Stout in Economics, Finance and Accounting (SEFA)

DOI: https://doi.org/10.61424/b6prv293
Journal Homepage: https://stoutjournals.org/index.php/SEFA

Numerical-Reasoning-Guarded Financial Statement Agents: Auditable Ratio Analysis, Anomaly Detection, and Selective Filing-Consistency Explanations

Clara Zhao
Computer Science, Arizona State University, Tempe, AZ, USA
Corresponding Author: Clara Zhao, **E-mail**: clara.zhao.dev@gmail.com

ARTICLE INFORMATION

Submitted: March 19, 2026
Accepted: July 13, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Financial statement agents; XBRL; numerical reasoning; ratio analysis; anomaly detection; LightGBM; calibration; selective prediction; evidence verification.

ABSTRACT

Financial-statement agents must resolve accounting context before they calculate or explain a ratio. This study evaluated a guarded architecture on the U.S. Securities and Exchange Commission (SEC) Financial Statement Data Sets for 2026 Q1. The archive contained 6,169 submissions, 3,690,955 numeric facts, 91,794 tag-version records, and 733,134 presentation rows. Accession-, statement-, period-, unit-, segment-, and co-registrant-level filtering produced a 5,436-filing ratio cohort and a strict cohort of 3,612 original 10-K filings. A deterministic balance-sheet rule identified 461 observable reconciliation exceptions; this is a filing-consistency review signal, not fraud, default, bankruptcy, or an accounting adjudication. Across 8,400 ratio questions, an unguarded first-fact agent attained 20.51% tolerance-based numeric match, whereas the context-aligned XBRL engine and independent Decimal calculator attained 100%. The guarded agent released 7,827 answers (93.18% coverage); all met the numeric-match tolerance and were supported by re-queried evidence tuples. On a time-ordered test set of 759 filings, LightGBM achieved PR-AUC 0.6780 (95% bootstrap CI [0.6025, 0.7421]), ROC-AUC 0.8495, F1 0.6643, and expected calibration error 0.0125. Isolation Forest achieved PR-AUC 0.1968, close to the 0.1884 test prevalence. Strict context resolution raised numeric match from 20.51% to 100%, and tuple/formula checks verified every released answer. Calibrated supervised screening improved prioritization of reconciliation exceptions using independent filing features.

1. Introduction

Corporate financial statements combine consequential information with a reporting structure that is demanding for automated analysis. A dependable current ratio requires the correct filing, entity, period, statement, unit, taxonomy concept, and dimensional context. Consolidated and segmented values may coexist, and related standard tags are not always interchangeable. XBRL makes these distinctions machine-readable but still requires interpretation. XBRL 2.1 defines concepts, contexts, units, and linkbase relationships, while EDGAR rules govern their use in U.S. public-company disclosures (U.S. Securities and Exchange Commission [SEC], 2026a; XBRL International, 2003). Financial automation is therefore a data-resolution, computation, and provenance problem as much as a language problem.

Language models make this problem more visible: fluent prose does not demonstrate correct arithmetic, and a filing-level link does not establish support for a number. FinQA formalized questions requiring evidence selection and explicit operations (Chen et al., 2021). Program-aided methods improve reliability by assigning arithmetic to executable code (W. Chen et al., 2023; L. Gao et al., 2023), but an executor cannot determine whether operands share the correct entity, duration, or statement. Safeguards are needed before calculation, during execution, and before release.

The SEC Financial Statement Data Sets provide a strong empirical setting. Their SUB, NUM, TAG, and PRE tables link filings, facts, taxonomy metadata, and presentation relationships. The SEC lists the 2026 Q1 release at 81.31 MB and preserves registrants' as-filed values (SEC, 2026b). Context fields retain the accession, concept, period, duration, unit, segment, co-registrant, and value

used. Because this is not a clean rectangular database and filing-quality issues have been documented, validation remains necessary (Debreceny et al., 2010).

Financial ratios support distress and reporting-risk research. Altman (1968) studied bankruptcy, Dechow et al. (2011) examined misstatement risk, and later work compared fraud classifiers (Bao et al., 2020; Cecchini et al., 2010; Perols, 2011). The quarterly archive contains none of those adjudicated outcomes. An unusual ratio or XBRL inconsistency therefore cannot be relabeled as misconduct.

The study defines anomaly detection as review prioritization. Its observable target is a balance-sheet reconciliation exception: the normalized difference between assets and liabilities plus selected equity exceeds fixed relative and absolute thresholds. Such an exception can reflect structured-data inconsistency, concept choice, or unusual entity structure; it does not prove that audited statements are erroneous. Isolation Forest ranks rare multivariate profiles without labels (Liu et al., 2008), while LightGBM estimates association with the exception from independent operating and filing-structure features (Ke et al., 2017). Target components, their direct ratios, and assets-denominated features are excluded to prevent algebraic reconstruction.

The architecture combines four controls: an XBRL engine resolves standard consolidated USD facts on the proper statement; a Decimal calculator executes recorded formulas; Isolation Forest and LightGBM produce review scores; and a verifier releases a card only when cited tuples exist and reproduce the result. It applies tool-using language-model ideas (Schick et al., 2023) under accounting and verification constraints.

Numerical reliability is tolerance-based agreement with a deterministic reference. Screening uses PR-AUC under imbalance (Saito & Rehmsmeier, 2015), while evidentiary reliability requires cited tuples and formulas to reproduce claims (T. Gao et al., 2023). Brier score and ECE assess probability quality (Guo et al., 2017), and selective prediction permits abstention (Geifman & El-Yaniv, 2017). Conformal methods offer separate distribution-free guarantees (Angelopoulos & Bates, 2023); this experiment measures empirical selective risk, not conformal coverage.

The research question is whether context-aware retrieval, deterministic calculation, and verification improve numerical and evidentiary reliability, and whether independent filing features are associated with observable reconciliation exceptions. The contribution is an auditable chain from SEC facts to ratios, calibrated review scores, and selectively released explanations. It supports inspection but does not replace accounting judgment or legal fact-finding.

2. Literature Review

2.1 Structured reporting and accounting context

Under XBRL 2.1, fact meaning depends on concept, entity, period, unit, and dimensions (XBRL International, 2003). EDGAR adds filing and validation conventions (SEC, 2026a), while the U.S.-GAAP taxonomy and Data Quality Committee resources define current concepts and rules (Financial Accounting Standards Board, 2026). Ratios must therefore begin with context resolution: consolidated and segment values cannot be pooled, instant and duration facts require different period logic, and familiar labels do not make custom concepts standard totals.

Calculations 1.1 specifies consistency tests for declared summation relationships (XBRL International, 2023). Neither passing nor failing alone establishes economic correctness or material error. Structured availability likewise does not guarantee data quality (Debreceny et al., 2010). Deterministic rules should establish reproducibility; statistical models can then describe review-order associations.

2.2 Financial screening, calibration, and decision policy

Outcome names must remain precise. Altman (1968) studied bankruptcy, Dechow et al. (2011) studied material misstatements, and Cecchini et al. (2010), Perols (2011), and Bao et al. (2020) used fraud-related labels. Recent applied studies make the same distinction through task-specific targets: calibrated default tiers, probability-of-default workflows, extreme-imbalance fraud detection, and going-concern prediction each depend on a defined outcome and cannot be treated as interchangeable labels (Y. Chen et al., 2023; Y. Chen et al., 2024; Jin et al., 2024a, 2024b). A deterministic reconciliation rule therefore cannot inherit the semantics or evidentiary standards of default, fraud, bankruptcy, or misstatement outcomes.

Isolation Forest ranks easily isolated observations but does not explain their rarity (Liu et al., 2008). LightGBM supplies a supervised complement that accommodates nonlinear effects, missingness, and interactions (Ke et al., 2017). Time ordering and leakage control remain essential: if target components or their direct ratios enter the features, a classifier can reconstruct the

label. Comparable model-risk concerns appear in calibrated recruiter screening and explainable credit scoring, where probability quality, subgroup behavior, and refusal or recourse mechanisms matter alongside ranking accuracy (Jin, 2025a; Xin, 2026b). Precision-recall curves expose retrieval and false-alert trade-offs under imbalance (Saito & Rehmsmeier, 2015). Brier score, ECE, and reliability diagrams are also needed because ranking does not guarantee meaningful probabilities (Guo et al., 2017). Selective classification withholds uncertain cases (Geifman & El-Yaniv, 2017); conformal prediction provides a distinct held-out framework under explicit assumptions (Angelopoulos & Bates, 2023). Privacy-robust uplift modeling, marketing-mix optimization, and earlier LLM-assisted incrementality research extend this decision-oriented view by separating predictive fit from the policy consequences of acting on a score (Bai, Wang, et al., 2026; Bai & Wu, 2026; Mu et al., 2023).

Offline policy evaluation reinforces the same separation between prediction and action. Bandit studies evaluate whether a policy should be selected under logged feedback rather than assuming that a higher model score implies a better intervention (Mu et al., 2025; Ye et al., 2026). Related recommendation work uses behavioral retrieval, review-grounded explanations, and customer representations to preserve the evidence behind a ranked action (X. Chang et al., 2026; Xin, 2026a, 2026e). Profit-aware GPU admission illustrates the same general principle in another operational domain: costs, asymmetric errors, and the content of a decision memo must be evaluated together (S. Zhao et al., 2026).

Financial-agent research increasingly combines these statistical controls with accounting evidence. Accounting-aware RWA retrieval, macro-market fusion, constrained budgeting, and evidence-constrained monitoring show why financial conclusions must retain both numeric inputs and the surrounding disclosure or economic context (Y. Chen et al., 2025; Y. Lu et al., 2025; H. Zhou & K. Zhang, 2025; S. Zhou et al., 2026). Numerical guardrails over SEC and FRED data directly motivate executable formula, unit, window, and citation checks (Z. Li et al., 2026), while financial-search reranking and multi-regulation legal RAG show that retrieval quality and jurisdictional scope remain part of the answer (J. Li & Zhou, 2026; S. Meng et al., 2026). Disclosure review cards and financial-risk dashboards add a presentation requirement: a reviewer should see the source, calculation, uncertainty, and action boundary rather than only a generated conclusion (K. Zhang et al., 2025; Z. Li et al., 2025).

2.3 Numerical reasoning, retrieval, and provenance

FinQA showed that financial questions require fact retrieval and explicit operations (Chen et al., 2021). Program-aided language models and Program-of-Thoughts separate reasoning from arithmetic through executable programs (W. Chen et al., 2023; L. Gao et al., 2023). Once operands, signs, dates, and durations are fixed, deterministic Decimal arithmetic suits ratio calculation.

Execution can still use a wrong formula or unsupported operands. Toolformer demonstrated API invocation (Schick et al., 2023), and verifier research showed the value of checking mathematical outputs (Cobbe et al., 2021). XBRL permits direct row re-query, context checks, and recomputation. The present design therefore treats the calculator as one controlled component rather than as proof that the selected facts were appropriate.

Retrieval-augmented generation grounds text in external evidence (Lewis et al., 2020), but financial evidence must preserve accession, tag, version, date, duration, unit, segment, co-registrant, and value. Citation research separates coverage from correctness (T. Gao et al., 2023). Evidence-grounded DevOps RAG, bilingual incident triage, and word-level source attribution similarly distinguish retrieval from supported generation (B. Zhang et al., 2024; G. Liu et al., 2024; C. Li et al., 2024). Code-intelligence retrieval and budgeted multi-hop agents further show that finding a relevant item, preserving an evidence chain, and stopping within a resource budget are separate design problems (Y. Li, 2024; Su et al., 2025).

Abstention is especially important when retrieval does not cover the question. Evidence-calibrated unanswerable QA connects retrieval coverage to abstention and hallucination analysis, while deterministic provenance work checks whether claims can be traced to source units (Jin, 2025b; Zhong, Chen, et al., 2025). Scientific-search evidence cards make the same distinction visible to a human reviewer, and learned retrieval-quality prediction provides a complementary warning signal before generation (Jin, 2025c; R. Zhang et al., 2025). These ideas support the stricter rule used here: exact tuples and the recorded formula had to reproduce the number; a filing link alone was insufficient.

Operations research supplies useful stress cases because logs, alerts, and runbooks also require evidence-preserving summaries. LLM-assisted root-cause attribution, retrieved normal prototypes, and distributed-log incident cards link a concise narrative to concrete operational observations (G. Liu et al., 2023; J. Nie et al., 2026; Xin, 2025a). Retrieval-grounded HDFS narratives and cloud-native troubleshooting copilots add deterministic or command-safety checks after retrieval (B. Zhang et al., 2026; X. Sun et al., 2026). Cost-aware routing across log-analysis tasks illustrates that an agent may also need to choose how much computation to spend before explaining an incident (C. Li et al., 2026).

Agent reliability depends on the trajectory, not only the final sentence. Tool-use success forecasting makes intermediate failure states explicit (Zhong et al., 2026), while interpretable CloudTrail stage detection illustrates how ordered event evidence can constrain a security narrative (Bai, Chen, et al., 2026). Narrative-aware scientific claim verification and multilingual humanitarian RAG likewise evaluate evidence ranking and contextual coverage rather than fluency alone (Y. Chen & Xu, 2026; Su, Chen, & Qian, 2026). The same logic applies to filings: a ratio explanation is reliable only if every retrieval, alignment, arithmetic, and release step is auditable.

2.4 Guardrails, uncertainty, and distribution shift

Guardrails must remain effective under changing inputs. Continual red-teaming and behavior-level jailbreak defenses treat safety as an update and utility-preservation problem, not a one-time refusal prompt (D. Zheng et al., 2023; Zheng & Li, 2024). Evidence-based self-verification adds a second control after generation (Zheng, Zhang, & Geibel, 2024), and privacy/data-integrity risk cards expose prompt-injection consequences to a human overseer (Su, Rao, & Ma, 2026). Confidence-gated memory and differentially private preference learning show related needs to control what an agent writes, retrieves, updates, or forgets across sessions (Kuo et al., 2025; Sun et al., 2023).

Forecasting and infrastructure studies provide transferable evidence on calibration under drift. Capacity forecasting with calibrated uncertainty, cross-cloud transfer, conformal oversubscription envelopes, and noisy-neighbor residual fusion all separate a point estimate from an operational risk boundary (S. Chen et al., 2024; S. He et al., 2023; S. He et al., 2024; Nie & Zheng, 2024). Multi-horizon GPU forecasting and few-shot tenant forecasting further emphasize temporal horizons, cold start, and workload semantics (S. He et al., 2025; S. Zhao et al., 2025). Power-aware inventory planning and GPU-memory prediction show that resource decisions also require a forecast-to-action translation rather than an isolated error metric (He et al., 2026; Wang et al., 2025).

The methodological lesson is not that infrastructure and accounting are the same task. It is that uncertainty must be evaluated under the deployment distribution relevant to the decision. Approximately shared features can support transfer while acknowledging residual domain differences (Zhong, Pan, et al., 2025); uncertainty-aware fusion likewise requires calibrated component confidence before combining signals (Xin, 2025c). Hybrid-cloud deployment and visual capacity briefs connect model uncertainty to cost and reviewer-facing operating choices (G. Liu et al., 2025; Xin, 2025b). SSD health-state classification offers another example in which sequential representation and attention are assessed for a concrete risk category rather than treated as generic intelligence (Wen et al., 2025).

Time-series and anomaly applications also motivate chronological evaluation. Probabilistic bike-demand forecasting evaluates changing weather and seasonal regimes (Xin, 2026d), while self-supervised and conformal log-anomaly studies distinguish representation learning from alert control and evidence generation (Xin, 2026c, 2026f). Rank aggregation and synchronization research make uncertainty quantification explicit even when the estimator is mathematically structured (Zhong & Ling, 2024a, 2024b). These works support the present use of chronological splits, separate calibration data, and a selective-risk curve rather than a random split and a single threshold metric.

2.5 Human-facing evidence and oversight

Auditable computation still needs an interface that communicates evidentiary status. Medical-image explanation cards, patient-facing safety cards, and their benchmark variants place calibrated scores, warning states, and supporting evidence into a visible hierarchy (C. Li et al., 2025; Zhong, Wu, et al., 2025; B. Zhou et al., 2025). Their domain is different from corporate reporting, but the communication requirement transfers directly: a probability, source fact, generated explanation, and escalation instruction should not be visually conflated.

Evidence-card research in scientific search and e-commerce similarly makes ranking evidence inspectable rather than hiding it behind a recommendation (B. Zhang et al., 2025; Jin, 2025c). Structured visual briefs and text-grounded design-rationale interfaces show how metadata can be converted into editable decision explanations without losing the variables that motivated them (J. Mu et al., 2026; Wu et al., 2025). For finance, accounting-disclosure cards and RWA dashboards are the closest analogues because they organize source text, numeric claims, chart hierarchy, and uncertainty for review (K. Zhang et al., 2025; Z. Li et al., 2025).

Human-in-the-loop systems also need role-appropriate language. Therapist-facing retrieval and evidence-linked counseling cards demonstrate that the same retrieved material may require different presentation for the professional making a decision and the person receiving support (Y. Zhang & H. Zhang, 2025a, 2025b). In financial review, the analogous separation is between

an analyst-facing audit card and any downstream plain-language explanation. The evidence, formula, uncertainty state, and review boundary should stay fixed even when the wording changes.

3. Methodology

3.1 Data source and context alignment

The experiments used the SEC Financial Statement Data Sets release for 2026 Q1. The archive occupied 85,259,424 bytes and had SHA-256 hash d18c01c615da8f5cd46273b0dacaeee5b1163f4089171da0f84153a5f3c362f6. It contained 6,169 SUB records, 3,690,955 NUM facts, 91,794 TAG records, and 733,134 PRE rows. Restriction to annual and quarterly study forms retained 5,812 submissions from 5,620 CIKs. Filing dates spanned January 2 through March 31, 2026. "2026 Q1" identifies the release window; comparative facts from earlier fiscal periods remained contexts within their filing rather than independent 2026 observations.

Table 1. Dataset profile and analytical cohorts

Measure	Value	Unit
Archive compressed size	85,259,424	bytes
SUB records, all forms	6,169	rows
SUB records, study forms	5,812	rows
Distinct study CIKs	5,620	entities
TAG records	91,794	rows
NUM records	3,690,955	rows
PRE records, study forms	688,107	rows
Study filing-date start	2026-01-02	YYYYMMDD
Study filing-date end	2026-03-31	YYYYMMDD
Context-aligned ratio cohort	5,436	filings
Strict annual anomaly cohort	3,612	filings
Reconciliation exceptions	461	filings

Source: SEC Financial Statement Data Sets, 2026 Q1; study calculations.

Original 10-K filings formed 73.33% of the eligible submissions, followed by 10-Q at 16.72%, 20-F at 6.40%, and 40-F at 1.93%. The ratio analysis retained annual, quarterly, and specified foreign forms when a standard U.S.-GAAP concept and valid USD context existed. The supervised anomaly analysis was narrower and used only original 10-K filings.

Table 2. Eligible submissions by filing form

Form	Filings	Entities	Share
10-K	4,262	4,261	73.33%
10-Q	972	936	16.72%
20-F	372	372	6.40%
40-F	112	112	1.93%
10-K/A	47	44	0.81%
10-Q/A	30	22	0.52%
10-KT	8	8	0.14%
20-F/A	8	7	0.14%
40-F/A	1	1	0.02%

Note. Shares use 5,812 eligible submissions as the denominator.

NUM was joined to SUB by accession number. PRE confirmed primary-statement membership through the joint (adsh, tag, version) key, and TAG supplied the authoritative custom-tag indicator. The filters retained standard us-gaap facts with nonmissing values, uom=USD, and empty segments and coreg. Balance-sheet concepts required stmt=BS, inpth=0, qtrs=0, and ddate=SUB.period; income facts required IS placement and operating cash flow required CF placement. Duration had to equal the form's fiscal-period focus: four quarters for annual forms and one, two, three, or four for Q1–Q4. These rules excluded segment members, co-registrants, parentheticals, and mismatched periods.

Concepts followed a fixed priority map. Equity preferred StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest; revenue preferred RevenueFromContractWithCustomerExcludingAssessedTax; and the prioritized cash measure used CashCashEquivalentsRestrictedCashAndRestrictedCashEquivalents when present, otherwise CashAndCashEquivalentsAtCarryingValue. Facts were not added merely because labels were similar. Twelve ratio families were computed without accounting-value imputation. Flow-to-ending-stock ratios were annualized by 4/qtrs; their names distinguish them from ratios based on average balance-sheet stocks.

Table 3. Ratio definitions, availability, and empirical distributions

Ratio	Formula	n	Coverage	Median	P05	P95
Equity / ending assets	Equity / assets	5,212	95.88%	0.3798	-2.5014	0.9094
Annualized net income / ending assets	(Net income / assets) × (4 / duration quarters)	5,133	94.43%	0.0066	-2.1238	0.1640
Annualized operating cash flow / ending assets	(Operating cash flow / assets) × (4 / duration quarters)	5,131	94.39%	0.0157	-1.2186	0.2061
Liabilities / ending assets	Liabilities / assets	4,801	88.32%	0.5715	0.0393	3.3739
Current ratio	Current assets / current liabilities	4,135	76.07%	1.7612	0.0660	13.2722
Annualized asset turnover	(Revenue / assets) × (4 / duration quarters)	3,666	67.44%	0.5317	0.0027	2.1310
Prioritized cash measure / current liabilities	Cash / current liabilities	3,605	66.32%	0.5098	0.0106	7.2669
Profit margin	Net income / revenue	3,524	64.83%	0.0173	-21.8142	0.4160
Operating cash flow margin	Operating cash flow / revenue	3,511	64.59%	0.0827	-11.3928	0.5707
Operating margin	Operating income / revenue	3,029	55.72%	0.0209	-25.4940	0.3403
Receivables / current assets	Receivables / current assets	2,543	46.78%	0.2137	0.0050	0.6419
Inventory / current assets	Inventory / current assets	2,062	37.93%	0.2255	0.0088	0.6530

Note. Flow-to-stock ratios are annualized by 4/qtrs. Percentiles are computed over available filings.

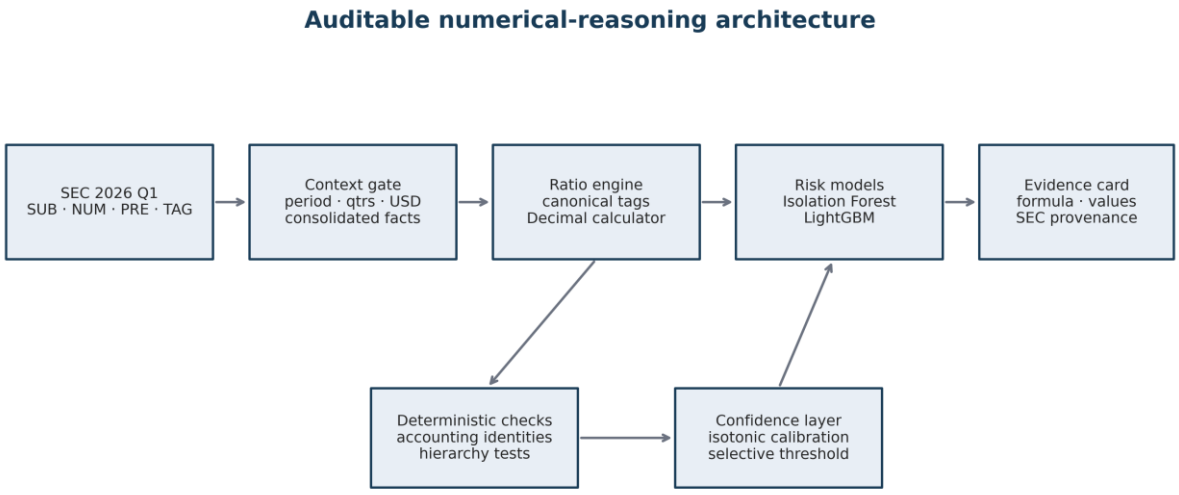

Figure 1. Guarded architecture from SEC facts to selectively released evidence cards
Note. Deterministic context and calculation checks precede statistical screening and explanation release.

At least one computable ratio and a valid filing date were required for the 5,436-filing ratio cohort. Availability ranged from 95.88% for equity to ending assets to 37.93% for inventory share. The anomaly cohort separately required an original 10-K, current-period assets, liabilities, and equity, and at least four observed model features; 3,612 filings remained. Figure 2 records both branches and their test blocks.

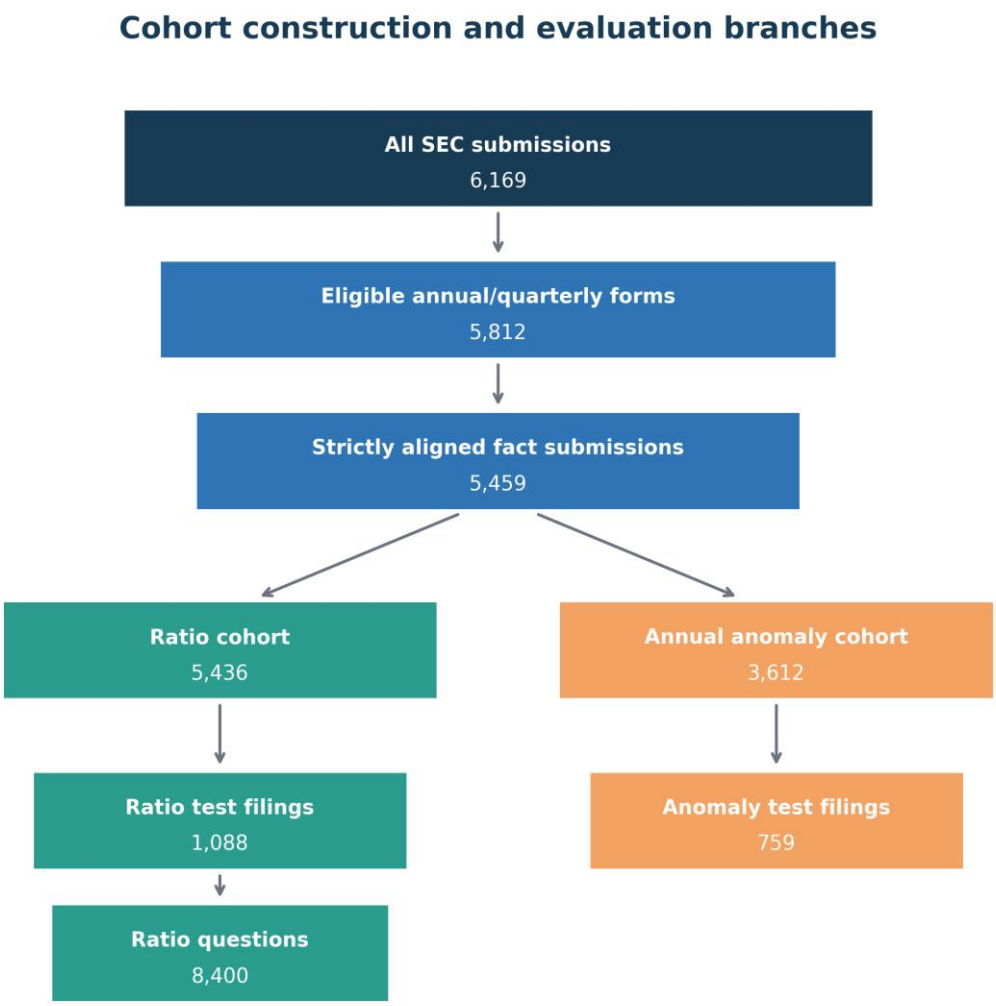

Figure 2. Cohort construction from eligible submissions to numerical questions and anomaly testing
Note. Ratio and anomaly branches apply different eligibility criteria.

3.2 Observable target and time-ordered evaluation

For filing i, reported assets were denoted by A_i, liabilities by L_i, and selected equity by E_i. The symmetric reconciliation residual was:

$$r_i = |A_i - (L_i + E_i)| / \max(|A_i|, |L_i + E_i|, 1).$$

The event indicator equaled one only when $r_i > 0.005$ and the absolute difference exceeded $1,000. The fixed rule identified 461 exceptions among 3,612 filings (12.76%). This outcome is a structured-data reconciliation signal. It is not a fraud, default, bankruptcy, enforcement, or financial-distress label.

The feature matrix excluded assets, liabilities, equity, their direct ratios, the residual, and every assets-denominated feature. Fourteen predictors remained: current ratio, cash ratio, profit margin, operating margin, operating-cash-flow margin, inventory share, receivables share, log absolute revenue, fact coverage, log presentation-line count, presentation-row custom-tag share from TAG, statement count, filing lag, and SIC division. For modeling only, current and cash ratios were clipped to [−20, 100]; the five margin/share features to [−20, 20]; filing lag to [−3,650, 3,650]; and SIC division to [0, 9]. Table 3 reports unclipped ratio distributions. Median imputation was fit on training data only.

Filings were ordered by SEC acceptance timestamp. Training contained 2,178 filings accepted from January 8 through March 11; calibration contained 675 filings accepted March 12–24; and testing contained 759 filings accepted March 25–31. Event prevalence rose from 9.69% to 15.85% and 18.84%, respectively. The shift was retained rather than erased through random sampling.

Table 4. Chronological model-evaluation splits

Split	Filings	Entities	Events	Event rate	Start	End
Train	2,178	2,178	211	9.69%	2026-01-08	2026-03-11
Calibration	675	675	107	15.85%	2026-03-12	2026-03-24
Test	759	759	143	18.84%	2026-03-25	2026-03-31

Note. Splits follow SEC acceptance timestamps; no filing appears in more than one block.

LightGBM, logistic regression, Isolation Forest, and a maximum robust-z baseline were compared. LightGBM used 300 trees, learning rate 0.03, 15 leaves, minimum child size 20, 0.90 row sampling on each boosting iteration, 0.90 feature sampling, L2 penalty 1.0, balanced class weights, and seed 20260729. Isolation Forest used 400 trees, at most 512 training observations per tree, and the same seed. Isotonic calibration used only the calibration block; F1-maximizing calibration thresholds were then frozen. PR-AUC was primary. Secondary metrics were ROC-AUC, F1, precision, recall, Precision@143 (K equaled the 143 test events), precision in the top 5% queue (38 filings), Brier score, and 10-bin ECE. PR-AUC confidence intervals used 200 filing-level bootstrap samples.

3.3 Numerical agents and evidence verification

The ratio test block contained 1,088 filings. Every finite ratio with absolute value at most 10^6 became a question-reference pair, producing 8,400 questions. The context-aligned engine computed the floating-point reference, while an independent Decimal program executed the recorded operands, operation, and duration. The guarded agent released an answer only when both contexts had one distinct candidate value after four-decimal normalization, both exact tuples were found again in NUM, the denominator was valid, Decimal recomputation matched the reference, and the absolute result did not exceed 100. The unguarded ablation still used standard U.S.-GAAP, consolidated, USD facts but selected the first values without statement, period, or duration resolution.

Tolerance-based numeric match required $|\text{prediction} - \text{reference}| \leq \max(10^{-8}, 10^{-6}|\text{reference}|)$. End-to-end match treated abstentions as failures. Structured numerical support required exact tuple re-query, formula reproduction, and agreement with the strict context-selected operands. Each numerical record retained accession, CIK, tag, version, date, duration, unit, segment, co-registrant, operand value, tolerance, formula, and verification fields. Each model-selected card additionally retained the calibrated score, threshold, three reconciliation facts, and both verification results.

4. Results and Discussion

4.1 Ratio coverage and numerical reliability

Equity to ending assets was available for 5,212 filings (95.88%), annualized net income to ending assets for 5,133 (94.43%), annualized operating cash flow to ending assets for 5,131 (94.39%), and liabilities to ending assets for 4,801 (88.32%). Current ratio coverage was 76.07%; receivables and inventory shares were available for 46.78% and 37.93%. Figure 3 shows all twelve ratios.

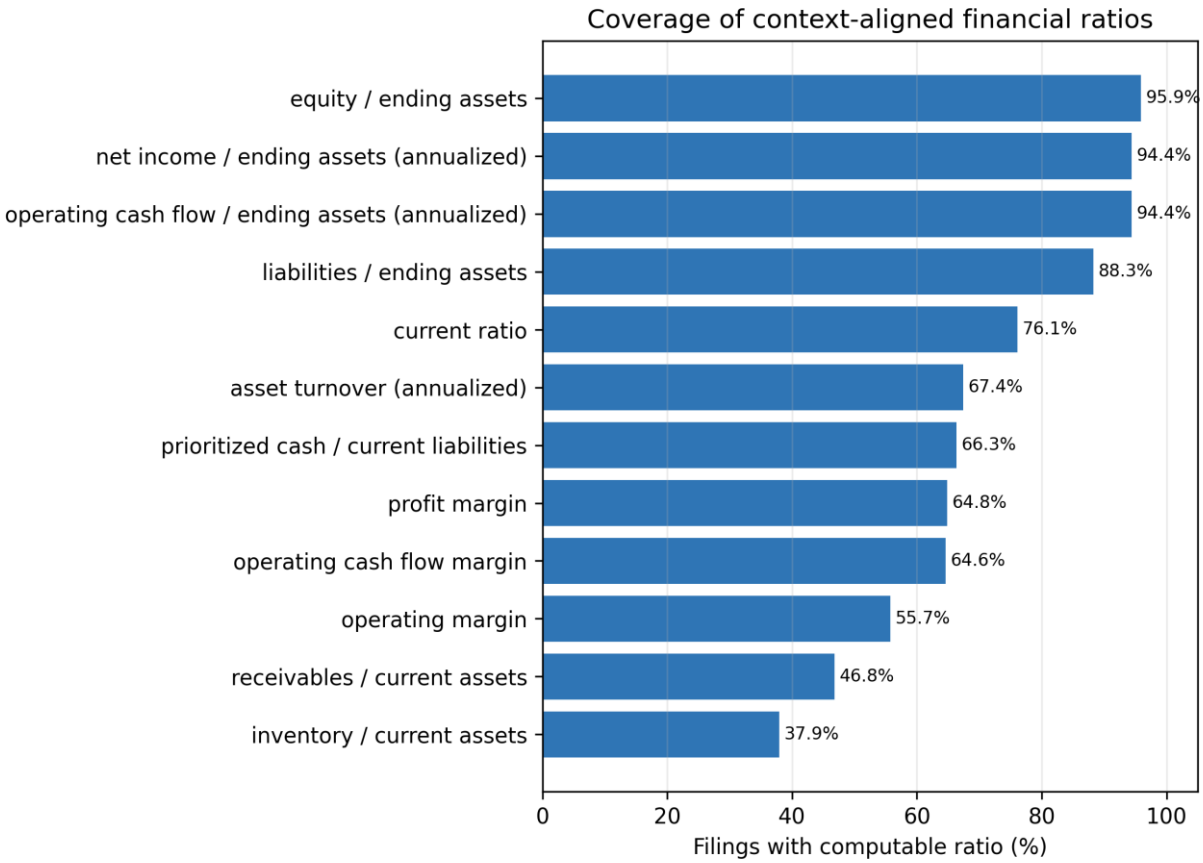

Figure 3. Availability of the twelve context-aligned financial ratios
Note. Coverage uses the 5,436-filing ratio cohort.

The raw distributions were heterogeneous. Median current ratio was 1.7612, with fifth and ninety-fifth percentiles 0.0660 and 13.2722. Median liabilities to ending assets was 0.5715 and its ninety-fifth percentile was 3.3739. Profit margin had median 0.0173 and fifth percentile −21.8142. Missing ratios were not replaced with zero, and an extreme filed ratio was not itself labeled a reconciliation exception.

The context-aligned engine and Decimal calculator answered all 8,400 questions and attained 100% tolerance-based match. Independent Decimal execution differed from the floating-point reference by a mean relative error of 1.85×10^{-19}. The guarded agent released 7,827 answers (93.18% coverage), all matched and structurally supported. The unguarded method answered 8,370 questions; match was 20.51% among answers and 20.44% end to end, with median relative error 0.2492. Its actual raw operands were re-queried, and 19.09% of answered outputs met the strict context-and-formula support criterion.

Table 5. Aggregate numerical-agent evaluation

Method	Answered	Coverage	Tolerance match | answered	End-to-end match	Median rel. error	Structured support
Context-aligned XBRL engine	8,400	100.00%	100.00%	100.00%	0.0000	100.00%
Decimal calculator agent	8,400	100.00%	100.00%	100.00%	0.0000	100.00%
Deterministically guarded agent	7,827	93.18%	100.00%	93.18%	0.0000	100.00%
Unguarded first-fact agent	8,370	99.64%	20.51%	20.44%	0.2492	19.09%

Note. Each method received 8,400 questions. Tolerance match | answered excludes abstentions; end-to-end match counts them as failures. Structured support requires exact tuple re-query, formula reproduction, and strict context agreement.

Guarded coverage was 100% for inventory and receivables shares, 99.85% for the cash ratio, 99.54% for current ratio, and 98.89% for liabilities to ending assets. It was 85.33% for equity to ending assets, 84.58% for annualized net income to ending assets, and 71.19% for profit margin. Among 573 abstentions, 460 reflected nonunique contexts and 113 exceeded the magnitude gate. Every released ratio family retained 100% match; unguarded end-to-end match ranged from 14.35% to 26.96%.

Table 6. Numerical reliability by ratio family

Ratio	Questions	Guarded coverage	Guarded end-to-end match	Unguarded end-to-end match
Annualized asset turnover	575	96.87%	96.87%	20.35%
Annualized operating cash flow / ending assets	963	99.79%	99.79%	20.98%
Annualized net income / ending assets	966	84.58%	84.58%	19.57%
Prioritized cash measure / current liabilities	669	99.85%	99.85%	14.35%
Current ratio	873	99.54%	99.54%	26.12%
Equity / ending assets	1,036	85.33%	85.33%	17.28%
Inventory / current assets	336	100.00%	100.00%	18.15%
Liabilities / ending assets	994	98.89%	98.89%	26.96%

Ratio	Questions	Guarded coverage	Guarded end-to-end match	Unguarded end-to-end match
Operating cash flow margin	540	92.59%	92.59%	18.52%
Operating margin	497	91.75%	91.75%	18.51%
Profit margin	538	71.19%	71.19%	19.70%
Receivables / current assets	413	100.00%	100.00%	19.13%

Note. Context-aligned and Decimal methods achieved 100% end-to-end tolerance match for every ratio family.

Independent Decimal execution preserved 100% match. The uniqueness and magnitude gates retained zero-error answers while reducing coverage by 6.82 percentage points; tuple and formula checks verified every released record.

4.2 Reconciliation-exception screening

LightGBM provided the strongest time-ordered discrimination. Test PR-AUC was 0.6780 (95% bootstrap CI [0.6025, 0.7421]), compared with a test prevalence of 0.1884, and ROC-AUC was 0.8495. At the frozen 0.3333 threshold, precision was 0.6866, recall 0.6434, and F1 0.6643. Precision@143 was 0.6503; precision among the top 38 scores was 0.8947. Brier score was 0.0911 and ECE 0.0125. Figure 4 shows the precision-recall curves.

Table 7. Time-ordered test performance for reconciliation-exception screening

Method	PR-AUC [95% CI]	ROC-AUC	F1	Precision	Recall	P@143	P@5%	Brier	ECE
LightGBM	0.678 [0.603, 0.742]	0.849	0.664	0.687	0.643	0.650	0.895	0.091	0.012
Logistic regression	0.291 [0.246, 0.351]	0.707	0.431	0.305	0.734	0.287	0.316	0.143	0.037
Isolation Forest	0.197 [0.170, 0.232]	0.526	0.328	0.197	0.979	0.119	0.079	0.152	0.013
Robust z-score	0.190 [0.167, 0.220]	0.505	0.318	0.190	0.993	0.126	0.105	0.153	0.011

Note. n = 759; events = 143; prevalence = 0.188. P@5% uses the 38 highest scores. Confidence intervals use 200 filing-level bootstrap samples.

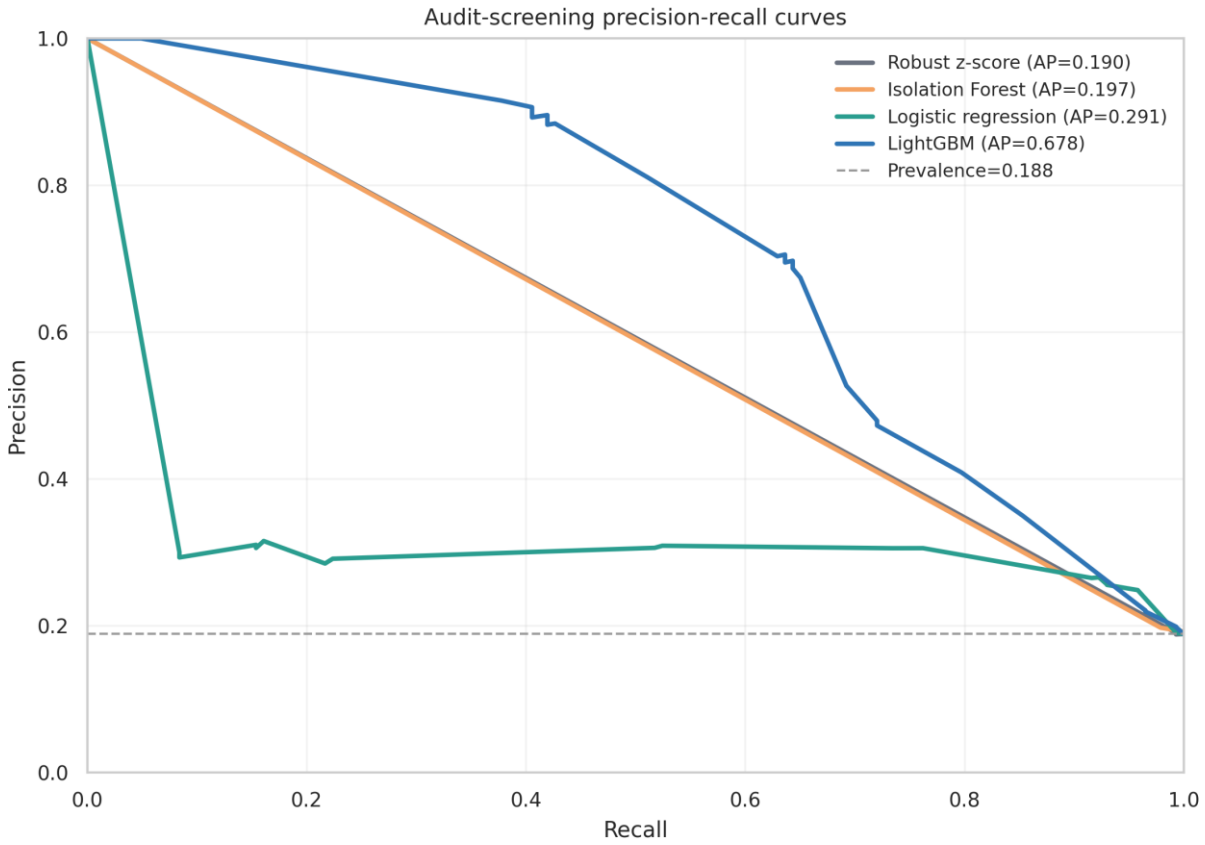

Figure 4. Precision-recall comparison on the chronological test block
Note. The dashed reference equals the 0.1884 test prevalence.

Logistic regression reached PR-AUC 0.2908 and ROC-AUC 0.7066; precision was 0.3052 at 0.7343 recall. Isolation Forest reached PR-AUC 0.1968 and ROC-AUC 0.5263; 0.9790 recall required 0.1969 precision. The robust-z baseline produced PR-AUC 0.1898 and ROC-AUC 0.5046. The unsupervised methods therefore formed broad, low-precision queues near the no-skill prevalence.

Isotonic calibration reduced LightGBM Brier score from 0.1199 to 0.0911 and ECE from 0.1119 to 0.0125. Logistic Brier score fell from 0.2539 to 0.1430 and ECE from 0.3238 to 0.0370. Figure 5 shows the reliability curves. Isolation Forest and robust z had low calibrated ECE (0.0130 and 0.0110) despite near-baseline ranking, so calibration and discrimination must be interpreted separately.

Table 8. Probability calibration results

Method	Probability stage	Brier score	ECE (10 bins)
LightGBM	Raw	0.1199	0.1119
LightGBM	Isotonic	0.0911	0.0125
Logistic regression	Raw	0.2539	0.3238
Logistic regression	Isotonic	0.1430	0.0370
Isolation Forest	Isotonic	0.1525	0.0130
Robust z-score	Isotonic	0.1531	0.0110

Note. Isotonic mappings were fit only on the chronological calibration block.

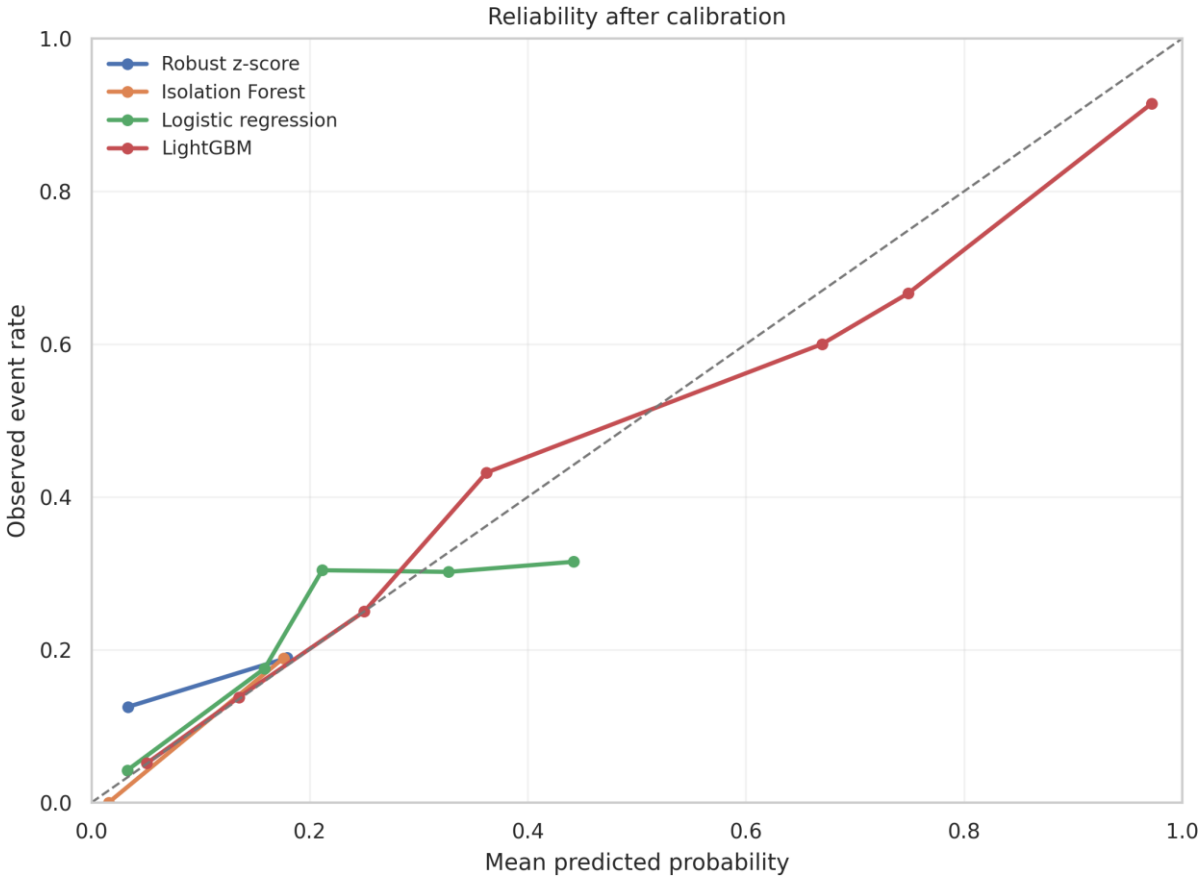

Figure 5. Reliability curves after chronological calibration
Note. Point size represents the number of test filings in each probability bin.

Figure 6 orders cases by 2|p − 0.5| and measures error with a 0.5 decision boundary; it is separate from the frozen 0.3333 operating threshold. LightGBM maintained the lowest selective risk across most coverage levels. This empirical pattern is limited to the observed window and calibration block of 675 filings and 107 events.

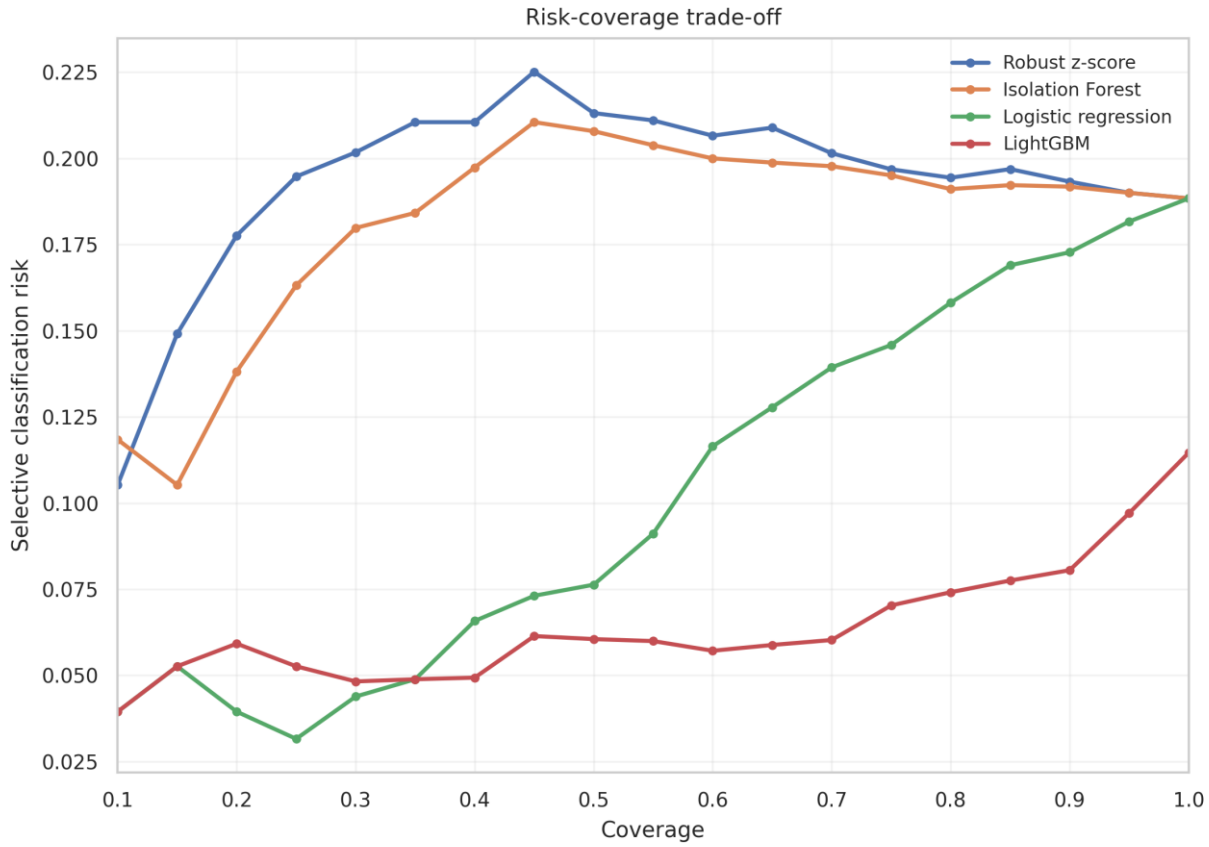

Figure 6. Selective risk as a function of retained coverage
Note. Cases are ordered by 2|p − 0.5| and classified at 0.5; lower risk means fewer retained errors.

4.3 Ablations, feature interpretation, and error analysis

The full 14-feature model had the numerically highest PR-AUC point estimate, 0.6780; metadata only reached 0.6747, with overlapping bootstrap intervals. Removing presentation features reduced PR-AUC to 0.5420, and ratios only reached 0.4100. Metadata-only precision in the top 5% was 0.9474, but the full model had higher F1, recall, Precision@143, and better Brier score and ECE. Metadata and filing-structure variables provided more discriminative information than the operating ratios alone.

Table 9. LightGBM feature-set ablation

Feature set	p	PR-AUC	ROC-AUC	F1	Precision	Recall	P@5%	Brier	ECE
Full feature set	14	0.678	0.849	0.664	0.687	0.643	0.895	0.091	0.012
Metadata only	7	0.675	0.852	0.627	0.714	0.559	0.947	0.094	0.043
No presentation features	10	0.542	0.781	0.595	0.571	0.622	0.737	0.108	0.028
Ratios only	7	0.410	0.757	0.496	0.425	0.594	0.579	0.132	0.046

Note. All variants use identical chronological splits, calibration logic, and test filings.

Presentation-line count was the most frequently used LightGBM feature, followed by presentation-row custom-tag share, revenue scale, operating-cash-flow margin, profit margin, and current ratio. These split importances describe model use, not causal effects. Figure 7 presents the full ranking.

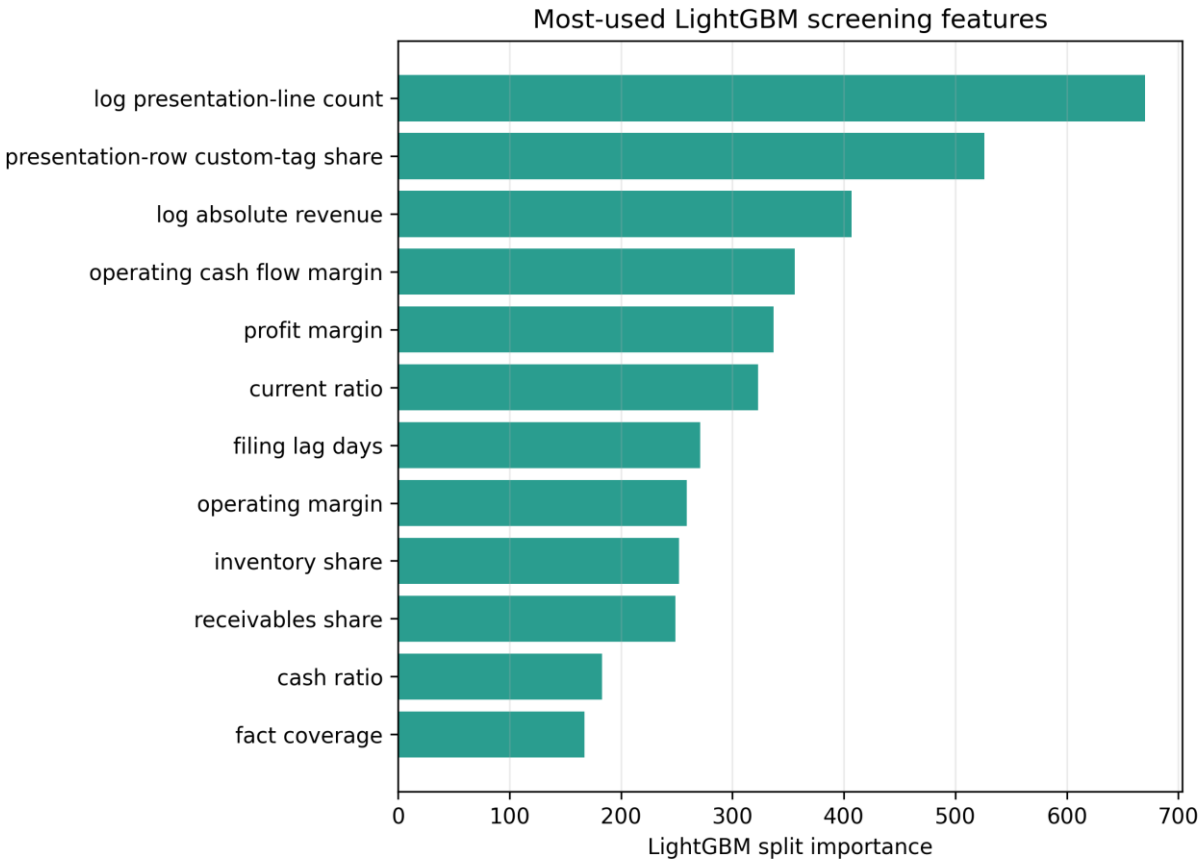

Figure 7. LightGBM split importance for the full feature set
Note. Split counts indicate model use and do not establish causal effects.

The test set contained 733 nonaccelerated filers and 142 of the 143 exceptions. Their PR-AUC was 0.6845, precision 0.6970, recall 0.6479, and ECE 0.0145. The large accelerated group had nine filings and one event; the accelerated group had 17 filings and no events. Table 10 reports all groups, while Figure 8 plots only groups with defined PR-AUC. The smaller samples cannot support stable comparisons.

Table 10. LightGBM test results by accelerated-filer-status group

Group	n	Events	Rate	PR-AUC	Recall	Precision	ECE
Large accelerated filer	9	1	11.11%	0.500	0.000	0.000	0.052
Accelerated filer	17	0	0.00%	—	0.000	0.000	0.136
Nonaccelerated filer	733	142	19.37%	0.684	0.648	0.697	0.014

Note. The two accelerated-filer groups are too small for stable comparative conclusions.

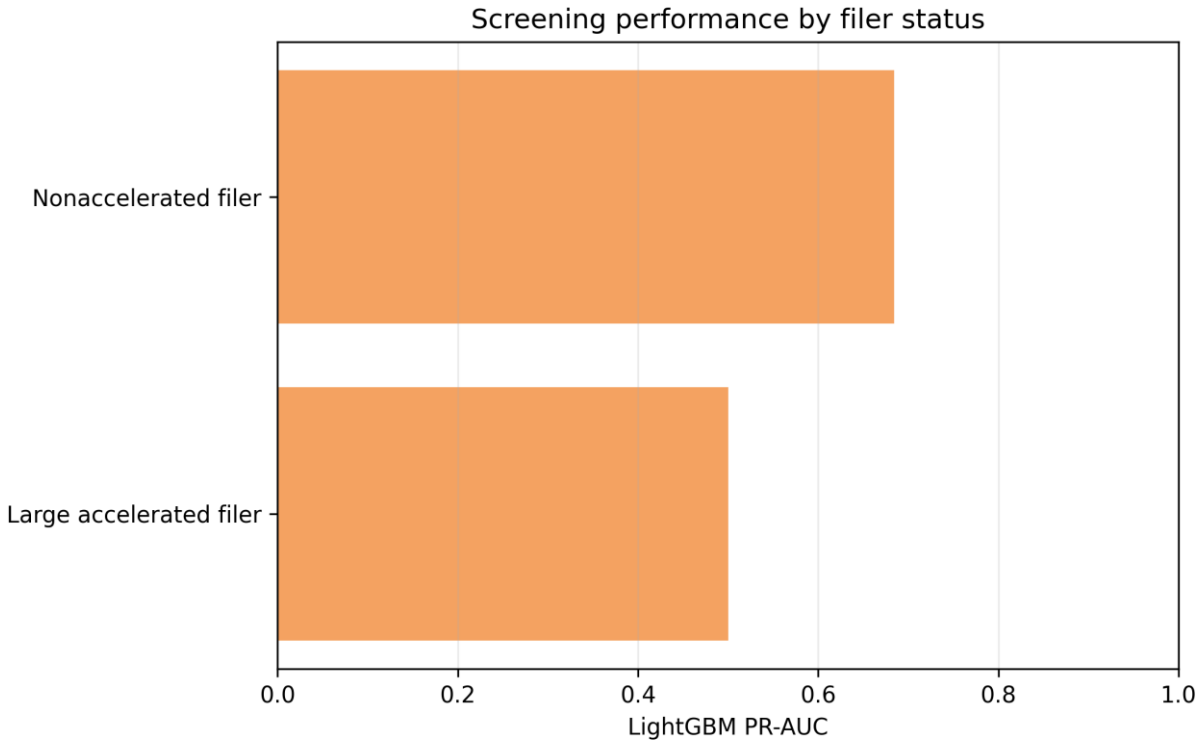

Figure 8. LightGBM PR-AUC by accelerated-filer-status group
Note. The plot includes only groups with defined PR-AUC; Table 10 reports all groups.

The target audit confirmed that all 461 positives arose from the component equation. No filing crossed the direct Assets versus LiabilitiesAndStockholdersEquity threshold, and the current-assets and current-liabilities hierarchy checks also returned zero. Two filings had nonpositive assets, but that diagnostic was not relabeled as a component reconciliation outcome. Original 10-K restriction made the amendment count zero by construction.

Table 11. Deterministic observable-flag audit

Observable check	Filings	Rule
Amended form	0	Form name ends in /A
Direct balance equation	0	|Assets - LiabilitiesAndEquity| / max(|Assets|, |Liabilities + Equity|, 1) > 0.5%
Component balance equation	461	|Assets - (Liabilities + Equity)| / max(|Assets|, |Liabilities + Equity|) > 0.5% and absolute residual > $1,000
Current-assets hierarchy	0	Current assets > 101% of total assets
Current-liabilities hierarchy	0	Current liabilities > 101% of total liabilities
Nonpositive assets	2	Assets ≤ 0

Note. Only the component balance equation defines the modeled outcome.

At the selected threshold, 574 of 616 nonexception filings were left out of the queue and 42 were false alerts. Of 143 exceptions, 92 were prioritized and 51 were missed. The resulting 134 cards comprised those 92 events and 42 nonevents; all 134 passed tuple and formula verification before release.

Table 12. LightGBM test confusion matrix at the frozen threshold

Observed outcome	Not queued	Queued
No observed exception	574	42
Observed exception	51	92

Note. The operating threshold was selected on calibration data and fixed before test evaluation.

Figure 9 shows score separation with remaining overlap. Because the target is a cheap deterministic function of three available facts, LightGBM is a secondary association experiment, not a substitute for computing the identity. Its value is descriptive prioritization; the deterministic reconciliation result remains authoritative.

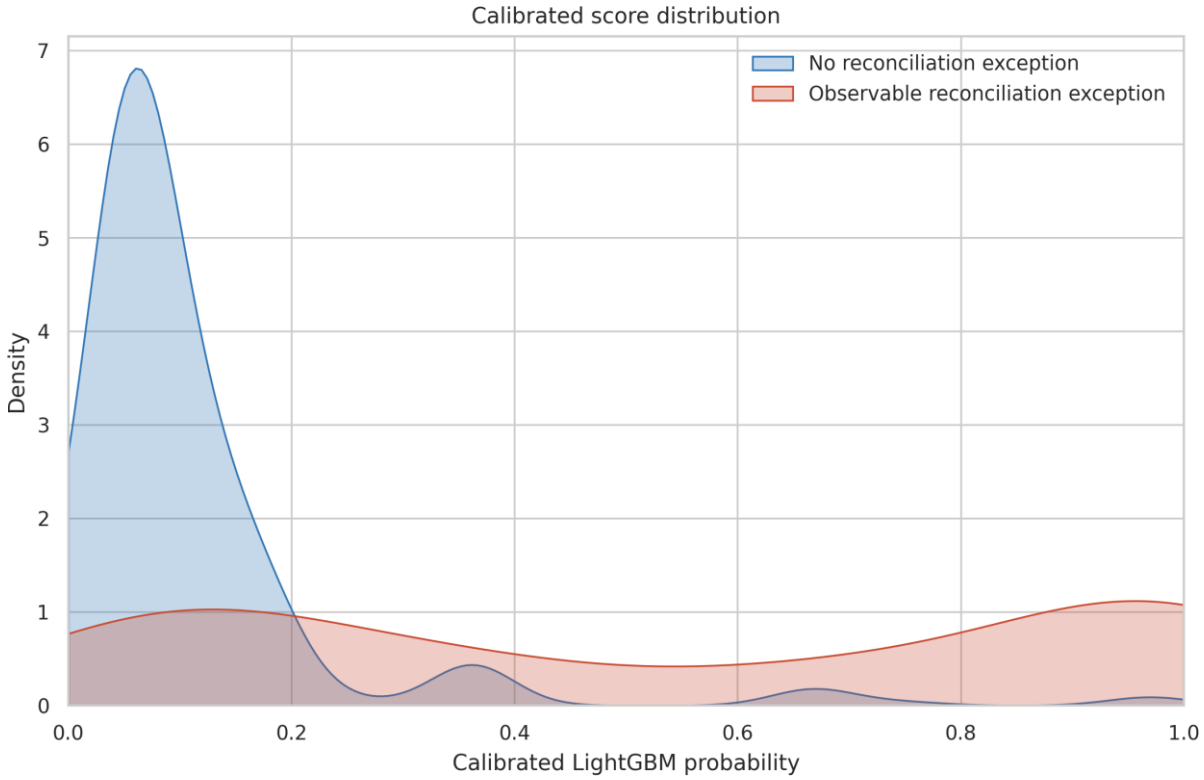

Figure 9. LightGBM calibrated-score distributions by observed test outcome
Note. The frozen decision threshold was 0.3333.

The complete archive-to-output run finished in 62.10 seconds and used 542.91 MB of peak resident memory. Archive extraction took 2.21 seconds, metadata loading 4.69 seconds, alignment and ratios 28.21 seconds, modeling 13.38 seconds, and numerical-agent evaluation 13.52 seconds on one machine using two model worker threads.

Table 13. Runtime and memory profile

Phase	Seconds	Peak RSS (MB)
Extract Archive	2.21	—
Load Metadata	4.69	—
Context Alignment And Ratios	28.21	—
Model Training And Evaluation	13.38	—
Numerical Agent Evaluation	13.52	—
Total	62.10	—
Peak resident memory	—	542.91

Note. Timings were recorded by the reproduction script on one machine; model estimators used two worker threads.

Together, the experiments demonstrate complementary safeguards. Context-aware calculation prevented errors caused by omitting statement, period, and duration resolution; evidence verification made released numbers reproducible; and supervised models quantified associations between independent filing features and reconciliation outcomes. None establishes misconduct or future failure. Taxonomy choice, entity structure, presentation, or extraction can produce a residual even when audited statements are not erroneous. Substantive judgment remains with the reviewer.

5. Conclusions

The SEC 2026 Q1 archive enabled an end-to-end evaluation of guarded financial-statement analysis. Strict context alignment produced 5,436 ratio filings and 3,612 original 10-K filings for reconciliation screening. The unguarded ablation attained 20.44% end-to-end numeric match; the context engine and independent Decimal calculator matched all 8,400 references. Deterministic verification retained tuple-supported answers at 93.18% coverage.

LightGBM achieved test PR-AUC 0.6780, ROC-AUC 0.8495, F1 0.6643, Brier score 0.0911, and ECE 0.0125 under a time-ordered prevalence shift. Isolation Forest and robust-z screening obtained high recall only through broad, low-precision queues. The full feature set had a numerically higher PR-AUC point estimate than metadata alone and delivered the best balance of threshold-based performance and calibration.

The conclusions are limited to one observable outcome and one quarterly release. A reconciliation exception is not fraud, bankruptcy, default, or a validated measure of corporate risk. The test block covered one week, prevalence changed over time, isotonic calibration used 107 events, and two filer groups were too small for stable comparison. The deterministic identity is simpler and authoritative once its three operands are available; model results describe association and queue ordering. Within these limits, calculation guards and evidence verification produced auditable numbers, while every selected review card preserved its source facts, formula, score, threshold, and verification state.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[2] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning, 16*(4), 494–591. https://doi.org/10.1561/2200000101

[3] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Information, Electrical and Electronics Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[4] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment. *Journal of Technology Informatics and Engineering, 5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468

[5] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronic Communication Systems, 6*(1). https://doi.org/10.24042/ijecs.v6i1.30533

[6] Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research, 58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

[7] Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science, 56*(7), 1146–1160. https://doi.org/10.1287/mnsc.1100.1174

[8] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[9] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/JACS.2024.40509

[10] Chen, W., Ma, X., Wang, X., & Cohen, W. W. (2023). Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=YfZ4ZPt8zd

[11] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

[12] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI Credit Card Clients Dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403

[13] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/JACS.2024.40407

[14] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

[15] Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B., & Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3697–3711). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.300

[16] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training verifiers to solve math word problems* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2110.14168

[17] Debreceny, R., Farewell, S., Piechocki, M., Felden, C., & Gräning, A. (2010). Does it add up? Early evidence on the data quality of XBRL filings to the SEC. *Journal of Accounting and Public Policy, 29*(3), 296–306. https://doi.org/10.1016/j.jaccpubpol.2010.04.001

[18] Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research, 28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

[19] Financial Accounting Standards Board. (2026). *2026 GAAP financial reporting taxonomy and Data Quality Committee rules taxonomy technical guide* (Version 2026). https://xbrl.fasb.org/resources/annualrelease/2026/GAAP_Financial_Reporting_Taxonomy_and_Data_Quality_Committee_Rules_Taxonomy_Technical_Guide.pdf

[20] Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., & Neubig, G. (2023). PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 10764–10799). PMLR. https://proceedings.mlr.press/v202/gao23f.html

[21] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6465–6488). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

[22] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4878–4887). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[23] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[24] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108

[25] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[26] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering, 5*(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

[27] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[28] Jin, J. (2025a). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[29] Jin, J. (2025b). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[30] Jin, J. (2025c). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[31] Jin, J., Huang, T., & Lu, S. (2024a). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI Credit Card Default Dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[32] Jin, J., Huang, T., & Lu, S. (2024b). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[33] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146–3154). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[34] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[35] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[36] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[37] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[38] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[39] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[40] Li, Y. (2024). Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[41] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[42] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[43] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

[44] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[45] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[46] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[47] Lu, Y., Zhou, H., & Zhang, Y. (2025). A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling. *Journal of Technology Informatics and Engineering, 4*(3), 493–520. https://doi.org/10.51903/jtie.v4i3.466

[48] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[49] Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards. *International Journal of Graphic Design, 4*(1), 192–208. https://doi.org/10.51903/ijgd.v4i1.3702

[50] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience-creative-channel policies. *Journal of Advanced Computing Systems, 3*(1), 31–48. https://doi.org/10.69987/JACS.2023.30103

[51] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 4*(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

[52] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[53] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112–123. https://doi.org/10.69987/JACS.2024.40409

[54] Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

[55] Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[56] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 68539–68551). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html

[57] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[58] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[59] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[60] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[61] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computer Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[62] U.S. Securities and Exchange Commission. (2026a, June). *EDGAR extensible business reporting language (XBRL) guide*. https://www.sec.gov/files/edgar/filer-information/specifications/xbrl-guide.pdf

[63] U.S. Securities and Exchange Commission. (2026b, March 31). *Financial statement data sets*. https://www.sec.gov/data-research/sec-markets-data/financial-statement-data-sets

[64] Wang, C., Wen, Z., Zhang, R., Xu, P., & Jiang, Y. (2025). GPU memory requirement prediction for deep learning task based on bidirectional gated recurrent unit optimization transformer. In *2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV)*. IEEE. https://doi.org/10.1109/AIVRV67401.2025.11350369

[65] Wen, Z., Zhang, R., & Wang, C. (2025). Optimization of bi-directional gated loop cell based on multi-head attention mechanism for SSD health state classification model. In *2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*. IEEE. https://doi.org/10.1109/ICECAI66283.2025.11171441

[66] Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. *International Journal of Graphic Design, 3*(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

[67] XBRL International. (2003, December 31). *Extensible Business Reporting Language (XBRL) 2.1* [Recommendation; errata corrected through February 20, 2013]. https://www.xbrl.org/specification/xbrl-2.1/rec-2003-12-31/xbrl-2.1-rec-2003-12-31%2Bcorrected-errata-2013-02-20.html

[68] XBRL International. (2023, February 22). *Calculations 1.1* [Recommendation]. https://www.xbrl.org/Specification/calculation-1.1/REC-2023-02-22/calculation-1.1-REC-2023-02-22.html

[69] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

[70] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195. https://doi.org/10.51519/journalisi.v7i3.1170

[71] Xin, Q. (2025c). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[72] Xin, Q. (2026a). Behavior retrieval plus response generation for interpretable conversational personalized recommendation. *International Journal of Electrical, Energy and Power System Engineering, 9*(2), 120–136. https://doi.org/10.31258/ijeepse.9.2.120-136

[73] Xin, Q. (2026b). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit Dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2), 215–231. https://doi.org/10.32664/j-intech.v14i02.2228

[74] Xin, Q. (2026c). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

[75] Xin, Q. (2026d). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. *Transport Findings*. https://doi.org/10.32866/001c.157499

[76] Xin, Q. (2026e). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. *Journal of Information and Technology, 14*(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

[77] Xin, Q. (2026f). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Science, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[78] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

[79] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

[80] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

[81] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

[82] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[83] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2511.19481

[84] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

[85] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

[86] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU Trace. *Journal of Technology Informatics and Engineering, 4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498

[87] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545

[88] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

[89] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/JACS.2023.30203

[90] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

[91] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

[92] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[93] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), iaae020. https://doi.org/10.1093/imaiai/iaae020

[94] Zhong, Z. S., & Ling, S. (2024b). Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2408.05944

[95] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[96] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[97] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[98] Zhou, H., & Zhang, K. (2025). News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015–2024 FRED panel. *Journal of Technology Informatics and Engineering, 4*(2), 487–501. https://doi.org/10.51903/jtie.v4i2.540

[99] Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized. *Journal of Technology Informatics and Engineering, 5*(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544