The American Journal of Engineering and Technology
ISSN 2689-0984 Volume 08 - 2026

Conformal Gating Of Extracted Factor Vectors In LLM Signal Pipelines Under Delayed Labels And Regime Shift: A Simulation Study

Wrick Talukdar
AWS AI & ML, IEEE CIS USA

Gowtham Shankar
AWS AI, IEEE USA

Prasad Prabhu
IEEE USA

Received: 11 Oct 2026 | Received Revised Version: 18 Oct 2026 | Accepted: 25 Oct 2026 | Published: 01 Nov 2026

Volume 08 Issue 11 2026 | DOI: 10.37547/tajet/Volume08Issue11-01

Abstract

Language model pipelines extract multi-dimensional factor vectors from financial text, but a trading desk needs to know when an extracted vector can be trusted. Audit labels arrive late, errors cluster in stress regimes, and error scale shifts after model or prompt updates. We evaluate a conformal gate on the extraction layer in a seeded simulation (300 firms, 60 quarters, 200 replications) with heavy-tailed errors, regime stress, a structural break and a two-quarter audit lag. Per-dimension 90% intervals cover the full eight-dimension vector only 40 to 45% of the time, while a max-score construction reaches 89.7% before the break with intervals 18 to 26% narrower than Bonferroni. Calibration frozen at go-live falls to 78.2% joint coverage after the break. Rolling, weighted and delayed-feedback adaptive calibration recover 88.0 to 89.9% on average, but no label-driven method protects the quarters just after a shift. As a risk control, the adaptive gates reject most signals after the break, accepting only 9 to 14%, yet the error rate among accepted signals is still about 20%, against 28% without a gate, and gating lowers risk-adjusted return in this design. A label-free drift monitor flags 78.5% of abrupt breaks in the quarter they occur, against 3.0% for label-based monitoring, but only 3.5% under gradual drift, so its advantage is specific to abrupt breaks. To test whether these mechanisms appear with a real extractor, we run a pilot that extracts seven financial fields from the earnings press releases of 492 large US companies from an S&P 500 earnings-call dataset with an open-weight model and scores each figure against the SEC XBRL-reported value. Extraction is accurate for most figures, 96.3% within 1% once the reference labels are corrected and verified. Fixing our choice of revenue tag and checking every large-error label against the press release show that 38 of the 81 figures first scored as large errors rested on wrong labels rather than misreadings. The one self-reported signal collected, the model's GAAP flag, is constant and uninformative, while disagreement across three models detects 59% of the remaining large errors, and 74% of those confirmed as extraction errors, at the cost of flagging 8.6% of correct figures, which is consistent with the view that a usable confidence signal must come from a check outside a single model's output, although richer self-reported signals were not tested. A Monte Carlo resampling calibrated to the pilot, with the detector's operating point held fixed, gives sampling intervals for the accuracy and large-error rates. The results suggest three design rules: report joint reliability with a max-score rectangle, gate on the composite the operator consumes, and size positions by a smoothed scale rather than by calibrated width. We place the gate in a deployment framework and state its limits.

Keywords: conformal prediction; large language models; signal verification; distribution shift; delayed feedback; model risk

© 2026 Wrick Talukdar,Gowtham Shankar,Prasad Prabhu. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The authors retain copyright and allow others to share, adapt, or redistribute the work with proper attribution.

Cite This Article: Wrick Talukdar, Gowtham Shankar, & Prasad Prabhu. (2026). Conformal Gating Of Extracted Factor Vectors In LLM Signal Pipelines Under Delayed Labels And Regime Shift: A Simulation Study. The American Journal of Engineering and Technology, 8(11), 01–23. https://doi.org/10.37547/tajet/Volume08Issue11-01

1. Introduction

Language models can turn earnings-call transcripts and filings into structured factor vectors that feed quantitative strategies. Ye et al. (2026) argue that reported alpha from end-to-end LLM trading agents should not be read as deployment evidence until it passes structural validity tests, including temporal integrity and predictive calibration. This paper addresses a narrower question that sits upstream of any trading rule. When is an extracted factor vector reliable enough to pass to a downstream operator, and how should that decision be calibrated when the conditions behind it keep changing?

Four features of production extraction make the standard recipe for conformal calibration unreliable.

The target is a vector. A pipeline emits several factor dimensions per firm-quarter. Intervals that are valid for each dimension separately can have much lower joint coverage, so a per-dimension gate overstates how often the whole vector is right.

Exchangeability fails. Earnings calls arrive in time order, error scale clusters in stress regimes, and a model or prompt update can change error behavior overnight. Split conformal guarantees assume calibration and test points are exchangeable (Vovk et al., 2005; Lei et al., 2018).

Labels are late and scarce. Ground truth for an extraction comes from an audit, which covers a small share of outputs and lags the extraction by one or more quarters. Methods that need immediate feedback cannot be applied unchanged.

The consumer matters. Singh and Rezaeimanesh (2026) report that a confidence map and the sizing scale fitted to it are not portable alone, and that the gain from matching them shrinks under an adaptive volatility controller. A gate should therefore be evaluated together with the operator that uses it, not only by coverage.

Conformal methods for language models exist, mostly for question answering and generation (Quach et al., 2024; Wang et al., 2026), and conformal methods for time series and distribution shift are well developed (Gibbs & Candès, 2021; Barber et al., 2023; Zaffran et al., 2022). We did not find an evaluation that combines these ingredients for gating extracted factor vectors under audit lag, regime stress and a structural break, together with a downstream operator. That combination is the subject here.

Contributions.

A gate specification that defines the coverage target and the label source, and distinguishes a factor-level joint score from a composite-level score matched to the quantity the operator consumes (Section 3).

A comparison of four calibration protocols under audit lag: frozen, rolling-window, exponentially weighted, and rolling with delayed-feedback adaptive miscoverage (Sections 3.5 and 4). The algorithms are known; the contribution is their evaluation in this setting.

A seeded simulation study with 200 replications, sensitivity runs over label lag, audit rate and adaptation rate, and robustness runs that relax five assumptions of the base design: independent errors, an abrupt break, a correctly specified scale model, random error-free audits and a return model in which extraction error carries no direct cost (Sections 3.9, 3.10, 4.6 and 4.7). All code is provided as supplementary material.

Results that include negative findings: in our simulation the gate lowers risk-adjusted return, and sizing positions by conformally calibrated width underperforms sizing by the raw scale estimate. A label-free drift monitor detects an abrupt simulated break far earlier than label-based coverage monitoring, but not a gradual drift.

A small pilot on public earnings press releases of large US companies, scored against SEC XBRL-reported actuals with the label lag observed from filing dates (Sections 3.11, 4.8 and 4.9).

A three-component deployment framework, three design rules distilled from the results, and a replication protocol for real filings (Sections 5 and 6).

Scope. The main results come from simulated data, and the simulation encodes assumptions about error structure that real extraction may not satisfy. Section 4.8 reports a small pilot on public earnings press releases that checks the extraction layer and candidate uncertainty signals with a real extractor and real labels. The paper does not claim real-world alpha, coverage or economic benefit, and Section 5.3 lists the threats to validity.

2. LITERATURE REVIEW

Conformal prediction beyond exchangeability. Split conformal prediction gives finite-sample marginal coverage when calibration and test data are exchangeable (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023). Tibshirani et al. (2019) handle covariate shift with likelihood-ratio weights, and Barber et al. (2023) bound the coverage loss under non-exchangeable data and justify weighted calibration, including exponentially decaying weights. Gibbs and Candès (2021) adapt the miscoverage level online from realized errors, Gibbs and Candès (2024) remove the need to tune the adaptation rate by aggregating several rates, and Angelopoulos et al. (2023) frame online calibration as a control problem. Xu and Xie (2021) and Zaffran et al. (2022) study intervals for dependent time series. Standard adaptive inference assumes the outcome of each issued interval is revealed before the next one. El Halabi and Brandt (2026) study adaptive conformal inference when every outcome arrives after a fixed delay, derive a long-run coverage bound that depends explicitly on the delay, and show that the effect of delay depends on how long residual information stays relevant. Our delayed-feedback variant is closely related but updates on a small audit sample under regime stress and a structural break, a setting those bounds do not cover. We use rolling, exponentially weighted and adaptive variants as baselines.

Multivariate conformal prediction. Messoudi et al. (2021) use copulas to build joint regions for multi-target regression that account for dependence between targets. Our max-score rectangle is the simplest member of this family. It is valid under exchangeability whatever the dependence, but it can be wider than a copula region when errors are strongly correlated. Section 4.7 reports its coverage under correlated errors but does not compare its width with a copula region.

Risk control and selective prediction. Bates et al. (2021) and Angelopoulos et al. (2024) control the expectation of a monotone loss, which is the right frame for a gate that accepts or rejects signals. Our selective-risk results are empirical and are not instances of those guarantees.

Conformal methods for language models. Quach et al. (2024) calibrate stopping and rejection rules for generated answer sets. Wang et al. (2026) use layer-wise internal-representation scores as nonconformity scores for question answering and report advantages under cross-domain shift. These works concern answer correctness. We study numeric factor extraction with a continuous error and a downstream consumer.

Validity of LLM trading evidence. Ye et al. (2026) take the position that reported alpha from end-to-end LLM trading agents should not be treated as deployment evidence until it passes structural tests, and propose a modular alternative in which language models are not the final source of trading authority. Our work follows that modular stance but focuses on one verification layer. Singh and Rezaeimanesh (2026) show on FactSet news that confidence maps and exposure scales should be evaluated together, with a gain of 9.2 percentage points per year under frozen-scale control that falls to 1.6 under an adaptive volatility controller. We test a related coupling in simulation and reach a different conclusion about sizing (Section 4.4).

Gap. We did not find prior work that evaluates joint conformal coverage over an extracted factor vector under delayed audit labels, regime stress and a structural break, together with the sizing operator that consumes the result. We did not conduct a systematic review, so related work may exist that we have not seen.

3. MATERIALS AND METHODS

This section specifies the gate (Sections 3.1 to 3.8), the simulation used to evaluate it (Section 3.9), the robustness designs (Section 3.10) and a pilot on public filings (Section 3.11).

3.1 Setting, labels and coverage target

For firm f and quarter t the pipeline emits an extracted vector $\hat{z} \in \mathbb{R}^d$ ($d = 8$). The target is the *audited* value z^*, obtained by a reviewer who reads the source document, so coverage concerns extraction accuracy and not market outcomes. The extraction error is $e = \hat{z} - z^*$. A random share a of firm-quarters is audited, and the labels for quarter t arrive in quarter $t + L$. A pre-deployment labeling campaign covers the quarters before go-live. The coverage target is $P(z^* \in C(x)) \geq 1 - \varepsilon$, with $\varepsilon = 0.10$.

3.2 Scale model and scores

An observable uncertainty proxy v (for example a verbalized-uncertainty or self-consistency spread) feeds a scale model $\hat{\sigma}_i(v) = \exp(a_i + b_i v)$, fitted by regressing $\log|e_i|$ on v using only the pre-go-live labeled data. The normalized score of dimension i is $s_i = |e_i| / \hat{\sigma}_i$.

3.3 Joint factor-level interval

Three constructions are compared on the same rolling calibration set:

Per-dimension: each dimension receives its own $1 - \varepsilon$ quantile. The joint coverage is only bounded below by $1 - d\varepsilon$ by the union bound.
Bonferroni: each dimension receives a $1 - \varepsilon/d$ quantile.
Max-score: one score $s = \max_i s_i$ and one quantile q, giving the hyper-rectangle $\hat{z}_i \pm q\ \hat{\sigma}_i$. Under exchangeability the joint coverage is at least $1 - \varepsilon$.

3.4 Composite interval matched to the operator

The operator consumes a weighted composite $c = \beta' \hat{z}$ with weights β (summing to one). Its error is $e_c = \beta' e$ and its scale is $\hat{\sigma}_c = (\Sigma\ \beta_i^2\ \hat{\sigma}_i^2)^{1/2}$, which treats errors as independent. This is an approximation, and the calibrated quantile absorbs the resulting scale error. The score is $s_c = |e_c| / \hat{\sigma}_c$, and the interval half-width is $h = q_c\ \hat{\sigma}_c$.

3.5 Calibration protocols

The quantile q is the $\lceil (n+1)(1-\alpha) \rceil$-th smallest calibration score. Four calibration sets are compared (Table 1).

Table 1 Calibration protocols compared in this study

Protocol	Calibration set at quarter t
M0 Frozen	Pre-go-live labels only, never updated
M1 Rolling	Audited points from quarters $t - L - W + 1$ to $t - L$, with $W = 8$
M2 Weighted	All available audited points, weight $0.5^{age/4}$, with test-point mass at $+\infty$ (Barber et al., 2023)
M3 Rolling + delayed ACI	As M1 but at level $\alpha_t = \alpha_{t-1} + \gamma(\varepsilon - err_{t-L})$, where err_{t-L} is the miss rate of the intervals issued in quarter $t - L$ among that quarter's audited points. α_t is clipped to [0.01, 0.5] and $\gamma = 0.10$

M3 is the adaptive conformal inference method of Gibbs and Candès (2021) with the update applied *L* quarters late. El Halabi and Brandt (2026) give a long-run coverage bound for adaptive conformal inference with a fixed delay

when every outcome is eventually observed. M3 updates on a 10% audit sample, so we do not claim that bound carries over and treat coverage under M3 as an empirical result.

3.6 Gate and selective risk

A signal is accepted if $h \leq \tau$. The tolerance τ is set at go-live as the 75th percentile of M1 half-widths. A violation is $|ec| > \tau$. Marginal coverage does not imply a guarantee on the accepted subset, so we measure P(violation | accepted) directly and compare it with ε as a reference level. The comparison is a benchmark, not a certified bound.

3.7 Operators

Operator A (rank) goes long the top 20% and short the bottom 20% by c, equal weighted, among accepted firms. It uses c and the acceptance set only, so a monotone change to widths cannot change it except through acceptance. *Operator B (width-scaled)* sets weights $w_f = (c_f - \bar{c}) / (\lambda\ h_f)$ among accepted firms, with λ fixed so that mean gross exposure is one over the pre-break quarters under M1. Operator B under M0 equals sizing by $\hat{\sigma}_c$ alone, because q is constant over time.

3.8 Drift monitors

The *label-free* monitor runs a two-sample Kolmogorov–Smirnov test of the first differences of $\hat{z}$ in quarter *t* against the previous eight quarters, for each dimension, with a Bonferroni correction across dimensions at level 0.01. The *label-based* monitor applies a one-sided binomial test of the miss count among the audited points of quarter $t - L$ under the frozen intervals against ε, at level 0.01.

3.9 Simulation design

We use simulation because ground-truth audited values, controlled error structure and a known break are needed to measure coverage honestly, and because we had no access to licensed audit data. The simulated world is stylized. Its parameters were set by judgment before the full run and not changed afterward.

Models and data used. The main study runs no language model and uses no real dataset. Both the extractor and the data are simulated. Section 3.11 describes a separate pilot on public earnings press releases and SEC XBRL values. **Table 2** summarizes the components.

Table 2 Models and data used in the study

Item	What was used
Extraction model	A simulated extractor, not an LLM. Extraction error is generated as $e = \sigma \cdot T$ with T Student-t (5 d.f.), plus a 2% per-entry hallucination term (Table 3). There is no model name, version or prompt.
Uncertainty model	A log-linear scale model $\hat{\sigma}_i(v) = \exp(a_i + b_i v)$, fitted by regressing $\log\|e_i\|$ on an observable proxy v, using pre-go-live labeled data only (Section 3.2).
Calibration methods	Split conformal with a max-score joint score and a composite score, under four protocols: frozen, rolling, exponentially weighted, and rolling with delayed-feedback adaptive miscoverage (Section 3.5).
Dataset	A fully synthetic panel of 300 firms, 60 quarters and 8 factor dimensions, with AR(1) true factors, regime stress, a structural break and random audits (Table 3). Main study seeds 1000 to 1199, sensitivity study seeds 5000 to 5059.
Real data	None in the main study. The pilot (Section 3.11) uses public earnings press releases and SEC XBRL values.
Pilot extraction model	gpt-oss-120b (about 117B parameters, Apache 2.0), with Llama 3.3 70B and gpt-oss-20b as a cross-model panel. Fixed prompt and temperature-zero decoding (Section 3.11).
Pilot source text	Earnings press releases (8-K exhibit 99.1) for 492 companies, drawn from the S&P 500 Earnings Transcripts universe (Kurry, 2025), Hugging Face, MIT license. Calls from January 2024 to March 2025.
Pilot reference values	XBRL-tagged values from the SEC companyfacts API, matched by CIK and fiscal period; reported actuals only (revenue, gross profit, operating expenses, operating income, net income, diluted EPS, cash and equivalents).
Pilot label lag	Observed: SEC filing date of the matching 10-Q or 10-K minus the call date.

Table 3 gives the data-generating process.

Table 3 Data-generating process of the simulation

Component	Specification
Panel	300 firms per quarter, 60 quarters, 8 factor dimensions
True factors z^*	AR(1) per firm and dimension with $\rho = 0.5$, unit variance, cross-dimension correlation 0.3
Extraction error	$e = \sigma \cdot T$, with T Student-t (5 d.f.) rescaled to unit variance
Error scale σ	$s_i \cdot \exp(0.5 \cdot u) \cdot m_t \cdot b_{t,i}$; s_i rises linearly from 0.25 to 0.45 across dimensions; $u \sim N(0,1)$ is latent difficulty
Observable proxy	$v = u + N(0, 0.5^2)$, available at extraction time and used by the scale model
Regimes	Two-state Markov chain, P(calm→stress) = 0.08 and P(stress→calm) = 0.25 per quarter; stress multiplies σ by 1.6 (m_t)
Structural break	From quarter 36: $\sigma \times 1.8$ and a bias of $+0.4\sigma$, both on dimensions 1 to 4 ($b_{t,i}$), simulating a model or prompt update
Hallucination	Each dimension with probability 0.02 receives extra error $N(0, 3^2)$
Audits	30% of firm-quarters in the pre-go-live campaign (quarters 0 to 11), 10% afterward, labels available L = 2 quarters later
Forward return	$r = k \cdot c^* + N(0,1)$, with k set so the cross-sectional IC of the true composite $c^* = \beta' z^*$ is 0.10
Composite weights β	(1.0, 0.8, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1) normalized to sum to one
Calibration	$\varepsilon = 0.10$, W = 8 quarters, half-life 4 quarters, $\gamma = 0.10$

Deployment covers quarters 12 to 59, of which 12 to 35 (24 quarters) precede the break and 36 to 59 (24 quarters) follow it.

Replications and uncertainty. The main study uses 200 independent seeded replications (seeds 1000 to 1199). All reported figures are means across replications with 95% intervals (±1.96 standard errors across replications). Coverage is evaluated on every firm-quarter because the audited truth is known in simulation. A separate sensitivity study uses 60 replications per setting (seeds from 5000), varying audit lag, audit rate and the adaptation rate γ.

Metrics. Joint coverage is the share of firm-quarters whose whole eight-dimension vector lies inside the rectangle. Composite coverage is the share whose composite error lies inside the composite interval. Worst-window coverage is the minimum four-quarter moving average after the break, averaged across replications. For the gate we report the acceptance rate and P(violation | accepted). For operators we report a per-quarter risk-adjusted ratio, the mean of quarterly long-short returns divided by their standard deviation over the 48 deployed quarters. It is not annualized and is expressed in units of the idiosyncratic return standard deviation, so it must not be read as a real Sharpe ratio. For drift monitors we report per-quarter false-alarm rates in calm and stress quarters before the break, and the probability of an alarm within *k* quarters of the break.

3.10 Robustness designs

To test whether the main findings depend on assumptions built into the base design, we specify five robustness designs, each changing one element of the base design and leaving the parameters of Table 3 otherwise unchanged. The designs were fixed before any were run, and all five are reported in Section 4.7.

R1 Correlated errors*:* the unit t-variates share a common factor, standing in for transcript or document quality, so that the error correlation across dimensions is $\rho e \in \{0.3, 0.6\}$. This tests the joint constructions and the independence approximation in the composite scale (Section 3.4).

R2 Gradual drift: instead of the abrupt break at quarter 36, the scale multiplier and bias on dimensions 1 to 4 rise linearly from quarter 30 to quarter 42 to the same end values ($\times 1.8$ and $+0.4\sigma$). The change is small in any single quarter, which tests both the calibration protocols and the label-free monitor.

R3 Misspecified scale model: the error scale depends on latent difficulty through a piecewise function (flat below the median of u, exp(u) above it), and the proxy is weaker, $v = u + N(0, 1.0^2)$, while the fitted scale model keeps its log-linear form.

R4 Error-linked returns: forward returns are $r = k \cdot c^* - \kappa \cdot (\hat{c} - c^*) + N(0,1)$ with $\kappa \in \{0.5k, k, 2k\}$, so a position taken on an overstated composite loses money. This removes the property of the base design in which a bad extraction only dilutes the signal.

R5 Targeted, noisy audits: half of each quarter's audits are drawn from the top quintile of $\hat{\sigma}c$, and audited values carry reviewer noise $N(0, (0.1\sigma)^2)$. Calibration weights audited points by inverse inclusion probability.

A sweep of the acceptance tolerance τ over the 50th, 75th and 90th percentiles of go-live M1 half-widths was also specified. Its results are in the supplementary material and are not reported here for space.

3.11 Real-data pilot: data sources and label construction

The pilot checks whether the qualitative findings of the simulation appear with a real extractor, real source text and real labels. It is small and does not estimate production performance. It combines two public sources: earnings press releases filed with the SEC as the text the extractor reads, and values reported to the SEC in XBRL as the reference against which extractions are scored.

Source text. The firm and period universe is the S&P 500 Earnings Transcripts dataset (Kurry, 2025), distributed on Hugging Face under the MIT license. It contains 33,362 earnings calls for 685 S&P 500 constituents and other large US companies, from 2005 to 2025, each giving the ticker, company name, fiscal year, fiscal quarter and call date. We use the version identified by Hugging Face commit f3ded37. The extractor does not read the transcript. For each call we fetch the earnings press release the company furnished to the SEC as exhibit 99.1 of the results 8-K filed within seven days of the call, and the extractor reads that release. The universe is every company with a call between January 2024 and March 2025, using its latest call in that window, with one company that is the authors' employer excluded, giving 521 companies. Of these, 492 had a ticker in the SEC ticker file, a results 8-K within seven days of the call and a matching 10-Q or 10-K. Of the 29 companies excluded, 17 had a ticker that is not in the SEC ticker file, 11 had no results 8-K within seven days of the call, and 1 had no matching 10-Q or 10-K.

Reference values. Labels are the XBRL-tagged values the filer reported in its 10-Q or 10-K, retrieved from the SEC. In the run reported here the values were pulled from the SEC companyfacts API, and the SEC Financial Statement Data Sets (Securities and Exchange Commission, n.d.-a) republish the same tagged values as quarterly files and can be used interchangeably. Each value carries its taxonomy tag, its reporting date and the number of quarters it covers. These values are reported by the filers themselves and the SEC does not guarantee their accuracy, so they are a reference standard rather than an independent audit.

Label alignment. Each call is matched to the filing for the quarter it discusses in three steps.

Firm. The ticker is mapped to a CIK with the SEC's ticker to CIK file (Securities and Exchange Commission, n.d.-b). Tickers absent from that file, typically because they changed or were reused over the sample period, are excluded rather than resolved by hand.

Period. The call is matched to the first 10-Q or 10-K whose period end date is the last fiscal quarter end before the call date. The fiscal year and quarter fields in the call data are not used as the key, because fiscal calendars differ across firms.

Value. Only reported actuals for the quarter can be checked against XBRL. We use seven items: revenue, gross profit, operating expenses, operating income, net income, diluted EPS and cash and equivalents, each with a fixed US GAAP tag list in priority order. For revenue the broad Revenues tag is preferred over RevenueFromContractWithCustomerExcludingAssessedTax, which excludes interest, lease and premium income

at banks, REITs and insurers (Section 4.8). This priority changed 46 revenue labels. Every label behind a large error (above 5%) under either tag priority, 86 in all, was then checked by hand against the company's press release, 81 of them without reference to the model's extraction (Section 4.8), and the evidence for each is in the supplementary material. Quarterly values are taken from tagged facts that cover one quarter. For fourth fiscal quarters, which 10-K filings usually tag only for the full year, the quarterly value of flow items such as revenue and net income is the annual value minus the nine-month year-to-date value from the third-quarter 10-Q. Per-share items are not additive across quarters, so fourth-quarter diluted EPS is used only where the 10-K tags a quarterly value, and is otherwise excluded. Where a value is later restated, the value in the original filing is the label, because that is what was available when the label arrived.

Guidance, segment detail and non-GAAP measures such as adjusted EPS usually have no XBRL value, so they are not extracted. The extractor is asked for GAAP figures only. Calls often lead with non-GAAP numbers, so returning an adjusted figure where the GAAP figure was asked for is scored as an extraction error. That is a realistic failure mode, distinct from the reference-label errors discussed in Section 4.8. Coverage in Table 10 is the share of reference figures that the extractor stated.

Label lag. The simulation imposes a lag of two quarters. In the pilot the lag is observed. The label for a call becomes available on the date its 10-Q or 10-K was filed with the SEC, and the lag is that date minus the call date. The observed lag has a median of 6 days and an interquartile range of 0 to 16 days, ranges from minus 1 to 89 days, and is negative for 2.8% of calls where the filing came first. Because labels arrive well within a quarter, the pilot tests the protocols under a short lag, which complements the longer lags in Section 4.6.

Audit rate. XBRL gives a label for every matched firm-quarter, while the simulation assumes that only a sample is audited. Because the pilot tests extraction accuracy and the uncertainty signal rather than audit sampling, we report it with all labels revealed.

Extractor and uncertainty proxy. The main extractor is the open-weight model gpt-oss-120b (about 117 billion parameters, Apache 2.0 license), served through Amazon Bedrock, run with a fixed prompt and temperature-zero decoding through a fixed tool-use JSON schema. Decoding at temperature zero was not fully reproducible. On a complete rerun, 2,562 of the 2,672 figures stated in both runs were identical. The other 110 differed by a median of 4.2%, 53 of them by more than 5%, and a further 139 figures were stated in only one of the two runs. Results are reported for the original run, whose outputs are in the supplementary material. The prompt asks for each item as stated in the release together with its scale (units, thousands, millions or billions), which the code converts to US dollars, or dollars per share for EPS, and for null when the item is not stated. Errors are signed relative errors against the XBRL value, $(\hat{z} - z^*) / |z^*|$. Accuracy is the share of extracted figures whose absolute relative error is at most 1%; figures the extractor did not state count against coverage rather than accuracy, and no reference value is zero, so the relative error is always defined. Large errors are those above 5%. The uncertainty proxy is the disagreement across three open-weight models that read the same release, gpt-oss-120b, Llama 3.3 70B and gpt-oss-20b, flagged when the spread of their extracted values for a figure is nonzero. The prompt, settings and extraction outputs are in the supplementary material, together with the raw model replies of a complete rerun.

Cross-model panel. The pilot covers a single reporting period, so it does not test a temporal model-version shift. Instead, each figure is read by three models, and their disagreement serves as the uncertainty signal. The set of three models was fixed before the results were inspected.

Contamination. Press releases from before the extractor's training cutoff may have been seen during pretraining, so the model could recall reported figures rather than read them. The main model's stated knowledge cutoff (June 2024; OpenAI, 2025) precedes most of the sample, so to check for leakage we also report accuracy on the 476 companies whose calls are dated after the cutoff, 96.4% against 96.3% for the full sample, and we note that extraction reads the release text directly rather than relying on recall.

Scope. The pilot runs no trading operator and uses no returns. It reports extraction coverage and accuracy, the observed label lag and the detection power of candidate uncertainty signals. No calibrated gate, shift test or monitor is run on the pilot; Section 4.9 evaluates only a simple cross-model rejection rule in simulation.

4. RESULTS

All figures are means across 200 replications and intervals are 95% across replications. The results describe the simulated world only.

4.1 Coverage after a structural break

Figure 1 plots joint coverage by quarter for the four calibration protocols.

Before the break all four protocols hold joint coverage close to the 90% target (89.3 to 90.1%, averaged over quarters 12 to 35). After it, the frozen calibration set averages 78.2%, its worst four-quarter window averages 66.4%, and in post-break stress quarters it falls to 61.7%. Rolling, weighted and delayed-feedback adaptive calibration average 88.0%, 88.5% and 89.9%.

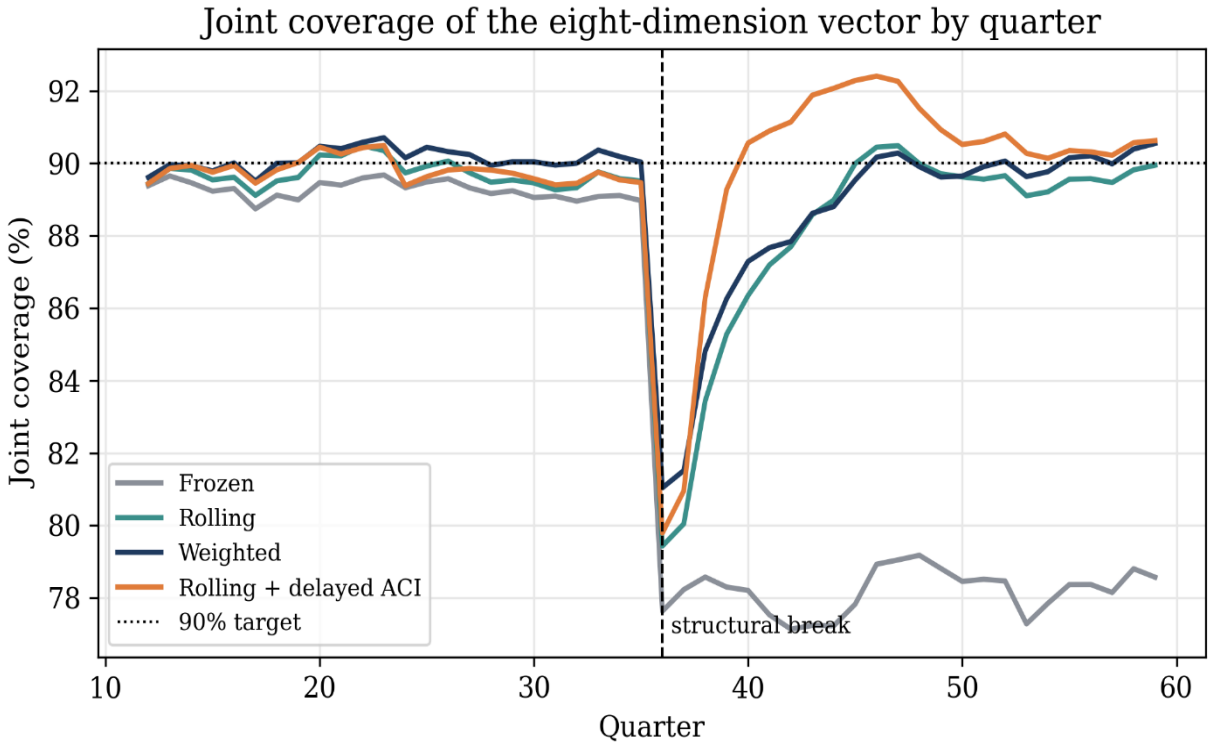


Figure 1. Joint coverage of the eight-dimension vector by quarter, mean of 200 replications. After the structural break at quarter 36, frozen calibration stays about 12 points under the 90% target, while delayed-feedback adaptation recovers in about four quarters. Simulation, seeds 1000 to 1199; deployment quarters 12 to 59.

The averages hide an important feature. In the break quarter and the next, all four protocols sit at roughly 78 to 82%, because no audit labels from after the break exist yet under the two-quarter lag. Delayed-feedback adaptation is back at 89.3% by the fourth affected quarter (quarter 39); the rolling window needs about ten quarters (90.0% at quarter 45) because old labels must age out. Adaptation then overshoots, holding 90.5 to 92.4% over quarters 43 to 52, so it is conservative for a period. Even the best protocol has an average worst four-quarter window of 80.6% (±0.8), and in post-break stress quarters it reaches 83.4% (91.6% in calm quarters). Composite-level coverage tells the same story. After the break it is 65.6% (frozen), 85.8% (rolling), 86.0% (weighted) and 89.6% (adaptive), with worst windows of 54.6%, 70.5%, 72.0% and 75.7%.

No label-driven method can protect the first two quarters after a shift. That is the motivation for the label-free monitor in Section 4.5.

4.2 Joint coverage over the factor vector

Table 4 compares the three constructions of Section 3.3 on the same rolling calibration set.

Table 4 Joint coverage and mean half-width of three interval constructions (rolling calibration, ε = 0.10)

Construction	Joint coverage, pre-break	Joint coverage, post-break	Half-width, pre-break	Half-width, post-break
Per-dimension, each at 90%	44.5% ± 0.5	40.1% ± 0.6	0.78 ± 0.01	1.04 ± 0.01
Bonferroni (ε/8 per dimension)	91.3% ± 0.1	90.7% ± 0.2	2.95 ± 0.04	3.51 ± 0.05
Max-score joint	89.7% ± 0.2	88.0% ± 0.2	2.19 ± 0.03	2.87 ± 0.03

Note. Half-width is the mean of $q\hat{\sigma}$ over firms and dimensions, in units of the factor standard deviation. Values are means ± 95% interval across 200 replications.

Each per-dimension interval is close to its own target (average marginal coverage 89.5% before the break, 87.8% after), yet the whole vector is covered less than half the time. A gate built from per-dimension intervals would therefore overstate how often a whole extracted vector is correct. The max-score construction reaches 89.7% before the break with half-widths 26% narrower than Bonferroni (18% narrower after the break). Bonferroni is conservative at 91.3%. The max-score construction is slightly below nominal, which is consistent with heavy-tailed errors and scale variation across regimes.

The absolute widths are large. In this simulation a 90% joint rectangle has a half-width above two factor standard deviations because the error distribution is heavy-tailed and 2% of entries are hallucinations. The factor-level rectangle is better read as an audit statement than as an informative forecast, which is why the gate in Section 4.3 acts on the composite the operator consumes.

4.3 The gate as selective risk control

The gate accepts a signal when the composite half-width is at most the tolerance τ, set at go-live to the 75th percentile of rolling half-widths (mean 0.38 composite units, about 0.60 standard deviations of the true composite). A violation is a composite extraction error larger than τ. Table 5 reports acceptance and violation rates.

Table 5 Acceptance rate and violation rate among accepted signals

Calibration	Accepted, pre-break	Violation among accepted, pre-break	Accepted, post-break	Violation among accepted, post-break
No gate	100%	8.8% ± 0.3	100%	28.1% ± 0.9
Frozen	76.4% ± 1.0	7.5% ± 0.3	76.5% ± 1.1	24.5% ± 0.8
Rolling	68.6% ± 2.2	7.4% ± 0.2	13.9% ± 1.4	20.4% ± 1.0
Weighted	68.0% ± 2.1	7.2% ± 0.2	13.2% ± 1.4	20.2% ± 1.1
Rolling + delayed ACI	64.5% ± 2.6	7.4% ± 0.2	9.2% ± 0.9	20.0% ± 1.1

Note. Values are means ± 95% interval across 200 replications. A signal is accepted when its half-width is within the go-live tolerance τ.

The frozen gate keeps accepting about 76% of signals after the break because its widths do not reflect the larger errors, and the violation rate among accepted signals rises from 7.5% to 24.5%, close to the 28.1% without a gate.

The adaptive gates widen their intervals and reject most signals, with acceptance falling to between 9 and 14%. They lower the violation rate among accepted signals to about 20% (20.0 to 20.4%), which lowers that rate by about 28% relative to no gate but stays well above the 10% reference. A likely reason is the bias term of the break, which shifts the errors of accepted low-uncertainty signals as well as the others; we did not test this. Before the break the violation rate among accepted signals is 7.2 to 7.5%, below the 10% reference, because accepted signals are the low-uncertainty ones.

Two cautions apply. The 10% level is a reference taken from ε, not a guarantee, since marginal coverage does not imply control of the accepted subset, and after the break no gate reaches it. And the price of the partial control is large. After the break, the best adaptive gate passes about one signal in ten.

4.4 Operators and the cost of gating

Figure 2 plots each operator's risk-adjusted ratio (Section 3.9) against the share of signals accepted, where the gate keeps the narrowest-interval signals within each quarter. The ordering by width is the same for every calibration protocol, so Operator A does not depend on the protocol.

The ratio rises with the acceptance share for every operator, so in this simulation gating costs performance. Accepting 60% of signals lowers Operator A's ratio from 1.44 to 1.13 (21%), and accepting 20% gives 0.66. We did not test the mechanism. The likeliest candidate is lost breadth, since fewer positions diversify less, and ranking by width carries limited information about return in this design. Section 4.7 tests a return model in which extraction error is costly (R4).

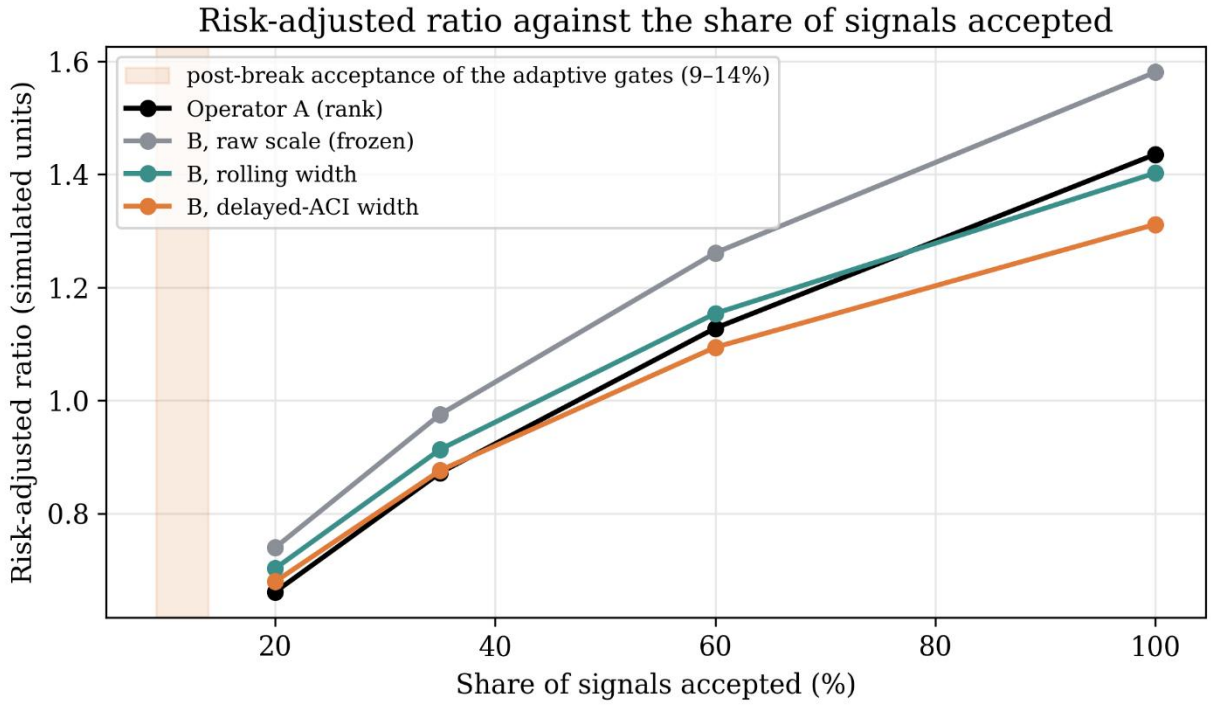


Figure 2. Risk-adjusted ratio against the share of signals accepted, for a rank operator and three width-scaled operators. The ratio rises with acceptance for every operator, and sizing by conformally calibrated width does worse than sizing by raw scale. The shaded band marks the post-break acceptance range of the adaptive gates. Simulation, 200 replications, all deployment quarters; ratio in simulated units.

At full acceptance, sizing by the raw scale estimate (Operator B with frozen calibration, which equals sizing by $\hat{\sigma}$ alone) gives 1.58 ± 0.03, the rank operator 1.44 ± 0.03, sizing by rolling-calibrated width 1.40 ± 0.03, and sizing by delayed-ACI width 1.31 ± 0.03. Sizing by a conformally calibrated width therefore underperforms sizing by the raw scale. A plausible explanation is that the calibration quantile varies over time and passes its sampling noise and its post-break growth into position sizes. We did not test that explanation either. Table 6 shows the exposure effect.

Table 6 Exposure drift after the break when the sizing constant is fixed before it (post-break value divided by pre-break value)

Sizing basis	Gross exposure ratio	Realized volatility ratio
Raw scale (frozen)	1.06 ± 0.00	1.06 ± 0.03
Rolling width	0.67 ± 0.01	0.71 ± 0.02
Weighted width	0.67 ± 0.01	0.72 ± 0.02
Delayed-ACI width	0.59 ± 0.01	0.65 ± 0.02

Note. Values are means ± 95% interval across 200 replications.

Because calibrated widths grow after the break, width-scaled sizing cuts gross exposure by 33 to 41% and realized volatility by 28 to 35% without any change to the sizing rule. That is the coupling between calibration and sizing in concrete form. Change the calibration and the exposure changes with it. Whether the automatic de-risking is desirable is a policy question, and in this simulation it coincided with a lower ratio.

Singh and Rezaeimanesh (2026) find on real FactSet news that matching a confidence map with its fitted scale improves return under frozen-scale control, with a smaller gain under an adaptive volatility controller. We do not reproduce a gain from matched sizing, but the designs differ (real news and nine open-weight models there, simulated extraction width here), so the two results are not in conflict. Together they suggest that conclusions about calibrated sizing depend on how the sizing rule consumes the calibration. A practical implication, tested only in simulation, is to use calibrated intervals for the gate and for reporting, and to size by a smoothed or fixed scale estimate, then evaluate the two together.

4.5 Drift monitors

Section 3.8 defines a label-free monitor, which tests changes in the extracted factors against the previous eight quarters, and a label-based monitor, which tests the miss rate of the frozen intervals on audited points. Figure 3 compares how quickly each raises an alarm after the break.

Before the break the per-quarter false-alarm rates are 0.9% (label-free) and 1.7% (label-based) in calm quarters, and 2.3% and 4.3% in stress quarters. Stress quarters recur and enter the reference window, so they rarely trigger the label-free monitor. The label-free monitor raises an alarm in the break quarter in 78.5% of replications and within four quarters in 85.0%. The label-based monitor does so in 3.0% and 57.0%, because audit labels from after the break arrive two quarters late and the test needs enough misses. The median first-alarm delay is 0 quarters for the label-free monitor and 4 quarters for the label-based one. Within 12 quarters the detection rates are 89.0% and 84.5%, and 7.0% and 3.0% of replications raise no alarm in the 24 post-break quarters

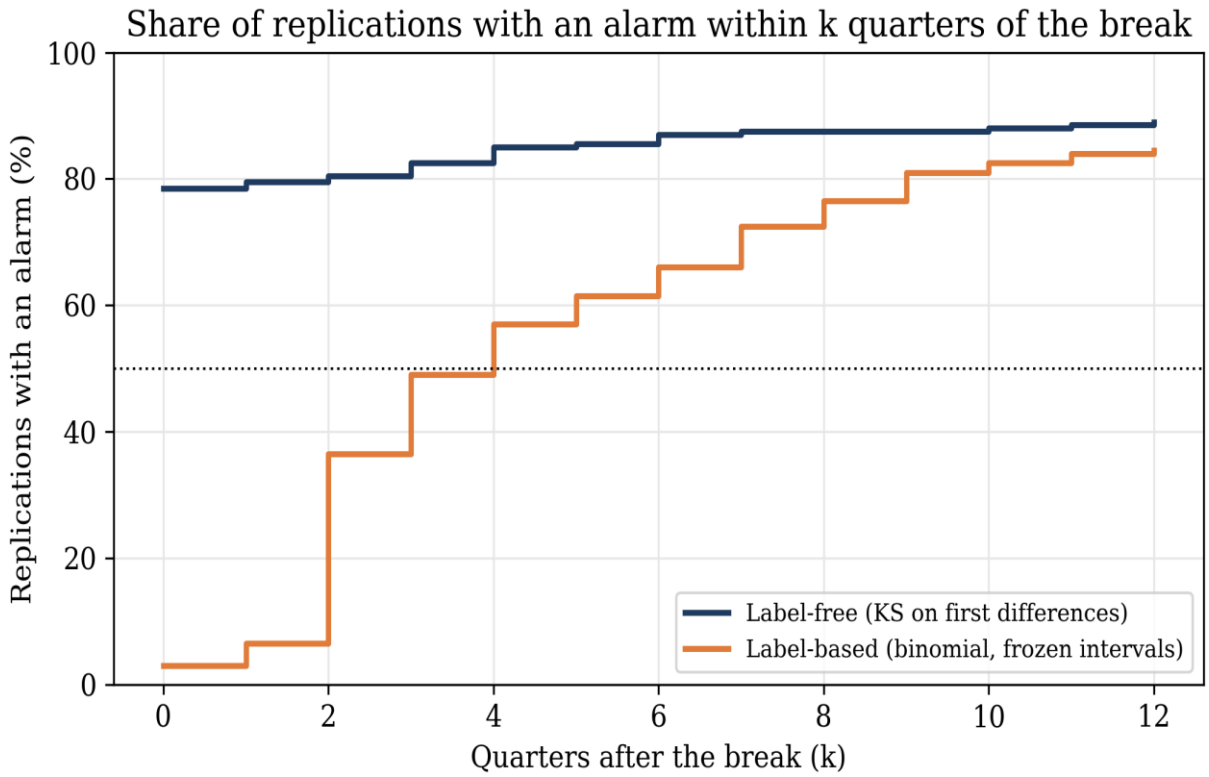


Figure 3. Share of replications with an alarm within *k* quarters of the structural break. The label-free monitor flags 78.5% of breaks in the quarter they occur, while label-based coverage monitoring needs about four quarters to reach 50%. Simulation, 200 replications; alarms counted from the break quarter (quarter 36).

The two monitors behave differently afterward. The label-free monitor alarms in 10.2% of post-break quarters, mostly early, because the new level enters its reference window. The label-based monitor alarms in 27.6% of them because the frozen intervals stay wrong. They are complementary. The label-free monitor gives early warning and the label-based monitor confirms persistent miscalibration.

A limit matters here. The label-free monitor detects changes in the distribution of extraction first differences, and the simulated break was built to change that (a larger error scale on four dimensions plus a bias). A break that left those differences unchanged would not be detected. The robustness run under gradual drift (R2, Section 4.7) shows this directly. When the same shift is spread linearly over twelve quarters, the label-free monitor alarms at quarter 36 in only 3.5% of replications (1.0% at drift onset in quarter 30), against 78.5% for the abrupt break. The monitor's early-warning advantage is therefore specific to abrupt changes.

4.6 Sensitivity

We varied the audit lag L, the post-go-live audit rate and the adaptation rate γ, one at a time, with 60 replications per setting (seeds 5000 to 5059, so the baseline row differs slightly from the main study). Table 7 reports post-break joint coverage.

Table 7 Post-break joint coverage by setting

Setting	Frozen	Rolling	Weighted	Delayed ACI	Worst 4-quarter window (ACI)
Baseline (L = 2, audit 10%, γ = 0.10)	78.2% ± 1.8	87.8% ± 0.4	88.3% ± 0.4	89.9% ± 0.3	80.0% ± 1.4
Audit lag L = 1	78.2% ± 1.8	88.5% ± 0.3	88.7% ± 0.3	89.8% ± 0.3	82.6% ± 1.0
Audit lag L = 4	78.2% ± 1.8	86.7% ± 0.5	87.7% ± 0.4	89.6% ± 0.4	75.4% ± 2.4
Audit rate 5%	78.2% ± 1.8	88.2% ± 0.5	89.0% ± 0.4	89.8% ± 0.5	78.4% ± 1.7
Audit rate 20%	78.2% ± 1.8	87.8% ± 0.3	88.0% ± 0.3	89.8% ± 0.3	81.0% ± 1.1
γ = 0.02	78.2% ± 1.8	87.8% ± 0.4	88.3% ± 0.4	88.6% ± 0.3	78.4% ± 1.5
γ = 0.05	78.2% ± 1.8	87.8% ± 0.4	88.3% ± 0.4	89.3% ± 0.3	79.1% ± 1.4
γ = 0.20	78.2% ± 1.8	87.8% ± 0.4	88.3% ± 0.4	89.8% ± 0.3	80.6% ± 1.5

Note. Values are means ± 95% interval across 60 replications. The last column is the average worst four-quarter window for delayed ACI. The frozen column is identical across rows by construction.

Average coverage under delayed ACI is stable, between 88.6% and 89.9%. Worst-window coverage is not. It is sensitive to the audit lag, falling from 82.6% at $L = 1$ to 75.4% at $L = 4$, and much less sensitive to the audit rate (78.4 to 81.0%) and to γ (78.4 to 80.6%). A deployment that cannot shorten its audit lag should therefore expect a longer stretch of undercoverage after a shift, whatever the calibration protocol, and should rely on the label-free monitor and a reduced acceptance policy in the meantime.

4.7 Robustness

Table 8 reports coverage and drift detection under each robustness design of Section 3.10, with 200 replications per design. The robustness runs use the same data-generating code and seeds (1000 to 1199) as the main study, so the base row equals the main-study values. Gate acceptance and violation, reported for the base design in Table 5, are omitted. The coverage conclusions are robust: delayed-ACI joint coverage after the break stays between 89.7% and 90.1% across every design, the worst four-quarter window between 80.0% and 81.9%, and max-score joint coverage stays between 87.6% and 88.5% and Bonferroni between 88.4% and 91.2%, close to the base design, with Bonferroni below target only under targeted, noisy audits (R5). Under correlated errors (R1) composite coverage under delayed ACI is 89.8% at both ρe = 0.3 and 0.6, against 89.6% in the base design, so the independence approximation in the composite scale does not reduce coverage, and max-score joint coverage is 88.0% and 88.1%. The one material change is detection under gradual drift (R2): the label-free monitor alarms in the break quarter in only 3.5% of replications, against 78.5% for the abrupt break, so the monitor's early warning is specific to abrupt changes. Table 9 reports Operator A under the error-linked return design R4, with signals kept by narrowest interval as in Figure 2. Error-linked returns lower the ratio at every acceptance level, but accepting fewer signals still lowers it for every κ, so the conclusion of Section 4.4 that gating costs performance holds when extraction error is costly. The cost of gating shrinks as κ grows: at 60% acceptance the ratio falls from 1.44 to 1.13 with κ = 0, against 0.79 to 0.74 with κ = 2k.

Table 8 Post-break coverage and drift detection under each robustness design

Design	Frozen joint post	ACI joint post	ACI worst 4q	Bonferroni post	Label-free alarm at break / ≤4q
Base design	78.2%	89.9%	80.6 ± 0.8	90.7%	78.5% / 85.0%
R1 correlated errors ρe=0.3	77.8%	89.9%	80.9 ± 0.8	90.8%	76.5% / 83.5%
R1 correlated errors ρe=0.6	78.3%	89.9%	81.1 ± 0.7	91.2%	77.0% / 85.0%
R2 gradual drift (q30–42)	79.4%	90.1%	81.9 ± 0.7	91.0%	3.5% / 20.5%
R3 misspecified scale	77.6%	89.7%	80.0 ± 0.7	90.3%	97.0% / 97.5%
R5 targeted, noisy audits	79.6%	89.7%	80.0 ± 0.8	88.4%	78.5% / 85.0%

Note. Means across 200 replications, with 95% intervals for the worst window. Post-break quantities are averaged over quarters 36 to 59 for every design, including R2, whose drift runs from quarter 30 to quarter 42. Alarms are counted from quarter 36; for R2 the label-free monitor alarms at drift onset in 1.0% of replications and within four quarters of it in 8.5%. Gate columns are omitted; Table 5 reports them for the base design.

Table 9 Operator A risk-adjusted ratio by share of signals accepted under error-linked returns (R4)

Return model	100% accepted	60% accepted	35% accepted	20% accepted
Base design (κ = 0)	1.44 ± 0.03	1.13 ± 0.03	0.87 ± 0.02	0.66 ± 0.02
R4, κ = 0.5k	1.28 ± 0.03	1.03 ± 0.03	0.81 ± 0.02	0.62 ± 0.02
R4, κ = k	1.11 ± 0.03	0.93 ± 0.03	0.74 ± 0.02	0.57 ± 0.02
R4, κ = 2k	**0.79 ± 0.03**	**0.74 ± 0.03**	**0.61 ± 0.02**	**0.48 ± 0.02**

Note. Ratio in simulated units, not annualized (Section 3.9). Means with 95% intervals across 200 replications. Signals are kept by narrowest composite interval within each quarter, as in Figure 2, and κ is in units of k.

4.8 Real-data pilot: extraction and uncertainty-signal validation

This section reports a pilot on real filings. Its purpose is to validate the extraction layer and the uncertainty signal that the gate depends on, not to reproduce the calibration, gate and monitor results of Sections 4.1 to 4.5, which require the multi-quarter design specified in Section 5.4. We extract the seven fields of Section 3.11 from the earnings press release filed as an exhibit to each company's results 8-K, and we score each extracted figure against the matching XBRL-reported value. The universe is every company in the dataset with an earnings call between January 2024 and March 2025, excluding one company that is the authors' employer, which yields 521 companies. Of these, 492 could be matched to a ticker, a results 8-K within seven days of the call and a periodic filing, and the remaining 29 are reported as not matched

(Section 3.11). The analysis covers 2,113 company-field pairs that have an XBRL reference value.

Coverage is high across all seven fields. The main model extracted 1,956 of the 2,113 reference figures, or 93%. Per field, coverage ranges from 84% for cash and equivalents to 99% for diluted earnings per share, with revenue at 95% and net income at 98% (Table 10). This is the coverage that the earnings-call transcript source did not reach. In a same-company comparison, press releases yielded about 4 scored figures per company against about 1 for transcripts, because releases print each figure in a labelled table row while calls state figures in speech and often omit them.

Extraction accuracy is high. Of the 1,956 extracted figures, 1,884, or 96.3%, are within 1% of the reference value, with a 95% interval of 95.4% to 97.2% computed by resampling companies rather than figures, so that figures from the same company stay together. The result is not sensitive to the tolerance: 95.6%, 97.0% and 97.8% of figures are within 0.5%, 2% and 5%, and scoring diluted EPS to the cent instead gives 96.4%. Counting unstated figures as failures, 89.2% of all 2,113 reference figures are extracted within 1%. These figures use corrected and verified reference labels. The first scoring preferred the RevenueFromContractWithCustomerExcludingAssessedTax tag, which excludes interest, lease and premium income at banks, REITs and insurers, so for those filers the reference was a narrower line than the total revenue the release reports. Preferring the broader Revenues tag and correcting the labels that the check below found wrong lifts revenue within-1% accuracy from 89.7% to 96.8% and the overall figure from 94.2% to 96.3%, with no change to the extractions. Restricting the sample to the 360 companies not used while developing the pipeline gives 96.2%, and in a complete rerun of all three models on the same releases, scored on the verified labels, Llama 3.3 70B and gpt-oss-20b reach 94.3% and 94.6%, against 96.0% for the main model in that rerun, so the accuracy is not specific to one model.

The error distribution is the point that matters for the gate. Of the 81 figures first scored as large errors, 38 rested on wrong labels. The revenue tag change removed 22 (the check below confirmed 21 of the new labels and could not verify one), and a check of every remaining large-error label against the company's press release, made without reference to the model's extraction, corrected 16 more. The check covered 86 labels in all: those 81, checked blind, and the five that the tag change itself turned into large errors, checked first against the release and the 10-K (four corrected, one confirmed). It found 48 labels correct, 20 wrong and 18 that the release does not settle, because it does not state the item for the quarter (12) or states more than one plausible candidate (6). The wrong labels came mostly from fourth-quarter values derived as annual minus nine-month year-to-date and from tags that name a different line than the release reports, such as a fee-income line or a non-GAAP segment profit. After the correction 44 figures (43 of the original 81 plus the Snap-on case below), 2.2% of those extracted, remain off by more than 5%, and 1.4% are off by 1 to 5%. Of the 44, 27 are confirmed extraction errors, including one that the narrower revenue tag had hidden (for Snap-on the model returned net sales where the release's total revenues include financial services revenue), and 17 rest on labels the release cannot verify. The confirmed extraction errors include definition mismatches, where the model returns an adjusted earnings per share figure or a broader or narrower line item than the one asked for. These are rarer on a press release, where GAAP and adjusted figures sit in separate labelled rows, but they still occur.

The large errors are what the gate must catch. The only self-reported signal collected is the model's GAAP flag, and it is constant: the model marked every one of the 1,956 figures as GAAP, as the prompt asks for GAAP figures, so the flag never fires and flags none of the errors. Verbalized confidence was not collected, and a sampling-based self-consistency proxy was tried during development and dropped because it showed no signal, so neither is evaluated here. A unit-scale check flags 4. Cross-model disagreement, the spread of the values the models return for the same figure (three models, or two when one does not state it), is the only signal with real power. On the verified labels it flags 26 of the 44 large errors (59%), and 20 of the 27 confirmed extraction errors (74%), while also flagging 162 of the 1,884 correct figures (8.6%). This is a partial and imperfect detector, not a solution. It is consistent with the view that a usable confidence signal must come from outside a single model's output, although richer self-reported signals were not tested, and it quantifies how far even a cross-model check falls short on this task.

Table 10 Extraction coverage and within-1% accuracy by field in the real-data pilot (main model, 492 companies)

Field	Reference figures	Extracted (coverage)	Within 1% of XBRL	Errors > 5%
Revenue	460	436 (95%)	96.8%	11
Net income	440	431 (98%)	93.7%	14
Operating income	385	356 (92%)	98.6%	3
Cash and equivalents	414	347 (84%)	96.3%	7
Gross profit	192	174 (91%)	98.3%	0
Diluted EPS	118	117 (99%)	97.4%	3
Operating expenses	104	95 (91%)	92.6%	6
All fields	**2,113**	**1,956 (93%)**	**96.3% [95.4, 97.2]**	**44**

Note. Coverage is extracted figures over reference figures. Within-1% accuracy is the share of extracted figures whose absolute relative error against the reference value is at most 1%. The all-fields interval is a 95% interval from 10,000 bootstrap resamples over companies. Figures use corrected and verified reference labels (Section 4.8): the Revenues tag preference and a check of every large-error label against the press release lower the revenue large-error count from 38 to 11 and the all-fields count from 81 to 44, with no change to the extractions.

4.9 Calibrated simulation

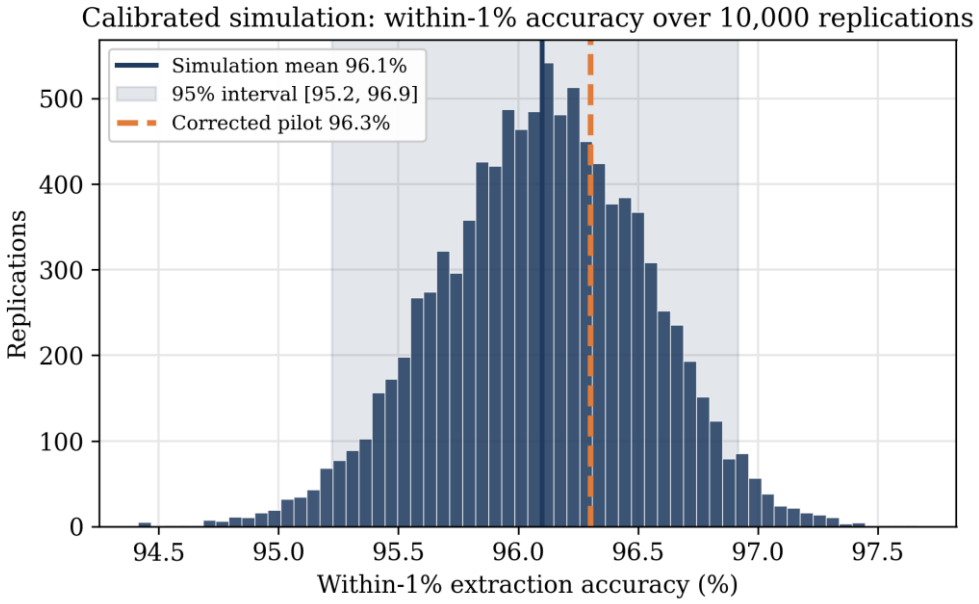
The pilot measures the extraction layer on 492 companies. To study how its error and gate behaviour vary across resampling, we run a Monte Carlo simulation whose parameters are calibrated to the pilot. This is a simulation, not a larger run of real companies. For each field we take the pilot per-field coverage, the share of figures in the within-1%, 1 to 5% and above-5% error bands, and the measured magnitudes of the large errors. We also take the operating point of the cross-model detector from the pilot, meaning its true-positive rate on large errors and its false-positive rate on correct figures. Each replication regenerates a pilot-sized sample of figures with these properties and reapplies the gate. We run 10,000 replications (Figures 4 to 6).

The simulated within-1% accuracy is 96.1%, with a 95% interval across replications of 95.2% to 96.9%, calibrated to the verified-label pilot. The pilot value of 96.3% lies inside this interval, which is a consistency check rather than an independent result. As expected from its inputs, the simulation reproduces the per-field pattern of the pilot, with operating expenses and net income the hardest fields and operating income and gross profit the easiest. Applying the cross-model gate lowers the rate of large errors among accepted figures from 2.2% to 1.0% while accepting about 90% of figures, so the cross-model signal removes about 59% of the large errors (the rate among accepted figures falls by about 55%) at a modest cost in coverage. It cannot remove large errors on which all three models agree, which include some of the unverifiable cases of Section 4.8, so the gate helps but does not solve the problem.

Table 11 Calibrated simulation summary, 10,000 replications of a 492-company pilot

Quantity	Mean	95% interval across replications
Within-1% accuracy	96.1%	95.2% to 96.9%
Large-error rate, ungated	2.2%	1.6% to 2.9%
Large-error rate, after cross-model gate	1.0%	0.6% to 1.5%
Accept rate under the gate	90.3%	89.0% to 91.5%

Note. A Monte Carlo simulation with per-field error rates, error magnitudes, coverage and detector operating point calibrated to the corrected-label 492-company pilot. It is not a run of 10,000 real companies. Seed 20261009. The within-1% bucket is drawn from a half-normal whose scale includes the 1 to 5% errors, which biases simulated accuracy slightly downward. Figures are drawn independently and the detector operating point is fixed, so the intervals ignore company clustering and uncertainty in the detector rates. Figures 4 to 6 are rendered from the same parameters.



Monte Carlo simulation calibrated to the corrected-label 492-company pilot. Not 10,000 real companies.

Figure 4. Distribution of within-1% extraction accuracy across 10,000 calibrated-simulation replications. The dashed line marks the corrected-label 492-company pilot value of 96.3%, which falls inside the simulated 95% interval. The simulation is calibrated to the corrected-label pilot and does not represent 10,000 real companies.

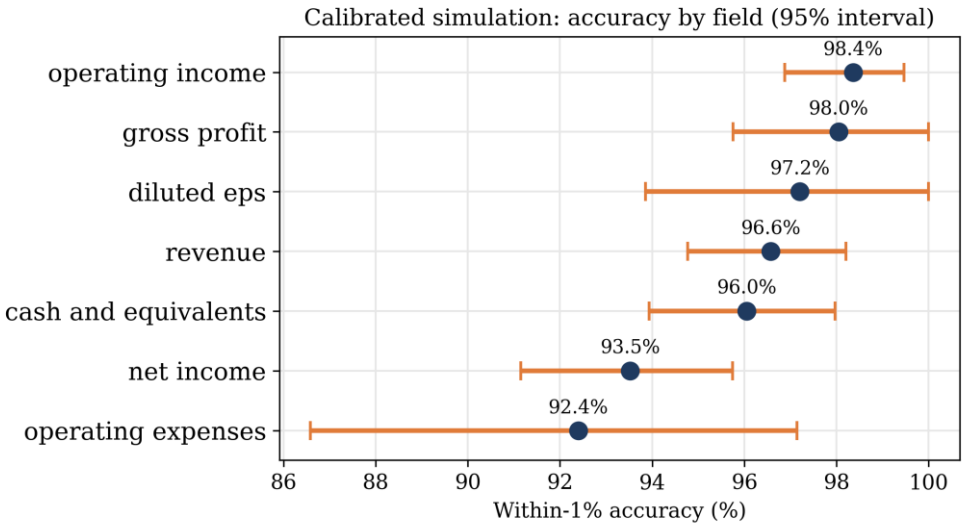
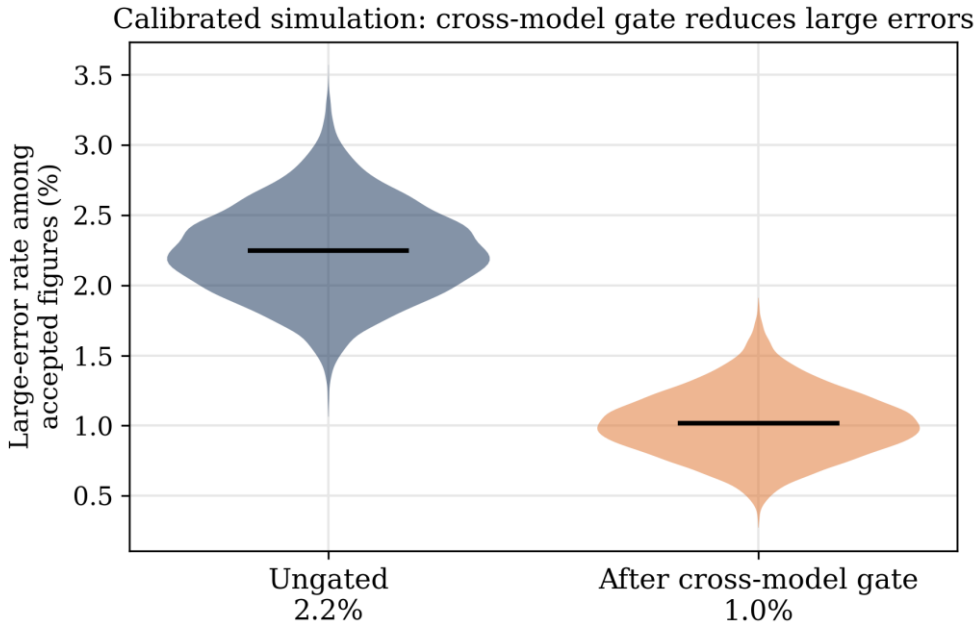


Figure 5. Simulated within-1% accuracy by field with 95% intervals. The pattern matches the pilot, with operating expenses and net income the hardest fields. The diluted EPS and gross profit intervals reach 100% because those fields have few figures and few errors.



Accept rate 90%. Calibrated to the corrected-label 492-company pilot.

Figure 6. Rate of large errors among accepted figures, before and after the cross-model gate, across 10,000 replications. The gate lowers the rate from 2.2% to 1.0% at an accept rate near 90%.

5. DISCUSSION

5.1 Deployment framework

The simulation evaluates component B and the detection part of component C. This subsection places it in a framework with two others (Figure 7). The framework states architectural requirements and does not prescribe a vendor or model.

Evidence control. A document is eligible for extraction only if its knowledge date precedes the decision time. The index records both the knowledge date and the ingestion date, so late or restated documents can be handled. Ye et al. (2026) list temporal contamination among the

confounders that reported alpha must rule out. Two limits apply. A date filter does not remove contamination that comes from the model's own pretraining, and an instruction telling the model to ignore what it knows is a request, not a control. We recommend point-in-time model versions where available and a leakage probe before go-live, for example testing whether the model can recall facts about events after its training cutoff for the entities in the evaluation set. This component is not evaluated here.

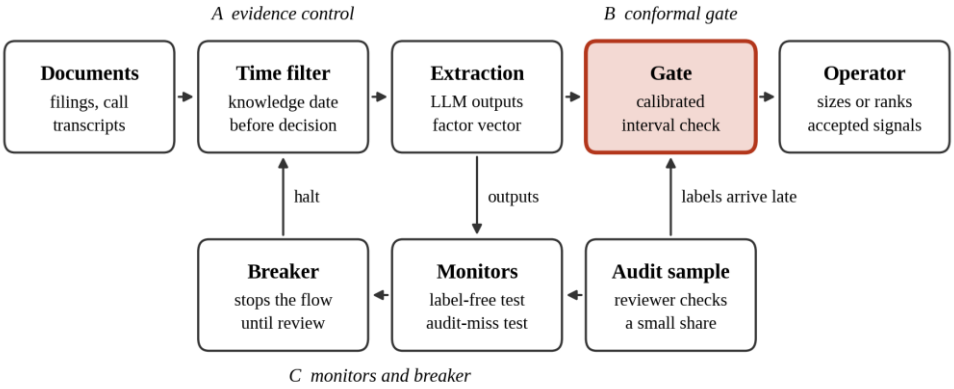


Figure 7. Deployment framework with three components around the extraction step: (A) evidence control, (B) the conformal gate, and (C) monitors and a circuit breaker. Component B and the detection part of component C are evaluated in simulation; component A and the breaker policy are not.

B.Conformal gate. The specification has six parts: an audit program with a stated sampling rate, reviewer protocol and expected lag; a scale model fitted on a pre-go-live labeled set; a joint max-score rectangle for audit reporting and a composite interval for the gate; a calibration protocol that handles label lag; an acceptance tolerance τ set before go-live and reviewed on a schedule; and a periodic report of coverage, acceptance rate and violation rate among accepted signals. In our simulation weighted and delayed-feedback adaptive calibration kept average coverage near the target, with the worst-window caveat of Section 4.1. Because calibrated widths moved position sizes (Section 4.4), sizing should use a smoothed or fixed scale estimate and be evaluated together with the gate.

Monitors and circuit breaker. Run both monitors. A label-free alarm should tighten τ or reduce the acceptance policy at once and trigger additional audits to shorten the label lag. A label-based alarm confirms persistent miscalibration. A breaker that halts signal emission needs a path independent of the model pipeline and a defined review procedure before restart. We simulated detection only, not the policy response.

Scope decisions. We omit three elements that appear in other proposals. Regime-dependent reordering of verification gates changes the outcome only if gates interact, because when every gate must pass the order affects latency and cost alone, and we have no evidence on it. Layered memory architectures for the extractor are outside the verification question studied here. Strategy constraints and signal-decay limits belong to portfolio risk management and are not part of extraction verification.

5.2 Supervisory context

Functions of the framework relate to existing expectations for algorithmic trading controls (Directive 2014/65/EU; Commission Delegated Regulation (EU) 2017/589), market access controls (17 C.F.R. § 240.15c3-5) and model risk management (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011). Component C supplies monitoring and a breaker, component B supplies

ongoing validation evidence for the extraction model, and component A supports the temporal integrity of inputs. This mapping is for discussion and is not a compliance claim, since whether a control meets an obligation is a legal and supervisory determination. We make no claim about how such pipelines are classified under Regulation (EU) 2024/1689.

5.3 Limitations and threats to validity

Mostly simulation. The data-generating process is stylized and its parameters were chosen by judgment. The numbers show how the methods behave in this world, not how they will behave in production. The pilot in Section 4.8 is small and covers only reported GAAP actuals that have XBRL values, not guidance, non-GAAP measures or judgmental factors such as tone. Its labels are filer-reported rather than independently audited, 17 of its 44 remaining large errors rest on labels the press release cannot verify (Section 4.8), its label lag is short, and some calls may predate the extractor's training cutoff.

Error structure. The base design has errors that are independent across dimensions and a scale model with the same functional form as the error process. The correlated-error and misspecified-scale designs (R1 and R3) are reported in Section 4.7. Coverage is robust under both, but real error dependence may take forms these designs do not capture.

Break design. The main break is abrupt and permanent. Under the gradual-drift design (R2, Section 4.7) the label-free monitor alarms at quarter 36 in only 3.5% of replications and at drift onset in 1.0%, so the monitor's advantage does not extend to slow drift. Intermittent regressions and breaks that leave extraction first differences unchanged remain untested.

Audits. The base design uses random, error-free audits. The targeted, noisy-audit design (R5, Section 4.7) keeps delayed-ACI coverage near 89.7% with inverse-probability weights; systematic reviewer disagreement remains untested.

Guarantees. The fixed-delay bound of El Halabi and Brandt (2026) assumes every outcome is eventually observed. Under audit sampling our delayed-feedback variant has no coverage guarantee, and the violation rate among accepted signals is an empirical measurement, not a bound.

Operators. The return models have no costs, turnover or capacity limits, and the ratio is not annualized. Under the error-linked return design (R4, Section 4.7) gating still lowers the ratio, though by less as the cost of extraction error grows. The operator conclusions in Section 4.4 remain specific to the return models studied and to sizing in proportion to inverse width.

Statistics. Intervals reflect variation across simulated replications, not real-world uncertainty. Sensitivity settings were varied one at a time, interactions were not examined, and no multiplicity correction was applied across the many comparisons reported.

Untested components. Point-in-time evidence control, leakage from model pretraining and the policy response to an alarm are not evaluated.

5.4 Full replication protocol on real filings

We specify the following so that the simulation's claims can be tested at scale, beyond the pilot in Section 4.8.

Collect model extractions for a set of firm-quarters and an audit sample with reviewer-verified values, recording audit dates so that the label lag is known.

Fix before looking at outcomes: the miscoverage level, window length, weighting half-life, adaptation rate, acceptance tolerance and the composite weights.

Fit the scale model on the pre-go-live labeled set only, and hold out later quarters.

Evaluate the four calibration protocols and the three joint constructions, reporting joint and composite coverage, worst four-quarter windows, acceptance and violation rates, and results split by regime where a regime indicator exists.

Run both drift monitors on the extraction history, with any known model or prompt change dates held back for scoring only.

Report operator results only with transaction costs and a pre-specified sizing rule, and report the gate's effect on the ratio, not only coverage.

The simulation code is provided as supplementary material so that each step can be run on synthetic data first.

6. CONCLUSION

Extracted factor vectors can be gated with conformal methods, but the details decide whether the gate means what it says. In a simulation with delayed audits, regime stress and a structural break, per-dimension intervals misstate the reliability of the vector (40 to 45% joint coverage against a 90% target), a calibration set frozen at go-live loses about eleven points of coverage after the break, and adaptive calibration restores average coverage but not the quarters just after a shift, which depend on the audit lag. As a risk control, the adaptive gates reject most

signals after the break but still leave about 20% violations among accepted signals, against 28% without a gate. In this simulation the gate did not raise risk-adjusted return. Calibrated widths should not set position sizes without joint evaluation of the gate and the operator. A label-free monitor closes much of the detection gap that audit lag leaves open for abrupt breaks, though not for gradual drift. Three design rules are suggested by these results, although smoothed-scale sizing and a composite gate were not compared with alternatives: report the reliability of the whole vector with a max-score rectangle rather than per-dimension intervals; gate on the composite the operator consumes, with an acceptance tolerance set before go-live and a label-free monitor to cover the audit lag; and size positions by a smoothed or fixed scale, not by calibrated width, evaluating the gate and the sizing rule together.

The headline results come from simulation, and real extraction may behave differently on every point listed in Section 5.3. The pilot in Section 4.8 tests only the extraction layer and the uncertainty signal, not the calibration, gate and monitor behaviour over time. It confirms that extraction is accurate for most figures. Correcting the revenue tag and verifying every large-error label against the press release raise accuracy to 96.3% and show that 38 of the 81 large errors first scored rested on wrong labels, leaving unchanged the finding that the model's own flag does not catch the errors that remain. The one self-reported flag collected catches none of the remaining large errors, and cross-model disagreement catches 59% of them while flagging 8.6% of correct figures. The calibrated simulation in Section 4.9 gives sampling intervals for the accuracy and large-error rates but adds no real evidence beyond the pilot. The robustness runs in Section 4.7 relax five assumptions of the base design, and the replication protocol in Section 5.4 states how to test the calibration and gate claims at scale on real filings.

Data and Code Availability

The simulation code, the robustness designs and the aggregated outputs are provided as supplementary material, with all seeds listed in Sections 3.9 and 4.7 and in Table 11. The supplementary code produces every result reported in Sections 4.1 to 4.9, Tables 4 to 11 and Figures 1 to 6. The cross-model disagreement of the original pilot run is supplied as a precomputed column, because that run's per-model extractions were not retained; a complete rerun with all three models' extractions is included. The pilot uses the S&P 500 Earnings Transcripts dataset (Kurry, 2025), available on Hugging Face under the MIT license, and public SEC data (the XBRL-tagged values and the company ticker file). The supplementary material lists the companies used, the ticker to CIK mapping and the excluded companies, the label verification record with the press-release evidence for each checked label, the matched accession numbers, the GAAP tag list, the prompt, the decoding settings and the extraction outputs, so that the pilot can be rebuilt from the public sources. Before release, the supplementary code must be scrubbed of any author-identifying content, including the hard-coded excluded-company ticker, to preserve anonymity. A repository link will be added after review.

References

1. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591.
2. Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2024). Conformal risk control. In Proceedings of the International Conference on Learning Representations (ICLR 2024).
3. Angelopoulos, A. N., Candès, E. J., & Tibshirani, R. J. (2023). Conformal PID control for time series prediction. In Advances in Neural Information Processing Systems (Vol. 36).
4. Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2), 816–845.
5. Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M. (2021). Distribution-free, risk-controlling prediction sets. Journal of the ACM, 68(6), Article 43.
6. Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR 11-7).
7. Commission Delegated Regulation (EU) 2017/589 supplementing Directive 2014/65/EU of the European Parliament and of the Council with regard to regulatory technical standards specifying the organisational requirements of investment firms engaged in algorithmic trading.
8. Directive 2014/65/EU of the European Parliament and of the Council on markets in financial instruments (MiFID II).
9. El Halabi, L., & Brandt, A. (2026). Adaptive conformal inference under delayed feedback:

Coverage guarantees and a delay-to-memory diagnostic (arXiv:2609.07251). arXiv. https://arxiv.org/abs/2609.07251

10. Gibbs, I., & Candès, E. J. (2021). Adaptive conformal inference under distribution shift. In Advances in Neural Information Processing Systems (Vol. 34).
11. Gibbs, I., & Candès, E. J. (2024). Conformal inference for online prediction with arbitrary distribution shifts. Journal of Machine Learning Research, 25. https://www.jmlr.org/papers/v25/22-1218.html
12. Kurry. (2025). S&P 500 earnings transcripts [Data set]. Hugging Face. MIT License. https://huggingface.co/datasets/kurry/sp500_earnings_transcripts
13. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094–1111.
14. Messoudi, S., Destercke, S., & Rousseau, S. (2021). Copula-based conformal prediction for multi-target regression. Pattern Recognition, 120, 108101.
15. OpenAI. (2025). gpt-oss-120b & gpt-oss-20b model card (arXiv:2508.10925). arXiv. https://arxiv.org/abs/2508.10925
16. Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T., & Barzilay, R. (2024). Conformal language modeling. In Proceedings of the International Conference on Learning Representations (ICLR 2024).
17. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
18. Securities and Exchange Commission. Risk management controls for brokers or dealers with market access, 17 C.F.R. § 240.15c3-5.
19. Securities and Exchange Commission. (n.d.-a). Financial Statement Data Sets. https://www.sec.gov/dera/data/financial-statement-data-sets
20. Securities and Exchange Commission. (n.d.-b). Company tickers [Data file]. https://www.sec.gov/files/company_tickers.json
21. Singh, R., & Rezaeimanesh, S. (2026). Ordinal gates, cardinal bets: Matching LLM confidence to the financial decision operator (arXiv:2609.00187). arXiv. https://arxiv.org/abs/2609.00187
22. Tibshirani, R. J., Barber, R. F., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. In Advances in Neural Information Processing Systems (Vol. 32).
23. Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer.
24. Wang, Y., Kuang, P., Han, X., Xu, K., & Wang, H. (2026). Beyond surface statistics: Robust conformal prediction for LLMs via internal representations (arXiv:2604.16217). arXiv. https://arxiv.org/abs/2604.16217
25. Xu, C., & Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In Proceedings of the 38th International Conference on Machine Learning.
26. Ye, Y., Han, J., Hu, A., Bu, J., Chen, Y., Wen, L., Mandic, D., Sun, D. D., Yinghui, X., & Xu, Z. (2026). The alpha illusion: Reported alpha from LLM trading agents should not be treated as deployment evidence (arXiv:2605.16895). arXiv. https://arxiv.org/abs/2605.16895
27. Zaffran, M., Féron, O., Goude, Y., Josse, J., & Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In Proceedings of the 39th International Conference on Machine Learning.