Detecting Accounting Fraud with AI

Oliver Le
Hamish Macalister
Ben R. Marshall
Nhut H. Nguyen*
Cameron Truong
Qiuhong Zhao
Nuttawat Visaltanachoti

Abstract

We examine whether a commercially developed artificial intelligence (AI) fraud score predicts future financial reporting failures. Using 49,267 U.S. firm-year observations from 2007 to 2024, we find that firms with higher AI fraud scores are significantly more likely to experience financial statement restatements, material (Big-R) restatements, SEC Accounting and Auditing Enforcement Releases (AAERs), and undisclosed SEC investigations over both one- and three-year horizons. The results remain robust after controlling for firm and auditor characteristics and continue to hold after accounting for the Dechow, Ge, Larson, and Sloan (2011) F Score and the machine-learning model of Bao, Ke, Li, Yu, and Zhang (2020), indicating that the AI Score contains incremental information beyond existing fraud prediction models. Using an auditor cost-benefit framework, we further show that the AI Score generates substantial economic value when deployed as a risk-ranking tool. Overall, our findings show that AI can complement existing fraud detection models and improve assessments of financial reporting risk.

Keywords: Accounting fraud; Artificial intelligence; Fraud detection; Financial reporting quality; Audit risk; Machine learning.

JEL Classification: M41, M42, C45

First version: 01 September 2026
This version: 15 September 2026

*Oliver Le: Monash Business School, Monash University. Email: oliver.le@monash.edu. Hamish Macalister: Transparently.AI. Email: hamish@transparently.ai Ben R. Marshall: Massey Business School, Massey University. Email: b.marshall@massey.ac.nz. Nhut H. Nguyen (Corresponding Author): Department of Economics and Finance, Auckland University of Technology. Email: nhut.nguyen@aut.ac.nz. Cameron Truong: Monash Business School, Monash University. Email: cameron.truong@monash.edu. Qiuhong Zhao: Texas A&M University. Email: qiuhong.zhao@tamucc.edu. Nuttawat Visaltanachoti: Massey Business School, Massey University. Email: n.visaltanachoti@massey.ac.nz.

Detecting Accounting Fraud with AI

Abstract

We examine whether a commercially developed artificial intelligence (AI) fraud score predicts future financial reporting failures. Using 49,267 U.S. firm-year observations from 2007 to 2024, we find that firms with higher AI fraud scores are significantly more likely to experience financial statement restatements, material (Big-R) restatements, SEC Accounting and Auditing Enforcement Releases (AAERs), and undisclosed SEC investigations over both one- and three-year horizons. The results remain robust after controlling for firm and auditor characteristics and continue to hold after accounting for the Dechow et al. (2011) F Score and the machine-learning model of Bao et al. (2020), indicating that the AI Score contains incremental information beyond existing fraud prediction models. Using an auditor cost-benefit framework, we further show that the AI Score generates substantial economic value when deployed as a risk-ranking tool. Overall, our findings show that AI can complement existing fraud detection models and improve assessments of financial reporting risk.

Keywords: Accounting fraud; Artificial intelligence; Fraud detection; Financial reporting quality; Audit risk; Machine learning.

JEL Classification: M41, M42, C45

This version: 15 September 2026

1 Introduction

Financial statement fraud remains one of the most significant threats to capital market integrity. Investors, creditors, regulators, and auditors rely on financial statements to allocate capital and monitor corporate performance. When those statements are intentionally manipulated, the consequences extend well beyond the firms involved. Fraudulent financial reporting erodes investor confidence, increases firms' cost of capital, distorts resource allocation, and undermines trust in financial markets (e.g., Guiso, Sapienza, and Zingales 2008). High-profile accounting scandals, including Enron, WorldCom, Luckin Coffee, and Wirecard, illustrate that even sophisticated governance mechanisms and external audits can fail to detect material misstatements before they inflict substantial economic losses.

Because accounting fraud is typically concealed until well after the reporting period, identifying firms with elevated fraud risk before misconduct becomes public is important. Early identification allows auditors to allocate audit effort more efficiently, enables regulators to prioritize investigations, and assists investors in assessing reporting risk. Consequently, a large body of literature has developed models to predict fraudulent financial reporting using publicly available accounting information. Beneish (1999) introduces the M-score, while Dechow et al. (2011) develop the F Score by combining multiple financial reporting characteristics associated with accounting manipulation. More recently, machine-learning techniques have substantially improved fraud prediction by exploiting nonlinear relationships among accounting variables (e.g., Cecchini, Aytug, Koehler, and Pathak 2010; Bao et al. 2020).

Despite these advances, existing fraud-prediction models face several limitations. First, most models are estimated only on detected fraud cases. However, a large share of manipulation is never uncovered, so models that condition on detected cases alone understate the true incidence of misreporting; Dyck, Morse, and Zingales (2024) estimate that, in an average year, approximately 41 percent of large public companies misrepresent their financial reports and about 10 percent commit alleged securities fraud. Second, models validated

on archival data face information leakage and survivorship bias. Information leakage arises when a model is trained on information that was not actually available when the prediction would have been made, such as figures from subsequently amended filings or restatements. Survivorship bias arises when knowledge of a firm's later inclusion in, or exclusion from, the sample is embedded in a score dated to an earlier period. Third, traditional accounting-based models rely on manually specified financial ratios and on linear or relatively simple nonlinear functional forms, which may limit their ability to capture complex interactions embedded in financial statements. Although recent machine-learning models improve predictive accuracy, they generally require researchers to construct and train proprietary models using large sets of accounting variables. Such approaches may not be readily accessible to practitioners and often rely exclusively on structured accounting data, potentially overlooking additional information embedded in firms' broader reporting environment.

This paper examines whether advances in artificial intelligence (AI) can improve the identification of firms that subsequently experience financial reporting failures. Specifically, we evaluate the predictive ability of the AI fraud score developed by Transparently.AI, a commercial AI platform designed to assess the likelihood of accounting manipulation using publicly available corporate information. The AI fraud score is not trained only on detected cases, as other models are. It uses evidence from forensic accounting, short-seller research, credit analysis, and fundamental equity analysis, and applies over 200 econometric and machine-learning sub-models organized into 14 risk clusters and aggregated into a company-level score.

Unlike conventional accounting-based fraud models that rely on a predetermined set of financial ratios, the AI system employs a proprietary framework that integrates a broad set of financial reporting signals to generate a firm-year fraud risk score. Given the score is generated independently of our empirical analyses, it provides an opportunity to evaluate whether modern AI techniques contain incremental information about future accounting fraud.

We examine four increasingly severe measures of financial reporting problems. The first measure includes all financial statement restatements, capturing both intentional and unintentional reporting errors. The second focuses on Big-R restatements that materially affect previously issued financial statements. The third measure uses SEC Accounting and Auditing Enforcement Releases (AAERs), which identify firms subject to formal SEC enforcement actions for accounting violations. The final measure captures undisclosed SEC investigations, providing a broader measure of regulatory scrutiny that extends beyond publicly announced enforcement actions. Examining multiple outcome measures allows us to assess whether the AI Score predicts both relatively common reporting failures and more severe instances of accounting fraud.

Using 49,267 firm-year observations from 2007 to 2024, we find that firms with higher AI fraud scores are significantly more likely to experience financial reporting problems in subsequent years. The AI Score positively predicts all four fraud measures over both one-year and three-year forecasting horizons. The results remain robust after controlling for a comprehensive set of firm characteristics, auditor characteristics, industry fixed effects, and year fixed effects. More importantly, the predictive ability of the AI Score remains economically and statistically significant after controlling for two leading fraud-detection models: the machine-learning model of Bao et al. (2020) and the F Score of Dechow et al. (2011). These findings suggest that the AI Score contains information about future reporting failures that existing accounting-based fraud measures do not capture.

Prediction accuracy alone does not necessarily imply practical usefulness. Auditors face an asymmetric decision problem in which failing to identify a fraudulent client may expose them to litigation, reputational damage, and client losses, whereas incorrectly identifying a non-fraudulent client increases audit costs through unnecessary investigation. Accordingly, we translate statistical prediction into economic value using the cost-benefit framework of Beneish and Vorst (2022). We show that although the economic value of the AI Score depends on assumed costs of false positives and false negatives, the AI Score generates substantial

net benefits when deployed as a risk-ranking tool targeting firms with the highest predicted fraud risk.

Our study contributes to three strands of literature. First, we contribute to the accounting fraud literature. Early fraud prediction models, including Beneish's (1999) M-score and Dechow et al.'s (2011) F Score, show that publicly available accounting information can identify firms with an elevated likelihood of financial misreporting. Subsequent studies employ machine-learning techniques to improve fraud detection by modeling complex nonlinear relationships among accounting variables (e.g., Cecchini et al. 2010; Bao et al. 2020). We advance this literature in two important ways. Rather than proposing another prediction model, we independently evaluate a commercially developed artificial intelligence system trained outside the academic setting. We also show that the AI Score provides incremental predictive power beyond the two leading academic fraud measures, indicating that modern AI systems extract information about future financial reporting failures that existing accounting-based and machine-learning models do not capture.

Second, we contribute to the emerging literature that views artificial intelligence as a measurement technology in accounting and finance. Recent studies demonstrate that AI can extract economically meaningful signals from large volumes of structured and unstructured data, enabling researchers to measure constructs that are difficult to observe directly. For example, Cohen, Lu, and Nguyen (2026) use AI to identify and forecast mutual fund managers' trading behavior, while Li, Mai, Shen, Yang, and Zhang (2026) employ generative AI to construct measures of corporate culture from analyst reports. We extend this literature by demonstrating that AI can also generate informative measures of firms' latent fraud risk. Our results show that AI-generated fraud scores provide incremental information beyond existing accounting-based and machine-learning fraud models, suggesting that AI captures dimensions of financial reporting quality that conventional empirical techniques struggle to identify. More broadly, our findings illustrate the growing role of AI as a powerful measurement tool for accounting and finance research.

Third, our study bridges the gap between academic fraud prediction models and the practical adoption of artificial intelligence. Most prior studies develop fraud prediction models within academic settings, leaving open whether AI systems used in practice provide reliable and economically meaningful signals. By independently evaluating a commercially developed AI platform, we provide evidence on the effectiveness of AI as implemented in real-world decision environments. Our findings therefore offer practical implications for auditors, regulators, investors, and lenders seeking to incorporate AI into fraud risk assessment.

The remainder of the paper is organized as follows. Section 2 describes the data, sample selection procedure, variable construction, and empirical methodology. Section 3 presents the empirical results. We first report descriptive statistics and baseline evidence on the AI Score's ability to predict future financial reporting failures over one- and three-year horizons. We then examine whether the AI Score provides incremental predictive power beyond existing fraud detection models and assess its economic usefulness through an auditor cost-benefit analysis. Section 4 concludes.

2 Data and Sample Selection

2.1 Sample Construction

Our sample consists of U.S. publicly listed firms covered by Transparently.AI from 2007–2024. We merge the AI Score with firm-level accounting information from Compustat, stock return data from CRSP, audit-related variables from Audit Analytics, and financial reporting outcomes from multiple regulatory databases. The final sample comprises 49,267 firm-year observations. Because the AI Score is not available for every publicly listed firm in every year, we determine the sample based on the availability of the AI measure.

To evaluate the predictive ability of the AI Score, we examine four measures of future financial reporting problems. The first, *Restatement*, equals one if the firm subsequently restates its financial statements and zero otherwise. Because some restatements arise from

unintentional reporting errors, we also consider *BigR*, which captures material restatements that require withdrawing previously issued financial statements. Our third measure, *AAER*, identifies firms subject to SEC Accounting and Auditing Enforcement Releases, representing formal enforcement actions involving accounting violations. Finally, *SEC* captures undisclosed SEC investigations that become publicly known only after regulatory action (e.g., Blackburne, Kepler, Quinn, and Taylor 2021). Collectively, these measures span a continuum from relatively broad reporting failures to severe regulatory enforcement actions.

2.2 AI Fraud Score

Our primary independent variable is the *AI Score*, obtained from Transparently.AI. The score is generated using a proprietary artificial intelligence framework that evaluates evidence from forensic accounting, short-seller research, credit analysis, and fundamental equity analysis. Approximately 200 econometric and machine-learning sub-models are organized into 14 risk clusters and aggregated into a company-level score. The cluster structure makes elevated readings attributable to observable categories of concern even though the individual features, ensemble weights, and calibration mechanics remain proprietary.

The score ranges from zero to one, with higher values indicating a greater predicted probability of fraudulent financial reporting and subsequent manipulation-related collapse. It is therefore not a calibrated probability of a restatement, AAER, or SEC investigation. The target construct is the consequential subset of manipulation associated with severe corporate failure. A central feature of the model is its point-in-time protocol. Each company-month uses only statements announced by that month-end; delisted and suspended firms remain in the historical universe; and the complete model is re-estimated on the information set then available. This design prevents look-ahead bias, whereby subsequent filings or future universe membership could enter a dated score. Appendix A explains the training universe, signal construction, aggregation, interpretation, and prior validation in detail. Because the underlying algorithm is developed independently of this study, our analysis focuses on

evaluating its predictive performance rather than constructing the model itself.

2.3 Benchmark Fraud Models and Control Variables

To assess whether the AI Score contains information beyond established fraud-prediction models, we benchmark it against the two leading academic measures. The first is the F Score of Dechow et al. (2011). We compute the F Score by applying the published coefficients of Dechow et al.'s (2011) Model 1, which combines RSST accruals, changes in receivables and inventories, the proportion of soft assets, the change in cash sales, the change in return on assets, and an indicator for securities issuance. The fitted probability is scaled by the unconditional probability of misstatement, so that values above one indicate above-average misstatement risk. Because the model requires lagged financial statement data, the F Score is available for a subset of our sample. Exact definitions of all inputs appear in Appendix Table A1.

The second benchmark, labeled BKLYZ in the tables, is the machine-learning measure of Bao et al. (2020). We generate the score by re-estimating their RUSBoost ensemble model with the authors' published code and data, using the 28 raw financial statement variables specified in their study. Following their design, the model for each test year is trained only on prior years, with a two-year gap between the training and test periods to ensure that no information unavailable at the prediction date enters the model; training observations belonging to serial fraud cases that overlap the test year are relabeled as non-fraud, following Bao et al.'s (2020) serial-fraud treatment. The resulting score is the ensemble's predicted fraud signal for each firm-year. Because the underlying training data end earlier than our sample period, the BKLYZ measure covers fewer observations than the AI Score or the F Score, as reported in Table 1.

All multivariate specifications include firm and auditor characteristics commonly used in the restatement and fraud-prediction literature. Firm characteristics comprise Size, Leverage, ROA, Loss, Total Accruals, Current Ratio, Inventory_ARec, Business Segment, M&A,

and Foreign, which capture scale, financial condition, performance, accrual intensity, and operating complexity. Auditor and reporting-environment characteristics comprise Internal Control, BigN, Going Concern, Litigation, Auditor Tenure, and NAS Ratio, which capture internal control quality, auditor type and tenure, going-concern risk, litigation exposure, and the extent of non-audit services. Continuous variables, including the three risk scores, are winsorized at the 1st and 99th percentiles. Appendix Table A1 defines every variable and identifies its source.

3 Empirical Results

This section evaluates whether the AI fraud score identifies firms that subsequently experience financial-reporting problems. The analysis proceeds in four steps. First, we describe the sample, the outcome variables, and the relation between the AI Score and established fraud-prediction measures. We then examine whether the score predicts reporting failures over one- and three-year horizons. The next step pits the AI Score directly against the machine-learning measure of Bao et al. (2020) and the F Score of Dechow et al. (2011). The last step translates predictive classifications into an auditor cost-benefit framework. Taken together, the empirical results examine three distinct properties of the score: predictive validity, incremental information content, and economic usefulness.

3.1 Sample Characteristics

Table 1 reports 49,267 firm-year observations spanning 2007–2024. The sample covers 2009–2023; the endpoint years contain relatively few observations. The industry distribution is also uneven. Business-equipment firms account for 16,905 observations, or approximately 34.3 percent of the sample, and chemical firms account for another 9,453 observations, or 19.2 percent. Manufacturing and the residual "Other" category contribute 5,041 and 5,100 observations, respectively, while the Money and Utilities categories contain no observations.

Including industry and year fixed effects in the multivariate tests is therefore important because the raw sample composition varies materially across both dimensions.

The AI Score has a mean and median of 0.567 and 0.566, respectively, and a standard deviation of 0.231, indicating substantial cross-sectional variation, with values ranging from 0.146 to 0.956 and an interquartile range of 0.366 to 0.768. Thus, the score is not concentrated near a single value and varies enough to distinguish firms along the predicted-fraud dimension. Approximately 7.5 percent of firm-year observations subsequently experience a financial statement restatement, while material Big-R restatements occur in approximately 2 percent of observations. AAERs are relatively rare (0.4 percent of the observations within the AAER coverage window), whereas undisclosed SEC investigations occur in approximately 9.5 percent of the observations for which the investigation data are available. These frequencies illustrate the importance of considering multiple measures of financial reporting failures that differ in both prevalence and severity. Ordinary restatements are relatively frequent but may include unintentional errors, Big-R restatements capture economically material reporting failures, AAERs capture a narrow set of cases that result in public SEC enforcement, and the undisclosed-investigation measure captures a broader set of regulatory concerns. Consistent results across these measures would therefore be difficult to attribute to the idiosyncrasies of any single fraud proxy.

The remaining descriptive statistics indicate that the sample contains many firms with characteristics commonly associated with reporting risk: 43.5 percent report a loss, average return on assets is -12.8 percent, 8.6 percent report an internal-control weakness, and 7.3 percent receive a going-concern opinion. At the same time, 72.4 percent are audited by a Big N auditor. These features reinforce the need for multivariate specifications to distinguish the incremental content of the AI Score from readily observable financial distress and audit-risk characteristics.

The AI Score is observed for all 49,267 Table 1 observations, compared with 41,677 for the F Score and only 14,962 for BKLYZ. Coverage is therefore a potentially important practical

advantage of the AI Score, but it also means that cross-model comparisons must be made on common samples.

3.2 Baseline Predictive Validity

3.2.1 One-Year-Ahead Reporting Outcomes

Table 2 examines whether the AI Score measured in year t predicts reporting outcomes in year $t+1$. The fully specified models include the set of firm and auditor controls reported in the table, industry and year fixed effects, and standard errors clustered by firm. Across all four outcomes, the coefficient on the AI Score is positive and statistically significant at the 1 percent level. The coefficient is 0.069 for any restatement ($t = 4.76$), 0.040 for Big-R restatements ($t = 5.50$), 0.014 for AAERs ($t = 3.80$), and 0.219 for undisclosed SEC investigations ($t = 8.53$). The evidence therefore extends from relatively broad measures of reporting error to material restatements and regulatory outcomes.

The magnitudes are economically meaningful relative to the low unconditional incidence of these events. Using the Table 1 standard deviation of 0.231, a one-standard-deviation increase in the AI Score corresponds to an approximately 1.59-percentage-point increase in the probability of any restatement, a 0.92-percentage-point increase in Big-R risk, a 0.32-percentage-point increase in AAER risk, and a 5.05-percentage-point increase in the probability of an undisclosed SEC investigation. Relative to the Table 1 outcome means, these effects are approximately 21 percent, 47 percent, 90 percent, and 53 percent, respectively. The relative calculation for AAERs is particularly large because AAERs are rare; in absolute terms, the implied increase remains about one third of a percentage point. These standardized magnitudes should be interpreted as approximations because the descriptive and regression samples reported in Tables 1 and 2 are not identical.

The adjusted R^2 ranges from 0.007 for AAERs to 0.072 for undisclosed SEC investigations. These values are unsurprising for rare-event outcomes and should not be read as evidence that the score deterministically classifies fraud. More relevant is that the AI Score is

reliably associated with future outcomes after conditioning on financial performance, accruals, firm complexity, internal-control weaknesses, auditor characteristics, and fixed effects. This temporal ordering and conditioning support a predictive interpretation. They do not, however, establish a causal effect of the score itself; the score is a diagnostic signal of latent reporting risk.

The control-variable estimates also provide a useful benchmark. Internal-control weaknesses are positively associated with next-year restatements and undisclosed SEC investigations, while Big N auditors are negatively associated with restatement, Big-R, and undisclosed SEC outcomes. Merger activity is positively associated with restatements and Big-R outcomes. These patterns indicate that the models capture familiar dimensions of reporting and audit risk. The AI Score remains significant in their presence, suggesting that it aggregates additional information rather than merely proxying for a single observable risk factor.

3.2.2 Persistence Over a Three-Year Horizon

Table 3 replaces the one-year outcome with an indicator for whether the event occurs during years $t+1$ through $t+3$. The coefficient on the year-t AI Score remains positive and significant at the 1 percent level for every outcome: 0.138 for any restatement ($t = 4.962$), 0.083 for Big-R ($t = 5.970$), 0.030 for AAERs ($t = 3.716$), and 0.320 for undisclosed SEC investigations ($t = 7.944$). The number of observations falls to 31,343 and 30,836 for the restatement outcomes because constructing a complete forward window requires additional future data, and further to 19,305 for AAERs and 14,623 for SEC investigations because the full outcome window must also fall within each source's coverage. Depending on the outcome, the models contain approximately 3,300–4,700 firm clusters.

The longer-horizon coefficients are larger than their one-year counterparts, as expected when the dependent variable captures the occurrence of an event over three years rather than in one year alone. The coefficient increases by a factor of 2.00 for restatements, 2.08 for Big-R, 2.14 for AAERs, and 1.46 for undisclosed SEC investigations. A one-standard-deviation

change in the AI Score corresponds to approximately 3.18, 1.91, 0.69, and 7.38 percentage points, respectively. The adjusted R^2 also rises to 0.098 for restatements, 0.118 for Big-R, and 0.151 for undisclosed SEC investigations; it remains lower for AAERs at 0.027. The AI signal therefore does not dissipate immediately. Instead, it identifies reporting risk that can materialize over several subsequent reporting cycles.

The horizon comparison is informative for how the score may be used. A one-year prediction is most relevant to the immediate audit-planning decision, whereas the three-year result suggests that the score also captures slower-moving features of financial reporting quality. The evidence is consistent with persistent latent risk rather than a signal tied only to an isolated filing. Nevertheless, the larger three-year coefficients partly reflect the mechanical increase in cumulative event incidence, so they should not be interpreted as evidence that the model becomes intrinsically more accurate as the forecast horizon lengthens.

3.3 Incremental Information Beyond Established Fraud Models

Table 4 evaluates whether the AI Score contributes information beyond BKLYZ and the F Score. Panel A reports one-year-ahead results, and Panel B uses the $t+1$ through $t+3$ horizon. Odd-numbered columns pair the AI Score with BKLYZ; even-numbered columns pair it with the F Score. All models retain the control variables, industry and year fixed effects, and firm-clustered standard errors.

For the one-year horizon in Panel A, the AI Score is positive and statistically significant in every head-to-head specification. In the BKLYZ common sample, its coefficients are 0.066 for restatements, 0.053 for Big-R, 0.018 for AAERs, and 0.255 for undisclosed SEC investigations. The first estimate is significant at the 5 percent level and the remaining estimates at the 1 percent level. BKLYZ is positively associated with restatements and Big-R, but is not statistically significant for AAERs or undisclosed SEC investigations. In the larger F Score common sample, the AI coefficients are 0.071, 0.040, 0.011, and 0.213, all significant at the 1 percent level. The F Score is positively associated with restatements, Big-R, and

AAERs, but negatively associated with the undisclosed SEC measure. Importantly, the AI coefficients in the F Score sample are close to the Table 2 estimates of 0.069, 0.040, 0.014, and 0.219. The baseline results are thus not materially altered by conditioning on a widely used accounting-based fraud score.

Panel B yields the same conclusion over the three-year horizon. In the BKLYZ common sample, the AI coefficients are 0.120 for restatements, 0.082 for Big-R, 0.034 for AAERs, and 0.326 for undisclosed SEC investigations. In the F Score common sample, the corresponding coefficients are 0.137, 0.082, 0.025, and 0.320. Every AI coefficient is statistically significant at the 5 percent level or better. BKLYZ becomes positively significant for all four cumulative outcomes, while the F Score is positive for restatements, Big-R, and AAERs; its negative one-year association with undisclosed SEC investigations does not persist at the three-year horizon. As in Panel A, the AI coefficients in the F Score sample closely track the Table 3 baseline estimates of 0.138, 0.083, 0.030, and 0.320.

These tests establish incremental predictive content, not necessarily statistical dominance. Raw coefficients cannot be compared directly across the three scores because the measures have different scales and distributions. Moreover, the BKLYZ specifications contain only about 12,000–14,000 observations, whereas the F Score specifications contain approximately 13,000–36,000 observations. The most defensible interpretation is that the AI Score includes a component of fraud risk not captured by either benchmark. This inference is strengthened by the low or negative correlations between these variables and by the stability of the AI coefficient when the F Score is added. A claim that one model strictly outperforms another would require common-sample comparisons using scale-free classification metrics or formal tests of incremental fit. Table 4 nevertheless addresses the central concern that the AI system merely reconstructs existing financial-statement models: it does not.

3.4 Auditor Cost-Benefit Analysis

Statistical significance does not determine whether a fraud signal is economically useful. Auditors face asymmetric errors. A false negative can expose the auditor to litigation, client losses, and reputational damage, whereas a false positive can prompt costly additional audit procedures and investigation. Table 5 applies the Beneish and Vorst (2022) framework to the AI Score, BKLYZ, and F Score and reports the benefit from true positives, the cost of false positives, and their difference. All monetary amounts in the table are in millions of dollars.

3.4.1 Restatement Classifications

In Panel A, the reported low-threshold AI rule uses a cutoff of 0.28 over 71,636 observations. It identifies 5,926 true positives and 51,823 false positives. With a $16.36 million benefit per true positive and the reported false-positive cost, aggregate benefits are $96.949 billion, costs are $24.092 billion, and net benefits are $72.858 billion.

More selective AI rules reduce the investigation burden and the number of detected events. At 0.80, the score yields 721 true positives, 9,751 false positives, and $10.783 billion of net benefit. At 0.90, it yields 226 true positives, 3,271 false positives, and $3.504 billion of net benefit. Under the Table 5 assumptions, all three AI rules are positive, but the large false-positive counts show why the score is more defensible as a triage mechanism than as a stand-alone fraud label.

At their reported thresholds, the F Score and BKLYZ also generate positive net benefits. The baseline F Score rule produces $42.616 billion and BKLYZ produces $24.105 billion. Aggregate comparisons are confounded by evaluation-sample size. Dividing net benefit by reported N gives approximately $1.02 million per observation for the low-threshold AI rule, $0.82 million for BKLYZ, and $0.70 million for the F Score. These normalizations are descriptive and do not replace common-sample decision analysis.

3.4.2 Big-R Classifications

Panel B reports Big-R classifications. At the 0.37 cutoff, the AI Score identifies 2,196 true positives and 46,571 false positives. Aggregate true-positive benefits are $40.099 billion, false-positive costs are $19.095 billion, and net benefits are $21.004 billion. At 0.80, net benefits decline to $5.642 billion; at 0.90, they are $2.008 billion. The stricter rules again trade substantially fewer investigations for fewer detected events and lower aggregate value under the reported parameters.

For Big-R restatements, the baseline F Score and BKLYZ rules no longer generate positive net benefits: the F Score rule is approximately break-even (−$1 million) and the BKLYZ rule produces a net cost of $2.057 billion. Normalized by N, the low-threshold AI rule yields approximately $0.29 million per observation, compared with approximately zero for the F Score and −$0.07 million for BKLYZ

Table 5 supports a conditional economic conclusion. Under its benchmark assumptions, each reported AI threshold produces positive aggregate net benefits for both outcomes. The permissive cutoffs maximize the reported aggregate benefit but flag most of the evaluation sample; higher cutoffs materially reduce false positives and workload while retaining positive estimated value. An auditor's preferred rule must therefore reflect capacity constraints and locally relevant error costs.

4 Conclusions

We examine whether future financial reporting failures can be predicted by a commercially developed artificial intelligence (AI) fraud score. Our results can be summarized as follows: First, the AI Score predicts four distinct reporting outcomes, including material restatements and SEC-related measures, after controlling for conventional financial and audit-risk variables. Second, the predictive relation persists over a three-year horizon, suggesting that the score captures a durable component of reporting risk. Third, the AI Score remains informative

when BKLYZ or the F Score enters the same regression. This incremental result matters because the score is useful academically and operationally only if it adds information beyond established models. Fourth, the cost-benefit analysis shows that incremental prediction can translate into economically meaningful audit value, but that the conclusion is sensitive to the assumed cost of undetected fraud and to the threshold used to trigger additional work.

The results support using the AI Score as a risk-ranking and audit-triage tool rather than a dispositive fraud classification. Fraud and enforcement events are rare, the regression fit is necessarily modest for some outcomes, and permissive thresholds can generate many false positives. A practical implementation would use the score to allocate attention, for example, to determine which clients, accounts, or disclosures merit enhanced procedures, while retaining auditor judgment and corroborating evidence. This interpretation is also consistent with the empirical design: the reported associations demonstrate predictive content, not that a high AI Score proves misconduct or causes subsequent enforcement.

Finally, the evidence highlights the distinction between statistical and decision usefulness. The AI Score contains a reliable signal across outcomes, horizons, and benchmark models. Under the Beneish and Vorst (2022) cost parameters, the AI Score ≥ 0.80 rule retains a positive net benefit while flagging about 15 percent of firm-years rather than the 81 percent flagged by the permissive cutoff. This finding provides a concrete economic channel for the paper's contribution: AI can complement established accounting-based fraud models by improving the ranking of future reporting risk, and auditors can convert that ranking into value when they deploy it selectively.

References

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

Bao, Y., B. Ke, B. Li, Y. J. Yu, and J. Zhang. 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

Beneish, M. D. 1999. The detection of earnings manipulation. *Financial Analysts Journal* 55 (5): 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Beneish, M. D., and P. Vorst. 2022. The cost of fraud prediction errors. *The Accounting Review* 97 (6): 91–121. https://doi.org/10.2308/TAR-2020-0068

Bertomeu, J., E. Cheynel, E. Floyd, and W. Pan. 2021. Using machine learning to detect misstatements. *Review of Accounting Studies* 26 (2): 468–519. https://doi.org/10.1007/s11142-020-09563-8

Blackburne, T., J. D. Kepler, P. J. Quinn, and D. Taylor. 2021. Undisclosed SEC investigations. *Management Science* 67 (6): 3403–3418. https://doi.org/10.1287/mnsc.2020.3805

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

Cohen, L., Y. Lu, and Q. H. Nguyen. 2026. Mimicking finance. NBER Working Paper 34849. National Bureau of Economic Research. https://doi.org/10.3386/w34849

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

Dyck, A., A. Morse, and L. Zingales. 2024. How pervasive is corporate fraud? *Review of Accounting Studies* 29 (1): 736–769. https://doi.org/10.1007/s11142-022-09738-5

Guiso, L., P. Sapienza, and L. Zingales. 2008. Trusting the stock market. *The Journal of Finance* 63 (6): 2557–2600. https://doi.org/10.1111/j.1540-6261.2008.01408.x

Li, K., F. Mai, R. Shen, C. Yang, and T. Zhang. 2026. Dissecting corporate culture using generative AI. *The Review of Financial Studies* 39 (1): 253–296. https://doi.org/10.1093/rfs/hhaf081

Macalister, H. 2023. Predicting corporate collapse with accounting quality: Transparently risk scores as predictors of corporate collapse. White paper. Singapore: Transparently Pte Ltd.

Macalister, H. 2024. Predicting time series stock returns with accounting quality: Transparently risk scores as predictors of future stock returns. White paper. Singapore: Transparently Pte Ltd.

Ohlson, J. A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 18 (1): 109–131. https://doi.org/10.2307/2490395

Table 1: Descriptive Statistics

Variable	N	Mean	STD	Min	p25	p50	p75	Max
Risk Score								
AI Score	49,267	0.567	0.231	0.146	0.366	0.566	0.768	0.956
F Score	41,677	0.984	0.786	0.082	0.456	0.802	1.287	5.183
BKLYZ	14,962	1.556	0.483	0.546	1.199	1.561	1.912	2.620
Manipulation								
Restatement	49,267	0.075	0.263	0.000	0.000	0.000	0.000	1.000
BigR	49,267	0.020	0.139	0.000	0.000	0.000	0.000	1.000
AAER	33,594	0.004	0.060	0.000	0.000	0.000	0.000	1.000
SEC	27,122	0.095	0.293	0.000	0.000	0.000	0.000	1.000
Control Variables								
Size	49,267	6.440	2.300	0.536	4.900	6.461	7.996	11.771
Leverage	49,267	0.248	0.256	0.000	0.028	0.194	0.378	1.367
ROA	49,267	-0.128	0.457	-2.994	-0.124	0.016	0.066	0.309
Loss	49,267	0.435	0.496	0.000	0.000	0.000	1.000	1.000
Total Accruals	49,267	-0.097	0.213	-1.512	-0.111	-0.056	-0.019	0.279
Current Ratio	49,267	3.409	4.278	0.000	1.273	2.034	3.610	27.620
Inventory_ARec	49,267	0.203	0.175	0.000	0.060	0.164	0.302	0.741
Business Segment	49,267	2.343	2.166	1.000	1.000	1.000	3.000	11.000
M&A	49,267	0.539	0.499	0.000	0.000	1.000	1.000	1.000
Foreign	49,267	0.680	0.467	0.000	0.000	1.000	1.000	1.000
Internal Control	49,267	0.086	0.280	0.000	0.000	0.000	0.000	1.000
BigN	49,267	0.724	0.447	0.000	0.000	1.000	1.000	1.000
Going Concern	49,267	0.073	0.261	0.000	0.000	0.000	0.000	1.000
Litigation	49,267	0.356	0.479	0.000	0.000	0.000	1.000	1.000
Auditor Tenure	49,267	8.538	6.012	1.000	3.000	7.000	13.000	24.000
NAS Ratio	49,267	0.188	0.258	0.000	0.011	0.093	0.255	1.425

This table reports descriptive statistics for the sample of 49,267 U.S. firm-year observations over 2007–2024. The sample excludes financial firms (SIC 6000–6999) and utilities (SIC 49). *AI Score* is the Transparently.AI manipulation-risk score; *F Score* is the Dechow et al. (2011) F Score; *BKLYZ* is the machine-learning score of Bao et al. (2020). *F Score* and *BKLYZ* are available for subsets of the sample because of their respective data requirements. *Restatement*, *BigR*, *AAER*, and *SEC* are indicators for financial statement restatements, material (Big-R) restatements, SEC Accounting and Auditing Enforcement Releases, and undisclosed SEC investigations, respectively, measured in fiscal year t. *AAER* and *SEC* are summarized over their source-coverage windows (through fiscal 2019, $N = 33{,}594$, and through fiscal 2017, $N = 27{,}122$, respectively). The regression samples in Tables 2 to 4 are subsets of this sample because the forward-looking dependent variables require the following fiscal year to be observed. All variables are defined in Appendix Table A1.

Table 2: The Impact of AI Score on Reporting Quality in Year $t+1$

	Restatement	BigR	AAER	SEC
AI Score	0.069***	0.040***	0.014***	0.219***
	(4.760)	(5.502)	(3.798)	(8.532)
Size	0.002	0.000	0.002***	0.040***
	(1.573)	(0.183)	(3.765)	(12.521)
Leverage	0.018**	0.007	0.001	−0.035**
	(2.367)	(1.563)	(0.613)	(−2.519)
ROA	0.008	0.003	−0.002*	−0.017
	(1.152)	(0.765)	(−1.948)	(−1.604)
Loss	−0.005	−0.005**	−0.002**	−0.021***
	(−0.960)	(−2.096)	(−2.012)	(−2.701)
Total Accruals	−0.011	−0.013*	0.004	0.000
	(−0.976)	(−1.812)	(1.612)	(0.004)
Current Ratio	−0.001**	0.000	−0.000	−0.001
	(−2.412)	(0.170)	(−0.020)	(−1.102)
Inventory_ARec	0.005	0.016*	0.020***	0.002
	(0.332)	(1.959)	(3.285)	(0.092)
Business Segment	0.004***	0.001	−0.000	0.001
	(3.474)	(1.521)	(−0.322)	(0.296)
M&A	0.024***	0.009***	0.001	−0.007
	(6.181)	(4.624)	(1.134)	(−0.969)
Foreign	−0.013***	−0.012***	−0.002	0.004
	(−2.740)	(−4.477)	(−1.332)	(0.505)
Internal Control	0.022***	0.005	0.001	0.018*
	(3.371)	(1.410)	(0.518)	(1.673)
BigN	−0.016***	−0.013***	−0.002	−0.027***
	(−2.854)	(−4.375)	(−1.430)	(−3.080)
Going Concern	−0.013*	−0.006	−0.001	−0.001
	(−1.959)	(−1.242)	(−0.832)	(−0.110)
Litigation	−0.005	−0.000	0.001	0.017
	(−0.842)	(−0.020)	(0.269)	(1.465)
Auditor Tenure	0.000	−0.000	−0.000	0.001
	(0.891)	(−1.141)	(−1.097)	(1.474)
NAS Ratio	0.001	0.005	0.001	−0.011
	(0.111)	(1.242)	(0.394)	(−1.034)
Industry FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Adj. R^2	0.024	0.014	0.007	0.072
N	42,231	42,231	28,055	22,192

This table reports OLS regressions of financial reporting outcomes in fiscal year $t+1$ on the AI Score and control variables measured in fiscal year t. The dependent variables are indicators for *Restatement*, *BigR*, *AAER*, and *SEC* (undisclosed SEC investigation) in fiscal year $t+1$. All specifications include industry and year fixed effects. t-statistics based on standard errors clustered by firm are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. All variables are defined in Appendix Table A1.

Table 3: The Impact of AI Score on Reporting Quality in Years $[t+1, t+3]$

	Restatement	BigR	AAER	SEC
AI Score	0.138***	0.083***	0.030***	0.320***
	(4.962)	(5.970)	(3.716)	(7.944)
Size	0.001	0.000	0.004***	0.058***
	(0.243)	(0.287)	(3.843)	(12.708)
Leverage	0.032**	0.018**	0.006	−0.064***
	(2.016)	(2.131)	(1.101)	(−2.840)
ROA	0.016	−0.000	−0.003	−0.019
	(1.163)	(−0.044)	(−1.070)	(−1.007)
Loss	−0.000	−0.008*	−0.005*	−0.022*
	(−0.059)	(−1.959)	(−1.885)	(−1.895)
Total Accruals	−0.028	−0.014	0.007	−0.007
	(−1.359)	(−1.068)	(1.285)	(−0.235)
Current Ratio	−0.002**	0.000	−0.000	−0.000
	(−2.253)	(0.546)	(−0.056)	(−0.064)
Inventory_ARec	0.045	0.045***	0.040***	−0.019
	(1.535)	(2.636)	(3.117)	(−0.469)
Business Segment	0.005***	0.001	−0.000	0.000
	(2.982)	(0.862)	(−0.782)	(0.119)
M&A	0.046***	0.015***	0.004*	−0.002
	(6.232)	(4.166)	(1.672)	(−0.241)
Foreign	−0.022**	−0.022***	−0.004	0.011
	(−2.348)	(−4.203)	(−1.196)	(0.902)
Internal Control	0.033***	0.008	0.001	0.029*
	(2.590)	(0.981)	(0.256)	(1.650)
BigN	−0.030***	−0.029***	−0.006	−0.051***
	(−2.676)	(−4.651)	(−1.628)	(−3.483)
Going Concern	−0.021	−0.007	−0.002	−0.011
	(−1.424)	(−0.749)	(−0.579)	(−0.456)
Litigation	−0.015	0.001	0.003	0.024
	(−1.160)	(0.164)	(0.670)	(1.325)
Auditor Tenure	0.001	−0.001*	−0.000	0.001
	(0.942)	(−1.947)	(−1.491)	(1.000)
NAS Ratio	0.001	0.010	0.002	−0.017
	(0.085)	(1.483)	(0.576)	(−1.126)
Industry FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Adj. R^2	0.098	0.118	0.027	0.151
N	31,343	30,836	19,305	14,623

This table reports OLS regressions of financial reporting outcomes over fiscal years $t+1$ through $t+3$ on the AI Score and control variables measured in fiscal year t. All specifications include industry and year fixed effects. t-statistics based on standard errors clustered by firm are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. All variables are defined in Appendix Table A1.

Table 4: Robustness Checks

	Restatement		BigR		AAER		SEC	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Reporting Quality in Year $t+1$								
AI Score	0.066**	0.071***	0.053***	0.040***	0.018***	0.011***	0.255***	0.213***
	(2.245)	(4.431)	(3.511)	(5.164)	(2.842)	(3.004)	(7.304)	(7.868)
BKLYZ	0.028***		0.011**		0.004		0.009	
	(2.967)		(2.408)		(1.518)		(0.779)	
F Score		0.008***		0.005***		0.003**		−0.009**
		(3.094)		(3.125)		(2.278)		(−2.311)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.028	0.024	0.019	0.015	0.009	0.008	0.069	0.068
N	13,838	35,819	13,838	35,819	13,838	24,370	13,838	19,396
Panel B: Reporting Quality in Years $[t+1, t+3]$								
AI Score	0.120**	0.137***	0.082***	0.082***	0.034***	0.025***	0.326***	0.320***
	(2.543)	(4.530)	(3.709)	(5.670)	(3.214)	(3.072)	(7.229)	(7.405)
BKLYZ	0.052***		0.014*		0.008**		0.024*	
	(3.454)		(1.899)		(2.166)		(1.695)	
F Score		0.019***		0.009***		0.007***		−0.006
		(4.235)		(3.306)		(2.766)		(−0.891)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R^2	0.047	0.093	0.032	0.106	0.017	0.015	0.091	0.146
N	12,220	26,899	12,086	26,477	12,040	16,873	12,064	12,779

This table reports OLS regressions that add the benchmark fraud scores to the baseline specifications. Odd-numbered columns pair the AI Score with BKLYZ (Bao et al. 2020); even-numbered columns pair the AI Score with the F Score (Dechow et al. 2011). Panel A measures the outcomes in fiscal year $t+1$ as in Table 2; Panel B measures the outcomes over fiscal years $t+1$ through $t+3$ as in Table 3. Sample sizes differ across columns because each specification is estimated on the subsample for which the respective benchmark score is available. All specifications include the control variables from Table 2, industry fixed effects, and year fixed effects. t-statistics based on standard errors clustered by firm are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. All variables are defined in Appendix Table A1.

Table 5: Auditor Cost-Benefit Analysis

Rule	Threshold	N	True Positives (TP)	False Positives (FP)	Litigation ($M)	Client Loss ($M)	Reputation ($M)	TP Benefit ($M)	Total TP ($M)	FP Cost ($M)	Total FP ($M)	Net Benefit ($M)
Panel A: Restatement												
AI Score	0.28	71,636	5,926	51,823	0.02	16.34	0	16.36	96,949	0.46	24,092	72,858
AI Score ≥ 0.80	0.80	71,636	721	9,751	0.02	16.34	0	16.36	11,796	0.10	1,013	10,783
AI Score ≥ 0.90	0.90	71,636	226	3,271	0.02	16.34	0	16.36	3,697	0.06	193	3,504
F Score	0.61	60,628	4,390	34,434	0.02	16.34	0	16.36	71,820	0.85	29,204	42,616
F Score ≥ 1.00	1.00	60,628	2,688	20,838	0.02	16.34	0	16.36	43,976	0.84	17,419	26,557
F Score ≥ 1.85	1.85	60,628	699	4,801	0.02	16.34	0	16.36	11,436	0.54	2,605	8,831
BKLYZ	1.29	29,226	2,439	16,845	0.02	16.34	0	16.36	39,902	0.94	15,797	24,105
Panel B: BigR												
AI Score	0.37	71,636	2,196	46,571	0.05	18.21	0	18.26	40,099	0.41	19,095	21,004
AI Score ≥ 0.80	0.80	71,636	367	10,105	0.05	18.21	0	18.26	6,701	0.10	1,059	5,642
AI Score ≥ 0.90	0.90	71,636	121	3,376	0.05	18.21	0	18.26	2,209	0.06	201	2,008
F Score	0.60	60,628	1,732	37,424	0.05	18.21	0	18.26	31,626	0.85	31,627	-1
F Score ≥ 1.00	1.00	60,628	1,071	22,455	0.05	18.21	0	18.26	19,556	0.83	18,679	878
F Score ≥ 1.85	1.85	60,628	317	5,183	0.05	18.21	0	18.26	5,788	0.55	2,836	2,952
BKLYZ	0.88	29,226	1,031	25,341	0.05	18.21	0	18.26	18,826	0.82	20,884	$-2,057$

This table applies the auditor cost-benefit framework of Beneish and Vorst (2022) to classifications based on the AI Score, the F Score, and BKLYZ. Panel A evaluates *Restatement* outcomes and Panel B evaluates *BigR* outcomes, both measured in fiscal year $t+1$ as in Table 2. A firm-year is flagged when its score is at or above the threshold shown. N is the number of firm-year observations for which the respective score is available, and the $t+1$ outcome is observed. Litigation, Client Loss, and Reputation are the per-true-positive components of the auditor benefit, taken from Beneish and Vorst (2022); TP Benefit is their sum. FP Cost is the per-observation cost of a false positive, comprising incremental audit effort and expected discovery costs estimated following Beneish and Vorst (2022). Total TP, Total FP, and Net Benefit aggregate these amounts across all flagged observations. All monetary amounts are in millions of 2016 dollars. Each score is evaluated on every firm-year for which that score and the $t+1$ outcome are observed, so N differs across scores. This follows Beneish and Vorst (2022), whose evaluation samples likewise differ across models with each model's data requirements, and it reflects how an auditor would apply a score to every client for which it is available rather than only to clients with a complete set of regression controls. Restricting the table to the Table 2 estimation sample (42,231 firm-years; the reduction arises mainly because audit-fee data begin in fiscal 2007) leaves the inferences unchanged: all three AI Score rules retain positive net benefits for both outcomes, the AI Score continues to produce the largest aggregate net benefit of the three scores, and the baseline F Score and BKLYZ rules for *BigR* produce net costs. Inferences are also unchanged when the 0.80 and 0.90 cutoffs are replaced by the 80th and 90th percentiles of the AI Score (0.75 and 0.85), which flag 20 and 10 percent of firm-years.

A The Transparently Manipulation Risk Score

A.1 Construct and Relation to Prior Models

This study uses the Transparently Risk Score, produced by the Transparently Risk Engine (TRE), a commercial machine-learning system developed by Transparently Pte Ltd in Singapore. The system entered commercial production in August 2021. It is designed to detect accounting manipulation and the severe corporate failure that can follow it. This objective is economically relevant because a substantial share of corporate misconduct is never observed directly (Dyck et al. 2024).

The TRE builds on, but does not re-estimate, any single published model. Traditional classifiers such as the Beneish (1999) M-score and Dechow et al. (2011) F Score identify manipulation footprints using a parsimonious set of accounting measures. Distress models such as Altman (1968) and Ohlson (1980) target related failure outcomes. Later work applies machine learning to broader predictor sets (Bao et al. 2020; Bertomeu, Cheynel, Floyd, and Pan 2021). The TRE instead combines a large library of forensic-accounting, short-seller, credit, and fundamental-analysis constructs within a proprietary ensemble.

A.2 Data Universe and Signal Construction

Inputs are annual and interim financial-statement fundamentals and associated reporting metadata for listed companies, sourced from LSEG (Refinitiv), together with market data used to define corporate-failure outcomes. Banks and insurers are excluded because their reporting structures require different specifications. Companies with fewer than three years of financial history or insufficient inputs are also excluded. The estimation panel includes active, delisted, and suspended firms, reducing survivorship bias.

From normalized fundamentals, the system constructs candidate signals associated with aggressive revenue recognition, depreciation choices, bad-debt provisioning, inventory and cash-flow anomalies, sales acceleration, off-balance-sheet activity, accrual management, real-

activities management, smoothing, financial strength, bankruptcy risk, and statistical outliers. The objective is to reproduce at scale aspects of the diagnostic behavior of forensic accountants, activist short sellers, credit analysts, and equity analysts (Macalister 2023; Macalister 2024).

A.3 Estimation, Aggregation, and Explainability

Approximately 200 sub-models exploit cross-sectional and time-series variation using econometric and machine-learning methods, with an emphasis on random forests. The sub-models are calibrated against an internally curated labeled set of confirmed manipulation and failure cases. Manipulation labels are checked against external evidence such as enforcement actions, court decisions, and audited restatements; training and evaluation samples are mutually exclusive in the provider's validation design.

Sub-model outputs are grouped into 14 published clusters: accruals management, asset quality, cash quality, credit quality, income quality, investment activity, gearing risk, corporate governance, growth, margins, business manipulation, smoothing activity, working-capital issues, and valuation. Cluster scores are combined dynamically into the overall risk score. Because the provider publishes the cluster readings and principal contributing factors, users can observe the categories driving a company-month score even though the feature definitions and ensemble weights are proprietary.

A.4 Score Definition and Interpretation

The score is bounded on the unit interval and targets the joint probability of accounting manipulation and corporate failure resulting from that manipulation over approximately two to three years. The provider's public collapse definition is a decline of at least 80 percent from the stock's historical peak in both absolute terms and relative to the relevant jurisdiction benchmark. This market-outcome definition captures economically catastrophic events, including adverse delistings, without conditioning on bankruptcy or an enforcement action.

Two interpretive limits follow. First, the score targets manipulation associated with severe failure, not every form of earnings management. Second, prior validation indicates a nonlinear relation in which discrimination is concentrated in the upper tail. A score is therefore a probabilistic risk signal, not a finding of fraud, and its use requires an explicit decision threshold. Our regressions test whether the continuous score nevertheless ranks distinct future reporting outcomes; Table 5 separately shows how threshold choice changes classification economics.

A.5 Point-in-Time Score History

The research data are a point-in-time monthly score history beginning in December 1999 and covering approximately 80,000 active, delisted, and suspended stocks globally, subject to the exclusions above. For each month-end, inputs are restricted to statements announced on or before that date. A company with a December fiscal year-end that reports on March 15, for example, contributes those statements to the March month-end score. Fiscal-period end dates do not substitute for public availability dates.

The complete model (not only the company inputs) is re-estimated using the pooled data available at each month-end. Accordingly, a score can change even when a company releases no new statement because the contemporaneous cross-section and fitted system have changed. The pre-production history from December 1999 through July 2021 simulates the score that would have been produced under the dated information constraint. Scores from August 2021 onward are contemporaneous production outputs and are not backfilled with later information.

A.6 Prior Validation

Macalister (2023) evaluates the score against manipulation-related corporate collapse in a global sample of approximately 60,000 stocks and 640,000 stock-years from 1994 through 2023. In the provider's out-of-sample tests, 97 percent of stocks scoring above 0.90, 86 percent

above 0.80, and 79 percent above 0.70 subsequently meet the failure definition, compared with approximately 2 percent of stocks scoring 0.10 or below. Classification rates are similar when the score is measured one, two, or three years before failure.

Macalister (2024) evaluates the monthly point-in-time history for 61,306 companies from January 2000 through October 2024, totaling approximately 7.5 million company-months. The median benchmark-relative return spread between the lowest- and highest-risk deciles is 26.7 percent over 12 months and 52.6 percent over 36 months. The relation appears across regions and remains after market-capitalization and subperiod splits. Factor regressions also report significant long-short alpha after controls for market, size, value, momentum, profitability, and investment factors.

Those studies validate the score against collapse and return outcomes used by the provider. The current paper asks a distinct question: whether the same score predicts later restatements and SEC-related outcomes. Because those events were not the system's target labels, the present tests assess external construct validity rather than in-sample recovery.

A.7 Proprietary Boundary, Data Access, and Researcher Independence

The disclosures above identify the input classes, universe, target construct, label-verification standard, point-in-time protocol, cluster structure, and prior validation. Transparently does not disclose the feature definitions, individual specifications, ensemble composition and weights, calibration mapping, or labeled training cases. These elements constitute the protectable commercial methodology. The score can therefore be interpreted and externally tested, but it cannot be reconstructed from the paper.

Replication requires the dated score extract rather than reconstruction of the engine. The authors obtained the point-in-time history from Transparently Pte Ltd under a research license. Transparently has agreed to make the same extract available to qualified researchers under a standard research data-use agreement. One coauthor is affiliated with Transparently.AI;

this affiliation and the proprietary data source make transparent reporting of the score's construction, validation domain, and empirical limitations especially important. The outcome tests and interpretations in this paper should be evaluated against the reported tables and, where available, independently replicated.

Table A1: Variable Definitions

Variable	Definition	Source
Panel A: Dependent variables		
Restatement	Indicator equal to one if the firm's financial statements for fiscal year t are subsequently restated, and zero otherwise. Restatement periods are matched to fiscal years using the restated period's overlap with the firm's fiscal year.	Audit Analytics
BigR	Indicator equal to one if the restatement of fiscal year t is disclosed in a Form 8-K Item 4.02 filing (non-reliance on previously issued financial statements), and zero otherwise.	Audit Analytics
AAER	Indicator equal to one if fiscal year t is subject to an SEC Accounting and Auditing Enforcement Release, and zero otherwise.	Audit Analytics
SEC	Indicator equal to one if the firm was under an SEC investigation open during any part of fiscal year t, and zero otherwise. These investigations were undisclosed at the time they occurred and were obtained through FOIA requests.	Blackburne et al. (2021)
Panel B: Risk scores		
AI Score	Point-in-time manipulation-risk score produced by the Transparently Risk Engine, bounded between zero and one. The firm-year value is the mean of the monthly scores over the calendar year matched to the firm's fiscal year. Construction, training universe, and validation are described in Appendix A.	Transparently.AI; Macalister (2023) and Macalister (2024)
F Score	The F Score from Model 1 of Dechow et al. (2011).	Dechow et al. (2011)
BKLYZ	Predicted fraud score of Bao et al. (2020).	Bao et al. (2020)
Panel C: Control variables		
Size	Natural logarithm of total assets (AT).	Compustat
Leverage	Total debt scaled by total assets: (DLC + DLTT)/AT.	Compustat
ROA	Income before extraordinary items scaled by total assets: IB/AT.	Compustat

Continued on next page

Table A1 (continued)

Variable	Definition	Source
Loss	Indicator equal to one if income before extraordinary items (IB) is negative (or, when IB is missing, if net income is negative), and zero otherwise.	Compustat
Total Accruals	Income before extraordinary items (cash flow statement, IBC) less operating cash flow (OANCF), scaled by total assets.	Compustat
Current Ratio	Current assets scaled by current liabilities: ACT/LCT.	Compustat
Inventory_ARec	Inventory plus accounts receivable scaled by total assets: (INVT + RECT)/AT.	Compustat
Business Segment	Number of business segments reported in the Compustat Segments database; set to one when segment data are missing.	Compustat
M&A	Indicator equal to one if the firm reports acquisition activity (AQC $\neq 0$) in year t or either of the two preceding years, and zero otherwise.	Compustat
Foreign	Indicator equal to one if the firm reports foreign operations (nonzero foreign currency adjustment, pretax foreign income, or foreign income taxes: FCA, PIFO, TXDFO, or TXFO), and zero otherwise.	Compustat
Internal Control	Indicator equal to one if a material weakness in internal control is reported under SOX Section 302 or 404, and zero otherwise.	Audit Analytics
BigN	Indicator equal to one if the firm's auditor is a Big 4 (Big N) auditor, and zero otherwise.	Audit Analytics
Going Concern	Indicator equal to one if the audit opinion for fiscal year t contains a going-concern modification, and zero otherwise.	Audit Analytics
Litigation	Indicator equal to one if the firm operates in a high-litigation-risk industry (SIC codes 2833–2836, 3570–3577, 3600–3674, 5200–5961, and 7370), and zero otherwise.	Compustat
Auditor Tenure	Number of consecutive years the incumbent auditor has audited the firm.	Audit Analytics
NAS Ratio	Non-audit fees scaled by audit fees.	Audit Analytics