# WHEN FINANCIAL FACTS CHANGE: AUDITING LABELS FOR POINT-IN-TIME FINANCIAL LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Historical financial data are versioned observations rather than immutable rows: later annual reports can change values reported for an earlier period. This creates two hazards for machine learning—invalid revision-derived labels and predictors that silently contain future information. We construct accession-paired observations from the U.S. SEC's public as-filed XBRL data and test whether comparative changes or explicit restatement dimensions recover a regulatory error-correction construct. Across fiscal years 2011–2023, 31,688 of 354,650 matched core facts change, affecting 22.24% of 47,918 company-years. Explicit restatement dimensions cover only 6.17% of changed facts and are absent in the first five years, so their absence is not a stable negative label. We use the SEC's `DocumentFinStmtErrorCorrectionFlag` as a filing-level reference in two temporal cohorts whose protocols were frozen before label access. Any comparative change has precision 17.26%/12.68% and recall 57.43%/57.78%; the explicit-dimension proxy has precision 66.67%/49.21% and recall 39.60%/34.44% for fiscal years 2023/2024. Neither proxy meets the frozen precision/recall thresholds for recovering the filing-wide flag in either year. A retrospective fiscal-year benchmark is not a deployable timing test: its cohort conditions on future availability and its training labels can postdate test prediction times. We reconstruct first-filing cohorts of 3,961 and 4,158 companies without future eligibility filters, preserve unknown labels, and require training-label maturity at each prediction cutoff. The audit distinguishes comparative revision from regulatory error correction and identifies additional cohort and training-label timing requirements for financial ML. Complete score coverage is distinct from complete reference-label coverage. Recovering an earlier training cohort restores score coverage to all 4,158 test companies. On the 3,606 cases with observed reference flags, the predeclared weighted boosting baseline achieves AUROC 0.591. A separate retrospective class-weight sensitivity improves its Brier score but still does not beat the training-prevalence reference on that metric. The 552 unknown labels limit both evaluations.

## 1 INTRODUCTION

Machine-learning pipelines commonly treat a historical financial-statement row as if it had one stable value. In practice, a downloaded row may encode a later presentation of an earlier period. The change can arise from an error correction, discontinued operation, reorganization, accounting-policy change, or reclassification. A database that overwrites the original value silently moves information backward in time. This matters for prediction, backtesting, and for constructing labels from apparent "restatements."

SEC data make this problem directly observable. The quarterly Financial Statement Data Sets retain filing accession numbers and state that numeric facts are presented without change from as-filed reports (U.S. Securities and Exchange Commission, 2026a). Consecutive annual filings can therefore be paired: the first filing supplies the original fact, and the next annual filing supplies a later comparative presentation of the same period. The SEC also introduced a tagged annual-report checkbox, `DocumentFinStmtErrorCorrectionFlag`, whose definition explicitly concerns restatements due to error correction (U.S. Securities and Exchange Commission, 2023; 2025). It provides a recent filing-level reference for auditing revision-derived proxy labels.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We treat this flag as the reference construct, not as unquestioned ground truth. Its filing-wide scope is broader than our ten-fact proxies; measuring that mismatch is part of the audit, while the flag's own error rate remains unidentified.

We test whether a numerical revision is a valid label for error correction. The analysis uses an accession-paired observation that separates fiscal valid time from filing time and a reproducible 2011–2024 panel. Two proxy rules are tested against the SEC correction flag in sequential cohorts under thresholds fixed before label access. Both fail, including the seemingly specific restatement-dimension rule. We retain a historical fiscal-year comparison and an overwritten-feature control, while identifying their additional cohort and label-maturity limitations. A revised first-filing census and monthly training cutoffs separate these requirements. The supplement includes cohort rows, predictions, protocols and scripts.

The failed gates are part of the evidence. In these data, a numerical change is not automatically an error, the correction flag is not a fraud label, and the overwritten-feature comparison does not establish leakage inflation. The measured failure rates and information-timing protocol can inform future financial learning work.

## 2 Related work

**Financial revisions and point-in-time data.** The public SECDatabase schema already archives data-point revisions and their filing dates (SECDatabase.com, 2019); merely storing version history is therefore not novel. An SSRN working paper by Lyle et al. (2026) uses proprietary Compustat Snapshot vintages to show that historical firm-period observations are repeatedly adjusted and that vintage choice changes conclusions for many accounting anomalies. Another recent SSRN working paper develops a bitemporal SEC framework for leakage-safe equity research (Wang, 2026). Our contribution is a public, accession-level audit of two numerical proxies against the recent correction checkbox, followed by a monthly label-maturity evaluation. The checkbox-recovery task has filing-level scope.

Revision-aware learning is also studied outside corporate reporting. Back2Future learns from public-health backfill trajectories (Kamarthi et al., 2022); MACROCAST excludes revised macroeconomic values from real-time training (Carriero et al., 2026); evolving-observation models formalize feedback that changes after it is first observed (Bar-On & Mansour, 2025). These works motivate preserving the vintage axis, but do not resolve whether a corporate fact change is an error-correction label.

**Misstatement prediction and financial benchmarks.** Financial restatement and misstatement prediction has a long literature using accounting, governance, audit, and market data (Dutta et al., 2017; Bertomeu et al., 2021; Zhang Parker et al., 2025). Those targets are not interchangeable with the comparative-change event studied here. Recent arXiv XBRL and audit benchmarks evaluate numerical verification or enforcement-grounded LLM judgment (Panda, 2026; Liu et al., 2026). Our task is earlier in the pipeline: it audits the labels and timestamps supplied to a learner. This follows the broader warning that leakage can invalidate machine-learning-based science even when the downstream model is implemented correctly (Kapoor & Narayanan, 2023).

## 3 Data and study design

For registrant $i$ and fiscal period $y$, let $x_{iy}^{(t_0)}$ denote a fact as first observed in filing $t_0$ and $x_{iy}^{(t_1)}$ its comparative presentation in the next annual filing $t_1$. A revision event is a difference between these two filed observations; the later filing's correction flag is $z_i^{(t_1)}$. Proxy validity asks whether a revision rule recovers $z$ through precision and recall. Point-in-time prediction instead estimates $z_i^{(t_1)}$ from information available at $t_0$. Replacing $x^{(t_0)}$ by $x^{(t_1)}$ violates that information set and is used only as a contamination control.

### 3.1 Accession-paired facts

For fiscal year $y$, we take Q1 of $y+1$ and select the earliest non-amended Form 10-K with fiscal-year marker FY for each CIK. From Q1 of $y+2$, we select the earliest non-amended Form 10-K for fiscal

year $y + 1$. We retain CIKs present in both sets. This Q1 design favors a fixed, reproducible cutoff and omits annual reports filed later in the year.

We predeclare ten US-GAAP concepts: Assets; Liabilities; Stockholders' Equity; Liabilities and Stockholders' Equity; Revenues; Revenue from Contracts with Customers Excluding Assessed Tax; Operating Income/Loss; Net Income/Loss; Cash and Cash Equivalents; and Net Cash from Operating Activities. The set spans the balance sheet, income statement, and cash-flow statement while retaining broad cross-firm coverage, and was fixed before label inspection. Numeric rows must have no co-registrant or segment dimension. Instant concepts require `qtrs=0`; flow concepts require `qtrs=4`. The original fact must end at the original filing's reporting period. We match on CIK, local tag, period end, quarter count, and unit. Conflicting duplicate keys are excluded.

For matched values $v_o$ and $v_l$, the change indicator is

$$C = \mathbf{1}\left[|v_l - v_o| > \max\{10^{-6}, 10^{-9}\max(|v_o|, |v_l|)\}\right]. \tag{1}$$

For each changed fact, we separately search the later accession for the same fact key with a dimensional context whose flattened segment contains `Restatement=`. This is evidence about the filing's presentation; it is not assumed to be an error or material misstatement.

The main panel covers fiscal years 2011–2023 using fourteen Q1 compact archives. A frozen feasibility pilot on 2023 required at least 1,000 registrants, 500 changes, and five tags with 25 changes; all ten tags qualified. A subsequent proposed historical label gate required explicit dimension evidence in ten years. It failed, closing that labeled formulation before any model was fit. We retain the failure because it reveals label nonstationarity.

### 3.2 SEC CORRECTION FLAG

For fiscal years 2023 and 2024, we extract the boolean DEI correction flag from the exact later accession in the SEC detailed archives (U.S. Securities and Exchange Commission, 2026b). The 2023 cohort uses 2025 Q1; the 2024 replication uses the January–March 2026 files. Invalid, missing, or conflicting flags are excluded from metrics. We audit two company-level proxies:

$$P_{\text{any}} = \mathbf{1}[\text{at least one of ten facts changes}], \tag{2}$$
$$P_{\text{dim}} = \mathbf{1}[\text{at least one changed fact has a Restatement dimension}]. \tag{3}$$

The protocol requires at least 50 positive flags and requires $P_{\text{dim}}$ to achieve 80% precision and 50% recall before it may be used as a proxy label. These operational minima are a screening rule rather than a hypothesis-test cutoff. The 2023 thresholds were frozen before reading flags; the identical 2024 replication was frozen before downloading any 2026 data.

### 3.3 RETROSPECTIVE BASELINE SPECIFICATION

After reporting both proxy audits, we fixed the design of the prediction task. The historical model was trained on FY2023 and evaluated once on FY2024; this fiscal-year split alone does not enforce training-label availability. Its target is whether the next annual filing has a positive correction flag. Inputs are the ten original USD facts divided by absolute Assets, clipped to $[-100, 100]$, plus log absolute Assets, missing indicators, and two-digit SIC. Training medians impute missing values; unseen SIC categories are ignored.

We fix L2 logistic regression and histogram gradient boosting (HGB) with balanced training weights. HGB uses 200 iterations, learning rate 0.05, 15 leaves, L2 penalty 1, and minimum leaf size 20. We also fit each model using a deliberately contaminated representation in which the same historical facts come from the next filing. No future flag, change indicator, company ID, text, or return is a feature. Average precision (AP) is primary; we also report AUROC, Brier score, and ten equal-frequency-bin calibration error. A predeclared leakage claim requires at least 20% relative AP inflation and a positive paired 2,000-resample bootstrap interval.

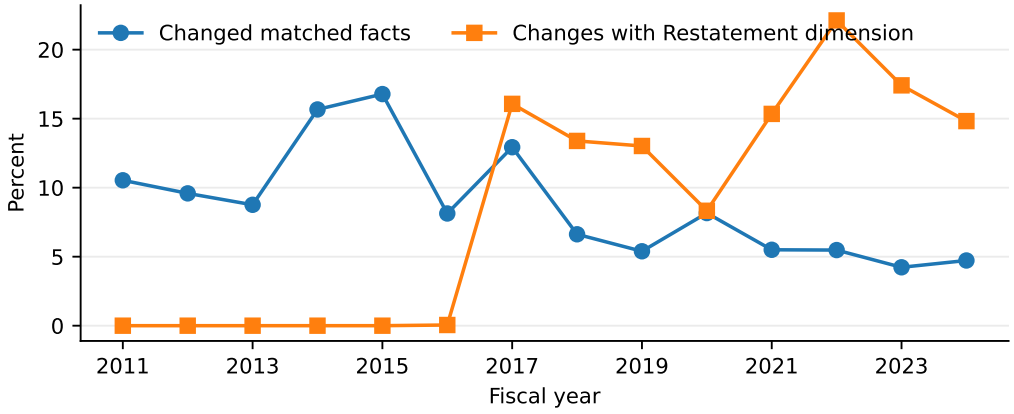


Figure 1: Matched-fact change rate and share of changes with an explicit Restatement dimension. The 2011–2023 panel and separately frozen 2024 replication use the same extraction rule.

Table 1: Proxy validity against the exact later-filing correction flag.

| FY | Proxy | TP | FP | FN | Precision | Recall |
|---|---|---|---|---|---|---|
| 2023 | Any comparative change | 58 | 278 | 43 | .1726 | .5743 |
| 2023 | Restatement dimension | 40 | 20 | 61 | .6667 | .3960 |
| 2024 | Any comparative change | 52 | 358 | 38 | .1268 | .5778 |
| 2024 | Restatement dimension | 31 | 32 | 59 | .4921 | .3444 |

# 4 Results

## 4.1 Revision frequency and dimensional evidence over time

The 2011–2023 panel contains 354,650 matched facts and 47,918 company-years. There are 31,688 changed facts (8.94%); 22.24% of company-years contain at least one change. Median nonzero relative changes range from 0.36% for Assets to 7.66% for contract revenue across tags. The rate is not monotone over time: it reaches 16.78% in 2015 and 4.23% in 2023.

Only 1,954 changed facts (6.17%) have explicit restatement-dimension evidence. Figure 1 illustrates this nonstationarity. The evidence is absent in 2011–2015, appears once in 2016, then becomes substantial in 2017. Encoding absence as a negative label would therefore make the label mechanism a function of calendar time and filing practice.

An independent filing-level check confirms the accession matching. CSX reported 2023 Assets of USD 42.408 billion in accession 0000277948-24-000010. Its next 10-K presents USD 42.212 billion and also includes dimensional facts for the previously reported USD 42.408 billion and a negative USD 196 million restatement adjustment. The supplement parses both official instance documents and asserts these values.

## 4.2 Proxy validity in two cohorts

Table 1 and Figure 2 report the audits under the frozen protocols. The reference has 101 positives among 3,536 valid flags in FY2023 and 90 among 3,592 in FY2024. Flag coverage exceeds 99.4% in both cohorts. The broad proxy finds about 57.5% of positives but produces 278 and 358 false positives; its precision is only 17.26% and 12.68%. The explicit dimension is more specific, but its precision falls from 66.67% to 49.21%, while recall remains below 40%. It fails both thresholds in both years. Post-hoc 95% Wilson intervals for its precision/recall are [.541,.773]/[.306,.494] in 2023 and [.373,.612]/[.254,.447] in 2024; every upper endpoint remains below the corresponding frozen threshold.

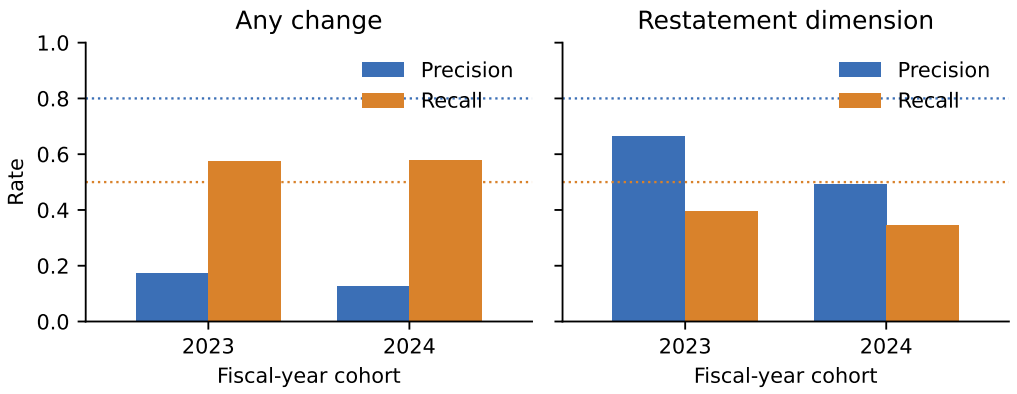


Figure 2: Proxy precision and recall in two sequential cohorts. Dotted lines show the frozen 0.8 precision and 0.5 recall requirements.

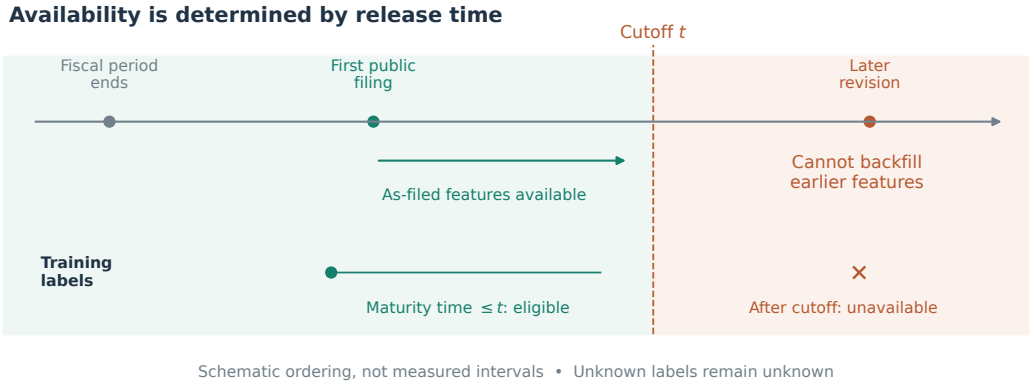


Figure 3: Availability contract (schematic). Fiscal period identifies what a fact describes; a filing release identifies when it becomes available. Features must be available at the prediction cutoff, and every label used for training must already be mature at that cutoff. The label check is a constraint, not an assertion that all labels are observed. Unknown outcomes remain unknown.

False positives and false negatives arise for different reasons. A false positive for $P_{\text{any}}$ is expected when a retrospective presentation changes for a reason other than error correction. A false negative can occur because the filing-wide checkbox concerns a fact outside our ten-item face-statement set. The explicit dimension narrows the event but does not eliminate either scope mismatch. The confusion matrices measure recovery of this filing-wide reference under the frozen thresholds. They do not estimate the flag's error rate or establish whether a changed target-year fact was itself erroneous. That distinction would require linking the disclosure to the affected period and fact in the full filing.

**Why feature timing alone is insufficient.** A historical fiscal-year comparison used first-filing features but required future comparative observations for cohort eligibility and did not enforce training-label maturity. Its overwritten-feature contrast also missed the predeclared inflation criterion. Appendix A retains those estimates; the construction below addresses cohort and training availability before evaluating the learner.

## 5 Availability-aware cohort and prediction design

**Cohort before future annotation.** We rebuild eligibility from the earliest accepted 10-K/FY for each company and target fiscal year in the following calendar year's Q1 bulk files. No future filing, comparative match, or observed label is required. The resulting FY2023/FY2024 cohorts contain

Table 2: Revised March-only, scored-and-annotated subset ($n = 1,734$). Intervals use 2,000 company bootstrap resamples. The constant prediction is the matured training prevalence, not the test prevalence.

| Model | AP (95% CI) | AUROC (95% CI) | Brier |
|---|---|---|---|
| LR | .0325 [.0206,.0636] | .5157 [.4279,.6000] | .1341 |
| HGB | .0324 [.0205,.0644] | .4811 [.3844,.5845] | .0405 |
| Constant | .0265 [.0190,.0340] | .5000 | .0258 |

3,961/4,158 companies, compared with 3,555/3,606 in the historical intersection cohorts. Nine and thirteen companies with all ten target facts missing are retained. Only US-GAAP USD facts at the original report period are used, excluding co-registrant and segmented facts; conflicting same-key values become missing. SIC and missingness come from the original accession. Deleting, corrupting, or adding future annotations leaves the entire first-filing payload hash unchanged. These are Q1 filing cohorts, not an all-market census. Acceptance timestamps reconstruct availability from later downloaded bulk files and do not audit historical real-time feed publication.

**Future labels and attrition.** We independently scan all eligible next-year Q1 annual filings and their text facts, then left-join the correction flag to the frozen census. FY2023 has 3,545 observed labels (101 positive), 397 companies without an eligible next-Q1 filing, 17 missing flags, and two conflicting flags. FY2024 has 3,606 observed labels (90 positive), 538 without the next filing, 13 missing flags, and one conflicting flag. Unknown labels remain unknown. The target remains the next filing's disclosure, not verified correction of a particular target-year fact. The historical proxy-validity analyses above remain conditional on their original matched cohorts.

**Training-label maturity.** A fiscal-year split also leaks information when a training label is published after a test company's prediction time. We therefore fit at the start of each test calendar month using only FY2023 labels accepted strictly before that cutoff, without changing the fixed LR/HGB hyperparameters. A model is fit only with at least 100 training companies and five examples of each class; this gate and the monthly rule were recorded before fitting. At January 1, 2025 there are no available training labels. At February 1, there are 62 labels with one positive. At March 1, there are 1,946 with 47 positives. Thus the 81 January and 1,949 February test companies remain unscored; all 2,128 March companies receive predictions. Of these, 1,734 have observed labels, including 46 positives. All denominators are retained.

Table 2 does not establish reliable ranking ability: both model AUROC intervals include 0.5, and the constant has lower Brier loss. These are conditional estimates for March companies with observed labels. They do not measure utility for unscored months or unobserved labels, and cannot be compared causally with the historical table because both cohort and training information changed. Broader earlier training cohorts, annotation semantics, and selection effects remain research requirements.

### 5.1 Unknown labels and score-selection audit

We retain every company, including labels that cannot be observed and months that cannot be scored. Of the 4,158 FY2024 companies, 3,606 have observed labels (90 positive) and 552 have unknown labels. Without assumptions about the latter, the full-cohort positive prevalence can range from 2.16% to 15.44%; this is an identification bound, not a confidence interval or a model-performance bound.

The preceding FY2023-only training protocol scores only 2,128 companies; 1,734 of these have observed labels (46 positive). The 2,030 unscored companies include 1,872 observed labels (44 positive). Unknown labels occur in 18.52% of scored companies versus 7.78% of unscored companies. Median log-one-plus absolute assets are 19.28 and 22.07 in scored and unscored groups, respectively, among nonmissing values. Thus score eligibility selects a materially different filing-time and size distribution. Unknown-label companies also have smaller median log assets than observed-label companies (19.20 versus 20.95). These are descriptive differences, not evidence that label missingness is ignorable. The artifact records all feature quartiles, missingness, industries, filing months, and the score/label cross-tabulation without filling unknown labels with negatives. The preceding restricted-

training metrics are conditional on the 1,734 scored and annotated cases; the extension below removes the unscored-month restriction.

**Recovered earlier-label coverage.** The SEC taxonomy records the correction flag in DEI 2022q4[1]; we separately verify actual filing coverage from the complete 2024Q1 notes archive. A new FY2022 first-filing cohort contains 4,644 companies, selected without future information. Its next-Q1 annual filings provide 3,591 observed correction labels (119 positive), while 1,002 companies have no corresponding next-Q1 filing and 51 lack the flag. Joining these labels leaves all first-filing fields byte-identical. Missing flags remain unknown.

### 5.2 Earlier-training extension with complete score coverage

Before examining this extension's results, we froze the same features, models, hyperparameters, cutoff rule, and 2,000 bootstrap replicates. Adding the FY2022 cohort to the candidate training pool yields 3,591/3,653/5,537 mature training labels at the January/February/March 2025 cutoffs, with 119/120/166 positives. All 4,158 test companies can now be scored. Evaluation includes the 3,606 with observed labels (90 positive); the remaining 552 are retained as unknown.

Table 3: Expanded earlier-training evaluation on all 3,606 annotated test companies. Intervals are pointwise 95% company bootstrap intervals. The monthly training-prevalence baseline uses only labels available before each cutoff.

| Model | AP | AUROC [interval] | Brier |
|---|---|---|---|
| Logistic | 0.03111 | 0.52557 [0.46439, 0.58295] | 0.22348 |
| Gradient boosting | 0.04075 | 0.59097 [0.52651, 0.65088] | 0.08753 |
| Training prevalence | 0.02425 | 0.48053 [0.42975, 0.53378] | 0.02438 |

The full-cohort scores cannot be directly compared with the earlier 1,734-case subset. On those identical annotated March cases, boosting AP changes from 0.03239 to 0.08079 (paired difference 0.04841, interval [0.01486, 0.12233]) and AUROC from 0.48111 to 0.63998 (difference 0.15886, interval [0.05273, 0.26737]). Its Brier score worsens from 0.04052 to 0.08082. Thus the earlier-data extension repairs score availability and shows limited ranking improvement, while balanced training still yields poor probability calibration. All logistic and prevalence comparisons, including adverse results, are retained. These intervals are not corrected for multiple metric comparisons. The audit independently reconstructs label eligibility at every cutoff, checks all saved predictions and point metrics, and retains the unchanged test cohort. Acceptance times remain retrospective availability proxies; the current SEC bulk archive is not an independently archived live data feed.

## 6 Retrospective class-weight sensitivity

We repeat the three monthly logistic and histogram-gradient-boosting fits with the frozen features, cutoff dates, preprocessing, and hyperparameters, removing only balanced class weights. All 4,158 predictions retain the original identifiers and cutoffs; 3,606 observed outcomes contain 90 positives. Predictions for the remaining 552 companies are retained, but their unknown outcomes are excluded from label-dependent metrics. This uses the already examined test set and is not independent confirmation. A separate implementation verifies tie-aware average precision, rank-based AUROC, and Brier arithmetic, with 2,000 paired CIK-cluster bootstrap replicates.

| Model / weighting | AP | AUROC | Brier |
|---|---|---|---|
| weighted/logistic | 0.03111 | 0.52557 | 0.22348 |
| weighted/HGB | 0.04075 | 0.59097 | 0.08753 |
| unweighted/logistic | 0.03196 | 0.52312 | 0.02493 |
| unweighted/HGB | 0.03746 | 0.60262 | 0.02502 |
| prevalence | 0.02425 | 0.48053 | 0.02438 |

[1] https://www.sec.gov/xbrl/site/doc/releasenotes-2023.pdf

Removing weights substantially improves probability calibration relative to weighted fitting, but neither unweighted model beats the monthly training-prevalence predictor in Brier score. The unweighted HGB minus prevalence Brier difference is 0.000634 (paired 95% CI [0.000166, 0.001079]), whereas its AP difference is 0.01321 [0.00549, 0.03111]. Better ranking does not imply better probability estimates. These retrospective results quantify the class-weight sensitivity on observed labels. They do not identify performance for the 552 companies with unknown outcomes or validate the correction flag against period-specific filing evidence.

**Scope of the disclosure target.** The correction checkbox is a filing-wide event. Identifying a new correction to a particular fiscal period requires its first disclosure and affected facts; repeated mentions are not automatically new events. Missing flags or incomplete filing coverage remain unknown.

## 7 Frozen decision sequence

The supplement preserves five sequential decisions: the one-year feasibility gate passed; the explicit-evidence pilot passed; the predeclared 13-year label coverage gate failed; the FY2023 proxy-label gate failed; and that failure replicated on FY2024. Only after the replication was reported did we freeze and open the FY2024 prediction test. Failed gates were not relaxed, and no locked model result was used to choose a later parameter.

**Ranking and calibration answer different questions.** Average precision evaluates the ordering of companies with observed labels, whereas Brier score evaluates the probabilities assigned to those labels. Removing balanced class weights can change calibration substantially without changing the feature availability problem. The monthly prevalence reference also varies across scoring cutoffs, so it is not one constant score on the pooled cohort. All of these comparisons remain conditional on observed outcomes; their confidence intervals cannot quantify the uncertainty created by the 552 unresolved labels. The missingness and maturity restrictions therefore remain part of the result even when a ranking contrast is positive.

## 8 Data construction checks

The 2011–2023 output has no duplicate benchmark keys; explicit evidence is always a changed fact. The FY2023 slice of the independent multi-year builder exactly matches all 27,960 pilot rows. One conflicting compact accession is excluded in 2011. Correction-flag parsing finds no invalid boolean values; two FY2023 and one FY2024 accessions contain conflicting booleans and are excluded along with missing flags.

The extraction is deterministic and uses public SEC data. The supplementary material supplies the extraction code and derived observations. The SEC warns that its derived data may contain filer or extraction errors and are not substitutes for the full filings (U.S. Securities and Exchange Commission, 2026a;b); our filing-level CSX check addresses one example, not the entire corpus.

## 9 Limitations and implications

The compact panel deliberately covers ten core US-GAAP concepts, consolidated contexts, non-amended 10-Ks, and Q1 filing cutoffs. A broader concept set or late-filer design can change proxy recall. The correction flag is filing-wide, available only under the recent reporting regime, and does not identify which fact changed. It refers to error correction, not intent, fraud, enforcement, or investment materiality. These scope differences prevent treating the reported confusion matrices as estimates for fraud detection. Because the inputs contain only Q1 archives, this study cannot estimate the full-year share or characteristics of omitted late filers.

The SEC periodically reprocesses its derived archives. The analysis uses the September 2026 data snapshot; a future download may differ. Historical variation in taxonomy and filing practice can also contribute to the time pattern of dimensional evidence. The accession-level CSX check validates one join, not every historical observation.

The original prediction benchmark has two cohorts and fewer than 200 positives in total; the separately frozen extension adds one earlier training cohort but still has only one 90-positive test cohort. Registrants may recur across cohorts, as they would in year-ahead deployment, but company identifiers are never features. Balanced training produces poorly calibrated logistic probabilities, and no calibration split or hyperparameter search is attempted. The overwritten-feature comparison is underpowered; its bootstrap interval prevents the predeclared claim. Its purpose is to expose the information-set distinction, not to advertise a positive result.

Financial datasets should keep valid time (the fiscal period) separate from filing time. Revision events should retain accession provenance, and labels should follow the regulatory or accounting construct required by the task. A value difference can be a useful version event without being an error label. A restatement dimension can be useful evidence without providing a stable historical negative class. Models should be trained and tested on the version available at their declared decision time, with retrospective snapshots reserved for sensitivity audits. This template also applies when public-health counts, macroeconomic releases, or other scientific records are revised after first publication: preserve each release, separate a change event from the semantic target, and evaluate on the version available at the stated decision time (Kamarthi et al., 2022; Carriero et al., 2026).

### 9.1 Heterogeneity across financial concepts

**2011–2023 matched facts by concept. Median is over nonzero relative changes.**

| Concept | Matched | Changed | Rate | Median change |
|---|---|---|---|---|
| Assets | 47,461 | 5,419 | .114 | .0036 |
| Liabilities | 38,154 | 4,322 | .113 | .0063 |
| Stockholders' equity | 42,790 | 2,652 | .062 | .0152 |
| Liabilities and equity | 47,303 | 5,392 | .114 | .0036 |
| Revenues | 16,136 | 1,681 | .104 | .0310 |
| Contract revenue | 8,608 | 529 | .061 | .0766 |
| Operating income/loss | 33,668 | 3,968 | .118 | .0574 |
| Net income/loss | 42,088 | 2,329 | .055 | .0355 |
| Cash and equivalents | 41,448 | 1,941 | .047 | .0310 |
| Operating cash flow | 36,994 | 3,455 | .093 | .0173 |

**Uncertainty and the unit of resampling.** After all frozen decisions, we computed Wilson intervals for proxy precision and recall and 5,000 company-level nonparametric bootstrap intervals for model AP, AP minus resampled prevalence, and AUROC. These descriptive intervals were not used to select a model, feature, threshold, or claim gate. The supplement records the seed, predictions, script, and complete intervals in `uncertainty_result.json`.

The per-concept differences also matter for evaluation design. Pooling all matched facts gives greater weight to concepts with more observations; it does not estimate an equally weighted average across concepts or companies. A changed cash-flow fact and a changed balance-sheet fact need not represent the same economic event. The table records numerical changes under the matching rule, not confirmed error corrections. Consequently, neither a large change rate nor a large median relative change can substitute for the missing fact-level event labels. This distinction remains necessary when reporting uncertainty around a precisely estimated proxy rate.

Company-level resampling addresses sampling variation among observed registrants, not uncertainty about future annotations of unlabelled cases.

## 10 Conclusion

Public SEC XBRL data expose financial-statement version history, but they do not identify why a value changed. In a 13-year panel, changes are common while explicit restatement evidence is historically nonstationary. In two subsequent cohorts, both numerical proxy labels fail against the SEC's error-correction checkbox. The historical fiscal-year baseline does not support a deployable timing claim: future-dependent cohort selection and immature training labels require a new evaluation. Its overwrite contrast does not isolate a causal leakage effect. The earlier-training extension removes unscored months and provides modest ranking evidence, but unknown labels and poor probability calibration remain. For point-in-time learning, defining the observation comes before fitting the model.

## Reproducibility Statement

The main-text reproducibility and integrity section describes accession provenance, fact keys, availability rules, and retained data-construction records. The first-filing and label-maturity revision specifies prediction cutoffs, training-label eligibility, and treatment of unknown outcomes. Retained analysis scripts and summaries distinguish historical intersection-cohort comparisons from the revised audits. Reproducing the construction requires the identified public filing versions; unresolved labels remain missing and must not be silently converted into negatives.

## Ethics statement

The study uses public corporate filings and makes no allegation about any company beyond the contents of its filing. Numerical changes and model scores must not be interpreted as audit, legal, fraud, or investment judgments. The released artifact preserves accession provenance so that cases can be checked against official filings.

## AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

## References

Yogev Bar-On and Yishay Mansour. Non-stochastic bandits with evolving observations. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *PMLR*, pp. 204–227, 2025. URL https://proceedings.mlr.press/v272/bar-on25a.html.

Jeremy Bertomeu, Edwige Cheynel, Eric Floyd, and Wenqiang Pan. Using machine learning to detect misstatements. *Review of Accounting Studies*, 26(2):468–519, 2021. doi: 10.1007/s11142-020-09563-8.

Andrea Carriero, Davide Pettenuzzo, and Shubhranshu Shekhar. Macrocast: A vintage-consistent time series foundation model for real-time macroeconomic forecasting, 2026. URL https://arxiv.org/abs/2606.28670.

Ila Dutta, Shantanu Dutta, and Bijan Raahemi. Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, 90:374–393, 2017. doi: 10.1016/j.eswa.2017.08.030.

Harshavardhan Kamarthi, Alexander Rodríguez, and B. Aditya Prakash. Back2future: Leveraging backfill dynamics for improving real-time predictions in future. In *International Conference on Learning Representations*, 2022.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Zhiwei Liu, Yueru He, Qing Ou, Tianlei Zhu, Xiaorui Guo, Xueqing Peng, and Sophia Ananiadou. Auditfraudbench: Benchmarking audit judgment in detecting fraudulent misstatements, 2026. URL https://arxiv.org/abs/2606.08345.

Matthew R. Lyle, Federico Siano, and Teri Lombardi Yohn. Re-adjusted financial statement data: Challenges and implications for empirical research. SSRN working paper, 2026.

Silu Panda. Finverbench: Benchmark validity and calibration in large language model financial statement verification, 2026. URL https://arxiv.org/abs/2605.29586.

SECDatabase.com. Sec xbrl financial statement dataset. https://github.com/secdatabase/SEC-XBRL-Financial-Statement-Dataset, 2019.

U.S. Securities and Exchange Commission. Release notes for sec taxonomies 2023. https://xbrl.sec.gov/doc/releasenotes-2023.pdf, 2023.

U.S. Securities and Exchange Commission. Edgar xbrl guide. https://www.sec.gov/file/xbrl-guide-2025-09-08, 2025.

U.S. Securities and Exchange Commission. Financial statement data sets. https://www.sec.gov/data-research/sec-markets-data/financial-statement-data-sets, 2026a. Accessed 2026-09-11.

U.S. Securities and Exchange Commission. Financial statement and notes data sets. https://www.sec.gov/data-research/sec-markets-data/financial-statement-notes-data-sets, 2026b. Accessed 2026-09-11.

Xindi Wang. Point-in-time data integrity in quantitative equity research: A reproducible framework for eliminating look-ahead bias. SSRN working paper, 2026.

Chanyuan Zhang Parker, Lanxin Jiang, Soohyun Cho, and Miklos A. Vasarhelyi. Predicting material misstatements using machine learning. *The Accounting Review*, 100(6):225–262, 2025. doi: 10.2308/TAR-2024-0035.

Table 4: Historical FY2024 intersection-cohort test; training-label maturity was not enforced. Lower is better for Brier and ECE; higher is better otherwise.

| Model | Representation | AP | AUROC | Brier | ECE |
|---|---|---|---|---|---|
| Logistic | Original features | .0376 | .5588 | .2140 | .3699 |
| Logistic | Overwritten | .0382 | .5559 | .2127 | .3692 |
| HGB | Original features | .0522 | .6361 | .0682 | .1260 |
| HGB | Overwritten | .0679 | .6342 | .0657 | .1237 |

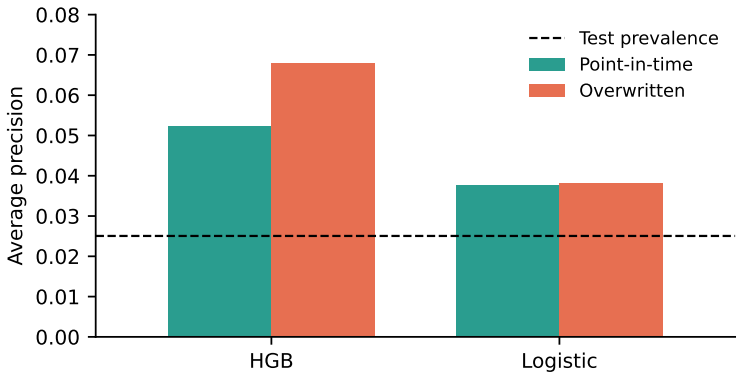


Figure 4: Historical FY2024 comparison; not a deployable timing evaluation. Overwritten facts come from the same later filing that supplies the label and are not deployable.

# A Retrospective Prediction Contrast

## A.1 Fiscal-Year Comparison and Timing Limitations

The historical intersection-cohort test has prevalence 0.0251, which is the AP of a random ranking. Original-feature HGB reaches AP 0.0522 and AUROC 0.6361; logistic regression reaches AP 0.0376 and AUROC 0.5588 (Table 4). A post-hoc company-level bootstrap gives an HGB AP interval of [.0359,.0901] and an AP-minus-prevalence interval of [.0131,.0642]. These intervals describe sampling uncertainty and do not alter the frozen leakage gate. This comparison does not instantiate a strict information-timing protocol. Eligibility used future matched observations, and some training labels were not public at test filing time. The numbers remain retrospective diagnostics and cannot establish deployable predictive performance. They use only ten core facts and one training cohort.

The overwritten HGB has 30.06% higher AP than its original-feature counterpart, but the paired absolute difference is 0.0157 with 95% bootstrap interval $[-0.0160, 0.0621]$. Logistic inflation is 1.40%, with interval $[-0.0015, 0.0057]$. Neither passes the frozen leakage gate, so these data do not establish an inflation effect. Because cohort eligibility and training-label maturity were also compromised, this comparison does not isolate the effect of feature overwriting. A new chronological construction is required before interpreting that effect.

The following figure retains the fiscal-year comparison whose training-label maturity was not enforced. It is historical audit evidence, not a deployable timing result.