Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI Disclosure Credibility and Corporate Financial Misconduct: Evidence from Green Digital Infrastructure in China

huan chen

157792684090163.com

Guangdong University of Technology https://orcid.org/0009-0006-7507-1403

Research Article

Keywords: green digital infrastructure, financial misconduct, AI disclosure, disclosure-action gap, machine learning, staggered difference-in-differences

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11836714/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

AI Disclosure Credibility and Corporate Financial Misconduct: Evidence from Green Digital Infrastructure in China

Abstract

Can green digital infrastructure discipline corporate misconduct, and can discrepancies between corporate AI talk and observable AI action reveal residual financial risk? We address these questions by linking 33,356 firm-year observations for 4,662 Chinese listed firms in 259 cities to the staggered rollout of the National Green Data Center program, enforcement records, annual-report AI disclosures, digital-capital investment, AI patents, audit information, innovation, and city conditions. A cohort-stacked difference-in-differences design indicates that program exposure is associated with a 1.71 percentage-point reduction in next-year punitive violations, equivalent to 50.5% of the treated firms' pre-treatment risk. The estimate remains negative with direct controls for three overlapping digital policies. Re-estimation around official list publication gives +0.54 percentage points for 2018 ($p = 0.336$) and -1.25 percentage points for newly exposed cities on the 2021 list ($p = 0.014$). We then construct an expanding-window AI disclosure-action benchmark that distinguishes narrative intensity from observable AI investment and patenting. Positive unexplained AI language is associated with subsequent recorded misconduct under two-stage bootstrap inference. Rolling OLS residuals and raw AI intensity provide comparable in-sample fit. In later-period classification tests, raw AI talk and the disclosure-action gap provide modest additional ranking information beyond financial variables, while the integrated model identifies 31.4% of recorded violations in the highest-risk decile. The findings connect digital infrastructure with corporate financial discipline and characterize the information in conditional AI language for risk-oriented monitoring.

Keywords: green digital infrastructure; financial misconduct; AI disclosure; disclosure-action gap; machine learning; staggered difference-in-differences

JEL classification G30 G38 M41 O33 Q55

1 Introduction

Digital infrastructure has become part of the financial-control environment. Cloud computing, standardized data centers, machine-readable records, and automated analytics alter how firms store transactions, reconcile accounts, monitor exceptions, and communicate with auditors and regulators. Yet the same infrastructure also expands the volume and complexity of data that managers can selectively present. This duality creates a problem for corporate finance. Does public investment in computing capacity improve financial integrity by making operations more observable, or does it mainly encourage firms to talk about artificial intelligence without producing verifiable changes in technology or conduct?

The distinction is economically important because AI has become both a productive input and a disclosure label. Measures based on patents capture invention, measures based on employee skills capture organizational capability, and annual-report keywords capture managerial communication. These objects need not move together. Babina et al. (2024) show that substantive AI investment can support firm growth and product innovation, while Cao et al. (2023) demonstrate that managers change corporate language when machines increasingly read filings. Dogan and Teklie (2026) find that AI disclosure became widespread in Item 1A of S&P 500 10-K filings after ChatGPT and that greater AI text volume was associated with lower lexical distinctiveness among disclosing firms. A rise in AI vocabulary therefore cannot by itself establish adoption. For investors and regulators, the actionable object may be the distance between what a firm says and what can be observed in its investment and innovation records.

This paper combines a staggered infrastructure program with a time-ordered measure of the gap between AI disclosure and observable AI action. The setting is China's National Green Data Center program. Official documents require participating centers to establish operation and maintenance systems, assign managerial responsibility, monitor operational indicators, retain reporting data, and undergo evaluation. These requirements may improve the local information and monitoring environment. We define treatment as the first year in which a listed firm's city enters the program, so the design estimates the effect of local exposure rather than verified firm-level adoption.

Our empirical architecture links 64,286 financial firm-years to annual-report AI text, book-value AI software and hardware investment, AI patents, general innovation measures, financial misconduct records, audit opinions, firm location, city characteristics, physical climate risk, provincial adaptive capacity, and two city-level policy panels. The main integrated sample contains 33,356 firm-years for 4,662 nonfinancial A-share firms in 259 cities from 2012 through 2022. We split violation records that list multiple conduct years into separate firm-conduct-year observations. This produces negative as well as positive firm-years from the complete financial panel and avoids conditioning the analysis on firms that were eventually sanctioned.

The causal analysis uses a cohort-stacked difference-in-differences design. For the 2015 and 2020 cohorts, each stack compares treated firms with firms in cities untreated through the end of the event

window. Stack-by-firm and stack-by-calendar-year fixed effects prevent already-treated units from serving as inappropriate controls. Standard errors are clustered at the city level, where treatment varies. This design addresses weighting and comparison-group contamination in traditional staggered two-way fixed-effects estimators documented by Goodman-Bacon (2021), Sun and Abraham (2021), Callaway and Sant'Anna (2021), and Baker et al. (2022).

The cohort-stacked estimate indicates a 1.71 percentage-point decline in next-year punitive violations, with a city-clustered standard error of 0.43 percentage points. This magnitude equals 50.5 percent of the treated groups' 3.38 percent pre-treatment risk. The supported lead estimates are 0.32 percentage points at event year -3 (95 percent CI: -0.66 to 1.30) and 0.37 percentage points at event year -2 (95 percent CI: -0.45 to 1.18), with a joint p value of 0.677. A model with a treated-specific linear trend produces a post-treatment estimate of -1.36 percentage points (SE = 0.52), and a bounded-trend calculation allowing 0.5 percentage points of annual differential drift yields a point interval from -2.71 to -0.71 percentage points and a conservative 95 percent interval from -3.55 to 0.13. The negative association persists across several program-entry specifications. Estimates around formal publication differ by list, and the drift analysis shows how the precision of the program-entry estimate depends on differential trends.

The decline is concentrated in punitive cases and a narrower fraud label. Announcement-year enforcement and modified audit opinions show weaker evidence. The available data record city exposure and firm reporting outcomes, while the role of firm internal controls, computing-service use and local enforcement remains open.

The policy effect is not accompanied by a comparable technology-production response. We find no robust average increase in AI patenting or AI investment intensity. AI disclosure intensity rises in the conventional fixed-effects model, but the corresponding event study rejects parallel pretrends. The policy also does not significantly reduce positive abnormal AI talk. Together, these findings distinguish the observed financial-discipline effect from measurable growth in firm-owned AI assets or patents; outsourced computing, vendor services, ERP adoption, cloud expenditure, and audit-process changes remain outside the available measures.

We then measure the AI talk-action gap. For every target fiscal year from 2012 onward, a gradient-boosting model fitted only on earlier fiscal years estimates normal AI disclosure intensity from contemporaneous AI investment, AI patents, R&D, innovation efficiency, firm fundamentals, city digital conditions, industry, location, and time. The positive residual identifies firm-years with more AI language than predicted by observed action, financial characteristics and location. Annual held-out-year R-squared rises from 0.399 in 2012 to 0.562 in 2022, showing that the benchmark captures meaningful structure while leaving economically relevant residual variation.

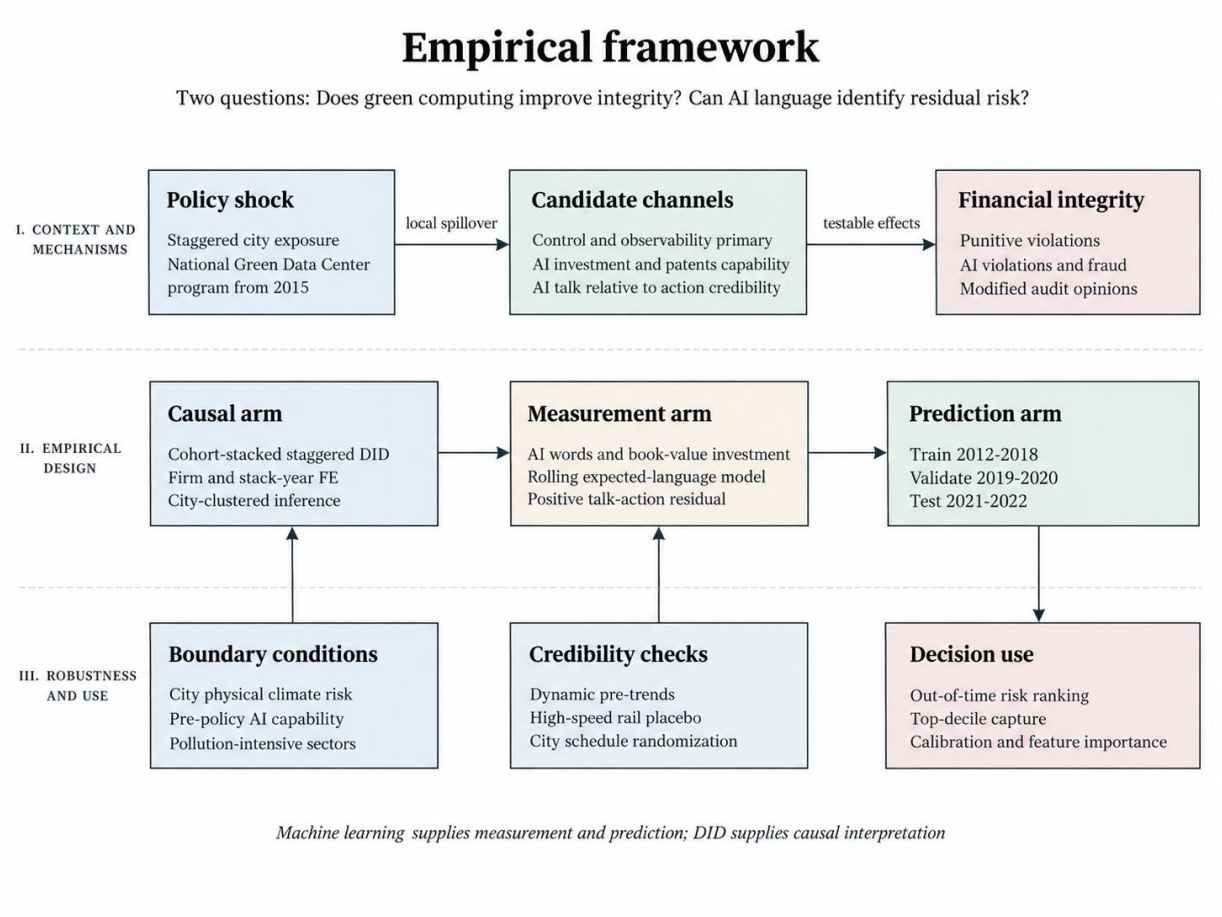
Positive abnormal AI talk is associated with misconduct assigned to the following conduct year. In firm and year fixed-effects regressions, a one-unit increase in the winsorized positive residual is associated with a 2.71 percentage-point increase in any next-year recorded violation and a 1.62

percentage-point increase in recorded punitive violations. Evaluated at the residual's sample standard deviation of 0.368, the associations are approximately 1.00 and 0.60 percentage points. The two-stage percentile intervals exclude zero. These estimates describe conditional associations in the archived data. The residual records language intensity not reconciled by the observed benchmark; it can also reflect omitted intangible investment, outsourcing, implementation timing, or measurement error.

We next examine whether AI disclosure improves the ranking of later-period misconduct risk. Holding the algorithm, fiscal-year splits, preprocessing, class weighting, categorical controls, and calibration protocol constant, the financial-only model records a test AUC of 0.736 and average precision of 0.082. Adding raw AI talk raises these values to 0.739 and 0.089; adding the positive disclosure-action gap yields 0.738 and 0.088; and adding both yields 0.738 and 0.089. Paired bootstrap intervals for the increments include zero, indicating modest but statistically imprecise ranking gains. The broader information model records an AUC of 0.734, average precision of 0.091, and top-decile capture of 31.4 percent, compared with 27.4 percent for the financial-only model. Because the broader model also contains climate, innovation, city, and policy variables, the capture difference cannot be attributed to AI alone.

The paper contributes to three related areas. First, it links public digital infrastructure to corporate financial discipline, extending evidence on information technology beyond productivity and innovation. Second, it develops a time-ordered disclosure-action benchmark that separates AI narrative intensity from observable investment and patenting and quantifies its association with subsequent recorded misconduct alongside simpler language measures. Third, it integrates policy evaluation, disclosure measurement, and risk ranking while preserving their distinct empirical roles: the stacked design estimates the effect of local program exposure, the residual measures language unexplained by observed action, and the classification exercise evaluates later-period risk ranking.

Figure 1 Empirical framework



2 Institutional Setting and Hypothesis Development

2 1 The National Green Data Center Program

China's listed-company setting is useful for studying the intersection of infrastructure, disclosure, and enforcement. A-share firms issue standardized annual reports, report detailed fixed and intangible asset notes, and face sanctions from securities regulators, stock exchanges, and related public authorities. Cities also differ in computing infrastructure, industrial composition, financing capacity, environmental exposure, and administrative resources. A city-level program can alter the local service environment faced by many firms, but city residence alone does not demonstrate that a firm used a designated center or obtained better computing services. The treatment is therefore an exposure measure with potentially incomplete take-up.

The National Green Data Center pilot plan was issued in March 2015 by the Ministry of Industry and Information Technology, the National Government Offices Administration, and the National Energy Administration. The plan framed data centers as energy-intensive infrastructure and set targets for efficiency, technology diffusion, and standardized management. Its design went beyond replacing equipment. Pilot units were expected to establish operation and maintenance systems, define managerial responsibility, monitor energy and environmental indicators, maintain data for periodic reporting, and undergo evaluation. The first national list covered data centers in finance, telecommunications, internet

services, manufacturing, energy, and public institutions. Subsequent annual lists were selected through self-assessment, third-party evaluation, local recommendation, expert review, and public announcement.

The program involved several administratively distinct stages. The Pilot Work Plan (MIIT Joint Energy Conservation [2015] No. 82) was issued on March 18, 2015. Applications from prospective pilot regions were due by May 31, 2015, and the selected regions subsequently organized applications from individual pilot units. Fourteen pilot regions were announced on September 28, 2015 under MIIT Joint Energy Conservation Letter [2015] No. 475. Following regional recommendation and expert review, 84 individual data centers were selected as pilot units on December 21, 2015 under MIIT General Office Joint Energy Conservation [2015] No. 163. Pilot selection was followed by creation, monitoring, and evaluation. Evaluation applications were due by February 29, 2016, followed by approximately one year of continuous monitoring. After self-evaluation, local preliminary review, expert evaluation, and public disclosure, 49 centers were formally included in the first National Green Data Center list under Announcement [2018] No. 6, dated January 24, 2018 and published on February 1, 2018. Pilot selection in 2015 and formal designation in 2018 therefore represent different stages of the initial program cycle. Supplementary Table S1 summarizes this institutional chronology.

The subsequent annual-list procedure differed from the initial pilot process. For the 2020 administrative cycle, candidate centers completed self-evaluation and independent third-party evaluation before recommendation by the relevant provincial authorities or designated national organizations. The proposed list was publicly disclosed on December 2, 2020 and opened for public comment on December 3, 2020. The final list of 60 centers was formally determined on January 12, 2021 and published on January 27, 2021 under Announcement [2021] No. 2. We therefore distinguish the administrative-cycle label from the application, evaluation, public-disclosure, and formal-announcement dates.

Treatment is defined at the city level rather than by direct firm participation. The policy dataset records each city's administrative code, earliest program-entry year, pilot-city status, and staggered treatment indicator. The baseline treatment date is the earliest recorded year of entry into the National Green Data Center program process, not the final-certification date of every center in the city. Supplementary Tables S1-S3 report the administrative chronology, the complete city schedule used in the analysis, and publicly verifiable center-to-city matches. When public records do not disclose a center-specific application or evaluation date, the tables report the applicable program deadline or announcement date and leave the center-specific field unavailable. The baseline comparison targets local program exposure. A separate analysis dates exposure from the public release of the two focal final lists.

Program placement is purposive. We code city exposure to Broadband China, the National Smart City geographical pilots and the National Big Data Comprehensive Pilot Zones from published designation lists. We assign district and county designations to their parent prefectures and use the first

public announcement year. The CSRC (2016) reports an enforcement expansion during 2015. Stack-by-year effects absorb changes common to firms, and alternative province-by-year and industry-by-year effects address heterogeneous shocks at those levels. Section 5.4 reports the estimates and remaining limits.

2 2 Digital Infrastructure and Financial Integrity

Digital infrastructure may affect reporting integrity through several channels. Shared computing services can lower the cost of centralized enterprise systems, automated reconciliation, exception detection, and auditable records. Standardized data-center operations may also expand local technical expertise and vendor capacity, while structured monitoring can make records easier to inspect. The program's operating, reporting, monitoring, and evaluation requirements provide an institutional basis for these channels. Distinguishing among them requires firm-center contracts, vendor links, cloud spending, ERP implementation, or audit-process data.

The effect should be more visible for punitive violations than for a broad misconduct indicator. The broad label combines economically different cases, including general accounting mistakes, omissions, misleading statements, and residual categories. Some reflect interpretation disputes or disclosure failures that better infrastructure cannot prevent. By contrast, our punitive label requires a fine, confiscation, market ban, or another substantive punishment code, or a disclosed monetary penalty. It concentrates on conduct for which regulators found consequences serious enough to impose a material sanction. We therefore state the first hypothesis in outcome-specific form.

H1 Local exposure to the green data-center program reduces subsequent punitive financial violations.

2 3 AI Capability and Disclosure Credibility

An alternative view emphasizes productive capability. More local computing capacity may relax constraints on data-intensive innovation and increase firm investment in AI software, servers, automated equipment, or patentable algorithms. Research using labor-based AI investment measures finds that AI adoption is associated with product innovation and firm growth (Babina et al., 2024). Green data centers might similarly lower the fixed cost of experimentation. Yet public infrastructure can also substitute for firm-owned assets, and a city-level program may affect control systems faster than invention. Patents are a delayed and selective measure, while book-value investment omits outsourced cloud services. Thus a null result for patents or owned AI capital would not imply that the infrastructure has no digital effect; it would identify the margin on which the effect is not observed.

H2 If the productive capability channel dominates, program exposure increases firm AI investment and AI patenting.

AI disclosure creates a separate information problem. Managers can use AI language to describe genuine implementation, to frame ordinary digitization as frontier technology, or to satisfy perceived investor demand. Textual disclosure research shows that wording contains information but also responds strategically to readers and litigation incentives (Li, 2010; Loughran and McDonald, 2011;

Cao et al., 2023). A raw keyword count cannot distinguish implementation from framing. We therefore condition expected AI language on direct measures of investment and invention. A positive residual is not automatically deceptive: it may capture outsourced services, early-stage projects, or proprietary applications absent from patents and balance sheets. But if the residual systematically precedes financial violations, it becomes a useful credibility signal.

There are two competing policy predictions for the residual. Better infrastructure could help firms implement what they discuss, narrowing the gap. Alternatively, the legitimacy and visibility of an AI-related public program could stimulate language more quickly than action, widening it. Because the residual is generated by a flexible model and policy designation is not randomized conditional on that model, we treat the relationship between the program and abnormal AI talk as a mechanism diagnostic rather than a primary causal hypothesis. Our third hypothesis concerns prediction rather than treatment.

H3 Positive AI language unexplained by observable AI action predicts a higher probability of subsequent financial violations.

Climate risk enters as a boundary condition and as a second textual domain. Physical extremes raise operational complexity, disrupt supply chains, and increase demand for forward-looking discussion. The annual-report climate-text measure follows a published Chinese dictionary and machine-learning approach (Du et al., 2023). City physical risk and provincial adaptive-capacity variables help distinguish communication driven by exposure from communication driven by managerial choice. The country climate series is excluded from the firm regressions because the sample contains only Chinese firms and therefore supplies no cross-sectional identifying variation. The city index is preferred to the province index because it aligns with the policy level and preserves local variation.

2 4 Related Literature and Contribution

The first related literature studies financial misreporting and its detection. Accounting-based models use accruals, profitability, financing needs, and market incentives to identify manipulation (Beneish, 1999; Dechow et al., 2011). Enforcement research shows that misconduct imposes large legal and reputational losses and that detection depends on a broad monitoring ecosystem rather than regulators alone (Karpoff et al., 2008; Dyck et al., 2010). Survey evidence further indicates that managers trade reporting choices against real operating costs and market expectations (Graham et al., 2005). We add an infrastructure margin to this literature. Instead of asking only which firm attributes predict misconduct, we test whether a city-level computing and monitoring program changes serious conduct.

A second literature examines disclosure as both information and strategic choice. Voluntary disclosure can reduce information asymmetry and financing costs, but managers also delay bad news and manage tone when incentives are strong (Diamond and Verrecchia, 1991; Healy and Palepu, 2001; Kothari et al., 2009). Financial dictionaries and machine-learning models extract information from complex narrative filings (Li, 2010; Loughran and McDonald, 2011), while linguistic cues can help

identify deceptive managerial discussion (Larcker and Zakolyukina, 2012). Cao et al. (2023) show that managers adapt language when machine readers become economically important. Our contribution is to benchmark a thematic disclosure against direct measures of the activity being discussed. The object is not the optimism or readability of AI text, but its intensity relative to observable AI action.

A third literature measures firm AI capability. Patents capture invention, worker resumes and vacancies capture human capital, and investment accounts capture recognized capital. Babina et al. (2024) use labor-based measures to link AI investment to innovation and growth. Each proxy has a different coverage error. Patents omit deployment and trade secrets; balance sheets omit outsourcing; labor data omit capital; text can reflect attention without implementation. We exploit rather than suppress these differences. Patents and book-value investment enter the expected-language model, and the residual is interpreted as disclosure not readily reconciled with those observable actions. The approach aligns with the emerging view that measurement choices determine what researchers mean by firm AI effort.

A fourth literature treats digital technology as a general-purpose input whose effects depend on complementary organization. IT capability is associated with firm performance, management practices help explain productivity gains from technology, and intangible adjustment can delay observable returns (Bharadwaj, 2000; Bloom et al., 2012; Brynjolfsson et al., 2021). This perspective implies that infrastructure may first alter organizational processes and information quality rather than immediately increase patents or capitalized AI assets. Direct adoption and process data would allow future research to distinguish these organizational responses.

Climate-finance research provides both context and measurement discipline. Investors price carbon exposure and regard climate risk as financially material, while firm-level text recovers heterogeneous exposure that standard databases often miss (Krueger et al., 2020; Bolton and Kacperczyk, 2021; Sautner et al., 2023). The Chinese annual-report climate dictionary developed by Du et al. (2023) is also associated with firms' cost of equity. We use physical risk and adaptive capacity to capture operational complexity and information demand, treating climate exposure as a boundary condition rather than a required mediator of the digital policy.

The empirical design separates causal estimation from risk classification. The stacked difference-in-differences analysis asks how treated firms would have evolved without local program exposure. The classification analysis asks whether patterns fitted in earlier fiscal years rank recorded violations in later fiscal years. Flexible machine learning serves the second objective, while the staggered design addresses the first. Model selection is based on the validation period, and the 2021-2022 test period remains locked. Because filing dates, historical source vintages, and label-availability dates are unavailable, the classification results describe later-period ranking within a fixed data vintage rather than a point-in-time forecasting system.

3 Data and Variable Construction

The empirical work begins with a financial panel of Chinese listed companies from 1999 through 2023. It contains revenue, growth, operating costs and profit, leverage, profitability, valuation, the separation between control and cash-flow rights, and audit opinion. We harmonize stock codes to six digits and define a firm-year using the fiscal year. The panel has 64,286 unique firm-years for 5,631 firms. We merge industry and location histories rather than assigning each firm a time-invariant headquarters because relocation and administrative code changes can otherwise contaminate city treatment.

We restrict the integrated analysis to 2012-2022. The starting year gives the rolling text models at least five prior years of training data, and the ending year supplies a next-year conduct outcome within the financial and policy coverage. Financial firms are excluded because their balance sheets, leverage, and regulatory environment differ from those of industrial firms. Observed ST or PT firm-years are removed, while firm-years with missing status are retained because missing status does not establish financial distress. Requiring annual-report climate text and city-policy coverage yields 33,356 firm-years for 4,662 firms in 259 cities. A robustness sample removes the text requirement.

3 1 Financial Violations

The enforcement archive has a file timestamp of September 11, 2026 and covers announcements from February 23, 2000 to May 1, 2025. Among records with a parsable conduct year, the latest announcement is April 30, 2025. The source does not include a separate vendor extraction-date field, so the archive timestamp identifies the data vintage available for replication. The extract contains 2,066 rows, 1,937 unique event identifiers, and 1,033 listed firms. Of these rows, 1,988 contain at least one parsable conduct year and 78 do not. Some records list several conduct years. We parse each four-digit year and deduplicate within event, producing 5,510 row-level expansions and 4,955 unique event-conduct-year-announcement combinations. Firm-year outcomes count unique event identifiers, preventing repeated violator rows from multiplying cases.

Because the source identifies conduct years rather than exact conduct dates, detection delay is measured in calendar years from the conduct year to the announcement year. Among 4,954 nonnegative unique event-conduct-year pairs, the median delay is three years, the 90th percentile is six years, and the 95th percentile is eight years. When delay is measured from each event's latest listed conduct year, the corresponding values are one, four, and five years across 1,859 events. One source-coded pair has a 2002 conduct year but a 2001 announcement; it is flagged as an anomalous negative lag and lies outside the 2012-2022 analysis period. Supplementary Table S6 reports the full cumulative lag distribution and the underlying counts.

Observed firm-conduct-year positives are merged into the complete listed-company financial panel. A zero indicates that the archive records no enforcement-linked violation for that firm and conduct year; it does not establish the absence of latent misconduct. This definition preserves the population at risk and avoids conditioning on eventual enforcement, while allowing false negatives

when misconduct remains undetected, unenforced, announced after the cutoff, or attached to one of the 78 records without a parsable conduct year. Those records do not enter conduct-year positives but remain available for the announcement-year outcome.

3 2 AI Action and AI Talk

AI action is observed in three ways. The patent archive contains 169,288 rows for 5,662 firms from 1990 through 2024 and distinguishes the listed entity, subsidiaries, affiliates, and group totals. We use only the group-total category; adding categories would double count subsidiary inventions already included in the total. AI patents sum invention and utility-model applications based on AI patent classification codes, and the analysis uses ln one plus applications. All records are marked post-listing, so the listing-stage field has no variation and is not used.

The AI investment file contains fixed-asset and intangible-asset amounts identified from financial-statement notes through terms for software, intelligent systems, information platforms, servers, automated equipment, data hardware, cloud computing, chips, and related categories. It provides original cost and book value. Our primary measure uses category two, book value, because it reflects the stock of recognized AI-related capital remaining on the balance sheet. AI investment intensity is total AI software plus hardware book value divided by total assets. The measure is available for 87.4 percent of main-sample firm-years. It can miss rented cloud capacity and outsourced algorithms, a limitation that makes the patent and text measures useful complements.

AI disclosure comes from the CSMAR AI word-frequency extract (AI_WordFreSta.xlsx). Its field dictionary defines the annual-report AI count as the total number of matches to the vendor's AI keywords. AI talk is the count per ten thousand report words, transformed as ln(1 + intensity). The source denominator uses jieba segmentation after the removal of images. The archive contains 64,469 valid records for 5,789 firms from 2007 through 2025. Eight duplicated firm-years occur in 2011. We retain the longer report, and all eight are outside the estimation window. The source extract includes a field dictionary and aggregate counts, with no term-by-term dictionary or matched sentences.

Supplementary Table S9 reports source-record and arithmetic checks. All 400 sampled disclosure fields and 700 sampled asset fields match the archived sources, and 22,610 stock equations and 22,665 addition equations reproduce within 0.01. A contextual reading of the two highest-intensity 2022 reports illustrates how company-specific descriptions of AI products and services coexist with general discussion (Table S14). These examples establish the types of language present in the reports. Assessing the vendor's tagging accuracy would require the matched sentences and full dictionary. The investment measure captures broad digital capital, and the language residual can reflect omitted intangible investment, outsourcing, implementation timing and measurement error.

3 3 Climate and City Variables

Climate information comes from three layers. The annual-report file counts 98 climate-risk terms, split into transition, acute physical, and chronic physical categories, and scales counts by report length. Transition words represent 98.07 percent of all climate terms and the correlation between total and

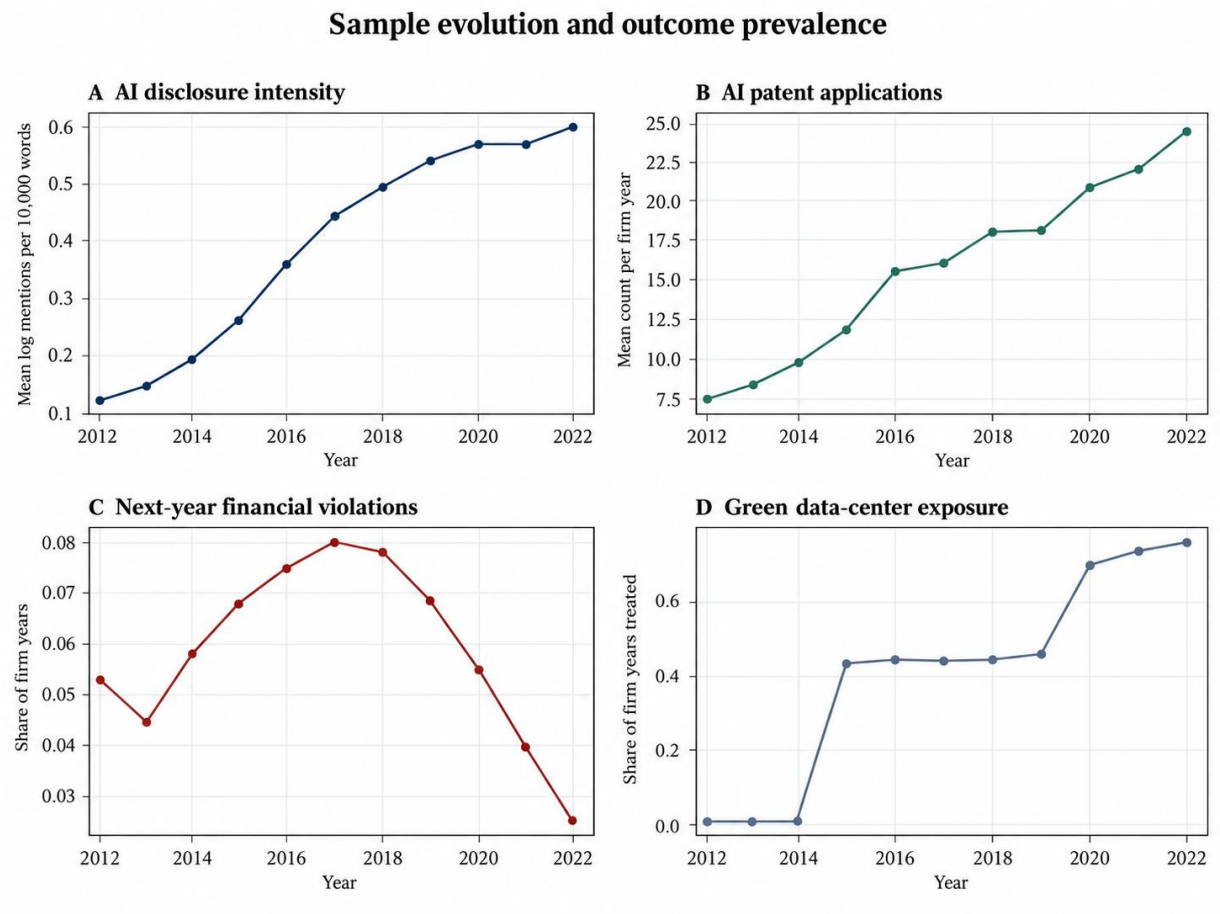
transition risk is 0.999. We therefore avoid interpreting the total and transition indices as independent constructs. The city panel supplies annual indices for extreme heat, extreme cold, extreme rainfall, drought, and an aggregate physical climate risk measure from 1993 to 2023. Administrative code changes create more codes than city names, so we match by code first and normalized city name second. Provincial vulnerability measures enter only where coverage is credible; variables with extreme missingness, such as the population-health index, are not mechanically combined into a principal-component score.

The policy panel identifies 87 treated cities among 297 cities with usable histories from 2000 through 2025. Thirty-two cities enter in 2015, 21 in 2020, 10 in 2021, 9 in 2022, 8 in 2023, and 7 in 2025. The 2012-2022 firm panel matches 259 cities. The stacked design uses a common event window from year -3 through year +2 for the 2015 and 2020 cohorts. The 2015 stack covers 2012-2017, and the 2020 stack covers 2017-2022. Event year -4 is available only for the 2020 cohort because the corresponding year for the 2015 cohort is 2011, outside the integrated sample. Cities treated after each stack's event window and never-treated cities serve as controls. Supplementary Table S4 reports the event-time support and cohort-specific counts.

Supplementary Table S2 reports the complete city schedule used in the analysis: 32 cities enter the 2015 administrative cycle, 21 in 2020, 10 in 2021, 9 in 2022, 8 in 2023, and 7 in 2025. Official center-level documents clarify the administrative stages and provide verifiable location matches. The empirical schedule remains tied to the city-level program-entry field, while unavailable center-specific dates remain unassigned.

City controls come from the raw sheet of a 300-city database rather than interpolated or forecast-filled series. The specifications use parsimonious measures related to program placement and local development: log GDP per capita, service-sector share, credit to GDP, and, where observed, internet users. A high-speed-rail panel covering 300 cities from 2000 through 2022 supplies a concurrent-infrastructure control and a falsification treatment unrelated to the program's computing content.

Figure 2 Sample evolution and outcome prevalence



4 Research Design

The causal estimand is the average effect of local program exposure on firms in the 2015 and 2020 cohorts during event years zero through two. Conventional staggered two-way fixed-effects regressions can assign negative weights or use already-treated units as controls when effects vary across cohorts or time. We therefore construct a separate stack for each cohort. Stack g contains firms in cohort g and firms in cities that remain untreated through g + 2. Both stacks retain the common event window from g - 3 through g + 2, with event year -1 omitted, so every coefficient in the principal event study is supported by both cohorts.

The principal misconduct outcomes are measured in year t + 1, whereas program exposure is recorded in year t. Accordingly, event year zero for the 2015 cohort relates program entry in 2015 to misconduct occurring in 2016. Likewise, event year zero for the 2020 cohort relates the 2020 administrative cycle to misconduct occurring in 2021. This timing convention prevents misconduct occurring before program entry from being classified as a post-treatment outcome. The distinct policy-stage dates are documented in Supplementary Table S1.

The average stacked specification is

$$Y_{it}^{g} = \beta\, \mathrm{StackPost}_{it}^{g} + \gamma' X_{it} + \alpha_i^{g} + \delta_t^{g} + \varepsilon_{it}^{g} \quad (1)$$

where StackPost equals one for cohort-g firms in event years zero through two. Alpha is a stack-by-firm fixed effect and delta is a stack-by-calendar-year fixed effect. X includes log revenue, revenue growth, leverage, return on assets, Tobin Q, book-to-market, and a modified-audit indicator except when the latter is the outcome. Continuous controls are winsorized at the first and ninety-ninth percentiles. Standard errors are clustered by city because the program varies at that level and firms within a city share the shock.

The dynamic specification replaces StackPost with indicators for event years -3, -2, 0, 1, and 2; event year -1 is the reference. Event year -4 is excluded because it corresponds to 2011 for the 2015 cohort but 2016 for the 2020 cohort. Supplementary Table S4 reports this support structure and the cohort-specific observation counts and effects. Identification requires that, conditional on the fixed effects and controls, treated and eligible control firms would have followed parallel counterfactual trends within each stack. The joint test of the two supported leads assesses compatibility with this restriction but cannot establish it. Supplementary Table S11 therefore reports both parametric and bounded differential-trend sensitivity analyses.

We concatenate the two cohort-specific samples after constructing each stack. In the baseline pooled regression, every complete stack-firm-year record receives unit weight; there is no additional cohort-level weighting. The 2015 and 2020 stacks contribute 13,211 (58.05 percent) and 9,545 (41.95 percent) records, respectively. Because a city can be an eligible clean control for more than one cohort, the 22,756 stack records correspond to 21,459 distinct source firm-years. Of these, 1,297 appear in both stacks: 725 are controls in both stacks, and 572 are controls in the 2015 stack before belonging to the treated 2020 cohort. These are intentional stack-specific comparisons, not duplicate source rows. Stack-by-firm and stack-by-calendar-year fixed effects keep their roles separate. To show that the pooled result is not driven by stack size or repeated controls, we additionally report equal-stack weights, which give each cohort stack one-half of total weight, and inverse-duplication weights, which give each underlying firm-year the same total weight across stacks. Because the within transformation and covariate distribution differ by stack, the pooled coefficient is not a simple arithmetic average of the cohort estimates.

$$Y_{it}^{g} = \sum_{k \in K,\, k \neq -1} \beta_k \, 1\{\text{EventTime}_{it} = k\} + \gamma' X_{it} + \alpha_i^{g} + \delta_t^{g} + \varepsilon_{it}^{g} \quad (2)$$

The principal sensitivity analyses retain the cohort-stacked estimator. They add stack-specific industry-by-year fixed effects, use the broader sample without the annual-report text screen, restrict controls to never-treated cities, shorten the event window, and shift treatment onset within the same common-calendar stacks. High-speed-rail exposure provides a placebo treatment in the same stacked sample and fixed-effect structure. A full-sample firm and calendar-year fixed-effects model is included as a complementary benchmark.

For the formal-publication analysis, we assign cities the public-release year of their first verified facility on the 2018 and 2021 focal final lists. The 2018 cohort includes the cities on the first list. The

2021 comparison excludes cities on the first list or in the verified 2015 pilot crosswalk from its treated group. Controls have no confirmed exposure on either focal list or in the pilot crosswalk and remain untreated through the window under the archived schedule. The 2018 window is -3 through +2, and the 2021 window is -3 through +1. This analysis compares the change around recognition, which may follow prior investment and selection.

4 1 Machine Learning Measurement

The AI disclosure benchmark uses temporal cross-fitting by fiscal year. For target fiscal year t, a histogram gradient-boosting regressor is trained on firm-years from 2007 through t - 1 and predicts ln(1 + AI mentions per ten thousand annual-report words) in t. Predictors include AI investment intensity and book value, AI patent and invention counts, R&D, innovation efficiency, firm fundamentals (including ROA and leverage), audit opinion, city economic and internet conditions, industry, province, and calendar year. Missing predictors are replaced with medians from the corresponding training window. This construction prevents later fiscal-year observations from entering an earlier benchmark. For compact notation, W_it denotes the predictor set.

Expected AI talk is the model estimate. Abnormal AI talk equals actual minus expected talk, and the positive disclosure-action gap is the maximum of this residual and zero. The directional measure focuses on unusually intensive AI language rather than sparse disclosure. Fitting each target year only on earlier fiscal years also prevents the post-ChatGPT vocabulary regime from being applied backward to earlier reports.

$$\mathbf{AITalk}_{it}^{\mathbf{pred}} = f_{\theta t}(W_{it})$$
$$\mathbf{Gap}_{it} = \mathbf{AITalk}_{it} - \mathbf{AITalk}_{it}^{\mathbf{pred}}, \quad \mathbf{PositiveGap}_{it} = \max\{\mathbf{Gap}_{it}, 0\} \tag{3}$$

To test **H3**, we regress next-year violations on the winsorized positive residual, AI patenting, AI investment intensity, an interaction between the residual and a standardized AI action index, financial controls, firm effects, and year effects. The action index is the mean of standardized AI patenting and investment intensity. The residual coefficient measures whether within-firm increases in unusually intensive AI language precede within-firm changes in violation risk after conditioning on observable action. Because omitted implementation, managerial shocks, or reverse anticipation may affect both language and conduct, we describe the coefficient as predictive association.

$$\mathbf{Violation}_{i,t+1} = \beta_1\,\mathbf{PositiveGap}_{it} + \beta_2\,\mathbf{AIPatents}_{it} + \beta_3\,\mathbf{AIInvestment}_{it} + \beta_4(\mathbf{PositiveGap}_{it} \times \mathbf{AIAction}_{it}) + \gamma' X_{it} + \mu_i + \lambda_t + \varepsilon_{it} \tag{4}$$

We re-estimate the outcome equation on a common sample using an expanding-window OLS residual, raw AI intensity, and a boosting residual computed without ROA, leverage or modified audit opinion. The financial variables remain controls in the outcome equation, and patents and investment enter each model directly. The alternatives use the same effects, controls and interaction structure. In Table 5, inference uses 399 cluster-pairs bootstrap draws, grouping whole firm histories by their earliest observed city. In each draw we refit the first-stage language models, regenerate the winsorized positive residual and standardized action index, and refit the outcome regressions. Supplementary Table S13 reports percentile intervals and the comparison specifications.

4 2 Later-Period Risk Classification

The classification analysis uses fiscal years 2012-2018 for training, 2019-2020 for validation, and 2021-2022 as a locked test period. Predictor year t is paired with conduct-year outcome t + 1. Fiscal-year ordering prevents later rows from entering earlier fitted models, while the fixed data vintage holds the source archive constant across comparisons. Financial, audit, text, and investment files provide fiscal-year or EndDate fields but not filing dates; patent, city, policy, and climate files do not record historical release vintages; and enforcement labels are reconstructed from later announcements. The resulting exercise evaluates later-period risk ranking rather than probabilities available at a stated historical forecast date. Supplementary Table S7 reports the timing structure and split counts.

We report area under the receiver-operating-characteristic curve, average precision, Brier score, and the fraction of all violations captured in the highest predicted-risk decile. Average precision is particularly informative because violations are rare: a model can have a respectable AUC while producing a large number of false positives. Calibration addresses a different question from ranking and matters if a regulator uses predicted probabilities to allocate inspections. Permutation importance is computed only in the held-out test sample and records the decrease in average precision when each raw feature is shuffled. Importance is predictive and conditional, not a causal mechanism.

5 Empirical Results

5 1 Sample Characteristics

Table 1 documents the sample path. Restricting the 64,286-row financial panel to 2012-2022 leaves 39,546 firm-years. Removing financial industries leaves 38,635 observations, and excluding observed ST or PT years leaves 37,768. The annual-report climate file matches 36,981 of these observations; intersecting this sample with city-policy coverage produces 33,356 firm-years. Because missing ST or PT status does not establish financial distress, only observed distress designations are excluded.

Table 2 shows that any next-year violation occurs in 5.71 percent of firm-years and a punitive violation in 2.42 percent. AI patenting is highly skewed: the median of ln one plus applications is zero, while the mean is 1.045. AI disclosure is also zero at the median across the full sample but rises quickly in later years. The mean AI investment intensity is 0.0057, with a standard deviation of 0.0188. Positive abnormal AI talk has a mean of 0.167 and a standard deviation of 0.368. The city physical-risk index is available for 26,487 observations; its mean is 27.30.

Figure 2 makes the timing visible. AI patent counts and AI disclosure rise over the sample while the next-year violation share varies non-monotonically. Program exposure grows as successive cities enter. The common secular growth of AI measures illustrates why cross-sectional correlations or an unstructured before-after comparison would be misleading. Firm and time effects, and especially clean not-yet-treated controls, are necessary to separate designation from the nationwide digital trend.

5 2 Main Treatment Effects

Table 3 presents the common-support cohort-stacked estimates. The post-treatment coefficient for recorded punitive violations is -0.0171 with a standard error of 0.0043 and p below 0.001. The estimate for all recorded violations is -0.0114 (p = 0.108). Recorded misreporting falls by 0.0115 (p = 0.085), while the narrower recorded fraud indicator falls by 0.0120 (p = 0.004). Enforcement announcements do not respond significantly (coefficient = -0.0072, p = 0.156). Conduct-year and announcement-year outcomes capture different stages of enforcement, and recent conduct labels are right-censored. We therefore use the contrast as an outcome-definition diagnostic rather than as proof of a particular detection mechanism.

Supplementary Table S4 separates the two treatment cohorts using the same controls, fixed effects, city clustering, and common event window. The 2015 stack contains 13,211 complete firm-year

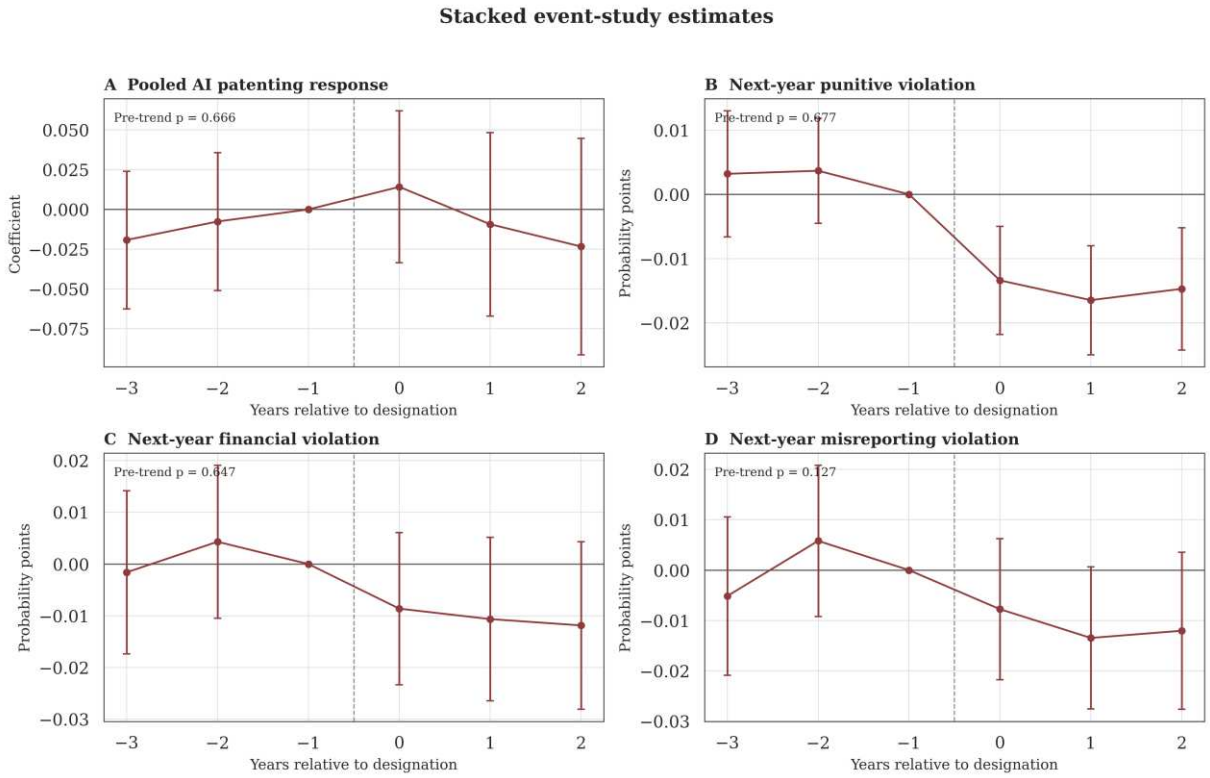
observations and yields an estimate of -0.0222 (SE = 0.0063, $p < 0.001$). The 2020 stack contains 9,545 observations and yields an estimate of -0.0096 (SE = 0.0055, $p = 0.084$). The pooled stack contains 22,756 observations and yields -0.0171 (SE = 0.0043, $p < 0.001$). Both cohort estimates are negative, with a smaller and less precise estimate for the later cohort.

Modified audit opinions fall by 0.0085 in the common-support stacked design ($p = 0.085$) and by 0.0150 in the full firm and year fixed-effects model ($p = 0.0068$). Audit opinions are not a direct measure of internal controls, and a cleaner opinion can sometimes reflect auditor accommodation. The direction of the estimate is consistent with improved reporting systems. The audit-opinion outcome provides complementary evidence, while the operational channel remains unobserved.

Figure 3 and Table 4 report the common-support dynamics. For punitive violations, the event-year coefficients before designation are 0.0032 at year -3 (95 percent CI: -0.0066 to 0.0130) and 0.0037 at year -2 (95 percent CI: -0.0045 to 0.0118); their joint pretrend p value is 0.677. These intervals permit economically meaningful differential trends and do not establish parallel trends. The coefficients become -0.0134 at event zero, -0.0165 at event one, and -0.0147 at event two, all significant at the 1 percent level. Supplementary Table S11 shows that the post estimate remains negative in a model with a treated-specific linear trend, but conservative bounded-trend uncertainty reaches zero under a 0.5 percentage-point annual drift. The timing is compatible with a local governance response, conditional on the maintained trend restriction.

The broad violation outcome displays similar but noisier timing. Its joint pretrend p value is 0.647. The event-zero, event-one, and event-two coefficients are -0.0086, -0.0106, and -0.0119, respectively, and none is statistically distinguishable from zero. Misreporting has a pretrend p value of 0.127, falls most at event one (coefficient = -0.0135, $p = 0.063$), and remains imprecise at event two. These results explain why the average broad-outcome estimate is less precise than the punitive estimate: treatment is most consistently associated with the subset of serious cases.

Figure 3 Cohort-stacked event-study estimates



5 3 AI Capability and AI Disclosure

AI patenting shows neither differential pretrends nor a post-treatment response in the common-support specification. Its pretrend p value is 0.666, and all event-zero through event-two coefficients are statistically indistinguishable from zero. The average stacked AI patent coefficient is 0.0023 with a standard error of 0.0240. AI investment intensity is positive in the stacked specification but insignificant, and the full-sample coefficient is also near zero. H2 is therefore not supported on the observable firm-owned capability margins.

Lower recorded punitive misconduct coincides with no measurable average increase in AI patents or book-value AI assets during the event window. The data leave outsourced cloud capacity, vendor services, ordinary software and changes in internal-control systems unobserved.

The conventional model associates treatment with a 0.0515 increase in log AI disclosure intensity, significant at 1 percent. However, the dynamic AI-disclosure pretrend test rejects with p equal to 0.006, so we do not interpret the increase causally. Positive abnormal AI talk does not fall significantly after designation. Climate abnormal talk likewise shows no treatment response. The strongest program-entry association is for punitive conduct. The estimates provide little evidence that the program narrows the disclosure-action gap.

Table 5 reports coefficients of 0.0271 for any next-year recorded violation and 0.0162 for a punitive violation. The respective two-stage bootstrap 95% intervals are [0.0079, 0.0455] and [0.0041, 0.0272]. AI investment intensity is also positively associated with both outcomes. The asset proxy includes general software and hardware, and the positive coefficients are compatible with transition costs, organizational complexity or investment undertaken in response to prior control concerns. The analysis cannot distinguish those explanations.

Both action interaction intervals include zero. A one-standard-deviation increase in the positive residual corresponds to approximately 1.00 percentage point for any recorded violation and 0.60 percentage points for a punitive violation. These conditional associations leave the roles of implementation quality and omitted forms of AI action open.

Supplementary Table S9 reports the first-stage disclosure benchmarks. The MD&A specification models a different language outcome. For the second-stage violation equation, Table S13 compares the boosting residual, rolling OLS residual and raw intensity after directly controlling for patents and investment. In both outcomes, the alternatives have slightly higher within-R-squared than the boosting model. Omitting ROA, leverage and modified audit opinion from the first stage yields positive residual coefficients of 0.0264 and 0.0157. The positive association persists across the benchmark choices. The fit comparisons support retaining the simpler language measures as comparators.

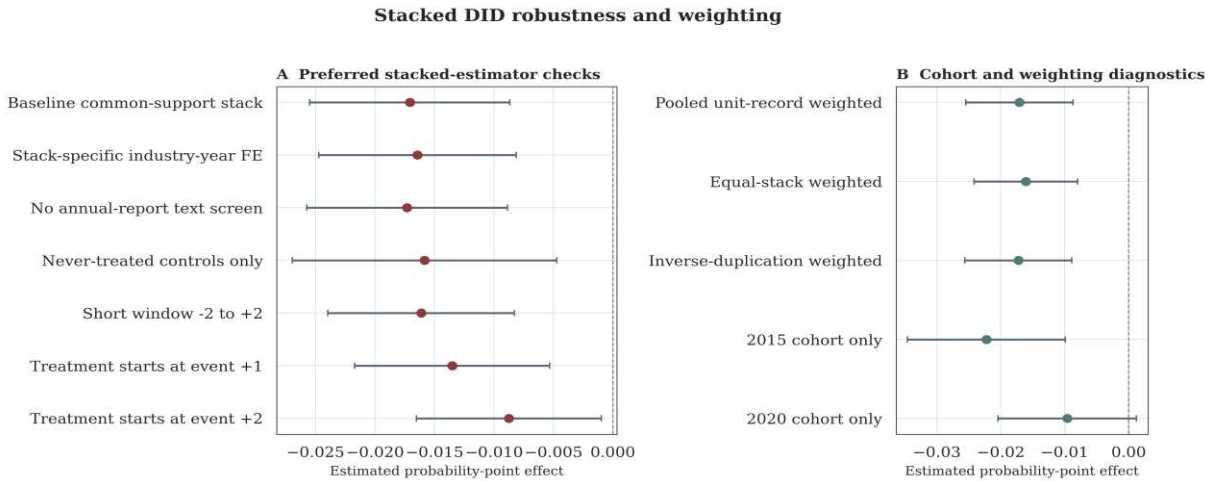
5 4 Robustness and Boundary Conditions

Table 6 and Figure 4 present robustness estimates from the common-support stacked design. The baseline punitive-violation estimate is -0.0171 (SE = 0.0043, p < 0.001). The estimate is -0.0164 with stack-specific industry-by-year fixed effects, -0.0173 without the annual-report text screen, -0.0159 when controls are restricted to never-treated cities, and -0.0161 in the shorter event window from -2 through +2; all four are statistically significant. Shifting treatment onset to event year +1 yields -0.0135 (p = 0.001), while shifting it to event year +2 yields -0.0088 (p = 0.028). Thus, the estimate attenuates as exposure is delayed, consistent with the timing sensitivity discussed above, but remains negative in these common-calendar stacks.

Trend-sensitivity diagnostics address a distinct identifying risk. A common-support stacked model with a cohort-treated relative-year trend estimates a post effect of -0.0136 (SE = 0.0052, p = 0.0099) and an annual treated differential trend of -0.0012 (SE = 0.0019, p = 0.5303). In the bounded-drift calculation, allowing a 0.005 annual differential drift yields a point interval from -0.0271 to -0.0071 and a conservative 95 percent interval from -0.0355 to 0.0013. Thus the point estimate remains negative, but conservative uncertainty includes zero. Table S11 reports the full scale calculation.

The climate, prior AI capability, and pollution-intensive industry interactions are exploratory boundary-condition analyses estimated in the auxiliary full-sample fixed-effects model; they are not robustness tests of the preferred stacked estimator. All three interactions are statistically insignificant, and we resist converting these nulls into subgroup stories. Physical climate risk remains useful in the prediction system and in the expected-disclosure benchmark, but it does not reliably amplify the policy effect.

Figure 4 Stacked DID robustness and weighting



The negative estimate persists under alternative weighting schemes. Equalizing the total weight of the two stacks gives -0.0161 (SE = 0.0041, $p < 0.001$), and inverse-duplication weighting gives -0.0172 (SE = 0.0043, $p < 0.001$). Estimated separately, the 2015 cohort yields -0.0222 (SE = 0.0063, $p < 0.001$) and the 2020 cohort yields -0.0096 (SE = 0.0055, $p = 0.084$). Both point estimates are negative, but the later cohort is smaller and less precise. Applying the high-speed-rail placebo treatment to the same stacked sample, controls, fixed effects, and city clustering produces 0.0013 (SE = 0.0079, $p = 0.867$). Supplementary Table S5 reports record multiplicity, weighting, and specification diagnostics. Conventional two-way fixed-effects estimates are retained as complementary comparisons.

Adding the Broadband China, Smart City and Big Data pilot indicators jointly gives -0.0164 (SE = 0.0042, $p < 0.001$). The estimate remains negative with stack-specific province-by-year or industry-by-year effects and when the 2015 predictor year is omitted. The estimate is -0.0169 ($p < 0.001$) when the pre-2015 city CSRC case rate is interacted with the post-2015 indicator. With the competing policies, the announcement-year coefficient is -0.0076 ($p = 0.133$) for all agencies, -0.0041 ($p = 0.308$) for the CSRC or its bureaus and -0.0034 ($p = 0.348$) for regional bureaus alone. These agency-based outcomes and the prior case rate capture realized enforcement records. Local enforcement effort remains unobserved. With the three policy controls, the modified-opinion estimate is -0.0059 ($p = 0.195$). The data lack comparable measures of firm internal-control quality, weaknesses or audit report lag. Supplementary Tables S8 and S12 report these diagnostics.

Recent conduct labels are not fully mature. The principal sample uses next-year outcomes from 2013 through 2023, and the latest dated announcement linked to a conduct year is April 30, 2025. The 2023 conduct outcome has a two-calendar-year announcement horizon that may omit later actions. Supplementary Table S6 varies the final event year for each cohort and includes controls untreated through the end of each window. For the 2015 cohort, whose post-treatment outcomes end in 2018 and have seven announcement years through 2025, the estimates are -0.0212 for event year zero only, -0.0214 through event year +1, and -0.0222 through event year +2; all have p below 0.001. For the 2020 cohort, the corresponding estimates are -0.0069 ($p = 0.144$), -0.0094 ($p = 0.070$), and -0.0096 ($p =$

0.084), with announcement horizons of four, three, and two years for the latest included outcomes. The mature 2015 cohort therefore receives greater evidentiary weight, while the pooled and 2020 estimates may be attenuated by right censoring.

The treatment year marks entry into the program process rather than final certification. For the initial cohort, pilot-unit selection occurred in December 2015, followed by monitoring, evaluation, and formal recognition through 2018. For the 2020 administrative cycle, public disclosure occurred in December 2020 and the final announcement followed in January 2021. Because the principal outcomes measure misconduct in year t + 1, the baseline timing links program entry to conduct in the subsequent year. Shifting treatment onset to event year +1 attenuates the common-support estimate to -0.0135 (SE = 0.0042, p = 0.001), and shifting onset to event year +2 yields -0.0088 (SE = 0.0040, p = 0.028). The gradual attenuation indicates that the estimated effect is associated with the program-entry period rather than a single universal certification date.

Using actual official publication dates and verified host cities gives +0.0054 (SE = 0.0056, p = 0.336) for 2018 and -0.0125 (SE = 0.0050, p = 0.014) for the new 2021 comparison. The first list provides little evidence of an additional decline around formal recognition. The 2021 result is negative over a shorter window and its last conduct year has only two calendar years of follow-up. The publication-date estimates differ across cohorts and do not establish that the earlier program-entry association reflects infrastructure becoming operational. A separate diagnostic for the program-entry design reassigns the city schedules and rebuilds the preferred stacks in 499 permutations (two-sided p = 0.044, Figure A1). The comparison is descriptive given purposive placement. Table S12 gives the samples, city mappings and estimates.

6 Later-Period Violation Risk Classification

Table 7 evaluates model performance in fiscal years outside those used for fitting and selection. Gradient boosting records the highest validation average precision, 0.212, with a validation AUC of 0.767. On the common set of 8,457 test observations, the financial-only model has an AUC of 0.736 and average precision of 0.082. Adding raw AI talk yields 0.739 and 0.089; adding the positive disclosure-action gap yields 0.738 and 0.088; and adding both yields 0.738 and 0.089. Supplementary Table S10 reports paired uncertainty. The point estimates indicate modest gains, although all incremental intervals include zero.

Logistic regression records a test AUC of 0.708 and average precision of 0.086. Random forest has the highest test AUC, 0.738, but lower average precision, 0.085. Validation average precision determines model selection. Gradient boosting is therefore the designated model for the test-period comparisons.

The nested feature blocks isolate the contribution of AI information. Relative to the financial baseline, raw AI talk raises test AUC by 0.0026 and average precision by 0.0074. The disclosure-action gap raises AUC by 0.0014 and average precision by 0.0061. Adding both raises AUC by 0.0012, average precision by 0.0072, and top-decile capture by 1.08 percentage points. The broader AI-action block, which contains the inclusive digital-capital proxy, does not improve AUC. The full information model improves average precision and top-decile capture but also contains climate, city, innovation,

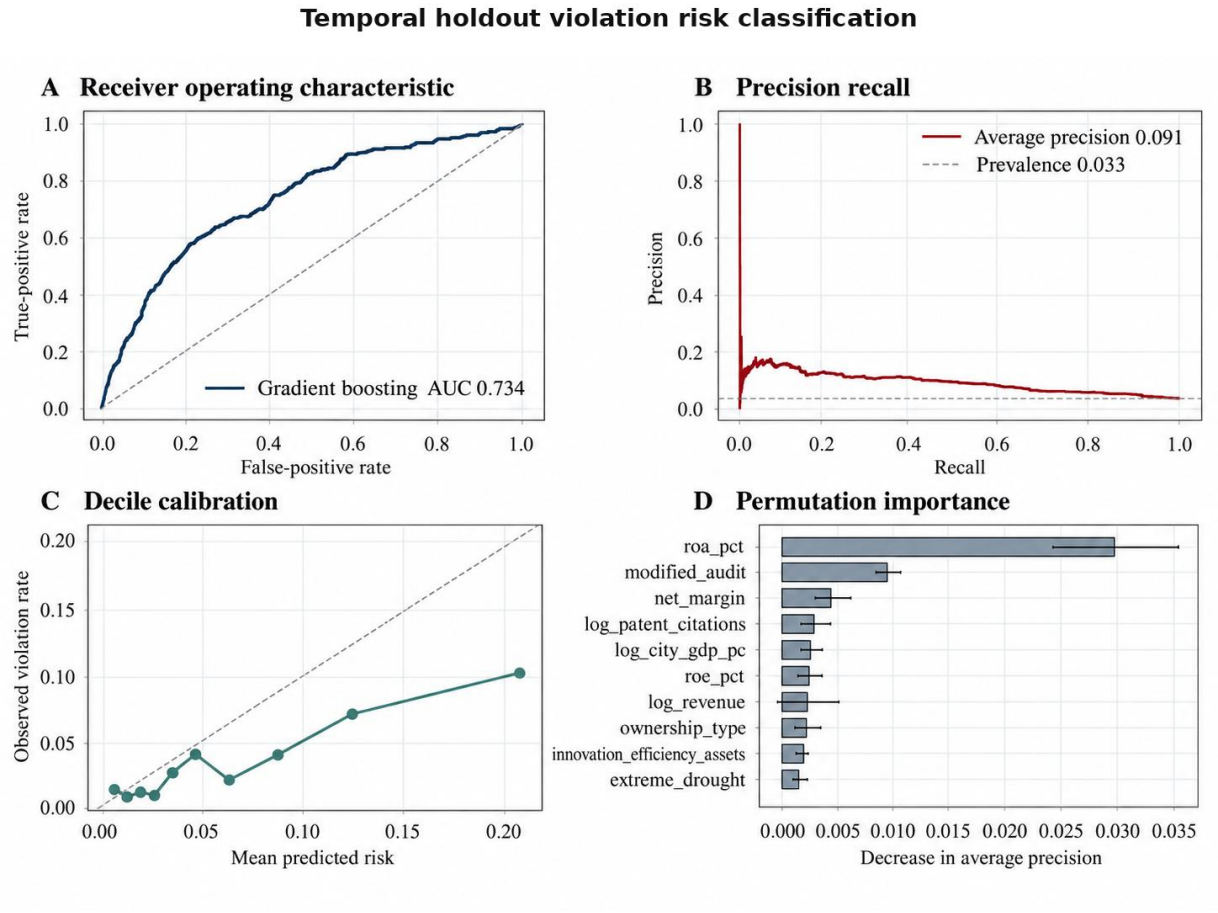
and policy variables. Thus, AI disclosure supplies modest incremental ranking information, whereas the integrated model's larger capture rate reflects the broader information set.

Figure 5 presents ROC, precision-recall, calibration, and risk-decile diagnostics. The ranking curves show useful but imperfect discrimination, and observed violation rates are broadly increasing across calibrated risk deciles. The full model's test Brier score is 0.032759, compared with 0.032499 for a constant based on validation prevalence. The ranking and probability scores therefore support different uses. In addition, validation Brier scores use raw class-weighted probabilities before Platt fitting, whereas test Brier scores use Platt-calibrated probabilities. Table 7 distinguishes these probability bases, and Supplementary Table S10 reports the constant-probability comparisons.

Permutation importance in Table 8 and Figure 5 is led by return on assets and modified audit opinion, followed by net margin, patent citations, city GDP per capita, ownership, revenue, and innovation efficiency. Individual AI fields have limited marginal importance after conditioning on correlated financial and innovation variables. This result complements rather than duplicates the panel association: permutation importance measures the loss in test-sample ranking when one feature is shuffled, whereas the fixed-effects regression relates within-firm changes in the disclosure-action residual to subsequent recorded violations.

The later-period ranking exercise shows how an integrated information set can concentrate recorded violations within a smaller review group. The highest-risk decile captures 31.4 percent of test-period violations, compared with 27.4 percent under the financial-only model. Operational use would additionally require filing timestamps, historical source vintages, labels frozen at each training date, a prespecified label-maturation rule, and prospective recalibration. The present results therefore support risk-oriented screening research rather than automated inspection or sanction decisions.

Figure 5 Later-period violation risk classification



7 Discussion and Implications

City program entry is associated with fewer serious recorded financial violations, while estimates around formal listing differ by cohort. Program documents describe operating, monitoring and reporting requirements for participating centers. Firm use of those services and changes in firms' own internal controls remain unobserved. Modified audit opinions provide a separate reporting outcome, with imprecise estimates after accounting for competing policies.

Positive residual AI language is associated with misconduct assigned to the following conduct year. The residual conditions on reported digital capital, patents and financial characteristics. Simpler measures have comparable within-sample fit, and the later-period ranking gains are modest. The public report examples illustrate how high vocabulary counts can coexist with specific implementation claims.

The policy and disclosure findings describe distinct dimensions of financial risk. Local program exposure predicts lower punitive misconduct, whereas positive unexplained AI talk identifies residual risk across firms. The first relationship concerns the local infrastructure environment; the second concerns inconsistency between corporate narrative and observable implementation. Their coexistence helps separate governance effects from technology production and disclosure-based screening.

The scope of inference follows from the data architecture. Program placement is purposive, and unobserved city-specific policies may remain after controlling for the three observed digital programs

and allowing for province-specific shocks. Local differences in enforcement effort remain unmeasured. Treatment captures city exposure rather than firm-level use of a designated center. Enforcement labels can arrive several years after conduct, so recent outcomes remain right-censored and a zero denotes no recorded violation in the archive. The investment dictionary captures broad digital capital and can miss outsourced services, while the vendor archive provides aggregate counts. The contextual examples come from separately consulted public reports. Supported pre-treatment coefficients are imprecise, and the bounded-trend analysis shows that sufficiently large differential drift can make the conservative interval include zero. These features limit mechanism attribution and the precision of the causal interpretation.

Several extensions follow directly from these boundaries. Firm-center contracts, cloud expenditure, vendor relationships, ERP implementation, auditor office location, and digital-audit adoption could identify take-up and organizational channels. A point-in-time prediction study would require annual-report filing dates, historical financial, patent, city, and enforcement vintages, and a prespecified label-maturation rule. Sentence-level disclosure data would support human semantic validation and models that distinguish verifiable implementation claims from general AI discussion.

For policy evaluation, the findings suggest that digital-infrastructure programs should be assessed not only by energy efficiency and computing output but also by measurable financial-governance outcomes. Recording client firms, service contracts, cloud usage, vendor relationships, process changes, and overlapping local digital policies would make it possible to distinguish infrastructure availability from actual take-up and to test which organizational channels influence reporting integrity.

8 Conclusion

City entry into the National Green Data Center program is associated with a 1.71 percentage-point reduction in next-year punitive violations. This estimate persists after controlling for three overlapping digital programs. Estimates around formal listing differ across the 2018 and 2021 cohorts, and the firm-level mechanism remains open.

The positive AI language residual is associated with subsequent recorded misconduct under two-stage bootstrap inference. Rolling OLS and raw intensity give comparable in-sample fit. In the later-period tests, the integrated model captures 31.4 percent of recorded violations in its highest-risk decile.

Tables

Table 1 Sample construction

Sample step	Firm years
Financial firm-year observations 1999 to 2023	64,286
Restricted to 2012 to 2022	39,546
Nonfinancial industries	38,635
Exclude observed ST or PT years	37,768
Annual-report climate disclosure available	36,981
Covered by the city policy panel main sample	33,356

Notes The base is the firm-level financial panel. Missing ST or PT status is not treated as evidence of distress; only observed ST or PT firm-years are excluded. The main sample additionally requires annual-report climate text so that the same panel supports the integrated analyses. The no-text-screen robustness uses the broader policy-covered sample.

Table 2 Descriptive statistics

Variable	N	Mean	SD	P25	Median	P75
Next-year financial violation	33,203	0.0571	0.2320	0.0000	0.0000	0.0000
Next-year punitive violation	33,203	0.0242	0.1536	0.0000	0.0000	0.0000
Next-year misreporting violation	33,203	0.0487	0.2153	0.0000	0.0000	0.0000
Green data-center treatment	33,356	0.4568	0.4981	0.0000	0.0000	1.0000
AI patenting ln 1 plus count	33,356	1.0454	1.4276	0.0000	0.0000	1.7918
AI disclosure ln 1 plus mentions per 10000	32,846	0.4217	0.6877	0.0000	0.0000	0.5633
Positive abnormal AI talk	32,846	0.1669	0.3681	0.0000	0.0000	0.1309
AI investment intensity	29,162	0.0057	0.0188	0.0005	0.0018	0.0052
Climate disclosure ln 1 plus mentions per 10000	33,356	2.5923	0.7679	2.0783	2.5060	2.9944
Positive abnormal climate talk	33,356	0.2604	0.4125	0.0000	0.0282	0.3898
Physical climate risk	26,487	27.2997	6.9830	23.1988	27.1546	30.7042
Revenue ln RMB 10000	33,350	12.2184	1.5608	11.1689	12.0748	13.1167
Leverage percent	33,356	42.5022	43.5587	24.8209	40.6052	57.1651
Return on assets percent	33,356	2.9587	29.2131	1.2757	3.7635	6.9221
Revenue growth percent	30,963	475.8918	76508.3899	-4.1333	9.4979	25.8087
Tobin Q	32,120	2.1857	4.9699	1.2424	1.6131	2.3310
Book to market	32,120	0.6205	0.2545	0.4290	0.6200	0.8049

Notes Continuous regression variables are winsorized at the first and ninety-ninth percentiles; this table reports raw sample values. AI talk is normalized by annual-report word count. AI investment intensity equals identified AI-related software and hardware book value divided by total assets.

Table 3 Cohort-stacked difference in differences

Outcome	Post treatment	N	Cities	Within R squared
Punitive violation	-0.0171*** (0.0043)	22,756	256	0.0102
All financial violations	-0.0114 (0.0070)	22,756	256	0.0069

Outcome	Post treatment	N	Cities	Within R squared
Misreporting	-0.0115* (0.0066)	22,756	256	0.0059
Fraud	-0.0120*** (0.0042)	22,756	256	0.0084
Enforcement announcement	-0.0072 (0.0050)	22,756	256	0.0081
Modified audit	-0.0085* (0.0049)	22,835	256	0.0709
AI patents	0.0023 (0.0240)	22,835	256	0.0276
AI investment	0.0003 (0.0004)	19,873	252	0.0027

Notes Each stack compares the 2015 or 2020 cohort with cities untreated through the end of the relevant event window. Both stacks use the common window from event year -3 through +2, and the treatment indicator equals one in event years zero through two. Models include stack-by-firm and stack-by-calendar-year fixed effects and the baseline financial controls. Standard errors are clustered by city. *, **, and *** denote significance at the 10, 5, and 1 percent levels.

Table 4 Dynamic treatment effects

Event year	Punitive	All violations	Misreporting	AI patents
-3	0.0032 (0.0050)	-0.0016 (0.0081)	-0.0051 (0.0080)	-0.0192 (0.0221)
-2	0.0037 (0.0042)	0.0043 (0.0075)	0.0058 (0.0076)	-0.0076 (0.0221)
0	-0.0134*** (0.0043)	-0.0086 (0.0075)	-0.0077 (0.0071)	0.0142 (0.0244)
1	-0.0165*** (0.0043)	-0.0106 (0.0081)	-0.0135* (0.0072)	-0.0093 (0.0294)
2	-0.0147*** (0.0049)	-0.0119 (0.0083)	-0.0120 (0.0079)	-0.0233 (0.0347)
Pretrend p	0.677	0.647	0.127	0.666

Notes Event year minus one is omitted. Coefficients and city-clustered standard errors are reported. The final row gives the joint chi-squared test for the supported pre-treatment event years -3 and -2. Event year -4 is excluded from the principal common-support table because it is unavailable for the 2015 cohort. The same controls and stacked fixed effects as Table 3 are included.

Table 5 AI talk action gap and subsequent violations

Variable	Any violation next year	Punitive violation next year
Positive abnormal AI talk	0.0271 (0.0101)	0.0162 (0.0058)
AI patents	0.0057 (0.0031)	0.0012 (0.0015)
AI investment intensity	0.8697 (0.4184)	0.5379 (0.1845)
Talk gap by AI action	0.0056 (0.0097)	-0.0061 (0.0051)
Firm years	26,372	26,372
Firm fixed effects	Yes	Yes
Year fixed effects	Yes	Yes

Notes The positive residual comes from the rolling gradient-boosting model. AI action is the mean of standardized patenting and investment intensity. Parentheses report standard errors from 399 cluster-pairs bootstrap draws that refit the disclosure benchmark and outcome regression. Supplementary Table S13 reports percentile intervals and conventional city-clustered results for the comparison models.

Table 6 Preferred stacked DID robustness and weighting diagnostics

Stacked specification	**Estimate**	**N**
Baseline common-support stack	-0.0171*** (0.0043)	22,756
Equal-stack weighted	-0.0161*** (0.0041)	22,756
Inverse-duplication weighted	-0.0172*** (0.0043)	22,756
Stack-specific industry-year FE	-0.0164*** (0.0042)	22,756
No annual-report text screen	-0.0173*** (0.0043)	22,884
Never-treated controls only	-0.0159*** (0.0057)	16,920
Short window -2 to +2	-0.0161*** (0.0040)	19,484
Treatment starts at event +1	-0.0135*** (0.0042)	22,756
Treatment starts at event +2	-0.0088** (0.0040)	22,756
High-speed rail placebo	0.0013 (0.0079)	22,756

Notes All rows use the punitive-violation outcome and the common-support cohort stacks unless the row states otherwise. The baseline gives each complete stack-firm-year record unit weight. Equal-stack weighting gives the 2015 and 2020 stacks equal total weight; inverse-duplication weighting gives each underlying source firm-year equal total weight across stacks. Models include the baseline financial controls, stack-by-firm fixed effects, stack-by-calendar-year fixed effects, and city-clustered standard errors. The high-speed-rail row applies that placebo treatment within the same stacked sample and specification. *, **, and *** denote significance at the 10, 5, and 1 percent levels.

Table 7 Later-period classification performance

Model	Split	AUC	Average precision	Brier	Top decile capture	Events
Logistic	Validation raw	0.742	0.194	0.238	0.352	423
Logistic	Test calibrated	0.708	0.085	0.032	0.321	277
Random forest	Validation raw	0.765	0.178	0.162	0.352	423
Random forest	Test calibrated	0.738	0.085	0.033	0.318	277
Gradient boosting	Validation raw	0.767	0.212	0.176	0.378	423
Full information	Test calibrated	0.734	0.091	0.033	0.314	277
Financial only	Test calibrated	0.736	0.082	0.033	0.274	277
Financial plus raw AI talk	Test calibrated	0.739	0.089	0.033	0.274	277
Financial plus AI disclosure-action gap	Test calibrated	0.738	0.088	0.033	0.282	277
Financial plus raw AI talk and gap	Test calibrated	0.738	0.089	0.033	0.285	277

Notes Models are fitted on fiscal years 2012-2018, selected or calibrated on 2019-2020, and evaluated once on 2021-2022. Validation Brier scores use raw class-weighted probabilities before the Platt calibrator is fitted; test Brier scores use probabilities calibrated on the validation period, so the two are not directly comparable. The financial-only and nested AI rows hold the gradient-boosting algorithm, preprocessing, categorical controls, class weighting, calibration protocol, and test observations constant. The full-information row also contains climate, innovation, city, and policy variables, so its difference from the financial baseline is not an AI-specific increment. Predictor year t is paired with recorded conduct in t + 1. The exercise evaluates later-period classification within a fixed data vintage; Supplementary Tables S7 and S10 report the information-timing structure and incremental-performance comparisons.

Table 8 Permutation importance in the test sample

Feature	Mean decrease in average precision	Standard deviation
roa pct	0.0300	0.0059
modified audit	0.0097	0.0011
net margin	0.0046	0.0017
log patent citations	0.0030	0.0014
log city gdp pc	0.0027	0.0011
roe pct	0.0025	0.0012
log revenue	0.0024	0.0029
ownership type	0.0023	0.0012
innovation efficiency assets	0.0021	0.0002
extreme drought	0.0017	0.0007
ai invention log	0.0016	0.0002
log climate mentions	0.0011	0.0004

Notes Importance is computed by permuting one raw feature at a time in the held-out 2021 to 2022 sample for the selected gradient-boosting model. Values near zero indicate little marginal ranking information conditional on other features; they do not imply causal irrelevance.

Appendix Data and Reproducibility Notes

The data architecture is designed to preserve provenance. The financial panel is the master firm-year skeleton. Industry and location histories supply city and province identifiers. Climate text, AI text, patents, investment, innovation, and audit fields are merged one-to-one on stock code and fiscal year. City policy, climate, and macroeconomic variables are merged many-to-one on standardized six-digit city code and year, with normalized city names used only as a physical-risk fallback. Violation positives are merged after event-year expansion. Every merge validates the expected key cardinality.

The country climate-risk files are not used in the estimations because the sample contains domestic Chinese firms and country-year variation would be absorbed by calendar effects. Province climate risk is retained as an aggregation check but city physical risk is preferred in the integrated panel. Provincial vulnerability measures with acceptable coverage enter the expected climate-disclosure model; fields with extreme missingness are omitted rather than imputed into a composite score. The annual-report

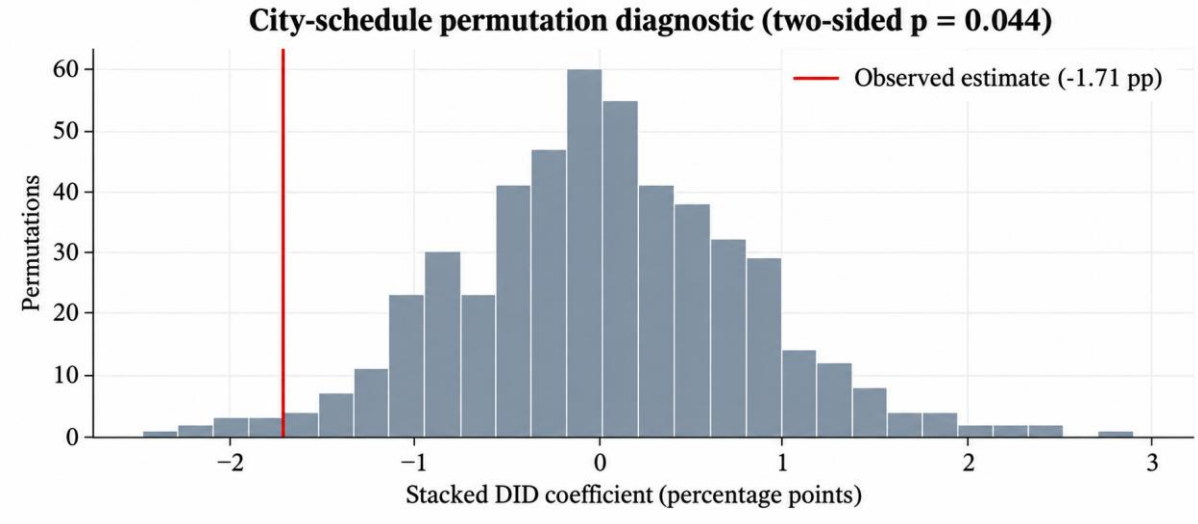
total climate index is not entered beside transition risk because the two are almost identical in the source data.

The rolling AI benchmark uses a fixed random seed and an expanding fiscal-year training window. Annual R-squared values are 0.399, 0.418, 0.422, 0.428, 0.409, 0.451, 0.506, 0.528, 0.527, 0.545, 0.562, and 0.550 for 2012 through 2023. The corresponding target-year sample grows from 2,354 to 4,879 firm-years. Each value is computed on target-year observations excluded from the model's earlier-year fitting sample and therefore measures generalization across fiscal years.

The violation classifier applies deterministic preprocessing within each fiscal-year split. Numeric values are median-imputed from the relevant training data. Industry and ownership are one-hot encoded, with unknown categories permitted in later periods. Logistic regression standardizes numeric variables and uses balanced class weights. Random forest uses 350 trees, maximum depth 10, and minimum leaf size 25. Histogram gradient boosting uses a learning rate of 0.05, 260 iterations, 15 leaves, minimum leaf size 25, and L2 regularization of 1. Hyperparameters are selected on validation average precision. Platt calibration is fitted on validation predictions; monotonic calibration does not change the ranking metrics.

Rebuilding from the original inputs yields 64,286 merged firm-years and 33,356 analysis firm-years. The replication archive includes the scripts, codebook, derived sample, results, figures and the publication-date, policy and bootstrap analyses for Tables S12 and S13. It also records source-specific processing requirements: the violation Excel worksheet has an incorrect XML used-range marker and is read from its Stata version; AI patent categories are not summed across overlapping group and subsidiary rows; all patent observations carry a post-listing label; city administrative codes change over time; and the original climate-cleaning script contains a full-width exclamation mark in one rename statement.

Figure A1 City schedule permutations for the common-support stacked DID



Notes The histogram displays 499 reassignments of complete city schedules with both stacks rebuilt in each draw. The observed coefficient is -1.71 percentage points. The two-sided permutation p-value of 0.044 compares the estimate with shuffled schedules.

References

Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. Annual Review of Economics 11, 685–725.

Babina, T., Fedyk, A., He, A.X., Hodson, J., 2024. Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics 151, 103745.

Baker, A.C., Larcker, D.F., Wang, C.C.Y., 2022. How much should we trust staggered difference-in-differences estimates. Journal of Financial Economics 144, 370–395.

Beneish, M.D., 1999. The detection of earnings manipulation. Financial Analysts Journal 55, 24–36.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates. Quarterly Journal of Economics 119, 249–275.

Bharadwaj, A.S., 2000. A resource-based perspective on information technology capability and firm performance. MIS Quarterly 24, 169–196.

Bloom, N., Sadun, R., Van Reenen, J., 2012. Americans do IT better US multinationals and the productivity miracle. American Economic Review 102, 167–201.

Bolton, P., Kacperczyk, M., 2021. Do investors care about carbon risk. Journal of Financial Economics 142, 517–549.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Brynjolfsson, E., Rock, D., Syverson, C., 2021. The productivity J-curve How intangibles complement general purpose technologies. American Economic Journal Macroeconomics 13, 333–372.

Callaway, B., Sant'Anna, P.H.C., 2021. Difference-in-differences with multiple time periods. Journal of Econometrics 225, 200–230.

Cao, S., Jiang, W., Yang, B., Zhang, A.L., 2023. How to talk when a machine is listening Corporate disclosure in the age of AI. Review of Financial Studies 36, 3603–3642.

Cengiz, D., Dube, A., Lindner, A., Zipperer, B., 2019. The effect of minimum wages on low-wage jobs. Quarterly Journal of Economics 134, 1405–1454.

Christensen, H.B., Hail, L., Leuz, C., 2021. Mandatory CSR and sustainability reporting Economic analysis and literature review. Review of Accounting Studies 26, 1176–1248.

China Securities Regulatory Commission, 2016. Briefing on securities enforcement in 2015 [in Chinese]. https://www.csrc.gov.cn/csrc/c100200/c1000407/content.shtml.

Dechow, P.M., Ge, W., Larson, C.R., Sloan, R.G., 2011. Predicting material accounting misstatements. Contemporary Accounting Research 28, 17–82.

Diamond, D.W., Verrecchia, R.E., 1991. Disclosure liquidity and the cost of capital. Journal of Finance 46, 1325–1359.

Du, J., Xu, X., Yang, Y., 2023. Does climate risk affect the cost of equity Evidence from textual analysis of Chinese listed companies' annual reports. Finance Review 15, 19–46.

Dyck, A., Morse, A., Zingales, L., 2010. Who blows the whistle on corporate fraud. Journal of Finance 65, 2213–2253.

Easley, D., O'Hara, M., 2004. Information and the cost of capital. Journal of Finance 59, 1553–1583.

Goldfarb, A., Tucker, C., 2019. Digital economics. Journal of Economic Literature 57, 3–43.

Dogan, B., Teklie, D.K., 2026. The diffusion of artificial intelligence in corporate disclosure: A decade of S&P 500 10-K filings, 2015–2025. Journal of Risk and Financial Management 19(7), 473. https://doi.org/10.3390/jrfm19070473.

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. Journal of Econometrics 225, 254–277.

Graham, J.R., Harvey, C.R., Rajgopal, S., 2005. The economic implications of corporate financial reporting. Journal of Accounting and Economics 40, 3–73.

Hassan, T.A., Hollander, S., van Lent, L., Tahoun, A., 2019. Firm-level political risk Measurement and effects. Quarterly Journal of Economics 134, 2135–2202.

Healy, P.M., Palepu, K.G., 2001. Information asymmetry corporate disclosure and the capital markets. Journal of Accounting and Economics 31, 405–440.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy 124, 1423–1465.

Karpoff, J.M., Lee, D.S., Martin, G.S., 2008. The cost to firms of cooking the books. Journal of Financial Economics 88, 193–215.

Kothari, S.P., Shu, S., Wysocki, P.D., 2009. Do managers withhold bad news. Journal of Accounting Research 47, 241–276.

Krueger, P., Sautner, Z., Starks, L.T., 2020. The importance of climate risks for institutional investors. Review of Financial Studies 33, 1067–1111.

Larcker, D.F., Zakolyukina, A.A., 2012. Detecting deceptive discussions in conference calls. Journal of Accounting Research 50, 495–540.

Leuz, C., Nanda, D., Wysocki, P.D., 2003. Earnings management and investor protection An international comparison. Journal of Financial Economics 69, 505–527.

Li, F., 2010. The information content of forward-looking statements in corporate filings A naïve Bayesian machine learning approach. Journal of Accounting Research 48, 1049–1102.

Loughran, T., McDonald, B., 2011. When is a liability not a liability Textual analysis dictionaries and 10-Ks. Journal of Finance 66, 35–65.

Mullainathan, S., Spiess, J., 2017. Machine learning An applied econometric approach. Journal of Economic Perspectives 31, 87–106.

Roychowdhury, S., 2006. Earnings management through real activities manipulation. Journal of Accounting and Economics 42, 335–370.

Sautner, Z., van Lent, L., Vilkov, G., Zhang, R., 2023. Firm-level climate change exposure. Journal of Finance 78, 1449–1498.

Sun, L., Abraham, S., 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics 225, 175–199.

Tetlock, P.C., 2007. Giving content to investor sentiment The role of media in the stock market. Journal of Finance 62, 1139–1168.

Ministry of Industry and Information Technology, National Government Offices Administration, National Energy Administration, 2015a. National Green Data Center Pilot Work Plan, MIIT Joint Energy Conservation [2015] No. 82.

Ministry of Industry and Information Technology, National Government Offices Administration, National Energy Administration, 2015b. Notice announcing the National Green Data Center pilot regions, MIIT Joint Energy Conservation Letter [2015] No. 475.

Ministry of Industry and Information Technology, National Government Offices Administration, National Energy Administration, 2015c. Notice announcing the National Green Data Center pilot units, MIIT General Office Joint Energy Conservation [2015] No. 163.

Ministry of Industry and Information Technology, National Government Offices Administration, National Energy Administration, 2018. National Green Data Center List (First Batch), Announcement [2018] No. 6.

Ministry of Industry and Information Technology and five other agencies, 2020. Notice on recommendations for National Green Data Centers (2020), MIIT General Office Joint Energy Conservation Letter [2020] No. 183.

Ministry of Industry and Information Technology and five other agencies, 2021. 2020 National Green Data Center List Announcement 2021 No 2.

Ministry of Industry and Information Technology and five other agencies, 2022. 2021 National Green Data Center List Announcement 2022 No 7.

Ministry of Industry and Information Technology and five other agencies, 2023. 2022 National Green Data Center List Announcement 2023 No 10.