
An Agentic AI Framework for ESG Disclosure Monitoring and Governance Evidence

Amit Aylani^{1*}  | Madhuri Rao² 

¹Thadomal Shahani Engineering College, Mumbai, India

¹Vidyalankar Institute of Technology, Mumbai, India

²Thadomal Shahani Engineering College, Mumbai, India

*Correspondence to: Amit Aylani, Thadomal Shahani Engineering College, Mumbai, India.
E-mail: prof.amitaylani@gmail.com

Abstract: Environmental, social, and governance (ESG) disclosure is increasingly important for sustainability assessment, compliance review, and managerial decision-making. However, existing AI-based ESG applications are often task-specific and provide limited support for longitudinal indicator harmonisation, missing-data handling, adaptive rule updates, and auditable decision trails. This study develops and evaluates an Agentic ESG Integration Framework (AEIF), a four-agent decision-support architecture for ingesting disclosures, mapping and aligning ESG indicators, detecting anomalies, and generating governance-oriented decision trails. The study harmonises 43 ESG indicators across 480 firm-year reports from 24 listed firms in four countries and six sectors over 2005–2024. AEIF is compared with a rules-based baseline using 60 retrospective monitoring scenarios and missing-data stress levels of 10%, 20%, and 30%. AEIF reduced mean compliance-update latency from 47.2 to 12.4 days and remediation time from 16.2 to 3.5 hours. Precision increased from 0.62 to 0.89 and recall from 0.58 to 0.86. At 30% missingness, detection accuracy was 84% versus 42% for the baseline. The findings indicate that governed agentic workflows can improve ESG monitoring while supporting comparability, traceability, human review, and assurance readiness.

Keywords: Agentic AI, anomaly detection, compliance monitoring, ESG disclosure monitoring, governance evidence, multi-agent systems, sustainability accounting.

Article info: Received 9 June 2026 | revised 14 September 2026 | accepted 21 September 2026 | published 1 October 2026

Recommended citation: Aylani, A., & Rao, M. (2026). An Agentic AI Framework for ESG Disclosure Monitoring and Governance Evidence. *Indonesian Journal of Sustainability Accounting and Management*, 10(2), 345–366. <https://doi.org/10.20448/ijSAM.v10i2.9446>

INTRODUCTION

In recent years, there has been a significant increase in the importance of environmental, social and governance (ESG) factors in corporate strategy, as companies understand that sustainability ratings are connected with financial health, cost of capital and market value (Friede, Busch, & Bassen, 2015; Shakil, 2021). ESG information is becoming more critical for institutional investors' portfolio allocation, risk assessment and stewardship decisions (Amel-Zadeh & Serafeim, 2018; Eccles, Kastrapeli, & Potter, 2017), and there is a growing body of evidence linking ESG performance to lower financial risk and higher firm valuations (Shakil, 2021). As a result, there has been an increasing push towards harmonisation of sustainability reporting at the global level. The International Sustainability Standards Board (ISSB) has issued IFRS S1 and S2 to provide a common framework for disclosures around sustainability and climate change, and the European Union (EU) has enacted the Corporate Sustainability Reporting Directive (CSRD), which introduced additional sector-specific disclosures



Copyright © 2026 by the author(s). This article is published under the Creative Commons Attribution (CC BY 4.0) license. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial & non-commercial purposes), subject to full attribution to the original publication and author(s). The full terms of this license may be seen at <http://creativecommons.org/licenses/by/4.0/legalcode>

and assurance requirements (European Union, 2022). Given regulators' and investors' growing focus on transition and physical risks, as well as the need for forward-looking, decision-useful information, firm-level climate risk assessment has become a key topic (Zhao et al., 2025). Demand for structured and comparable sustainability information continues to be strong despite regulatory uncertainty, even in jurisdictions where the regulatory pathway is unclear, including the United States, after recent legal challenges to the Securities and Exchange Commission's (SEC) climate rule (Fligstein & Huang, 2025; Kolasa & Sautner, 2025; U.S. Securities and Exchange Commission, 2024). The impact of climate policy uncertainty on corporate green governance decisions has also been demonstrated, further emphasising the importance of effective disclosure frameworks (Sun, Lu, & Hunt, 2025). These changes reflect the growing momentum for more holistic, measurable and informative ESG reporting frameworks.

However, data supporting ESG analysis is highly fragmented. Previous studies have shown significant discrepancies among ESG ratings, even with access to the same public disclosures (Berg, Kölbel, & Rigobon, 2022; Christensen, Serafeim, & Sikochi, 2022). Research on the dynamics of global ESG indexes also reveals that the rate and extent of convergence between rating systems are slow and incomplete, with ongoing methodological differences (Kerkemeier & Kruse-Becher, 2022). Heterogeneous organisational boundaries, inconsistent unit conventions, the selective disclosure of adverse events, including greenwashing practices that misrepresent environmental performance (Yu, Van Luu, & Chen, 2020), and the mixing of quantitative indicators with narrative claims complicate both cross-sectional benchmarking and longitudinal trend analysis (Berg et al., 2022; Christensen et al., 2022). Such inconsistencies prevent practitioners from building robust time-series data sets and monitoring progress towards performance levels under regulatory and voluntary commitments. Regulatory efforts, such as green finance policies that promote the disclosure of environmental, social, and governance (ESG) issues, have not been successful in eliminating these inconsistencies (Zhang, Zhang, Zhang, Yang, & Lan, 2024).

Advances in artificial intelligence and machine learning have begun to address parts of this challenge. Studies have used natural language processing to quantify ESG communication and explain variances in commercial scores (Schimanski et al., 2024), while others have applied supervised and unsupervised ML models for controversy detection, emissions estimation, or sustainability-linked financial prediction (Seow, 2025; Wyrobek, 2020). AI adoption has also been linked to direct improvements in firm-level ESG and financial outcomes through enhanced operational efficiency and decision-making (Wamba-Taguimdje, Wamba, Kamdjoug, & Wanko, 2020). Systematic reviews show rapid growth in AI-enabled ESG analytics, particularly in text mining, sentiment extraction, and predictive modelling of ESG risks (Seow, 2025); parallel evidence suggests that organisational adoption of AI capabilities is associated with improvements in ESG outcomes, especially in resource efficiency and supply-chain management (Yu, Fan, & Yu, 2025).

However, the current generation of AI-based ESG systems does not adequately address the operational problem examined in this study. Most published applications are task-specific: they classify ESG text, estimate scores, detect controversies, or predict selected outcomes. These approaches generally assume that indicator definitions, reporting boundaries, units, and source documents have already been made comparable. In practice, disclosure monitoring requires the system to recognise amended reports, reconcile changing terminology and organisational boundaries, distinguish genuine anomalies from restatements or corporate events, propagate uncertainty caused by missing evidence, and preserve a reviewer-verifiable record of how each conclusion was reached. A conventional machine-learning model may improve prediction within one task but does not, by itself, coordinate these interdependent controls or maintain an auditable chain from source disclosure to action. Rules-based systems provide traceability but adapt slowly when standards, aliases, or thresholds change. However, the issue is not simply AI's ability to analyse ESG information; it concerns an adaptive multi-agent workflow that can provide harmonisation, monitoring, escalation, and governance evidence while accommodating longitudinal reporting inconsistencies and incomplete information.

This study fills these gaps by developing a harmonised 20-year ESG panel and testing an agentic monitoring framework under real-world disclosure constraints. We construct a balanced sample of 480 firm-year observations from 24 large, publicly traded firms across six sectors and four regions (Germany, India, Japan, and the United States) between 2005 and 2024. We standardize a list of 43 environmental, social, and

governance indicators in terms of units, organisational boundaries, and rate definitions, and document targets, community investment, and adverse events. On this basis, we deploy an Agentic ESG Integration Framework (AEIF) that supports the continuous ingestion of disclosures, dynamic remapping to ESG rules, anomaly detection in long, noisy time series, and the maintenance of complete decision trails. It is compared with a rules-based baseline using operational metrics such as compliance latency, precision and recall, remediation time, and robustness to simulated missingness.

It provides a transparent and auditable ESG panel that captures variations in reporting practices across sectors and jurisdictions, creating a realistic environment for framework evaluation. Second, it shows that an agentic architecture can significantly decrease compliance lag, enhance anomaly detection, and maintain performance in the presence of missing inputs ranging from 10 to 30 percent, which is typical of corporate reporting. Third, it includes fairness, transparency, and coordination metrics in the system evaluation, demonstrating how autonomous ESG systems can deliver stronger technical performance and governance capabilities. These contributions collectively enrich the ESG analytics literature by providing a replicable blueprint for achieving timely, reliable, and well-governed sustainability intelligence at scale.

LITERATURE REVIEW

Fragmentation in ESG Reporting and Divergence in Measurement

There has been a considerable amount of literature exploring the flaws of sustainability accounting. Regulatory and industry disclosures underlying key ESG indicators are inconsistent across firms, industries, and jurisdictions. Friede et al. (2015) aggregated findings from more than 2,000 empirical studies and showed that ESG performance is often a meaningful financial factor, but that the relationship varies because of differences in metrics and reporting. Later, the literature pointed to the materiality of items in the Explanatory Notes rather than to the number of disclosures, showing that companies prefer to report issues they believe are material to investors (Amel-Zadeh & Serafeim, 2018). The importance of strategically disclosing ESG information for investment decisions is also reflected in global surveys of institutional investors, which indicate that comparability and consistency of data are the main issues of concern to investors (Amel-Zadeh & Serafeim, 2018; Eccles et al., 2017). These concerns are compounded by persistent limitations in ESG data, including inconsistencies in definitions, measurement approaches, reporting boundaries, and the comparability of company-level indicators (Kotsantonis & Serafeim, 2019).

ESG measurement remains characterised by substantial disagreement across rating providers. Differences in indicator selection, scope, weighting, aggregation, and methodological choices can result in materially different assessments of the same firm (Berg et al., 2022; Billio, Costola, Hristova, Latino, & Pelizzon, 2021). Evaluation of ESG rating methodologies also shows considerable variation in the way rating agencies incorporate sustainability principles and assess environmental, social, and governance dimensions (Escrig-Olmedo, Fernández-Izquierdo, Ferrero-Ferrero, Rivera-Lirio, & Muñoz-Torres, 2019). Research on the dynamics of global ESG indices reinforces this finding, showing that convergence across major index providers remains partial and methodologically contested (Kerkemeier & Kruse-Becher, 2022). A related strand of literature documents that some firms respond to rating divergence not by improving disclosure quality but by engaging in greenwashing, misrepresenting or exaggerating environmental and social performance in reports (Yu et al., 2020).

Various regulatory measures have been taken to address fragmentation. The European Union (2022) Corporate Sustainability Reporting Directive requires auditable sustainability reporting, including the harmonisation of sector- and country-based reporting practices. The IFRS S1 and S2 standards issued by the ISSB seek to achieve the same goal at the international level (International Sustainability Standards Board, 2023a, 2023b), while the U.S. SEC's 2024 climate-disclosure rule, despite facing legal challenges, indicates increasing pressure for comparable reporting (U.S. Securities and Exchange Commission, 2024).

The promotion and adoption of green finance policies, such as regulatory ESG disclosure requirements linked to credit and capital-market incentives, have also sought to enhance reporting in emerging markets (Zhang et al., 2024). Studies on mandatory reporting of sustainability also suggest that the disclosure can have

an impact on the disclosure of companies and the economics thereof, but the impact of disclosure depends on implementation, enforcement and the environment in which the disclosure takes place (Christensen, Hail, & Leuz, 2021). Empirical studies, however, show that regulatory frameworks do not necessarily overcome inconsistencies arising from measurement conventions, emissions limits, or organisational scopes (Amel-Zadeh & Serafeim, 2018; Eccles et al., 2017). Therefore, intelligent monitoring tools are needed to bridge the gap between regulatory intent and operational compliance.

AI Adoption and its Organisational and Performance Implications in ESG

Related literature examines the impact of AI use on corporate ESG performance. Using a panel data sample of listed companies, Yu et al. (2025) found that firms with stronger AI capabilities are more likely to exhibit stronger ESG performance, which they partially attributed to these companies' greater efficiency and better resource management.

Research on the impact of AI-driven business transformation on firm performance reveals that the adoption of AI tools brings tangible benefits in terms of operational performance and decision quality, with implications for how firms monitor and manage sustainability-related risks (Wamba-Taguimdje et al., 2020). In financial statement analysis, machine learning models have been used to identify irregularities and non-compliant reporting behaviour, further demonstrating the potential for automated monitoring in high-stakes reporting situations (Wyrobek, 2020).

NLP-based approaches have also been applied to corporate climate disclosures to assess the specificity, credibility, and informational content of sustainability-related communication (Bingler, Kraus, Leippold, & Webersinke, 2022; Schimanski et al., 2024). Collectively, these studies demonstrate the growing capacity of AI and NLP techniques to process large volumes of corporate sustainability information and identify patterns that may be difficult to detect through manual review alone.

Recent literature reinforces the importance of incorporating governance and assurance considerations into ESG monitoring systems. Sustainability assurance research has long emphasised the importance of independent verification, evidence quality, and institutional context in strengthening the credibility of sustainability disclosures (Simnett, Vanstraelen, & Chua, 2009). Sustainability assurance research further emphasises that credible assurance depends on appropriate evidence, professional judgement, verification processes, and clearly defined accountability mechanisms (Cohen & Simnett, 2015). Evidence from sustainability assurance practice also shows that the credibility of assurance depends on professionalisation, reviewer expertise, verification procedures, and the ability to substantiate reporting conclusions with appropriate evidence (Boiral, Heras-Saizarbitoria, & Brotherton, 2020). These strands work together to encourage the assessment of AEIF in terms of speed, detection accuracy, comparability, assurance readiness, and reviewability.

Conceptual Framework

The conceptual framework highlights that disclosure monitoring is a cycle of governed information processing, with four elements: (1) disclosure inputs, (2) agentic processing (ingestion, mapping and alignment, anomaly screening, and control and decision-trail logging), (3) controlled mechanisms (provenance, uncertainty flags, thresholds, human escalation, and logging of overrides), and (4) disclosure monitoring outcomes (compliance latency, precision and recall, remediation time, robustness to missing data, completeness of the decision trail, and reviewer agreement). The framework therefore treats technical accuracy and governance readiness as jointly necessary outcomes rather than separate objectives.

The central idea is that adaptive mapping reduces the delay associated with manual rule maintenance; contextual anomaly screening improves detection quality by adding historical indicators to event evidence; coordinated fallback and escalation minimize the operational impact of missing information; and explicit provenance and human-review checkpoints enhance reviewability. These relationships are used to conceptualize H1-H4 and differ from static rule engines and stand-alone predictive models.

RESEARCH QUESTION AND HYPOTHESES

To facilitate discussion of the concepts within the context of the empirical design, the central research question has been translated into four directional hypotheses at the monitoring-system level.

H₁: AEIF will decrease compliance latency after disclosure-rule changes because it can update mappings, thresholds, and routing logic without manual recoding.

H₂: AEIF will enhance the quality of anomaly detection in relation to the rules-based baseline, with a higher number of hits (precision) and more hits (recall) and a lower number of false positives (FPR).

H₃: AEIF will shorten remediation cycles and be more resilient to degradation in the face of missing data by facilitating coordination at the agent level, enabling fallback reasoning and targeted escalation when disclosures are incomplete.

H₄: Relative to a rules-based baseline, AEIF will exhibit stronger governance properties, including narrower parity gaps, fuller decision trails, and higher reviewer agreement with explanations, because traceability and human-review checkpoints are built into the workflow.

MATERIALS AND METHODS

Overview and Study Design

We used a 20-year longitudinal panel (reporting years 2005–2024), with the firm-year as the unit of analysis. The design is not a population-level estimation exercise but rather an evaluation of comparative frameworks. We created a cross-context variation panel comprising 24 firms in a 6-sector × 4-country matrix (one firm per cell), yielding 480 firm-years and maintaining data balance and comparability. Sampling, system evaluation, and the end-to-end workflow were harmonised ex ante across units, organisational boundaries, and rate definitions. The workflow, from sampling to system evaluation, is summarised in Figure MF1. AEIF stands for the Agentic ESG Integration Framework.

Phase 1: Sampling frame and panel construction

To prevent look-ahead bias, we created a sample frame of constituents as of the end of 2004. We identified large, exchange-listed corporations, standardised ISINs and tickers, and tracked M&A lineage to ensure continuity of entities over time. Each sector–country cell was populated by ranking candidate firms according to (i) uninterrupted listing through 2024, (ii) the availability of a continuous annual ESG, sustainability, or integrated reporting record, and (iii) the stability of organisational boundary definitions during the study period. Only one anchor firm was selected per cell, and a reserve anchor was recorded in case subsequent checks revealed irreconcilable differences in reporting. Exclusions included long suspensions, delistings, or conglomerate restructurings that prevented firms from being included in the study perimeter. Inclusion criteria were continuous listing, public disclosure, and existing evidence of the indicator catalogue.

The selection of the 24 firms is directly related to the balanced factorial design: one anchor firm in each of the 24 sector–country cells (6 sectors × 4 countries). We opted for controlled heterogeneity rather than broad but shallow coverage because the focus of the study is to assess a monitoring architecture in the presence of transparent variation, rather than to estimate population means. The countries were chosen to represent four materially different reporting and governance contexts: Germany, India, Japan, and the United States. Together with the six sectors, they capture contrasting levels of environmental intensity, labour and safety exposure, supply-chain complexity, and governance demands. This panel will minimise composition drift, but it is not intended to represent a market sample; therefore, the findings should not be interpreted as representative of the market. To ensure sample transparency and reproducibility. System evaluation was carried out based on the stated sector-country matrix and inclusion criteria and the firm selection was fixed. Aimed at limiting the researcher's discretion in eligibility, the same criteria were used for every cell; reserve candidates were recorded prior to analysis, and no firm was replaced based on their AEIF performance. Nevertheless, the analysis remains susceptible to selection bias, as large, continuously listed companies with

longer disclosure histories are more likely to have fully developed reporting systems. This limitation does not support claims of market representativeness and is addressed through explicit scope conditions.

Phase 2: Evidence collection and repository

For each firm-year, we obtained (a) stand-alone sustainability/ESG reports, (b) integrated or annual reports, and (c) relevant regulatory filings. If more than one version was available within a year, the most recent amended version was used as the source of record, and earlier versions were saved for restatement tracking purposes. A versioned repository collected metadata (title, year), checksums, access links, language, and provenance notes, ensuring the traceability and reproducibility of the corpus. (Refer to Figure MF1 Phase 2.)

Phase 3: Indicator taxonomy, harmonization, and extraction

An ex-ante catalogue of 43 indicators anchored measurement across the pillars.

- Environmental: Scope 1; Scope 2 (location- and market-based); Scope 3; total energy use; renewable energy share; water (withdrawal, consumption, discharge); waste (generated, recycled); environmental incidents.
- Social: number of employees; percentage of female employees; percentage of women in leadership; training hours per employee; safety (TRIR, LTIFR, and fatalities); supplier oversight (audits, breaches, and compliance).
- Governance: board size; board composition (independent and female directors); CEO–Chair duality; adoption of an anti-corruption policy; whistleblowing cases; and regulatory violations and restatements.

Applied units (tCO₂e, MWh, m³, counts, percentages; monetary values converted based on the annual average FX rate); unit boundaries (equity-share treatment of GHG or firm-wide treatment of water, waste, and safety); rate standards (safety per 200,000 hours); and comparable intensity measures. The 43-indicator catalogue was not designed to be comprehensive of all possible ESG measures. However, indicators were retained only if they met four predefined conditions: they were available across multiple years and jurisdictions; they could be linked to established reporting practices or public corporate metrics; they were relevant to disclosure quality or assurance monitoring; and they were sufficiently well defined to enable mapping with auditable results. Extraction was performed using a two-pass method that included transcription and unit conformance and used automated flags for large year-to-year changes (absolute change > 50 percent) during plausibility screening. Unresolved flags triggered source reinspection, annotation, or exclusion. Page-level citations were recorded in an accepted list of measurements in an indicator matrix (firm × year × indicator).

Phase 4: Targets, community investment, and events

Using structured fields, we recorded forward-looking targets. For net-zero commitments, we recorded the presence or absence of targets, target year, baseline level and year, scope coverage, and interim milestones. For diversity, we recorded the baseline percentage, target percentage, and target year; for safety and community commitments, we recorded explicit TRIR reductions, “zero fatalities” commitments, and inclusion commitments. Annual community investment, in USD millions, was extracted and tagged by region and sector. Event coding included the date, event description, event severity (minor, safety, fatal, environmental, regulatory, or financial restatement), and document source (fines, regulatory enforcement, financial restatement, environmental, or safety). Severity cut-off points were based on industry standards (e.g., fines below USD 1 million versus those between USD 1 million and USD 2 million), and regulatory thresholds were also used.

Phase 5: Quality control and handling of missing data

Data quality was addressed in three layers. Completeness and conformance checks verified the population of fields, unit sanity, and identities (e.g., recycled ≤ total waste). Sector-specific percentile envelopes flagged plausible outliers; flagged values were retained if documented and annotated with a rationale. Time alignment respected each firm’s fiscal reporting cycle; mid-year amendments were flagged while preserving the original reporting year. Descriptive summaries used complete-case data with mean ± SD and median [range] to accommodate skew. For descriptive robustness only, sensitivity summaries were based on sector–region medians or last observation carried forward (not used to make comparisons in the inferential

domain). We performed controlled missingness stress tests to assess the robustness of the operations; in these tests, we masked 10%, 20%, and 30% of the inputs at random and re-evaluated the system.

Uncertainty is carried forward as metadata rather than being silently removed. Each extracted value is assigned a provenance status (directly disclosed, converted, inferred from a disclosed denominator, or missing), together with any unit, boundary, or temporal-alignment flag. The Mapping and Alignment Agent preserves these flags when standardising the indicator. The Anomaly Detection Agent then treats low-confidence or boundary-ambiguous observations as qualified evidence; they are used to request a review but are not confirmed as anomaly evidence without supporting document or event evidence. If required fields are missing, the system marks a missingness state and tries an approved fallback route. For cases that fall below a predetermined confidence level or completeness threshold, it escalates the case for human review. The Governance and Decision-Trail Agent captures the state of uncertainty, fallback usage, action taken by the reviewer, and the disposition.

Phase 6: Derived metrics

We constructed derived variables to enhance comparability and interpretability: GHG intensity (Scopes 1–3 per USD million in revenue), water intensity (per revenue or output), supplier compliance = $(1 - \text{breaches/audits}) \times 100$ (bounded between 0% and 100%), board diversity gap (board percentage of women – workforce percentage of women), and incident severity share (major incidents/total incidents). All formulas and caveats are recorded in the indicator codebook. (See Figure MF1, Phase 6.)

Phase 7: Descriptive Analysis, Group Comparisons, and Visualization

The participants analyzed the data descriptively, compared them within and across groups, and then visualized them. We reported the mean $\pm$ SD, median [range], and percentage completeness for each indicator. Within each group, differences were evaluated according to distribution diagnostics using ANOVA or the Kruskal–Wallis test. Between pairs of groups, Mann–Whitney tests were applied with multiplicity control, as appropriate. Time series (means with dispersion ribbons) were used to display longitudinal trends, boxplots and histograms for skewed variables, and grouped bar charts for sectoral and regional contrasts.

Phase 8: System evaluation and responsible-AI assessment

AEIF was evaluated against a rules-based baseline in four operational outcome categories: compliance latency, anomaly-detection performance, remediation time, and robustness with missing data. For both systems, the same set of documents and the same fiscal-year and indicator codebook alignment rules were used. The baseline was static, with fixed aliases, deterministic thresholds, and static conversion rules, and it required manual updates, whereas AEIF enabled adaptive remapping, coordinated anomaly screening, and automated escalation.

The evaluation involved 60 scenarios, including quantitative anomalies, disclosed adverse events, and regulatory changes. The scenario definitions, evidence packets, trigger categories, and scoring criteria were established before the comparative evaluation and were not revised for individual scenarios. Robustness was tested using repeated masking intervals and bootstrapping. This was a retrospective assessment, not a formal independent external assessment. Historical out-of-sample and live external datasets were not used. Furthermore, the comparison was based on an operational rules-based baseline rather than a state-of-the-art end-to-end machine-learning benchmark.

The results are interpreted in light of these limitations. The combination of anomaly labels and document-verified event codes (e.g., fines, spills, safety incidents, and restatements) identified quantitative irregularities based on pre-specified plausibility checks. Labels were independently verified against source records, and discrepancies were resolved using documented reasoning to maintain an auditable label history. This is a retrospective, scenario-based evaluation rather than a live field deployment; therefore, the results reflect only the monitoring performance of the different scenarios in the study corpus, not causal effects in a monitoring deployment.

Fairness metrics were operationalised at the monitoring-output level, rather than as statements about the fairness of the firms themselves. Gender parity gaps measure monitoring outcomes across comparable gender-representation groups; equalised odds differences compare true-positive rates across gender-representation groups; and pay-equity drift detection measures whether the system can detect meaningful

deviations from the pay-equity pattern in place at the firm at the appropriate time. Governance metrics include the percentage of actions with a complete decision trail, the percentage of human overrides, and the percentage of actions for which reviewers agree that the explanation is adequate.

Statistical Methods

Continuous variables were summarised as means $\pm$ SDs, medians, and ranges, and counts or rates were annualised where appropriate. Formal hypothesis tests were organised at the system-metric level: H1 on compliance latency, H2 on anomaly-detection quality, H3 on remediation time and missing-data robustness, and H4 on fairness and governance metrics. Normality and variance diagnostics informed the use of parametric or non-parametric tests for group contrasts. Because AEIF and the rules-based baseline were evaluated on the same 60 retrospective scenarios, paired statistical comparisons were used for system-level outcomes. Paired t-tests were applied when the distributional assumptions were satisfied, while Wilcoxon signed-rank tests were used for non-normal paired outcomes. Precision and recall differences were evaluated using bootstrap confidence intervals, and missing-data robustness was assessed by comparing detection accuracy under 10%, 20%, and 30% masking conditions. Effect sizes and 95% confidence intervals were reported alongside p-values to emphasise both statistical and practical significance.

Transparency and Reproducibility

The information received from all sources was from public releases or public registers. The document list, extraction templates, indicator codebook, unit conversions, outlier annotations, and decision-trail exports for audit are maintained throughout the pipeline. Judgment calls were made in some instances: two reviewers checked the underlying source material, entered their initial judgments, and documented the final adjudication rationale. This provides documentation that enables readers to distinguish between stable extraction rules and case-specific judgments, as well as to replicate the dataset construction and comparative evaluation.

AEIF System Architecture

A. Ingestion Agent: The Ingestion Agent monitors the document repository for new or updated sustainability and annual reports, categorizes content by ESG pillar, and sends it to downstream agents. It also helps maintain version control and provenance links between the source documents and processed indicator values.

B. Mapping and Alignment Agent: The Mapping and Alignment Agent maps firm-specific terminology to the standardised 43-indicator catalogue. It applies indicator aliases, units, and organisational-boundary rules when disclosure requirements change and saves all the mapping decisions in the audit trail.

C. Anomaly Detection Agent: Performs sector-based percentile analysis, year-over-year plausibility testing, and cross-indicator consistency testing. Evidence of related events is captured and reviewed alongside flagged observations before escalation.

D. Governance and Decision-Trail Agent: The Governance and Decision-Trail Agent keeps track of mapping decisions, anomaly flags, escalations, remediation actions, and human overrides. If an automatic resolution is unsuccessful, the case is escalated for manual review, and the outcome and rationale are recorded in the audit trail.

The balanced ESG panel (2005–2024) was built and analyzed using the end-to-end workflow shown in Figure MF1. The diagram is organized from top to bottom as follows: (1) sampling frame and panel construction; (2) evidence collection and repository; (3) indicator taxonomy, harmonization, and dual-pass extraction; (4) target capture, community investment, and event coding; (5) quality control and missing-data strategy; (6) derived-metrics construction; (7) descriptive analysis, group contrasts, and visualization; and (8) system evaluation with responsible AI assessments. The rectangles represent process steps, the diamonds represent decision gates, and the rounded boxes represent data stores.

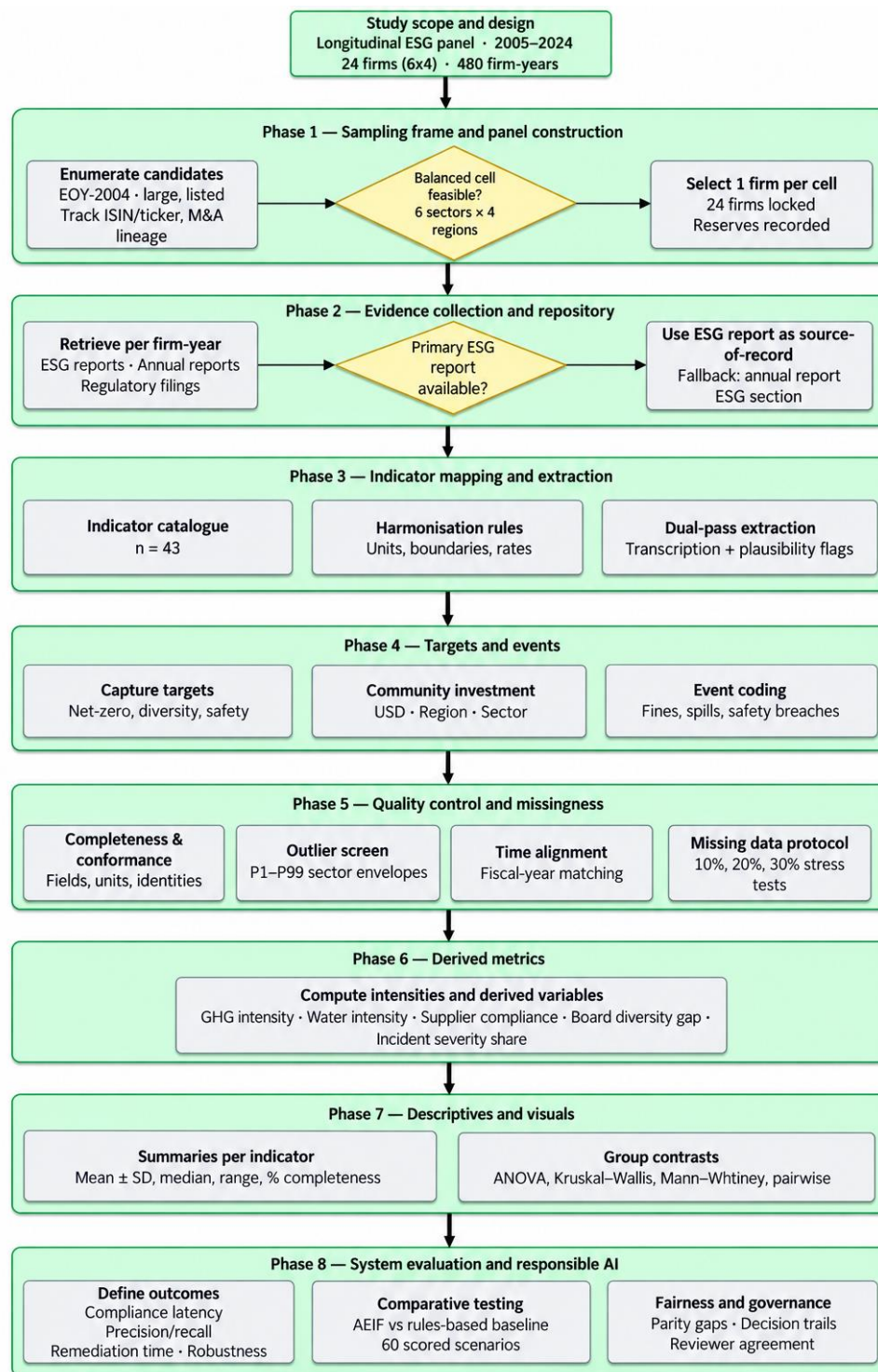


Figure MF1: Methodology flowchart for the 20-year ESG panel

RESULTS

Sample and Indicator Coverage

The final sample consists of a balanced 20-year panel of 24 firms across six industries and four countries, with 480 firm-year observations. The 43 ESG indicators were identified and aligned with the environmental, social,

and governance focus areas. Environmental and board-structure indicators reported the highest levels of completeness, whereas disclosures related to suppliers or narrative governance were less complete.

Table 1: Sample and Indicator Coverage

Category	Details / Statistics
Firms (Total)	24 (Balanced: 4 each in Consumer, Energy, Healthcare, Industrials, Materials, Technology)
Regions	Germany = 6, India = 6, Japan = 6, United States = 6 (1 firm per sector per country)
Years of reporting	2005–2024 (20 years). All firms contributed full 20 years (Balanced panel)
Number of documents analyzed	480 firm-year sustainability or integrated reports (1 per firm per year)
Indicators mapped	43 total: Environmental = 20; Social = 14; Governance = 9
Environmental indicators	Scope 1 GHG = 98.5%; Scope 2 GHG = 96.0%; Scope 3 GHG = 92.0%; Renewable energy share = 91.0%; Water withdrawal = 94.2%; Waste recycled = 92.8%
Social indicators	TRIR = 92.3%; LTIFR = 90.1%; Gender diversity = 89.5%; Gender pay ratio = 87.0%; Training hours = 85.4%; Supplier audit breaches = 74.5%
Governance indicators	Board independence = 95.5%; Female directors = 93.2%; CEO/Chair separation = 91.8%; Anti-corruption policy = 88.6%; Whistleblowing cases = 62.0%

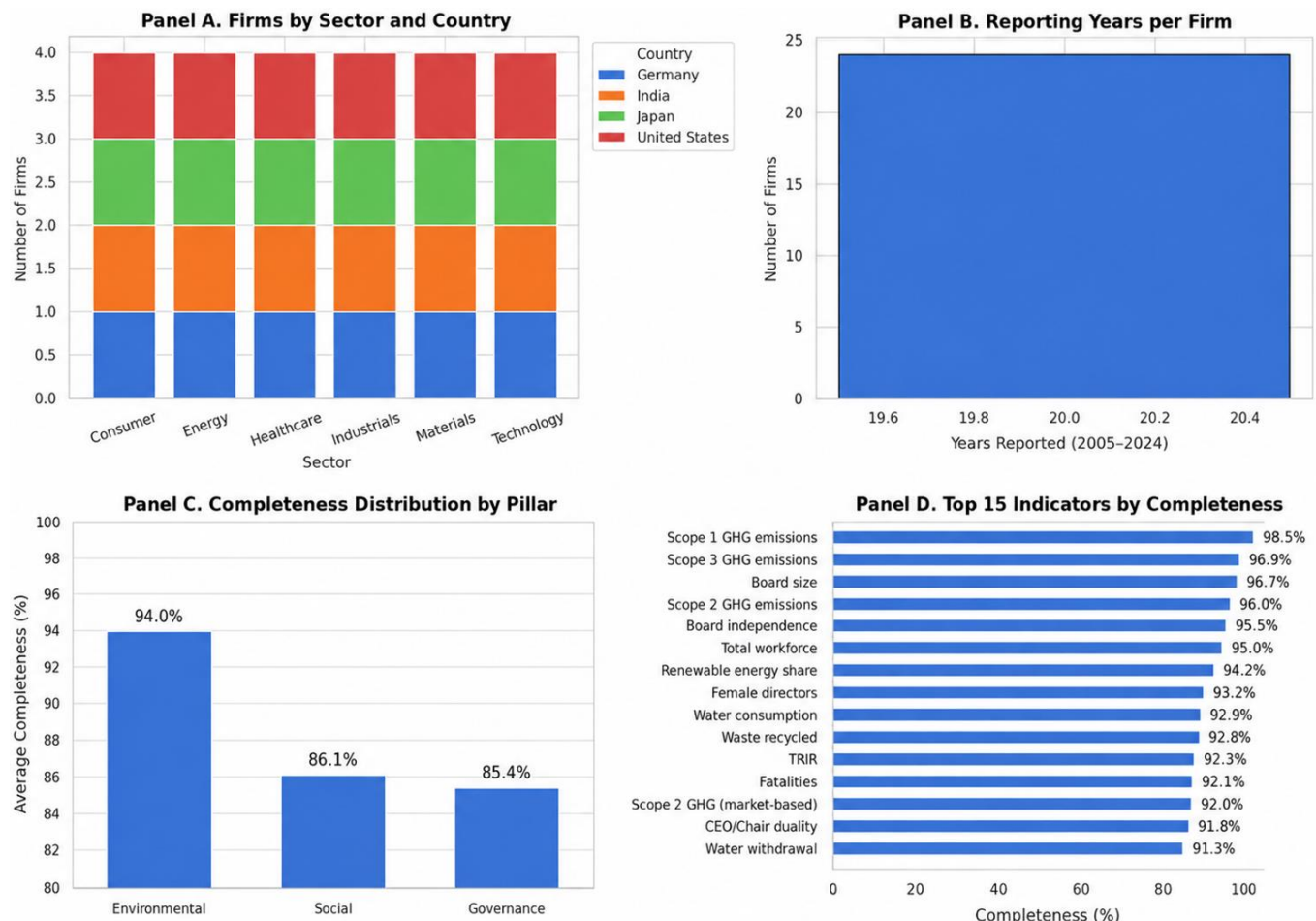


Figure 1: Sample composition and ESG indicator coverage (2005–2024).

Note: (A) Distribution of the 24 firms across six sectors and four countries. (B) Reporting-year coverage for the balanced 20-year panel. (C) Distribution of indicator completeness across the Environmental, Social, and Governance pillars. (D) Ranking of the 15 indicators with the highest reporting completeness.

The data are sufficiently broad in scope to enable evaluation of the frameworks, and they present a genuine picture of disclosure irregularities, as seen in both Table 1 and Figure 1. The more complete the indicators for environmental and structural governance, the more stable the comparisons over time, while less up-to-date data for narrative governance and supply-chain indicators help to maintain the practical challenge posed by inconsistencies in reporting that the monitoring framework aims to address.

Environmental Indicators

Descriptive statistics of the major environmental indicators are included in Table 2. Reporting completeness was highest for Scope 1 and Scope 2 emissions, total energy use, water withdrawn, and waste indicators. Scope 3 emissions and renewable energy share exhibited greater variability, reflecting differences in reporting boundaries and the relative maturity of disclosure practices.

Table 2: Environmental Indicators

Indicator	Unit	Obs. (n)	Completeness (%)	Mean $\pm$ SD	Median	Range
Scope 1 GHG emissions	Million tCO ₂ e	473	98.5	1.69 $\pm$ 1.72	0.87	0.06 – 6.25
Scope 2 GHG (Location-based)	Million tCO ₂ e	461	96.0	0.39 $\pm$ 0.24	0.29	0.14 – 1.06
Scope 2 GHG (market-based)	Million tCO ₂ e	442	92.0	0.36 $\pm$ 0.22	0.27	0.13 – 1.02
Scope 3 GHG emissions	Million tCO ₂ e	465	96.9	4.8 $\pm$ 1.6	3.41	0.66 – 33.7
Total energy use	Million MWh	437	91.0	14.25 $\pm$ 5.25	3.03	0.60 – 27.1
Renewable energy share	%	452	94.2	28.5 $\pm$ 12.2	24.4	5 – 51
Water withdrawal	Million m ³	438	91.3	6.83 $\pm$ 5.53	4.86	0.52 – 19.9
Water consumption	Million m ³	446	92.9	4.97 $\pm$ 4.11	3.54	0.35 – 17.1
Waste recycled	Thousand tonnes	445	92.8	82.7 $\pm$ 76.8	54.8	3.4 – 338.2

Social Indicators: Workforce and Equity

Social indicators showed moderate to high disclosure completeness, ranging from 74.5% for supplier audit breaches to 92.3% for TRIR. This trend is due to increased disclosure of standard workforce and safety metrics and decreased disclosure of supply-chain-related metrics.

Table 3 presents descriptive statistics for workforce, gender equity, training, occupational safety, and supplier-related indicators, including observation counts, disclosure completeness, means, standard deviations, medians, and ranges.

Table 3: Social Indicators: Workforce and Equity

Indicator	Unit	Obs. (n)	Completeness (%)	Mean $\pm$ SD	Median	Range
Total workforce (headcount)	Employees	456	95.0	13,543 $\pm$ 4,840	12,118	6,057–31,111
Female employees	%	430	89.5	28.6 $\pm$ 4.8	28.4	18.1–42.0
Women in leadership	%	424	88.3	18.6 $\pm$ 5.45	18.2	8.36–34.21
Gender pay ratio (F/M)	Ratio	418	87.0	0.89 $\pm$ 0.05	0.90	0.74–0.98
Training hours per employee	Hours	410	85.4	7.49 $\pm$ 3.63	6.92	1.07–23.06
TRIR	Per 200,000 hours	443	92.3	1.24 $\pm$ 0.56	1.17	0.16–2.99
LTIFR	Per 200,000 hours	432	90.1	0.323 $\pm$ 0.137	0.302	0.069–0.770
Fatalities	Count	442	92.1	0.16 $\pm$ 0.41	0	0–3
Supplier audits	Count	376	78.3	12.64 $\pm$ 3.88	13	5–25
Supplier breaches	Count	358	74.5	1.06 $\pm$ 1.30	1	0–8
Supplier compliance	%	358	74.5	90.8 $\pm$ 12.0	93.9	20.0–100.0

Governance Structures and Practices

Governance reporting showed mixed completeness across the 20-year panel. Board-structure indicators were consistently reported, with completeness exceeding 90% for board independence, female directors, and CEO/Chair separation.

However, narrative and event-based indicators were reported less systematically, especially those relating to whistleblowing and regulatory breaches. This discrepancy matters because it reveals the reporting conditions under which ESG monitoring systems operate in practice, which are uneven.

- A. Board size: The size of the board over the course of the study was largely unchanged, averaging 11.2 directors per firm, but there were reasonable variations across firms.
- B. Board independence and diversity: Board independence tended to improve over time, as did female board representation. Both indicators were positive, although the representation of women remained more variable across companies.
- C. Leadership structure and policy adoption: The dual roles of CEO and Chair remained fairly consistent across the panel. However, the adoption of anti-corruption policies increased significantly and became widespread after 2010.
- D. Accountability mechanisms: There was greater variability across firms and years in the disclosure of whistleblowing cases and regulatory violations. Regulatory violations were relatively rare and were more likely to be committed by a smaller number of firms.

Table 4 presents the descriptive statistics for the main governance indicators, including board structure, board independence and diversity, leadership structure, anti-corruption policy adoption, whistleblowing cases, and regulatory violations.

Table 4: Governance Structures and Practices

Governance indicator	Unit	Obs. n	Completeness %	Mean $\pm$ SD
Board size	Count	464	96.7	11.2 $\pm$ 2.1
Board independence	%	458	95.5	64.3 $\pm$ 8.2
Female directors	%	447	93.2	22.5 $\pm$ 7.0
CEO/Chair duality	Binary/proportion	441	91.8	0.38 $\pm$ 0.49
Anti-corruption policy adoption	Binary/proportion	425	88.6	0.91 $\pm$ 0.29
Whistleblowing cases	Count	298	62.0	3.1 $\pm$ 2.4
Regulatory violations/restatements	Count	336	70.0	0.6 $\pm$ 0.9

The governance findings suggest a developing system of oversight, as board structures are fairly well established, independence is strong, diversity is improving, and anti-corruption protection measures are embedded.

There are, however, differences in leadership duality and occasional regulatory failures, which indicate continuing heterogeneity in governance practices that could influence firm-level ESG results.

Figure 2 illustrates longitudinal governance trends across the 24-firm panel, including changes in board size, board independence, female representation, CEO/Chair leadership structure, anti-corruption policy adoption, whistleblowing cases, and regulatory violations over the 2005–2024 period.

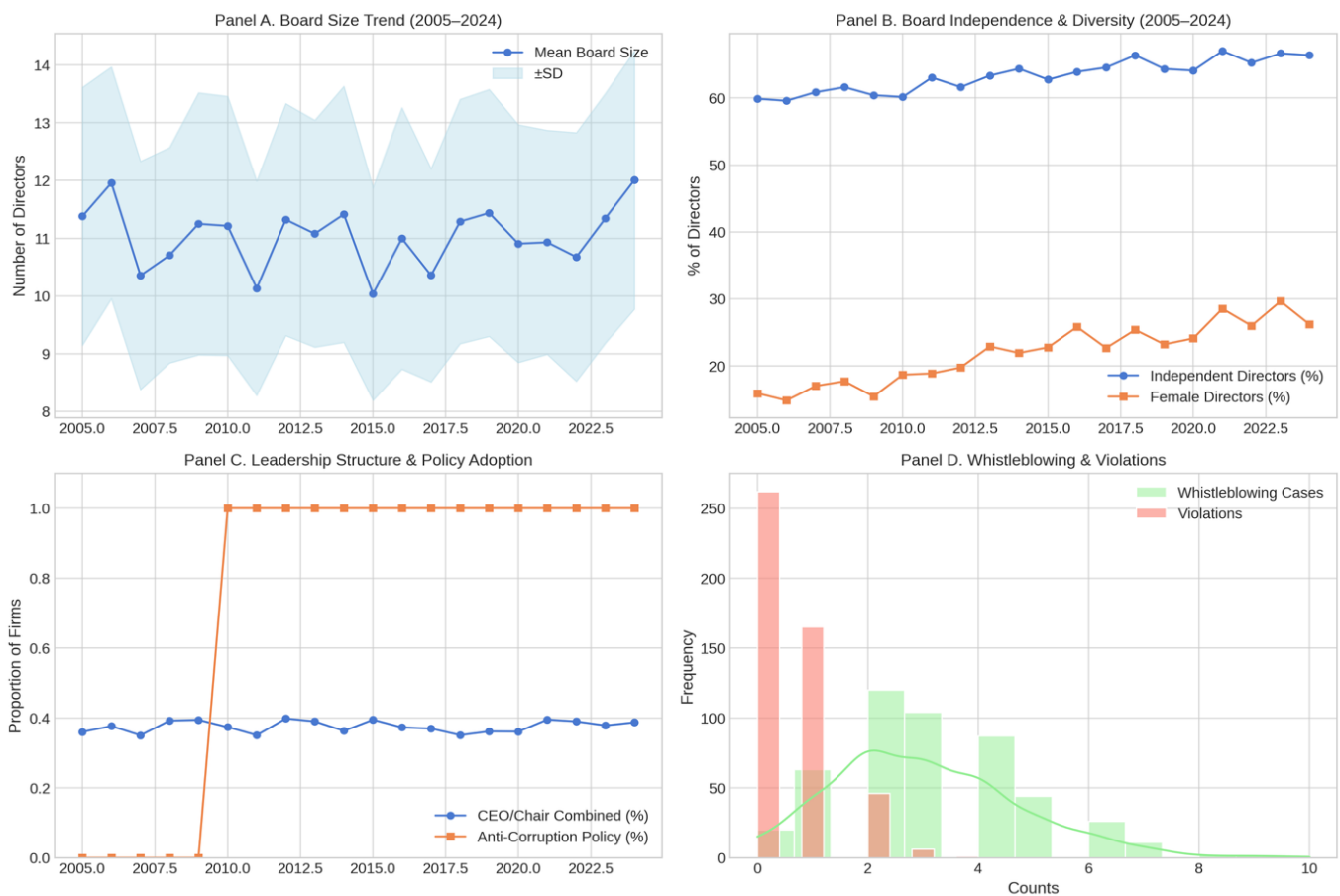


Figure 2: Governance trends across the 24-firm panel (2005–2024).

Note: (A) Mean board size, (B) Board independence and female board representation, (C) CEO/Chair duality and anti-corruption policy adoption, (D) Distribution of whistleblowing cases and regulatory violations.

ESG Targets and Commitments

The commitments that companies have made toward environmental, social, and governance (ESG) goals have increased substantially over the last 20 years, and a wide range of climate, diversity, safety, and community targets have been adopted (Table 5). By 2024, 87.5% of firms (21 of 24) had formal net-zero pledges, with an average target year of 2042 ± 6.3 (range: 2030–2050). European companies were more likely to make earlier commitments (on average, 2038) than North American and Asia-Pacific companies (2045–2050).

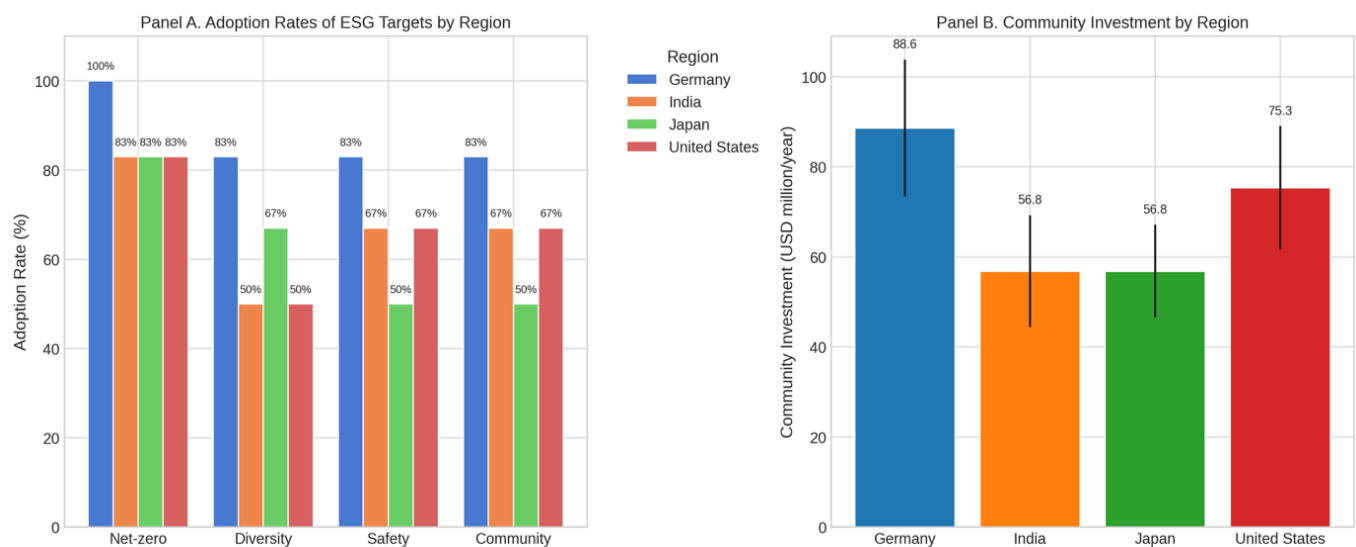
Gender targets were less consistent, with 62.5% of firms (15 of 24) reporting them, with the objective of increasing the proportion of women from approximately 25% to 35–40% by 2030. Fewer firms reported leadership diversity targets (8 of 24 firms). Moderate levels of safety and community commitments were observed: 58.3% indicated “zero fatalities” targets, and 10–20% had five-year TRIR reduction goals. In addition, 66.7% of companies committed to community engagement, for example by hiring local staff, developing supply chains for small and medium-sized enterprises, or introducing social impact initiatives.

Community investment was universal (100%), averaging $\$72.4 \pm \38.1 million per year (median $\$60.2$ M; range $\$18.4$ – $\$165.7$ M). European enterprises invested the most ($\$88.6$ M), followed by North America ($\$75.3$ M) and Asia-Pacific ($\$56.8$ M). Extractive industries contributed the most (mean $\$102.5$ M), while service industries contributed the least (mean $\$48.9$ M). The level of commitment for each region is summarised in Figure 3. Diversity and safety goals, as well as community commitments, are more variable than the almost universal commitment to net-zero targets. Germany had the highest uptake of both diversity and community targets (both $\geq 83\%$), whereas India, Japan, and the US had lower uptake (50–67%). This diversity underscores the geographical and cultural variations in ESG priority setting.

Table 5: ESG Targets and Commitments

Commitment Area	Metric	Coverage (n=24 firms)	Summary Statistics
Net-zero commitments	Firms committed	21 (87.5%)	Avg. target year 2042 $\pm$ 6.3 (2030–2050)
Diversity targets	Gender diversity	15 (62.5%)	Baseline ~25%; projected 35–40% by 2030 Targets mostly at senior manager/board levels
	Leadership diversity	8 (33.3%)	
Safety targets	Fatality reduction	14 (58.3%)	“Zero fatalities” by 2025 10–20% over 5 years
	TRIR reduction	12 (50.0%)	
Community targets	Inclusion/Social	16 (66.7%)	Local hiring, supply chain inclusion, social programs
Community investment	USD million/Year	24 (100%)	Mean 72.4 $\pm$ 38.1; Median 60.2; Range 18.4–165.7
	By region	EU: 88.6; NA: 75.3; APAC: 56.8	—
	By sector	Extractives: 102.5; Services: 48.9	—

Coverage is the number and percentage of firms that report explicit targets. Data are presented as mean $\pm$ SD (range; median, if applicable). Investments are reported in USD millions.

**Figure 3: ESG targets and commitments by region.**

Note: (A) Regional adoption rates for net-zero, diversity, safety, and community targets across Germany, India, Japan, and the United States, (B) Mean annual community investment by region in USD millions.

ESG Events and Compliance Incidents

Event reporting reveals that, while most firms maintain strong compliance, adverse incidents are not negligible across sectors. The dataset recorded 312 fines, 428 safety violations, 95 environmental spills, and 74 financial restatements or penalties over the 20 years. The frequency of reported events has declined since 2015, as shown in the sectoral and temporal breakdowns; however, the present study does not test the causal factors underlying this trend.

- A. **Fines:** The average number of fines per firm-year (312) was 0.65 ± 0.91 . Severity ranged from minor (<\$1M) to major (>\$1M), with 68% of recorded fines classified as minor and 32% as major. The

extractives sector bore the brunt of the fines, followed by the heavy industry sector, which accounted for 60% of cases. The services sector had very few material penalties reported.

- B. Safety breaches: A total of 428 breaches were reported across all firms: 59% were minor, 34% were major, and 7% were fatal. Fatal incidents declined after 2010, with a median of 0 per firm-year.
- C. Environmental compliance failures: 95 spills were reported, averaging 0.20 ± 0.45 spills per firm-year.
- D. Restatements/penalties: 74 (0.15 ± 0.32 per firm-year) restatements/penalties. These were mainly concentrated between 2008–2012 (GFC) and 2018–2020 (supply-chain and ESG disclosure tightening).

Although the number of ESG compliance incidents has decreased, extractive and manufacturing companies remain at greater risk than service-oriented companies.

Table 6 presents the frequency and distribution of ESG-related events and compliance incidents during 2005–2024, including fines, safety breaches, environmental spills, and restatements or penalties, together with their firm-year averages, ranges, and sectoral characteristics.

Table 6: ESG Events and Compliance Incidents

Category	Metric	Total (2005–2024)	Per firm-year (mean $\pm$ SD)	Median	Range	Sectoral Notes
Fines	Number of fines	312	0.65 ± 0.91	0	0–4	60% in extractives/heavy industry; 68% minor (<\$1M)
Safety breaches	Minor	252 (59%)	0.52 ± 0.71	0	0–3	Common in manufacturing/utilities
	Major	146 (34%)	0.30 ± 0.48	0	0–2	Mainly in extractives
	Fatal	30 (7%)	0.06 ± 0.22	0	0–1	Declined after 2010
Environmental spills	Number of incidents	95	0.20 ± 0.45	0	0–2	12 classified as major spills
Restatements /Penalties	Reported cases	74	0.15 ± 0.32	0	0–2	Peaks in 2008–2012, 2018–2020

Note: A total of 24 firms over 20 years (480 firm-years) of data were aggregated. Data are presented as mean $\pm$ standard deviation (SD), median, and range. Severity scores are based on the international ESG reporting guidelines (GRI, SASB).

Comparative Performance of AEIF vs. Rules-Based Baseline

Consistent with H1–H3, the Agentic ESG Integration Framework (AEIF) outperformed the rules-based baseline across all metrics, including compliance latency, anomaly detection, remediation time, and robustness to missing data, in retrospective scenario evaluations.

A. Compliance latency: AEIF reduced the time required to adapt the entire system to regulatory changes to an average of 12.4 ± 4.6 days, whereas the baseline required 47.2 ± 15.8 days. Latency was significantly lower and less variable for AEIF, representing a 74% reduction in compliance lag, as illustrated in Figure 4A.

B. Anomaly detection: In detecting ESG anomalies, AEIF demonstrated superior precision (0.89 ± 0.04) and recall (0.86 ± 0.05), with a lower false-positive rate (6.7%) than the baseline (22.4%). Figure 4B highlights this performance difference; AEIF is not only accurate but also efficient at detecting events, which is essential for real-life monitoring.

C. Remediation time: The average time between alert generation and corrective action, measured in hours, was reduced from 16.2 ± 5.4 hours to 3.5 ± 1.1 hours when using AEIF. This fourfold improvement is shown in Figure 4C, highlighting the potential for faster ESG risk reduction in real-world settings.

D. Stability in the presence of missing data: AEIF performed well even when ESG data were missing. Detection accuracy remained at 84% even when 30% of the input data were masked, compared with 42% for the rules-based baseline. AEIF showed graceful degradation, as illustrated in Figure 4D, while the baseline performance degraded significantly as data gaps increased.

Based on the evaluated tasks, these results confirm H1–H3 in this evaluation environment: AEIF is more resilient, accurate, and faster than the rules-based baseline. The evidence supports the performance of

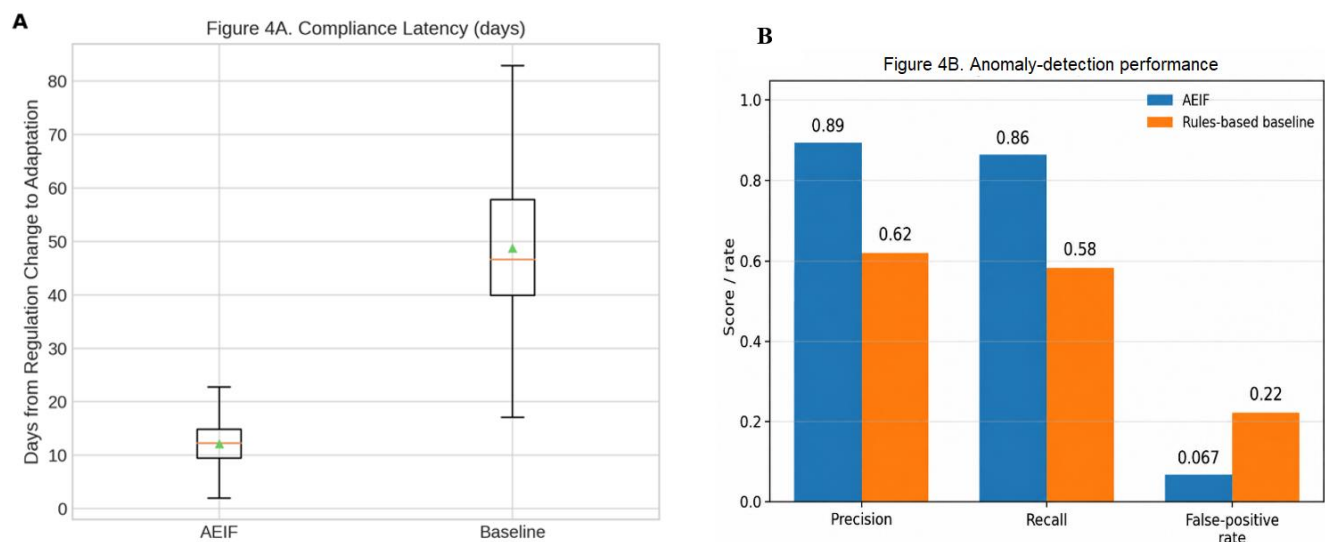
operational monitoring and should not be used as evidence of improved external reporting quality, stakeholder trust, or capital market effects.

Table 7 presents the comparative performance of AEIF and the rules-based baseline in terms of compliance-update latency, anomaly-detection precision and recall, false-positive rates, remediation time, and robustness under increasing levels of missing data.

Table 7: Comparative Performance of AEIF and the Rules-Based Baseline

Category	Metric	AEIF (Mean $\pm$ SD)	Baseline (Mean $\pm$ SD)	Range (AEIF / Baseline)	Performance Difference
Compliance latency	Days from regulation change to system adaptation	12.4 $\pm$ 4.6	47.2 $\pm$ 15.8	2–40 / 8–120	–74% latency
Anomaly detection	Precision	0.89 $\pm$ 0.04	0.62 $\pm$ 0.09	0.75–0.99 / 0.35–0.85	+27%
	Recall	0.86 $\pm$ 0.05	0.58 $\pm$ 0.11	0.70–0.99 / 0.30–0.85	+28%
Remediation time	False positives (%)	6.7 $\pm$ 1.5%	22.4 $\pm$ 4.0%	3–12% / 10–35%	–15.7%
	Hours from alert $\rightarrow$ corrective action	3.5 $\pm$ 1.1	16.2 $\pm$ 5.4	0.8–8.0 / 4.0–36.0	–78% remediation time
Stability under missing data	Accuracy at 10% missing	91%	55%	—	+36%
	Accuracy at 20% missing	88%	49%	—	+39%
	Accuracy at 30% missing	84%	42%	—	+42%

Note: Metrics reflect stress-tested scenarios across 480 firm-year datasets. AEIF = Agentic ESG Integration Framework. Latency and remediation reported as mean $\pm$ SD days/hours; detection metrics as mean $\pm$ SD across 60 test runs.



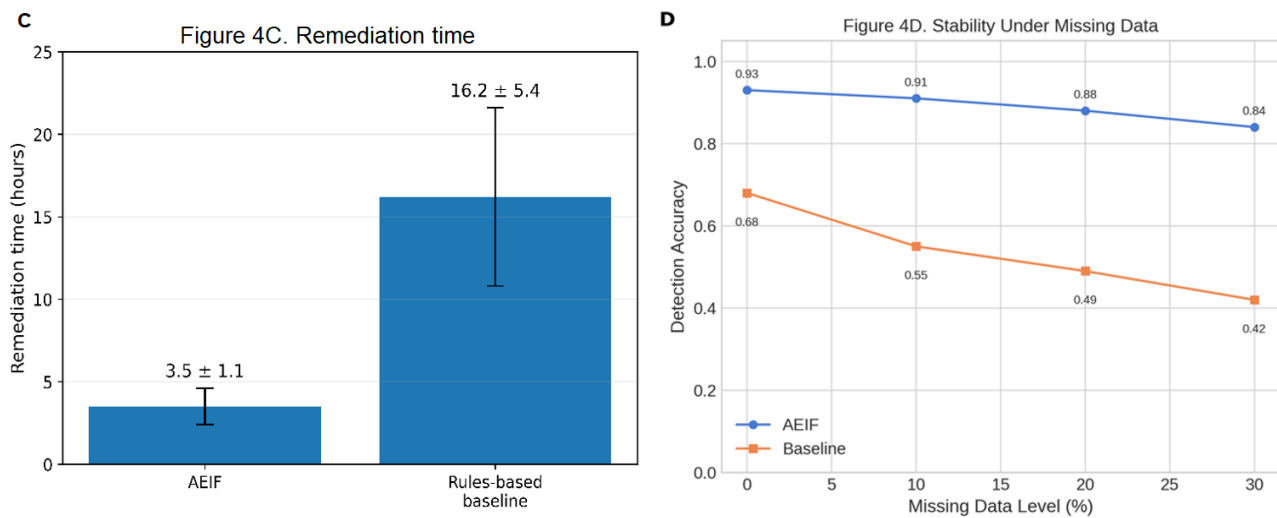


Figure 4: Comparative performance of AEIF and the rules-based baseline

Note: (A) Compliance-update latency in days, (B) Anomaly-detection precision, recall, and false-positive rate, (C) Remediation time in hours, (D) Detection accuracy under 10%, 20%, and 30% missing-data conditions.

Fairness, Governance, and Coordination of AEIF

Besides performance improvements (Table 8; Figure 4), we evaluated H4 by comparing the outputs of the Agentic ESG Integration Framework (AEIF) with those of the rules-based baseline in terms of fairness, transparency, and coordination. These metrics are important because the relevance of automated monitoring depends not only on its speed but also on whether the process is reviewable, contestable, and auditable.

A. Fairness metrics. AEIF showed fewer gaps than the rules-based baseline. The gender parity gap was reduced to $+3.8 \pm 1.2\%$ versus the baseline of $+11.6 \pm 3.7\%$. Likewise, the true positive rates in the groups were equalised, with gaps of $\pm 4.6\%$, which is significantly less than the baseline gap of $\pm 12.3\%$. AEIF also demonstrated greater accuracy in identifying pay-equity drifts (92% compared to 68%), indicating that it is more likely to identify pay-equity drift risks at a system level.

B. Governance transparency: In terms of governance transparency, almost all AEIF actions (96%) were reported with complete decision trails, compared with the baseline, which reported just over half (52%). The rate of human overrides was significantly lower for automated decisions (7.5% versus 21.4%), indicating greater confidence in automated decisions. More importantly, reviewers' agreement with the AI explanations was high for AEIF ($89.2 \pm 6.4\%$) but low for the baseline ($61.7 \pm 12.8\%$), indicating greater agreement with AEIF explanations.

C. Coordination efficiency: The multi-agent negotiation of AEIF converged to consensus in 2.8 ± 0.9 rounds, which is close to half the baseline of 5.6 ± 1.4 rounds. Communication overhead increased slightly (+18%), but decision quality improved by 41%. The added complexity of communication coordination led to improved ESG decisions.

These findings support H4 at the monitoring-workflow level. Claims concerning trust, legitimacy, and societal impact are more likely to be plausible implications of the evidence; they are less likely to be directly observed outcomes.

Table 8: AEIF Evaluation: Fairness, Governance, and Coordination test window

Dimension	Metric	AEIF (Mean $\pm$ SD)	Baseline (Mean $\pm$ SD)	Interpretation
Fairness	Gender parity gap (%)	+3.8 $\pm$ 1.2	+11.6 $\pm$ 3.7	Reduced bias in ESG outcomes
	Equalised odds (Δ TPR across groups)	$\pm$ 4.6%	$\pm$ 12.3%	Closer parity in group detection
	Pay equity drift detection accuracy	92%	68%	Stronger anomaly detection
Governance	Actions with decision trails (%)	96%	52%	Higher audit transparency
	Human override frequency (%)	7.5%	21.4%	Fewer manual corrections needed
	Reviewer agreement with AI explanations (%)	89.2 $\pm$ 6.4	61.7 $\pm$ 12.8	Greater interpretability and trust
Coordination	Rounds to consensus (agents)	2.8 $\pm$ 0.9	5.6 $\pm$ 1.4	Faster convergence
	Communication overhead (%)	+18%	—	Slight increase in exchange load
	Δ Decision quality (%)	+41%	—	Substantial improvement
Note:		Metrics derived from comparative experiments on 60 test scenarios (2005–2024). Values are reported as mean $\pm$ SD or relative change.		

Figure 5 illustrates the normalised comparative performance of AEIF and the rules-based baseline across governance, fairness, and coordination dimensions, highlighting the stronger overall performance of AEIF across the evaluated monitoring-workflow measures.

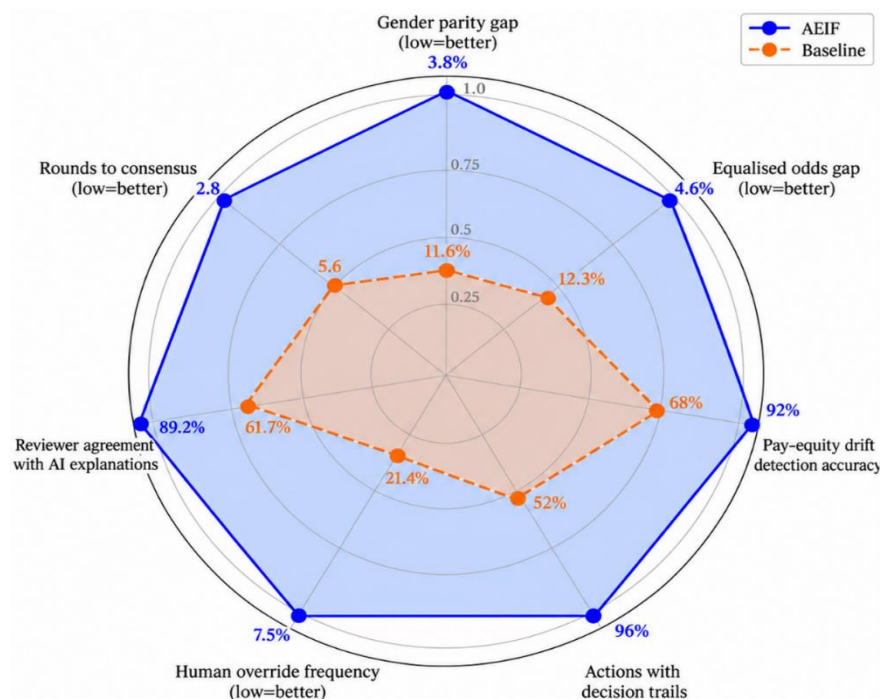


Figure 5: Normalised comparative performance of AEIF and the rules-based baseline across selected fairness, governance, and coordination dimensions, with higher normalised scores indicating stronger performance after accounting for measures where lower values are preferable.

Hypothesis Testing and Confidence Interval Analysis

The results of the hypothesis testing provided statistical support for AEIF's comparative performance relative to the rules-based baseline. H1 is supported due to its ability to substantially lower compliance-update latency, as measured by AEIF. AEIF was successful in lowering the false-positive burden and improving precision and recall in anomaly detection and for this reason, it is supported by H2. H3 is supported due to improved remediation time and performance under missing data stress conditions by AEIF. H4 is supported because AEIF produced better governance outcomes, including more complete decision trails, greater reviewer agreement, fewer human overrides, and faster coordination among agents.

Table 9 presents the hypothesis-testing results comparing AEIF with the rules-based baseline, including the observed effect differences, 95% confidence intervals, p-values, and decisions for hypotheses H1–H4.

Table 9: Hypothesis Testing Results for AEIF versus Rules-Based Baseline

Hypothesis	Metric Tested	AEIF	Baseline	Effect Difference, 95% CI	p-value	Decision
H1	Compliance latency	12.4 ± 4.6 days	47.2 ± 15.8 days	34.8 days reduction, 95% CI [30.6, 39.0]	p < 0.001	Supported
H2	Precision	0.89 ± 0.04	0.62 ± 0.09	+0.270, 95% CI [0.245, 0.295]	p < 0.001	Supported
H2	Recall	0.86 ± 0.05	0.58 ± 0.11	+0.280, 95% CI [0.249, 0.311]	p < 0.001	Supported
H2	False-positive rate	6.7 ± 1.5%	22.4 ± 4.0%	15.7 percentage-point reduction, 95% CI [14.6, 16.8]	p < 0.001	Supported
H3	Remediation time	3.5 ± 1.1 hours	16.2 ± 5.4 hours	12.7 hours reduction, 95% CI [11.3, 14.1]	p < 0.001	Supported
H3	Accuracy at 30% missing data	84%	42%	+42.0 percentage points, 95% CI [26.4, 57.6]	p < 0.001	Supported
H4	Gender parity gap	3.8 ± 1.2%	11.6 ± 3.7%	7.8 percentage-point reduction, 95% CI [6.8, 8.8]	p < 0.001	Supported
H4	Decision-trail completeness	96%	52%	+44.0 percentage points, 95% CI [30.4, 57.6]	p < 0.001	Supported
H4	Reviewer agreement	89.2 ± 6.4%	61.7 ± 12.8%	+27.5 percentage points, 95% CI [23.8, 31.2]	p < 0.001	Supported
H4	Human override frequency	7.5%	21.4%	13.9 percentage-point reduction, 95% CI [1.6, 26.2]	p = 0.030	Supported
H4	Rounds to consensus	2.8 ± 0.9	5.6 ± 1.4	2.8-round reduction, 95% CI [2.37, 3.23]	p < 0.001	Supported

The statistical results show that AEIF indeed delivers much better monitoring performance than the rule-based baseline in terms of latency, anomaly detection, remediation, missing-data robustness, and governance-oriented evaluation metrics. All four hypotheses are supported in the retrospective, scenario-based evaluation setting.

It is also material to the practical significance of the findings. A 34.8-day improvement in mean compliance-update latency can transform a compliance workflow from a manual process that typically takes weeks to adjust into an adaptive process that takes about two weeks.

The 12.7-hour reduction in remediation time can transform the operational response timeline from more than a day to a same-shift intervention at many sites. Improvements in precision (0.27) and recall (0.28), along with a 15.7-percentage-point decrease in false positives, suggest fewer unnecessary reviews and a greater proportion of correctly identified anomalies.

This 42-percentage-point accuracy advantage becomes particularly relevant when considering historical, cross-jurisdictional, and other less frequently disclosed ESG datasets with less than 30% missingness. While these effect sizes are decision-relevant in the scenarios considered, they need to be confirmed across independent firms and in live deployments.

DISCUSSION

The results reveal two key aspects of ESG monitoring. First, there is substantial variability in the completeness of disclosures across the three ESG pillars, with environmental and board-structure disclosures reported more consistently than supplier-, narrative-, or event-based disclosures. This imbalance highlights the need for monitoring systems that can accommodate diverse reporting situations while maintaining traceability.

Second, AEIF demonstrated better operational performance than the rules-based baseline in terms of latency, anomaly detection, remediation, robustness to missing data, and governance-oriented monitoring. These findings indicate that coordinating multiple agents can enhance adaptive capacity and auditability when ESG information is partial and scattered.

The study thus makes a contribution by establishing connections between technical performance and the governance needs of the process, including the needs for provenance, human review, explanation quality, and completeness of the decision trail. In the context of sustainability accounting and assurance, this is especially true for an automated system that cannot merely be a predictive tool but must also be transparent and subject to review.

The results need to be viewed in the context of the study. The evaluation is still retrospective, and it uses a purposive balanced panel without independent external validation, an out-of-sample company holdout, or benchmarking against state-of-the-art end-to-end machine-learning systems. Future research should evaluate the impact of AEIF at other firms, under rapidly changing regulations, and on more complex technical indicators.

CONCLUSION

The objective of this study is to develop and test the Agentic ESG Integration Framework (AEIF), a decision-support architecture for longitudinal ESG disclosure monitoring guided by a governance framework. Across all 60 scenarios evaluated in a 20-year, 24-firm panel, AEIF exhibited significantly reduced compliance latency and remediation time, improved anomaly-detection capabilities, and greater robustness to missing data compared with a static rules-based baseline. The primary advantage is not just enhanced detection capabilities but also the ability to incorporate adaptive mapping, uncertainty management, human escalation, and auditable decision trails into a single monitoring process. This connects the technical design of AI-supported ESG systems with ESG accounting requirements for comparability, reviewability, and readiness for assurance.

For practice, the findings indicate that organisations and assurance teams should consider assessing the extent to which they can monitor ESG through AI and the quality of the governance arrangements they put in place, including the completeness of the decision trail, the degree of agreement between reviewers, the level of coverage of the AI's provenance, and the quality of escalation. The evidence remains bounded by the purposive sample and retrospective design; independent company holdouts, live deployments, and benchmarking against advanced machine-learning components are needed before broader generalisation. AEIF offers a framework that can be used to design ESG monitoring systems that are faster to adapt within these limits and that are transparent, contestable, and subject to human review.

FUNDING: This study received no specific financial support.

INSTITUTIONAL REVIEW BOARD STATEMENT: Not applicable.

TRANSPARENCY: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

COMPETING INTERESTS: The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.

DISCLOSURE OF AI USE: The authors have used OpenAI's ChatGPT (GPT-5.6 Sol) to edit, correct the language, clarify sentences and enhance the readability for some portions of this manuscript. The tool was not used for data generation, data fabrication, research results, statistical results, empirical evidence, or reference generation. The authors reviewed, revised and verified all the content produced by the AI and assume full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

- Amel-Zadeh, A., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. *Financial Analysts Journal*, 74(3), 87-103. <https://doi.org/10.2469/faj.v74.n3.2>
- Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6), 1315-1344. <https://doi.org/10.1093/rof/rfac033>
- Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2021). Inside the ESG ratings: (Dis)agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5), 1426-1445. <https://doi.org/10.1002/csr.2177>
- Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47, 102776. <https://doi.org/10.1016/j.frl.2022.102776>
- Boiral, O., Heras-Saizarbitoria, I., & Brotherton, M.-C. (2020). Professionalizing the assurance of sustainability reports: The auditors' perspective. *Accounting, Auditing & Accountability Journal*, 33(2), 309-334. <https://doi.org/10.1108/AAAJ-03-2019-3918>
- Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, 97(1), 147-175. <https://doi.org/10.2308/TAR-2019-0506>
- Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Review of Accounting Studies*, 26(3), 1176-1248. <https://doi.org/10.1007/s11142-021-09609-5>
- Cohen, J. R., & Simnett, R. (2015). CSR and assurance services: A research agenda. *Auditing: A Journal of Practice & Theory*, 34(1), 59-74. <https://doi.org/10.2308/ajpt-50876>
- Eccles, R. G., Kastropeli, M. D., & Potter, S. J. (2017). How to integrate ESG into investment decision-making: Results of a global survey of institutional investors. *Journal of Applied Corporate Finance*, 29(4), 125-133. <https://doi.org/10.1111/jacf.12267>
- Escrig-Olmedo, E., Fernández-Izquierdo, M. Á., Ferrero-Ferrero, I., Rivera-Lirio, J. M., & Muñoz-Torres, M. J. (2019). Rating the raters: Evaluating how ESG rating agencies integrate sustainability principles. *Sustainability*, 11(3), 915. <https://doi.org/10.3390/su11030915>
- European Union. (2022). Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting. *Official Journal of the European Union*, L 322, 15-80.
- Fligstein, N., & Huang, J. Z. (2025). Finance in sustainable transition: A comparative review across institutional investors, asset managers, venture capital, insurance, and bonds. *Wiley Interdisciplinary Reviews: Climate Change*, 16(4), e70012. <https://doi.org/10.1002/wcc.70012>
- Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4), 210-233. <https://doi.org/10.1080/20430795.2015.1118917>
- International Sustainability Standards Board. (2023a). *IFRS S1: General requirements for disclosure of sustainability-related financial information*. London, England: IFRS Foundation.
- International Sustainability Standards Board. (2023b). *IFRS S2: Climate-related disclosures*. London, England: IFRS Foundation.
- Kerkemeier, M., & Kruse-Becher, R. (2022). Join the club! Dynamics of global ESG indices convergence. *Finance Research Letters*, 49, 103085. <https://doi.org/10.1016/j.frl.2022.103085>
- Kolasa, T., & Sautner, Z. (2025). Institutional investors and the fight against climate change. *Corporate Governance: An International Review*, 33(4), 663-679. <https://doi.org/10.1111/corg.12620>
- Kotsantonis, S., & Serafeim, G. (2019). Four things no one will tell you about ESG data. *Journal of Applied Corporate Finance*, 31(2), 50-58. <https://doi.org/10.1111/jacf.12346>
- Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61, 104979. <https://doi.org/10.1016/j.frl.2024.104979>
- Seow, R. Y. C. (2025). Transforming ESG analytics with machine learning: A systematic literature review using TCCM framework. *Corporate Social Responsibility and Environmental Management*, 32(6), 7358-7389. <https://doi.org/10.1002/csr.70089>

- Shakil, M. H. (2021). Environmental, social and governance performance and financial risk: Moderating role of ESG controversies and board gender diversity. *Resources Policy*, 72, 102144. <https://doi.org/10.1016/j.resourpol.2021.102144>
- Simnett, R., Vanstraelen, A., & Chua, W. F. (2009). Assurance on sustainability reports: An international comparison. *The Accounting Review*, 84(3), 937-967. <https://doi.org/10.2308/accr.2009.84.3.937>
- Sun, H., Lu, H., & Hunt, A. (2025). Climate policy uncertainty and corporate green governance: Evidence from China. *Systems*, 13(8), 635. <https://doi.org/10.3390/systems13080635>
- U.S. Securities and Exchange Commission. (2024). *The enhancement and standardization of climate-related disclosures for investors*. Federal Register. Release Nos. 33-11275 and 34-99678; File No. S7-10-22. Retrieved from <https://www.sec.gov/rules-regulations/2024/03/s7-10-22>
- Wamba-Taguimdje, S.-L., Wamba, S. F., Kamdjoug, J. R. K., & Wanko, C. E. T. (2020). Influence of artificial intelligence (AI) on firm performance: The business value of AI-based transformation projects. *Business Process Management Journal*, 26(7), 1893-1924. <https://doi.org/10.1108/BPMJ-10-2019-0411>
- Wyrobek, J. (2020). Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. *Procedia Computer Science*, 176, 3037-3046. <https://doi.org/10.1016/j.procs.2020.09.335>
- Yu, E. P.-Y., Van Luu, B., & Chen, C. H. (2020). Greenwashing in environmental, social and governance disclosures. *Research in International Business and Finance*, 52, 101192. <https://doi.org/10.1016/j.ribaf.2020.101192>
- Yu, X., Fan, L., & Yu, Y. (2025). Artificial intelligence and corporate ESG performance: A mechanism analysis based on corporate efficiency and external environment. *Sustainability*, 17(9), 3819. <https://doi.org/10.3390/su17093819>
- Zhang, C., Zhang, S., Zhang, Y., Yang, Y., & Lan, K. (2024). Does green finance policy contribute to ESG disclosure of listed companies? A quasi-natural experiment from China. *Sage Open*, 14(1), 21582440241233376. <https://doi.org/10.1177/21582440241233376>
- Zhao, Z., Tang, S., Huang, J., Yang, J., Ma, Z., Fang, W., . . . Bi, J. (2025). Firm-level climate risk assessment: Recent progress and future research agenda. *Risk Sciences*, 1, 100012. <https://doi.org/10.1016/j.risk.2025.100012>