:

the test calls the stamp with mocked news and a mocked judge, and any judge step that fails is now flagged.

Unattended means loud on failure, and "documented" is not "wired". The after-close run now flags any early exit, any failing test and any failed update; a health-check line reports stamp freshness. Any forward claim must name, in the same entry, the file that writes it, the health-check line that proves it alive, and the board row that counts it. Three "tracked" items in our own log had no code behind them before this rule was written. A reviewer found a fourth while this paper was being written: the text rows' second looks existed only in their verdict messages, not in the code. It was fixed the same day, before any row reached its first look. The same week, elsewhere in the lab, a failed decision-time quote fetch was stored as if it were data. Same class: a failed fetch written as a result.

News depth. On four sample nights spanning eight years, the share of held names with any headline in the prior three days was just over half in 2016, about a third in 2019, and most of them in 2022 and 2024. The feed is usable from about 2022. Section 6 says what that forbids.

5. What the stamps have shown so far

Small, and honestly measured. Every study below was pre-registered with its bar. "Seen" means the rows had been looked at before, and such looks never decide anything.

Catalyst fact. Passes on a real label (0.981 on 39 rows, Section 2). Its forward row counts the intraday strategy's chosen names: 3 of 40 at October 8, 2026, none flagged. The seen prior on past fills flips sign between the two years of our data: flagged fills did better than clear ones in one year and worse in the other, on very few flagged cases, which is why it is a prior and not a finding. Outcome: none yet.

Per-name overnight tone: small, and below its bar. The tone question was scored on every held name on every night over four years (July 2022 to July 2026; several thousand name-nights), with no errors. A seen window of a few hundred rows from the summer of 2026 was reported apart. On the unseen rows (about half supportive, two-fifths none, the rest against) the supportive-minus-none gap in the position's holding-period return was +12.9 basis points, against a pre-registered bar of 20. So the primary, pre-registered test failed, and the row became a counter. What it has is secondary evidence only: the order was monotonic (against at or below none), the gap was positive in 4 of 5 periods (-4.6, +11.6, +15.4, +11.8, +11.7 basis points),

and a night-block bootstrap put the probability of a positive gap at 0.997. A consistent sign, then, and a small size; not a finding. On a 1,000-row subsample re-read on the honest window, the gap fell from +19.1 to +7.0 (Section 4); the full sample has not been re-read. The seen window ran the other way, a gap of -13.3. An earlier seen look on 229 name-nights had suggested a larger effect (correlation +0.08); the full sample overturned it. That is the pattern the post-hoc rule warns about.

Pre-open news regime. The headline count reproduces it (Section 2). Counter only.

The market story before the overnight decision. The truncated run was withdrawn (Section 4). On the paged run over four years of nights, nothing cleared its pre-registered bars as a trigger. The one reading with a consistent sign fired on 39% of nights and fell to the firing-rate rule. News-conditioned sizing was tried in two shapes. A sector version had almost no input to work with (about 8% of nights stressed, a held name touched on about half of those) and was buried. A macro version reduced variance without improving the mean; it is held as a decade-blocked insurance lead, because a feed that starts in 2022 cannot meet our ten-year rule for overnight rules.

Refused decisions. Stamping the names the intraday strategy's own checks turned down showed a weak lean carried by a handful of recent cases. It is held for its forward count and nothing else.

Headlines during a trade. An exit idea built on headlines that land while a position is open was tested on the intraday strategy's historical fills. A minority of those fills had a headline land while the position was open, about a third of those against the position. The idea failed in the wrong direction: those trades did better, not worse. We do not know why. The sample is small, and our log's own reading is that a headline marks attention on the name rather than a direction; noise is just as likely.

The label outlived the question. The parsed earnings calendar now labels other studies. One refusal idea built on the calendar held on every sample that had been looked at and reversed sign on the one nobody had opened. The same calendar now serves as a risk receipt for the overnight strategy's exposure to report nights (figures withheld). A guidance-or-analyst-action stamp on several hundred rows found the input too rare on the names studied to size by (guidance raised 11, cut 3; upgrades 4, downgrades 3; targets raised 26, cut 13) and was buried. A gap-source reading (does company news, market news or no news explain the opening gap) found nothing stable across years.

Count of verdicts. Table 1 lists every use of the judge to date, 22 rows in all, failures included. A reader needs the full count to judge any single pass, because many

tries make a lucky pass likely (Harvey, Liu and Zhu, 2016).

Table 1. Every use of the text judge and its verdict (verdict words as the log uses them)

Study	What was asked	Sample	Verdict
Fact-check, proxy label	scheduled event	150 name-mornings	FAIL (0.737 vs 0.75)
Fact-check, real label	scheduled event	39 unopened rows	PASS (0.981)
Blind comparison with a chatbot	scheduled event	83 rows	CONTEXT (0.806 vs 0.761 on the proxy label; cannot rank the two)
Memorisation probe	all four catalyst questions, no headlines	6 + 4 mornings; refresh 4 + 3	PASS (0.03, tone none)
Drift panel baseline	tone, event	50 rows	stable (0.96, 0.011)
A refusal idea on the real label	calendar flag	three samples	BURIED (sign flipped unseen)
Overnight theme concentration	theme, one-theme membership	55 seen nights; 0 of 60 forward	seen context only; COLLECTING
Pre-open news regime	regime, risk-off	every session of the era	COUNTER (count baseline reproduces it)
Corporate-action watchdog	action, attention	wider pool, nightly	operations receipt
Scope versus the system's own refusals	scope	46 mornings	buried as a lens
Gap source	news driver	a few hundred rows	NOTHING (year flip)
Headlines during a trade	tone, in-trade window	the historical fills	BURIED (wrong direction)
Report-night exposure (overnight)	calendar flag	2025–26 name-nights	risk receipt (figures withheld)
Drift panel	tone, event	weekly 50 rows	BUILT
Guidance / analyst action	guidance, analyst action	several hundred rows	BURIED (too rare)
Per-name overnight tone	tone	several thousand name-nights	FAIL on the bar → COUNTER (+12.9; +7.0 on a re-read subsample)
Refused decisions	catalyst questions	several hundred rows	weak lean, forward count only
Market story before the overnight decision, unpagged	story, rates, risk-off	four years of nights	WITHDRAWN (feed truncated)
Market story before the overnight decision, pagged	story, rates, risk-off	four years of nights	NOTHING as a trigger; firing rate
Sector-stress sizing	nine sector questions	four years of nights, 4% touched	BURIED
Macro half-size	story, risk-off	four years of nights, 41% touched	risk-only, decade-blocked lead
Regime size-tilt harness study	regime	pre-registered	WITHDRAWN (Gate B)

Verdict words. PASS / FAIL: cleared or missed its pre-registered bar. COUNTER: failed its bar, still counted forward, used for nothing. BURIED: closed, not to be re-tested without new data. WITHDRAWN: cancelled because its input or a prior gate was faulty. NOTHING: no split cleared its bars. COLLECTING: forward count under way. CONTEXT: a comparison with no bar of its own. The five forward text rows' verdict rules have not yet been tested against random flags (Gate F). BUILT: infrastructure in place.

6. Limits, and what would change our mind

What we do not know, and the three things that would restart the count.

- **Small gate sample.** The real-label hearing is 39 rows with 23 event mornings. The chatbot comparison is one sample of 83 rows.
- **One news vendor, one judge vendor.** Headlines come from one data vendor's news API (Alpaca, 2026; Benzinga content); the judge is one commercial model (TypeSafe AI, 2026; Jev, model identifier jev-1.13.0). A vendor can retire a pinned version; the drift panel is our only warning. The gates themselves were built for one lab's constraints, and another lab may need different ones; the results here may not carry to another judge, feed or market.
- **News depth.** The feed is usable from about 2022 (Section 4). Our standing rule for the overnight strategy requires ten years of evidence before any rule

goes live. So no overnight rule built on text can qualify today, whatever it shows.

- **The judge says yes loosely.** On routine coverage it answers higher than the chatbot does (Section 1), so only the confident band is flagged: at 0.8 and above, 48% of proxy-flagged mornings and 3% of the rest in the first run. It half-noticed a deliberately swapped symbol, answering 0.51.
- **Thin coverage at every scale.** 67 of 150 calibration mornings (45%) had no headline in the window; 7 of 46 hearing rows; fewer than half of the live fills; about a fifth of the refused-decision rows; 19.7% of the name-nights in the tone study. A judge cannot read what was not printed, and "no headline" is a large and silent class.
- **The text-row judges are not yet tested against random flags.** Only the price row's judge is; the looks of every judge row are enforced in code since October 9 (Gate F).
- **The chatbot was not probed.** Its memory of events is unmeasured, and the comparison was scored against the proxy label, so its higher score cannot rank the two models.
- **No calibration analysis.** The flag bands (0.8 for events, 0.7 for sector stress, 0.6 for the watchdog) were set by judgement, the 0.8 from the first run's split on the proxy label. AUC cannot justify a band, and we have no reliability curve.
- **Headlines are untrusted text.** We have not tested adversarial, injection-style headlines. The typed interface limits what a planted line can do (a probability, or one word from a fixed list), and no live code reads the output, but a planted headline could still bias a stamp.
- **The gates are not yet validated as detectors.** The natural test is to put a general chatbot through all six and see which it fails. A standing symbol-swap probe would be a natural seventh gate.
- **Pre-registration rests on our own records.** Bars and questions are dated in a private research log and a private version-control history, so readers must take those timestamps on trust. Public timestamps (a registry, or published commit hashes) would make them checkable.
- **Reproducibility is limited by design.** The trading rules, the theme list, the sector descriptions and several counts are withheld to protect the live system. The vendors and the model identifier are real names, not anonymized.
- **A human label sheet was prepared and not scored.** Thirty real mornings were laid out for a human to label with the judge's answers hidden. The

human answers are not in the log, so we have no human-versus-judge agreement rate. It is the first thing a replication should add. Our prediction, stated before any human answer exists: agreement should be high on the event fact, as it was with the chatbot ($r = 0.90$), and lower on tone, where readers differ more.

- **What would change our mind.** A forward row reaching its pre-written bar at its named count (40 picks, 40 cases, 60 nights, 300 name-nights, 40 picks), read once there; a drift-panel failure, which moves every since-date and restarts every count; a later probe the judge fails.

7. Rules for someone doing the same

Ten rules, each paid for by one of the sections above.

1. **Fact before outcome.** Show the judge can read a fact against a real label before asking whether the fact predicts anything. Ours went from 0.737 under one setup to 0.981 under another (a real label, a longer window, empty rows dropped), so the gain cannot be pinned on the label alone; building the real label was still the work we could not skip.
2. **The dumbest baseline comes from the judge's own input.** Count the headlines. If the count is within a few hundredths of the judge, the judge added nothing. Ours cleared it by 0.092, a margin 39 rows cannot confirm.
3. **Probe for memory.** Famous mornings, headlines removed, real dates kept; the answer must equal the quiet controls.
4. **Pin the version and re-ask weekly.** The instrument's noise is the smallest effect you can see; a changed version restarts every count.
5. **Page the feed and state the window.** A capped feed turns "six hours" into "the last hour"; the window's clock is part of the question.
6. **Mind the firing rate.** A trigger on more than a quarter of observations is a size change in disguise.
7. **Test the forward judge against random flags** before it ships; read it at fixed counts with its power stated. We have done this for one row so far.
8. **A stamp test calls the stamp.** Searching the source for a function name tests nothing.
9. **Documented is not wired.** Name the file that writes the tracker, the health check that proves it alive and the row that counts it, in the same entry.
10. **The judge never touches a number.** No thresholds, no sizes, no edge validation, no parameter search.

Appendix A. The question texts

Every question lives in one source file; a changed question is a new forward item. Each question and each bar was written before the study that used it and recorded in a dated registry. Some questions are paraphrased here; the others are given as asked. Variable names in back-ticks are filled per call.

Catalyst stamp (the intraday strategy's chosen name; also the names its checks turned down), 18-hour window.

- *scheduled_event* (probability): "Do the `headlines` report or anticipate a SCHEDULED company event for `symbol` within the last day — an earnings release, guidance update, investor day, deal or merger announcement, regulatory or FDA decision, product launch?" True: a specific scheduled corporate event for this company is reported, imminent or just occurred. False: only general market commentary, analyst notes, or routine coverage. Flag at 0.8.
- *tone* (choice): "Taken together, what is the news tone for `symbol` itself (not the market)?" supportive / against / none.
- *scope* (choice): "Is the coverage of `symbol` about this company specifically or about a market-wide / sector-wide story it is mentioned in?" `stock_specific` / `market_wide`.
- *news_driver* (choice, paraphrased): what, if anything, in the `headlines` would explain a notable move in `symbol` today? `company_news` / `macro_or_market_news` / `no_meaningful_news`.

Overnight stamp (each held name), 72-hour window ending at that strategy's decision. A theme choice over nine sector-style themes plus "other" (the list is withheld), a yes/no on whether the name is currently trading as part of one named investment theme, and, as a separate call, the *tone* question above. The concentration count that makes a "one-theme night" is withheld. Bar for the tone row: a supportive-minus-none gap of 20 basis points.

Pre-open regime stamp (the market), 12-hour window.

- *regime* (choice): "Judging from the pre-open `headlines`, what is driving the US stock market this morning?" `macro_driven` (rates, the central bank, tariffs, geopolitics, economic data or a broad sell-off / rally dominate) / `earnings_driven` / quiet.
- *risk_off* (probability): "Is the tone of the pre-open `headlines` risk-off (fear, selling, hedging) rather than risk-on?"

Corporate-action watchdog (a wider pool), 24-hour window. An *action* choice over six classes: none; ticker change or rename; delisting or take-private; merger or acquisition announced; split or spin-off; bankruptcy or halt. And an *attention* yes/no, paraphrased: should a system that trades this name review it before its next order? Flag at 0.6.

Market-story stamp (the hours before the overnight decision), used in the paged and unpagged studies. A *story* choice over eight classes (central bank and rates; bonds and yields; inflation and economic data; tariffs and trade; geopolitics; earnings; AI and technology; quiet), a *rates_direction* choice (up / down / none), and the *risk_off* question above.

Sector-stress stamp. Nine yes/no questions, one per broad US equity sector, each asking whether the day's market headlines report sector-wide negative pressure; single-company items do not count. Flag at 0.7. The sector descriptions are withheld because they name companies.

Guidance and analyst stamp (a buried study). *guidance* (raised / cut / none) and *analyst_action* (upgrade / downgrade / target_raised / target_cut / none), each "within the last day".

Earnings-timing question (calendar cross-check only). "According to the `schedule_text`, when does `symbol` report?" before_open / after_close / unknown.

In all, 27 question slots and 25 distinct texts.

Appendix B. The probe and gate protocols

1. **Fact-check.** Stratify by the available label (all flagged plus a random draw of unflagged; ours 50 + 100), fix the window (18 hours), fix the bar ($AUC \geq 0.75$), run once. Report zero-headline rows separately. If the label is suspect, build a real one and hold a second, pre-registered hearing on rows nobody has opened (ours 23 + 23, 36 hours, 39 scored).
2. **Dumb baseline.** Compute the same AUC from the judge's own input alone (headline count). Require a margin; ours was 0.092 on 39 rows. For outcome studies, tercile the input count and check whether the judge's split survives inside each tercile.
3. **Memorisation probe.** Widely reported event mornings with headlines removed and real dates kept, plus quiet controls, each asked twice. Require the event rows to equal the controls (ours 6 + 4, then 4 + 3; all 0.03).
4. **Repeatability.** Re-ask identical input; record the spread (ours about ± 0.05 on confident answers and up to 0.07 on mid-range ones; on a 50-row panel 0.96 tone agreement and 0.011 mean change).

5. **Drift.** Pin the version; re-ask a fixed panel weekly; declare drift below a pre-set agreement (0.90); flag it; move every dependent since-date.
6. **Forward judge.** Give every row a permutation null (flagged mean at or below the 5th percentile of random same-size flag sets), a minimum flagged count (12), and fixed looks (60, then 120 and 180); pin by test that no more than 3 of 20 random sets integrate.
7. **Window and paging.** Fix each window's length and end clock as part of the question; page the feed (8×50) and filter boilerplate; report window sizes.

Use of AI tools

During the preparation of this work, the authors used TypeSafe's Jev (version jev-1.13.0), a typed text-classification model, in order to produce the headline readings this paper studies; every such reading was generated by the authors' own scripts and is reported as data. The authors also used Claude (Anthropic): Claude Fable 5.1 in order to help write the project's code and research log, run the logged analyses, serve as the blind comparison judge in Section 1 and draft this text, and Claude Opus 5.5 in order to review and revise the draft and check its numbers. The authors chose the research questions, set the gates and bars with the AI assistant (before the data they judged, except where Section 2 says otherwise), and made the keep-or-kill decisions. After using these tools, the output was reviewed, edited and verified by the authors, who take full responsibility for the content.

References

- Alpaca (2026). Historical news data. <https://docs.alpaca.markets/docs/historical-news-data> (accessed October 9, 2026).
- Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5), 458–471. <https://doi.org/10.1090/noti1105>
- Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2019). Data validation for machine learning. In *Proceedings of Machine Learning and Systems (SysML 2019)*.
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Glasserman, P., & Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. Working paper, arXiv:2309.17322.
- Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *Journal of Finance*, 72(4), 1399–1440.
- Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68. <https://doi.org/10.1093/rfs/hhvo59>
- Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. Working paper, arXiv:2304.07619 (revised 2025). <https://doi.org/10.2139/ssrn.4412788>
- Phipson, B., & Smyth, G. K. (2010). Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), Article 39. <https://doi.org/10.2202/1544-6115.1585>
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28* (pp. 2503–2511).
- TypeSafe AI (2026). Jev documentation. <https://docs.typesafe.ai/> (accessed October 9, 2026). Model identifier jev-1.13.0 as passed to the vendor's API.
- Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36* (Datasets and Benchmarks Track).