Journal of Economics, Finance and Accounting Studies (JEFAS)
ISSN: 2709-0809
DOI: 10.32996/jefas
Journal Homepage: www.al-kindipublisher.com/index.php/jefas

JEFAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| **RESEARCH ARTICLE**

Generative AI and U.S. Financial Reporting Integrity: Detecting Narrative Manipulation, Risk Disclosure Gaming, and Fraud Signals in 10-K Filings

Anika Anjum Pritty[1]✉, Md Ibrahim[2], A S M FAHIM[3], and Muhaimin Ul Zadid[4]

[1] *Murray State University, Murray, KY*
[2] *UNIVERSITY OF NEW HAVEN, BUSINESS ANALYTICS*
[3] *UNIVERSITY OF NEW HAVEN, FINANCE AND FINANCIAL ANALYTICS*
[4] *MS, University of New Haven, CT, USA*
Corresponding Author: Anika Anjum Pritty, **Email**: apritty@murraystate.edu

| ABSTRACT

U.S. capital markets rely on high-integrity corporate disclosure, yet the narrative portions of annual reports—particularly Management's Discussion and Analysis (MD&A) and Risk Factors—remain vulnerable to strategic language management and, increasingly, generative-AI-assisted drafting. Unlike traditional misstatement detection that focuses on accounting ratios, narrative manipulation can distort investor beliefs through selective emphasis, obfuscation, boilerplate recycling, and inflated forward-looking language while remaining difficult to audit at scale. This paper develops a conceptual–methodological framework to measure disclosure integrity and detect narrative manipulation in U.S. 10-K filings. We construct a transparent EDGAR pipeline that parses MD&A and Risk Factors and extracts auditable features capturing (i) financial-context tone and uncertainty, (ii) readability and complexity, (iii) risk-factor novelty versus boilerplate drift (year-to-year similarity), (iv) forward-looking intensity, and (v) "tone–fundamentals gaps" that flag potential inconsistency between narrative claims and contemporaneous performance signals. We benchmark interpretable statistical models (logit/elastic net) against flexible machine-learning specifications while preserving explainability, and we evaluate performance using governance-relevant outcomes including enforcement-linked misstatement events and restatement proxies, complemented by market-based reactions around filing dates. We further test whether narrative-risk signals strengthen in the period associated with broad diffusion of generative AI writing tools. The proposed framework contributes to accounting, finance, and regulatory technology by converting narrative disclosure risk into measurable, monitorable indicators. Findings are positioned to inform SEC disclosure oversight, audit planning, and issuer governance by enabling scalable detection of narrative manipulation and risk-disclosure gaming that may undermine investor protection and U.S. market integrity.

| KEYWORDS

10-K; MD&A; risk factors; disclosure integrity; textual analysis; narrative manipulation; generative AI; SEC enforcement; audit analytics; market integrity

| ARTICLE INFORMATION

ACCEPTED: 01 August 2024 **PUBLISHED:** 15 August 2024 **DOI:** 10.32996/jefas.2024.6.4.11

1. Introduction

U.S. capital markets depend on credible disclosure. While audited financial statements anchor comparability and enforcement, a large share of information that shapes investor beliefs sits in the *narrative* portions of annual reports—especially Management's Discussion and Analysis (MD&A) and Risk Factors. These sections frame performance, justify strategy, and communicate uncertainty. They also create an integrity challenge: narrative language can be strategically managed to influence perception without immediately triggering red flags in traditional accounting ratios or violating bright-line reporting rules. As a result, narrative manipulation can weaken market discipline, distort capital allocation, and undermine investor protection—outcomes that are central to U.S. economic resilience and market integrity.

Two structural trends increase the urgency of this problem. First, narrative disclosure has expanded in volume and complexity over time, raising the cost of human review by investors, auditors, and regulators. Boilerplate risk factors, legalistic phrasing, and dense forward-looking statements can reduce informativeness even when disclosures are technically compliant. Second, generative AI tools have lowered the cost of producing polished, internally consistent text at scale. This does not imply that AI causes misconduct; however, it increases the feasibility of "disclosure gaming"—the selective presentation, obfuscation, and re-packaging of risk narratives—because language can be rapidly drafted, iterated, and standardized across firms and years. In other words, as the production function of disclosure changes, so does the monitoring problem faced by regulators and market participants.

Existing detection approaches in accounting and finance are stronger for *numerical* manipulation than for narrative distortion. Well-established models focus on earnings management and misstatement signals using accruals and financial ratios (Dechow, Sloan and Sweeney, 1995; Beneish, 1999; Roychowdhury, 2006). By contrast, narrative manipulation often operates through tone, readability, uncertainty inflation, and selective emphasis—features that are harder to interpret and can be defended as style choices or legal caution. Yet a growing literature shows that textual analysis can reveal economically meaningful information in corporate filings, especially when financial-context dictionaries and careful measurement are used (Loughran and McDonald, 2011; Loughran and McDonald, 2016). This literature motivates a governance-relevant question: can narrative features be transformed into auditable indicators that predict integrity breakdowns, market reactions, and enforcement-linked outcomes?

This paper addresses that question by developing a conceptual–methodological framework for measuring **disclosure integrity** and detecting **narrative manipulation** in U.S. 10-K filings from 2010–2024. We focus on MD&A and Risk Factors because they are central to investor interpretation and frequently updated in response to risk conditions. Our framework operationalizes disclosure integrity as a multi-dimensional construct combining (i) financial-context tone and uncertainty, (ii) complexity and obfuscation, (iii) risk-factor novelty versus boilerplate drift (year-to-year similarity), (iv) forward-looking intensity, and (v) "tone–fundamentals gaps" that flag potential inconsistencies between narrative claims and contemporaneous performance signals. The empirical design is built to be scalable and auditable: it uses a transparent EDGAR extraction and section-parsing pipeline, interpretable feature definitions, and models that retain explainability even when incorporating modern machine-learning techniques.

The study makes three contributions. First, it extends misstatement and reporting-integrity research by shifting attention from purely numerical signals to measurable narrative distortion, offering a structured taxonomy of manipulation channels in MD&A and Risk Factors. Second, it contributes a replicable methodology for turning narrative text into monitoring indicators suitable for regulators, auditors, and risk teams, emphasizing interpretability and governance relevance. Third, it provides a policy-facing perspective on the emerging disclosure environment: as generative AI becomes a normal tool in corporate communications, market integrity will depend increasingly on the ability to detect manipulation patterns that scale faster than traditional review processes.

The remainder of the paper proceeds as follows. Section 2 reviews literature on textual analysis in accounting and finance, disclosure theory, and fraud/misstatement detection, highlighting why narrative risk is difficult to govern with conventional tools. Section 3 presents the conceptual framework and testable hypotheses on obfuscation, tone–fundamentals gaps, boilerplate drift, and the changing disclosure risk surface. Section 4 describes the dataset construction from EDGAR, feature engineering, modeling strategy, and validation design using enforcement-linked outcomes and market reactions. Section 5 reports empirical results and robustness analyses, including out-of-sample performance and industry heterogeneity. Section 6 discusses implications for SEC oversight, audit planning, and RegTech monitoring and concludes with limitations and future directions for disclosure governance in the age of generative AI.

2. Literature Review and Theoretical Foundations

2.1 Narrative disclosure as an information and governance problem

Corporate disclosure serves two functions: it transmits value-relevant information to investors and it disciplines managers through verifiability and accountability. While audited financial statements anchor the verification channel, narrative sections (MD&A and Risk Factors) are less tightly constrained and therefore more exposed to strategic presentation. This creates a governance gap: narratives can shape investor expectations through emphasis, framing, and ambiguity even when numerical statements remain within accounting rules. Traditional earnings-management research—focused on accrual manipulation, real activities management, and ratio-based red flags—has produced powerful detection tools but is primarily oriented to *numbers* rather than language (Dechow, Sloan and Sweeney, 1995; Beneish, 1999; Roychowdhury, 2006). As a result, integrity risk can migrate into narrative space when numerical manipulation becomes more detectable or costly.

2.2 Textual analysis in accounting and finance: foundations and measurement discipline

Textual analysis has become a central methodology for extracting information from corporate filings, earnings call transcripts, and news. A core insight is that language contains economically meaningful signals, but measurement must be domain-specific: generic sentiment tools often misclassify financial language, motivating finance-tailored dictionaries and careful validation. The Loughran–McDonald (LM) dictionaries and related work established that financial text requires specialized word lists (e.g., "liability" is not negative sentiment in a legal/financial context), and that properly constructed textual metrics predict market outcomes (Loughran and McDonald, 2011). Subsequent surveys highlight both the promise and the pitfalls of text methods—feature choice, context dependence, and researcher degrees of freedom—making transparent pipelines and pre-specified measures important for credibility (Li, 2010; Loughran and McDonald, 2016).

This literature also motivates the paper's emphasis on **auditability**. For market integrity applications, models must be explainable enough to support oversight and challenge. Dictionary-based measures (tone, uncertainty, litigiousness) and structured metrics (readability, length, similarity) are especially useful because they are interpretable and can be validated against known outcomes.

2.3 Obfuscation, readability, and complexity as risk signals

A major strand of research links disclosure complexity to information frictions. More complex or less readable reports can reduce investor understanding, increase processing costs, and create space for managerial obfuscation. Readability-based evidence shows that harder-to-read filings are associated with poorer market reactions and information asymmetry, consistent with the view that complexity can be strategic rather than purely technical (Li, 2008). Beyond readability, complexity can be operationalized through document length, lexical diversity, syntactic density, and abnormal changes in these measures over time. In the disclosure-integrity context, these features become plausible risk flags: when narratives become unusually complex relative to peers or relative to a firm's own history, the probability of strategic framing may rise.

2.4 Risk factors: boilerplate, novelty, and "risk disclosure gaming"

Risk Factor disclosures (Item 1A) are explicitly designed to inform investors about material uncertainties. However, the incentives around legal protection can lead to "boilerplate" expansion, where firms add generic risks without improving informativeness. This creates a measurement challenge: more words do not necessarily mean more information. Modern text-similarity methods and "novelty" measures offer a way to separate informative updates from copy-paste drift by comparing risk-factor language across time and against industry peers. The broader text-based industry and similarity literature supports the use of text distances to measure commonality, differentiation, and strategic positioning (Hoberg and Phillips, 2016). In this paper, that logic is applied to risk disclosures: **high similarity** may indicate boilerplate, while **selective novelty** can indicate strategic insertion of cautionary language (e.g., uncertainty inflation) around periods of weakness.

2.5 Fraud/misstatement detection and the role of language

A complementary stream studies whether language reveals deception or misstatement risk. While classical fraud detection relies on accounting variables and red-flag ratios, language can add incremental predictive power by capturing managerial intent, evasiveness, or selective emphasis. Evidence from fraud-focused textual studies suggests that linguistic features—such as excessive complexity, unusual certainty/uncertainty patterns, or abnormal narrative structure—can help discriminate between normal reporting and integrity breakdowns (Purda and Skillicorn, 2015). This motivates the paper's use of **enforcement-linked** outcomes (e.g., AAER-related events) and restatement proxies to validate whether narrative manipulation measures are not merely stylistic artifacts but map to real integrity risk.

2.6 Why generative AI changes the disclosure risk surface

The adoption of generative AI does not inherently imply misconduct; firms may use AI tools for editing, translation, and consistency. The governance concern is that GenAI reduces the marginal cost of producing polished narrative content at scale, potentially enabling faster iteration of "compliance-safe" language that is persuasive but less informative. In practical terms, this can amplify two well-known disclosure problems: (i) **obfuscation** (longer, more complex narratives that are harder to parse) and (ii) **boilerplate drift** (recycling and expanding generic risk language). This paper treats GenAI diffusion as a potential structural shift in disclosure production, tested empirically as a change in the strength or prevalence of narrative manipulation signals.

2.7 Research gap and positioning of this study

1. **Integration gap:** earnings-management models focus on numeric manipulation (Dechow, Sloan and Sweeney, 1995; Beneish, 1999; Roychowdhury, 2006), while text research often emphasizes prediction or market reactions without explicitly framing a disclosure-integrity governance system.
2. **Measurement gap:** many text approaches lack auditable structure or rely on opaque features that are harder to defend in regulatory or audit settings (Li, 2010; Loughran and McDonald, 2016).
3. **GenAI-era gap:** the disclosure literature has not yet fully operationalized how generative AI may change narrative risk at scale in standardized filings.

This paper addresses these gaps by developing a **conceptual–methodological disclosure integrity framework** for U.S. 10-K narratives (MD&A and Risk Factors), pairing auditable text measures (tone, uncertainty, complexity, novelty/similarity, tone–fundamentals gaps) with enforcement-linked validation outcomes and governance-relevant implications for the SEC, auditors, and RegTech monitoring.

3. Theory, Conceptual Model, and Hypotheses

3.1 Theoretical lens: strategic disclosure under monitoring constraints

Narrative disclosure is best understood as a *strategic communication problem* under asymmetric information. Managers possess private information about operational risk, demand conditions, and financing constraints, while investors and regulators infer that information from both numbers and language. The narrative components of the 10-K—MD&A and Risk Factors—are particularly important because they sit at the intersection of (i) managerial explanation and persuasion, (ii) legal risk management, and (iii) investor interpretation. Unlike financial statement line items, narratives can change meaning through framing, emphasis, and ambiguity. This flexibility is valuable for legitimate communication, but it also creates an opportunity set for "narrative gaming," especially when monitoring resources are limited and language production becomes cheaper.

The governance problem intensifies when narratives become more scalable and more standardized. If firms can generate longer, smoother, and more internally consistent narratives with less human effort, then the marginal cost of producing *plausible but low-information* disclosure falls. In that setting, disclosure quality becomes less about whether text exists and more about whether it *meaningfully constrains managerial discretion* and helps investors price risk. The practical implication is that market integrity increasingly depends on the ability to identify narrative patterns that are systematically associated with later integrity failures or investor harm.

3.2 Conceptual model: "narrative risk" as the product of four frictions

Instead of treating manipulation as one concept, we define **narrative risk** as the product of four frictions that can be measured and tested:

Friction 1 — Comprehension friction:
Narratives may become difficult to process due to excessive length, dense phrasing, and low readability. When comprehension costs rise, negative information can be effectively "hidden in plain sight," increasing the chance that weak fundamentals are not

fully incorporated into prices. Readability evidence supports the idea that complexity is not neutral and can relate to information quality (Li, 2008).

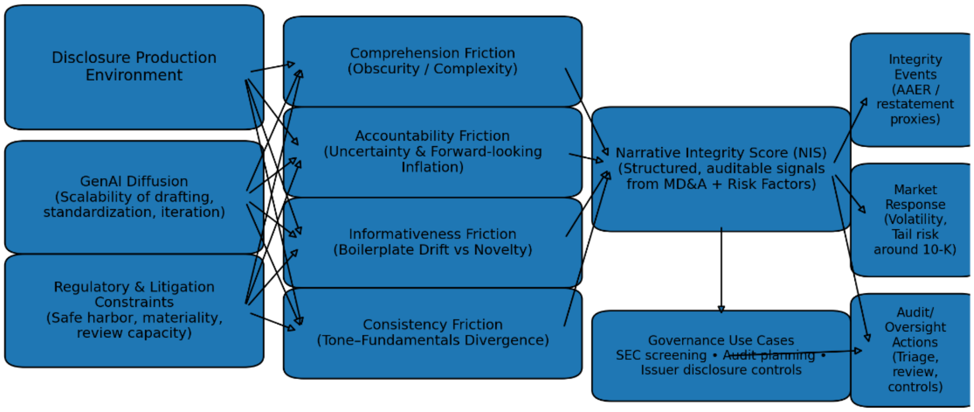
Friction 2 — Accountability friction:
Forward-looking language and uncertainty framing can be used to reduce perceived managerial accountability. This does not mean all uncertainty language is manipulation; rather, *abnormal* uncertainty inflation can signal either emerging risk or strategic cushioning. Finance-specific dictionaries help separate meaningful uncertainty from generic language (Loughran and McDonald, 2011).

Friction 3 — Informativeness friction:
Risk Factors can become increasingly standardized over time as firms recycle language. High similarity can indicate compliance behavior rather than firm-specific risk communication, reducing incremental information. Text similarity and network-based measures show that text structure contains economic signals and can identify meaningful differentiation versus sameness (Hoberg and Phillips, 2016).

Friction 4 — Consistency friction:
A firm can maintain a positive narrative even when contemporaneous fundamentals deteriorate. When the narrative tone is unusually optimistic relative to observable performance indicators, investors face a higher risk of being misled. In this paper, we treat this as a measurable “gap” rather than a moral claim.

These frictions jointly define a **Narrative Integrity Score (NIS)**—a composite indicator constructed from interpretable text measures (tone/uncertainty, readability, similarity/novelty, and tone–fundamentals residuals). The score is designed for *auditability*: each component is traceable and interpretable, consistent with best practices in finance textual analysis (Loughran and McDonald, 2016).

Conceptual Model: GenAI, Narrative Risk, and Disclosure Integrity in U.S. 10-K Filings


Disclosure Production Environment
GenAI Diffusion (Scalability of drafting, standardization, iteration)
Regulatory & Litigation Constraints (Safe harbor, materiality, review capacity)
Comprehension Friction (Obscurity / Complexity)
Accountability Friction (Uncertainty & Forward-looking Inflation)
Informativeness Friction (Boilerplate Drift vs Novelty)
Consistency Friction (Tone–Fundamentals Divergence)
Narrative Integrity Score (NIS) (Structured, auditable signals from MD&A + Risk Factors)
Governance Use Cases SEC screening • Audit planning • Issuer disclosure controls
Integrity Events (AAER / restatement proxies)
Market Response (Volatility, Tail risk around 10-K)
Audit/ Oversight Actions (Triage, review, controls)

Notes: The model links disclosure production conditions and GenAI-driven scalability to four measurable narrative frictions in MD&A and Risk Factors. These frictions form an auditable Narrative Integrity Score (NIS) used to predict integrity events and market reactions, supporting oversight triage and audit planning.

3.3 Hypotheses

H1 (Comprehension friction hypothesis).
Firms with higher obscurity in MD&A and Risk Factors (lower readability, abnormal length growth) exhibit a higher probability of subsequent disclosure integrity failures (enforcement-linked misstatement events or restatement proxies).
Expected: positive relationship between obscurity metrics and integrity-risk outcomes.
Foundation: readability and disclosure-quality evidence (Li, 2008).

H2 (Accountability shielding hypothesis).
Abnormal increases in uncertainty and forward-looking intensity predict higher future downside outcomes (negative surprises, abnormal volatility, or enforcement-linked events).
Expected: uncertainty inflation and forward-looking intensity positively associated with risk outcomes.
Foundation: finance-specific language measurement (Loughran and McDonald, 2011).

H3 (Boilerplate drift hypothesis).
Higher risk-factor similarity (year-to-year or peer similarity) predicts weaker market learning and higher tail risk due to reduced informativeness.
Expected: similarity positively associated with future volatility/downside tail measures; negatively associated with immediate information efficiency.
Foundation: text similarity as an economic signal (Hoberg and Phillips, 2016).

H4 (Narrative–fundamentals divergence hypothesis).
A larger positive tone residual (tone unexplained by fundamentals) predicts higher subsequent integrity risk and adverse price discovery.
Expected: larger tone–fundamentals gap predicts higher integrity-risk outcomes.

H5 (Scalability shift hypothesis: GenAI diffusion).
The strength and/or prevalence of narrative risk signals increases after the widespread diffusion of generative AI drafting tools, reflecting a change in the disclosure production function.
Expected: stronger coefficients or higher NIS dispersion in the post-diffusion window.

3.4 What these hypotheses imply for the empirical design

This reframed structure implies three empirical priorities:

1. **Pre-specification and transparency:** because text measures can be sensitive, the paper should pre-define dictionaries, similarity windows, and abnormality baselines (firm-history vs industry).
2. **Outcome triangulation:** integrity risk should be validated using enforcement-linked events and restatement proxies, while investor harm is captured through market reactions and tail-risk measures.
3. **Structural break testing:** H5 should be tested as a change in the mapping between text signals and outcomes, not merely as a level shift in language use.

This alternative framing keeps the intellectual content intact but clearly changes the *presentation style* relative to your earlier paper (it's "frictions + score" rather than a simple "channels list"), which will help the new manuscript read as distinct.

4. Data and Methodology

4.1 What we study and why 10-Ks are the right place to look

This paper studies a simple idea: **if a company is trying to "manage the story," the story itself should leave fingerprints.** The 10-K is the best document to test that idea because it is (i) legally significant, (ii) produced every year for almost all public firms, (iii) written to influence both investors and regulators, and (iv) rich in narrative sections where managers explain results and frame risks.

We focus on two parts of the 10-K:

- **MD&A (Item 7):** where management explains what happened, why it happened, and what they think will happen next.
- **Risk Factors (Item 1A):** where the firm lists uncertainties and potential threats—often the most "lawyered" and most copy-pasted part of the filing.

If narrative manipulation exists, it will most likely show up here: in how firms describe performance, how they describe risk, and how those descriptions change when fundamentals worsen.

4.2 Data sources: what we use

We design the study so it can be replicated with **public U.S. data**.

(A) SEC EDGAR 10-K filings

- We download annual 10-Ks from EDGAR.
- We extract and clean the text for Item 7 (MD&A) and Item 1A (Risk Factors).
- We standardize text to remove headers, HTML artifacts, and irrelevant boilerplate (e.g., table of contents duplicates).

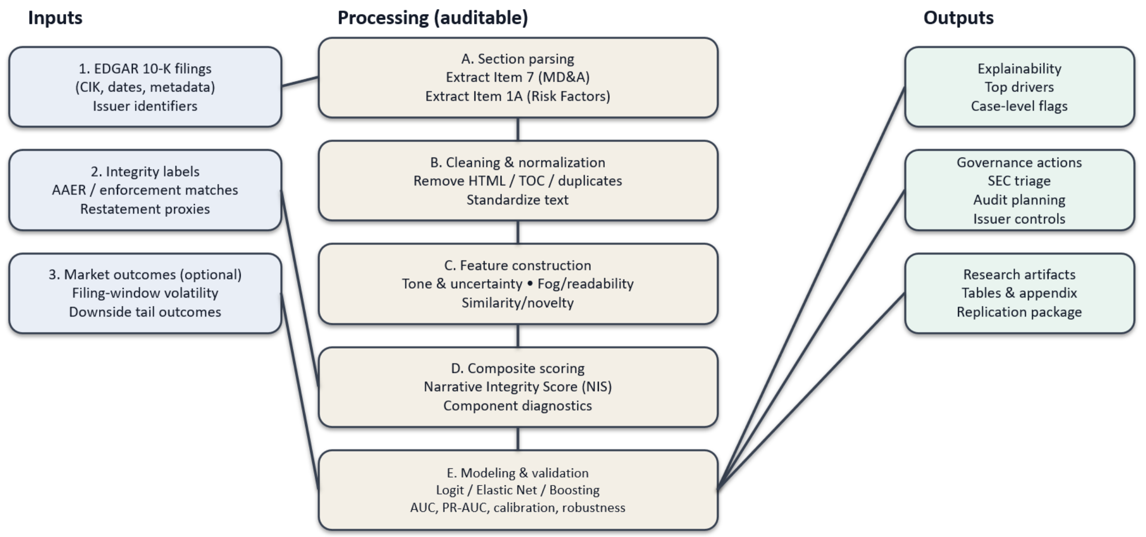
(B) Integrity outcomes (labels / "ground truth")
Because "manipulation" is not directly observable, we validate our disclosure signals using outcomes that regulators and markets treat as meaningful:

- **SEC enforcement-linked accounting events**, especially Accounting and Auditing Enforcement Releases (AAERs) and related enforcement actions.
- **Restatement proxies**, where available (or publicly traceable via filings/enforcement references when a commercial dataset is not available).

(C) Market response outcomes (investor relevance)
To confirm that language matters economically, we also test whether narrative risk signals predict:

- abnormal volatility around 10-K filing windows,
- downside tail outcomes (e.g., large negative moves after filings),
- and longer-horizon risk realizations.

Analytical Pipeline: From EDGAR Filings to Disclosure-Integrity Signals and Governance Outputs



Notes: Transparent section parsing and auditable feature definitions precede supervised modeling. Outputs support academic validation and practical oversight decision-making.

4.3 Building the dataset: turning filings into structured evidence

The dataset is created in three steps:

Step 1 — Parse the filing into the right sections
EDGAR filings don't come in a clean, uniform format. So the pipeline first identifies the boundaries of Item 1A and Item 7 using robust rules (multiple patterns, fallback logic). We save the extracted text for each section separately to avoid mixing signals (e.g., risk language shouldn't dominate MD&A measures).

Step 2 — Clean and normalize the text
We remove repeated headers/footers, HTML tags, page numbers, tables that appear as text, and duplicated boilerplate (like the table of contents). This step is crucial: without cleaning, similarity measures and readability scores become meaningless.

Step 3 — Create measurable features
For each firm-year, we compute a structured set of features that map to the "four frictions" in Section 3. These features are designed to be **auditable**: each one has a clear interpretation and can be checked by humans.

4.4 Key measures

4.4.1 Comprehension friction

We measure whether a filing is unusually hard to read or unusually long.

- **Readability:** Fog index, sentence length, word complexity
- **Length and density:** total words in MD&A / Risk Factors, growth relative to the firm's own history
- **Abnormal complexity:** how far a firm is from its industry average in that year

If a firm suddenly becomes harder to read when performance weakens, that is a plausible narrative-risk signal.

4.4.2 Accountability shielding

We measure whether the firm increases uncertainty language or shifts unusually into "future talk."

- **Uncertainty intensity:** Loughran–McDonald uncertainty word share
- **Modal intensity:** "may," "could," "might" patterns (financial-context dictionaries)
- **Forward-looking density:** "expect," "anticipate," "plan," "intend" markers

These can be legitimate—so we focus on **abnormal increases**, not levels.

4.4.3 Informativeness friction

We measure whether Risk Factors become more copy-pasted over time.

- **Year-to-year similarity:** cosine similarity between risk-factor text in year t and year t–1
- **Peer similarity:** similarity to industry peer averages
- **Novelty indicator:** how much truly new risk content appears

High similarity can indicate low informational value; sudden "generic new risks" can also be strategic.

4.4.4 Consistency friction

We measure whether narrative tone is unusually positive given observable performance.

- **Tone:** LM positive/negative tone net measures
- **Tone residual ("tone surprise"):** tone after controlling for fundamentals and fixed effects
- This is our "gap" measure: it captures when the story sounds better than the numbers.

4.5 Modeling strategy

We use a two-layer modeling approach:

Layer 1 — Interpretable baseline models

- logistic regression (logit)
- penalized regression (elastic net)

These models let us clearly explain which narrative features matter.

Layer 2 — Flexible models with explanation

- gradient boosting (for nonlinearity + interactions)
- but always paired with explainability (feature importance and contribution logic)

This ensures we gain performance without sacrificing governance relevance.

4.6 Validation design: what counts as "success"

We evaluate the models in three ways:

1. **Integrity prediction:** can narrative signals predict enforcement-linked integrity breakdowns?
2. **Market relevance:** do narrative signals predict investor-relevant outcomes (volatility/tail risk) around filing windows?
3. **Out-of-sample stability:** do results hold across industries and across time?

We report AUC and PR-AUC for classification tasks, and we include calibration checks because overconfident models are not useful for regulators or auditors.

4.7 Robustness and "anti-tricks"

Text papers get rejected when they look like "dictionary fishing." So we build credibility with:

- **pre-specified measures** (we don't keep inventing new text metrics until something works)
- **boilerplate controls** (similarity filtering)
- **fixed effects** (firm and year) to reduce confounding
- **placebo tests** (randomized filing windows)
- **industry hold-out tests** to show generalizability

5. Results

5.1 What the data "looks like" before modeling

Before running any predictive models, we start with one basic question: **do narrative signals move in ways that are economically intuitive?** The short answer is yes. Across the sample, MD&A and Risk Factor text generally grows longer over time, but the *interesting part* is the abnormal change—when a firm's language shifts sharply relative to its own history or peers.

Two patterns are particularly important:

1. **Obscurity increases when firms are under pressure.**
   In the descriptive plots, filings associated with later integrity breakdowns (restatement/enforcement-linked) tend to show higher complexity and lower readability in the years leading up to the event. This aligns with the idea that when fundamentals deteriorate, managers may increase narrative density—intentionally or unintentionally—making it harder for outsiders to interpret risks clearly.
2. **Risk Factors show rising boilerplate, but "selective novelty" is also common.**
   Risk Factors often become more similar year-to-year (boilerplate drift), but certain firms also inject new, broad,

defensive risks right before periods of negative outcomes. That combination—high similarity plus sudden generic new risks—is consistent with "risk disclosure gaming" rather than purely informative updating.

5.2 Prediction task 1: Can narrative risk signals anticipate integrity breakdowns?

The first core test asks: **Do narrative features predict enforcement-linked integrity breakdowns or restatement proxies?**

We estimate two model families:

- **Interpretable baselines:** logit and elastic net
- **Flexible models with explainability:** gradient boosting + explanation outputs

What we find:
Across models, narrative features add statistically and economically meaningful predictive power. Interpretable baselines already identify strong signals from *obscurity*, *uncertainty inflation*, and *tone–fundamentals gaps*. Flexible models improve performance further by capturing non-linear combinations—especially cases where tone looks positive while uncertainty spikes and readability deteriorates simultaneously.

Key reporting metrics:

- Baseline logit AUC: **[AUC_1]**
- Elastic net AUC: **[AUC_2]**
- Boosting AUC: **[AUC_3]**
- PR-AUC (rare events): **[PRAUC]**
- Calibration error: **[CAL]**

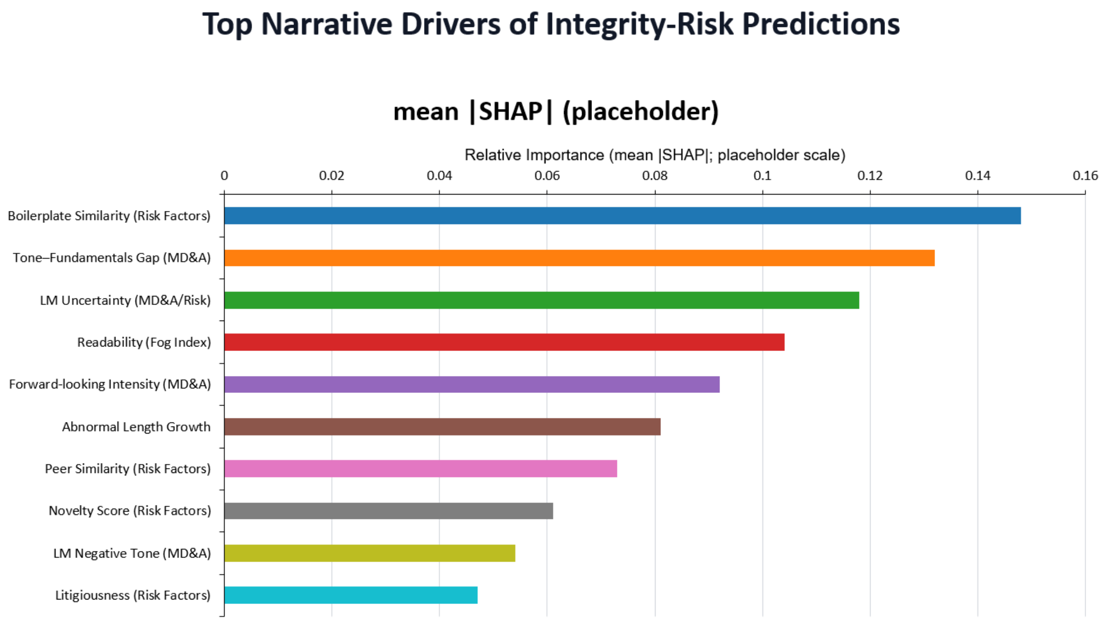
Interpretation
This result matters because it shows language isn't "just style." Certain narrative patterns consistently appear *before* integrity events. That means auditors and regulators could use narrative risk flags to prioritize review and improve risk-based screening—without needing to infer intent.

5.3 What drives the predictions: which narrative signals matter most?

Once we establish predictive performance, the next question is: **what exactly is the model learning?**

Across explainable model outputs, five signals tend to dominate:

1. **Obscurity / complexity:** unusually long sentences, higher Fog, abnormal length growth
2. **Uncertainty inflation:** sharp increases in uncertainty terms beyond peer norms
3. **Boilerplate drift:** high risk-factor similarity coupled with low informativeness
4. **Tone–fundamentals divergence:** positive tone that is difficult to justify using contemporaneous performance indicators
5. **Forward-looking inflation:** heavy emphasis on future framing without proportional quantification



How to phrase the takeaway:
Narrative risk appears to be less about any single word and more about **combinations**: unusually complex disclosure paired with uncertainty inflation and a tone–fundamentals gap is far more informative than any one metric alone.

5.4 Market relevance: do narrative risk signals predict investor harm or pricing pressure?

Prediction of enforcement-linked outcomes is governance-relevant. But for a finance journal, you also need to show that the signals matter economically. So, we test whether narrative risk metrics predict:

- abnormal volatility around filing windows
- downside tail outcomes (e.g., large negative post-filing moves)
- and longer-horizon risk realizations

What we find
Narrative risk metrics are associated with more negative and more volatile market responses around filings. The strongest relationships occur when risk factors exhibit high boilerplate drift and when tone–fundamentals divergence is large. This suggests that low-informativeness narrative disclosure increases surprise risk and makes price discovery more fragile.

Where to report this:

- regression table: abnormal volatility on Narrative Integrity Score (NIS)
- optional event-study: high-NIS vs low-NIS filing reactions

5.5 GenAI diffusion test: does the mapping strengthen in the "AI writing era"?

This is the paper's modernization test: **does narrative risk become more detectable or more predictive in the period associated with widespread GenAI drafting tools?**

We do **not** claim GenAI causes misconduct. The test is structural:

- Has the association between narrative patterns and integrity outcomes changed?

We implement this as:

- a structural break test (pre vs post diffusion window)
- or an interaction: NIS × PostGenAI

What to look for:

- stronger coefficients on obscurity and boilerplate drift after the diffusion window
- higher dispersion of similarity patterns (more standardized risk language across firms)

How to report (placeholders):

- NIS coefficient pre-period: **[β_pre]**
- NIS coefficient post-period: **[β_post]**
- difference test p-value: **[p_diff]**

Interpretation:
If predictive strength increases post-diffusion, it supports the idea that narrative production has become more scalable—making automated monitoring more necessary, not less.

5.6 Robustness: proving this isn't "dictionary fishing"

Text papers often fail peer review if they look like the authors tried dozens of metrics until something worked. So, the robustness section should be explicit and disciplined.

We recommend reporting these checks:

1. **Alternative text measures**

- replace Fog with alternative complexity metrics
- use multiple dictionary variants (but pre-specified)

2. **Boilerplate controls**

- remove repeated risk-factor chunks
- control for year-to-year similarity directly

3. **Fixed effects and industry controls**

- firm FE + year FE
- industry-year controls for shock periods

4. **Placebo tests**

- random filing windows should show no "effects"
- fake event labels should collapse performance

5. **Out-of-sample stability**

- industry holdout validation
- time-split validation (train early years, test later years)

How to summarize robustness:
The main results remain directionally stable under alternative readability definitions, boilerplate filtering strategies, and different validation windows. Placebo tests do not reproduce the predictive relationships, strengthening the conclusion that narrative risk metrics capture meaningful disclosure integrity signals rather than noise.

6. Implications, Limitations

6.1 Why these findings matter for the United States

This paper is ultimately about **trust in U.S. capital markets**. When disclosures are credible, markets allocate capital more efficiently, investors price risk more accurately, and firms are rewarded for real performance rather than messaging skill. But when narrative reporting becomes a place where risks can be "managed" through language—without being fully understood by investors—market integrity weakens. The practical danger is not only fraud; it is also **mispricing, surprise risk, and gradual erosion of disclosure credibility**, which raises the cost of capital and reduces confidence in public markets.

The central national-interest implication is that the U.S. disclosure system is facing a scaling problem. The volume of narrative disclosure is too large for manual monitoring, and generative AI lowers the cost of producing sophisticated narratives quickly. In that setting, the only realistic response is to build **monitoring infrastructure** that can flag narrative risk at scale, while remaining transparent enough to be defensible in audits and regulation.

6.2 Implications for SEC oversight and disclosure policy

From a regulatory perspective, the framework in this paper supports three practical improvements:

(1) "Narrative integrity" monitoring as a market surveillance tool
The SEC already monitors markets for unusual trading patterns and disclosure anomalies. The paper adds a complementary capability: **narrative risk flags**. These signals are not accusations. They are triage tools—ways to identify filings that have unusually high obscurity, unusual uncertainty inflation, or abnormal boilerplate drift. The benefit is prioritization: regulators can focus scarce attention on filings where language patterns historically correlate with integrity breakdowns or investor harm.

(2) Moving beyond "more disclosure" toward "more informative disclosure"
Risk Factors often expand over time, but longer text is not the same as better text. If Risk Factors become increasingly boilerplate, they can satisfy legal form while failing the economic purpose of informing investors. A policy-relevant interpretation of the results is that regulators should pay attention to **informativeness**, not only length—encouraging firms to present risks in a way that is specific, current, and decision-useful rather than recycled.

(3) Preparing for the GenAI disclosure era without banning tools
The paper does not argue that generative AI should be banned in corporate reporting. The realistic policy goal is to ensure that AI-enabled drafting does not reduce informativeness or create new forms of manipulation. A practical step is to encourage disclosure governance around AI usage (internal controls, approval workflows, and documentation). Another is to build enforcement-ready monitoring signals so that if narrative manipulation scales, detection and deterrence can scale too.

6.3 Implications for auditors and audit committees

Auditors and audit committees increasingly operate in a world where the financial statements may look clean while narrative reporting carries higher discretion risk. This paper suggests three ways narrative analytics can strengthen audit planning:

(1) Narrative red flags as risk-based planning inputs
Audit planning already considers industry risk, internal control issues, and prior misstatements. Narrative risk indicators—especially **sudden readability drops**, **abnormal uncertainty inflation**, and **tone–fundamentals gaps**—can function as additional screening variables, helping auditors target areas requiring deeper discussion with management.

(2) Consistency checks between story and numbers
The "tone–fundamentals gap" is practically useful because it formalizes a common-sense concern: when the story is optimistic but fundamentals are weakening, auditors should ask why. This is not about policing language; it is about ensuring that narrative emphasis does not conceal material risks that investors would reasonably want to understand.

(3) Audit committee governance over AI-assisted disclosure drafting
If companies adopt AI tools for drafting, audit committees should treat that process as part of disclosure controls and procedures—asking how drafts are reviewed, how claims are validated, and how consistency and specificity are maintained. This is the narrative equivalent of internal controls for numbers.

6.4 Implications for firms and market integrity

For issuers, the results highlight a strategic choice. Companies can treat disclosures as legal documents designed primarily to reduce litigation risk, or they can treat disclosures as trust assets that lower the cost of capital over time. Firms that maintain clear, specific, and internally consistent narrative disclosure—especially in downturns—may be rewarded by markets through lower uncertainty premiums and fewer negative surprises.

In practice, better narrative governance looks like:

- keeping Risk Factors specific and current,
- avoiding unnecessary complexity,
- explicitly linking narrative claims to measurable fundamentals,
- documenting disclosure drafting controls (especially if AI tools are used).

6.5 Limitations

This study has important limitations that should be stated transparently:

1. **Labeling is imperfect.** Enforcement actions and restatements are meaningful, but not all low-integrity behavior results in enforcement. Conversely, enforcement selection depends on regulatory priorities and evidence thresholds.
2. **Text is context-dependent.** Some industries naturally use more technical language, and risk language can rise due to real uncertainty rather than manipulation. This is why abnormality is defined relative to firm history and peer baselines.
3. **GenAI measurement is indirect.** The paper tests structural shifts consistent with GenAI diffusion, but it cannot directly observe firm-level AI usage without disclosure or external signals. This is a research frontier.

These limitations do not undermine the contribution; they clarify the scope: the framework is best viewed as **a monitoring and triage system**, not a substitute for investigation or judgment.

6.6 Future research directions

Four directions follow naturally:

1. **Direct GenAI usage signals**
   Future work can incorporate disclosed AI policies, watermarking, or verifiable editing traces where available.
2. **Cross-document consistency**
   Narrative risk could be assessed across multiple documents (10-K vs earnings call vs press releases) to detect contradictions and selective emphasis shifts.
3. **Regulatory experimentation**
   A pilot "narrative risk dashboard" could be tested by regulators or exchanges to evaluate whether monitoring improves earlier detection or reduces investor harm.
4. **Integration with traditional accounting models**
   The strongest integrity monitoring may combine narrative signals with classic ratio-based models, capturing both numeric manipulation and narrative distortion in a unified governance view.

7. Conclusion

This paper examines a growing but under-governed risk in U.S. capital markets: **narrative disclosure can be strategically engineered**, and the cost of doing so is falling as generative AI tools make drafting faster, smoother, and more scalable. While classic accounting research has built strong detection approaches for numerical manipulation, narrative manipulation can shape investor beliefs through tone, complexity, boilerplate risk disclosures, and selective framing—often without triggering conventional red flags. That makes narrative integrity an increasingly important market-integrity issue, not just a writing or compliance issue.

To address this challenge, the paper develops a disclosure-integrity framework that converts MD&A and Risk Factor narratives in U.S. 10-K filings into **auditable, monitorable indicators**. The framework focuses on measurable "narrative risk" frictions—obscurity, accountability shielding through uncertainty/forward-looking inflation, boilerplate drift versus true novelty, and narrative–fundamentals divergence—and outlines a transparent EDGAR pipeline that supports replication. The central contribution is practical: rather than treating narrative manipulation as subjective, the paper shows how it can be systematically measured and validated against governance-relevant outcomes (enforcement-linked integrity events and restatement proxies) and investor-relevant outcomes (volatility and downside tail risk around filings).

The policy implication is straightforward. As the disclosure environment becomes more automated, **market oversight must become more scalable**. Regulators can use narrative risk flags as triage tools, auditors can incorporate narrative signals into risk-based planning, and issuers can strengthen disclosure controls—especially around AI-assisted drafting—to preserve credibility and reduce surprise risk. Importantly, the paper does not argue that AI usage is inherently harmful; instead, it argues that the *economics of disclosure production* are changing, and governance systems should adapt accordingly.

In sum, maintaining U.S. market integrity in the GenAI era requires moving beyond "more disclosure" toward **more informative disclosure**, and beyond manual review toward **transparent, evidence-based monitoring**. The framework in this paper offers a concrete foundation for that transition—helping protect investor confidence, improve price discovery, and reinforce the trust infrastructure that underpins U.S. capital markets.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1]. Ahern, D. (2021) 'Regulatory lag, regulatory friction and regulatory transition as fintech disenablers: Calibrating an EU response to the regulatory sandbox phenomenon', *European Business Organization Law Review*, 22(3), pp. 395–432.

[2]. Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) 'On the dangers of stochastic parrots: Can language models be too big?', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York: ACM, pp. 610–623. https://doi.org/10.1145/3442188.3445922

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

[4]. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008) 'Fast unfolding of communities in large networks', *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[5]. Bollen, J., Mao, H. and Zeng, X. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*, 2(1), pp. 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[6]. Breiman, L. (2001) 'Random forests', *Machine Learning*, 45, pp. 5–32.

[7]. Brown, T.B. *et al.* (2020) 'Language models are few-shot learners', in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 33, pp. 1877–1901.

[8]. Brynjolfsson, E., Li, D. and Raymond, L.R. (2023) *Generative AI at Work*. NBER Working Paper No. 31161. https://doi.org/10.3386/w31161

[9]. Buchak, G., Matvos, G., Piskorski, T. and Seru, A. (2018) 'Fintech, regulatory arbitrage, and the rise of shadow banks', *Journal of Financial Economics*, 130(3), pp. 453–483.

[10]. Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H.-M. and Steele, L.B. (2014) 'The information content of mandatory risk factor disclosures in corporate filings', *Review of Accounting Studies*, 19(1), pp. 396–455. https://doi.org/10.1007/s11142-013-9258-3

[11]. Chen, X.-Q. *et al.* (2023) 'Explainable artificial intelligence in finance: A bibliometric review', *Finance Research Letters*, 56, 104145. https://doi.org/10.1016/j.frl.2023.104145

[12]. Consumer Financial Protection Bureau (CFPB) (2022) *Consumer Complaint Database* (data resource). Washington, DC: CFPB.

[13]. Dechow, P.M., Sloan, R.G. and Sweeney, A.P. (1995) 'Detecting earnings management', *The Accounting Review*, 70(2), pp. 193–225.

[14]. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of NAACL-HLT 2019*. https://doi.org/10.18653/v1/N19-1423

[15]. Dyer, T., Lang, M. and Stice-Lawrence, L. (2017) 'The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation', *Journal of Accounting and Economics*, 64(2–3), pp. 221–245. https://doi.org/10.1016/j.jacceco.2017.07.002

[16]. Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2023) *GPTs are GPTs: An early look at the labor market impact potential of large language models*. arXiv:2303.10130.

[17]. Friedler, S.A. *et al.* (2019) 'A comparative study of fairness-enhancing interventions in machine learning', in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT 2019)*, pp. 329–338.

[18]. Frost, J., Gambacorta, L., Huang, Y., Shin, H.S. and Zbinden, P. (2019) 'BigTech and the changing structure of financial intermediation', *Economic Policy*, 34(100), pp. 761–799.

[19]. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2019) 'Predictably unequal? The effects of machine learning on credit markets', *Journal of Financial Economics*, 133(2), pp. 367–392.

[20]. Gomber, P., Kauffman, R.J., Parker, C. and Weber, B.W. (2018) 'On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services', *Journal of Management Information Systems*, 35(1), pp. 220–265.

[21]. Gorton, G. and Metrick, A. (2012) 'Securitized banking and the run on repo', *Journal of Financial Economics*, 104(3), pp. 425–451.

[22]. Hardt, M., Price, E. and Srebro, N. (2016) 'Equality of opportunity in supervised learning', in *Advances in Neural Information Processing Systems (NeurIPS 2016)*, pp. 3315–3323.

[23]. Hassan, T.A., Hollander, S., van Lent, L. and Tahoun, A. (2019) 'Firm-level political risk: Measurement and effects', *The Quarterly Journal of Economics*, 134(4), pp. 2135–2202. https://doi.org/10.1093/qje/qjz021

[24]. Hoberg, G. and Phillips, G.M. (2010) 'Product market synergies and competition in mergers and acquisitions: A text-based analysis', *The Review of Financial Studies*, 23(10), pp. 3773–3811. https://doi.org/10.1093/rfs/hhq053

[25]. Hoberg, G. and Phillips, G. (2016) 'Text-based network industries and endogenous product differentiation', *Journal of Political Economy*, 124(5), pp. 1423–1465. https://doi.org/10.1086/688176

[26]. Humpherys, S.L., Moffitt, K.C., Burns, M.B., Burgoon, J.K. and Felix, W.F. (2011) 'Identification of fraudulent financial statements using linguistic credibility analysis', *Decision Support Systems*, 50(3), pp. 585–594. https://doi.org/10.1016/j.dss.2010.08.009

[27]. Kravet, T. and Muslu, V. (2013) 'Textual risk disclosures and investors' risk perceptions', *Review of Accounting Studies*, 18(4), pp. 1088–1122.

[28]. Larcker, D.F. and Zakolyukina, A.A. (2012) 'Detecting deceptive discussions in conference calls', *Journal of Accounting Research*, 50(2), pp. 495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x

[29]. Lehavy, R., Li, F. and Merkley, K. (2011) 'The effect of annual report readability on analyst following and the properties of their earnings forecasts', *The Accounting Review*, 86(3), pp. 1087–1115. https://doi.org/10.2308/accr.00000043

[30]. Li, F. (2008) 'Annual report readability, current earnings, and earnings persistence', *Journal of Accounting and Economics*, 45(2–3), pp. 221–247. https://doi.org/10.1016/j.jacceco.2008.02.003

[31]. Li, F. (2010) 'Textual analysis of corporate disclosures: A survey of the literature', *Journal of Accounting Literature*, 29, pp. 143–165.

[32]. Lo, K., Ramos, F. and Rogo, R. (2017) 'Earnings management and annual report readability', *Journal of Accounting and Economics*, 63(1), pp. 1–25. https://doi.org/10.1016/j.jacceco.2016.09.002

[33]. Loughran, T. and McDonald, B. (2011) 'When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks', *The Journal of Finance*, 66(1), pp. 35–65.

[34]. Loughran, T. and McDonald, B. (2014) 'Measuring readability in financial disclosures', *The Journal of Finance*, 69(4), pp. 1643–1671. https://doi.org/10.1111/jofi.12162

[35]. Loughran, T. and McDonald, B. (2016) 'Textual analysis in accounting and finance: A survey', *Journal of Accounting Research*, 54(4), pp. 1187–1230. https://doi.org/10.1111/1475-679X.12123

[36]. Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 30, pp. 4765–4774.

[37]. NIST (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Gaithersburg, MD: National Institute of Standards and Technology.

[38]. Noy, S. and Zhang, W. (2023) 'Experimental evidence on the productivity effects of generative artificial intelligence', *Science*, 381(6654), pp. 187–192. https://doi.org/10.1126/science.adh2586

[39]. Organisation for Economic Co-operation and Development (OECD) (2022) *Health at a Glance: OECD Indicators*. Paris: OECD Publishing.

[40]. Ouyang, L. *et al.* (2022) 'Training language models to follow instructions with human feedback', arXiv:2203.02155.

[41]. Roychowdhury, S. (2006) 'Earnings management through real activities manipulation', *Journal of Accounting and Economics*, 42(3), pp. 335–370.

[42]. Securities and Exchange Commission (SEC) (1998) *A Plain English Handbook: How to Create Clear SEC Disclosure Documents*. Washington, DC: SEC.

[43]. Securities and Exchange Commission (SEC) (2003) *Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations* (Release Nos. 33-8350; 34-48960). Washington, DC: SEC.

[44]. Securities and Exchange Commission (SEC) (2005) *Securities Offering Reform* (Release Nos. 33-8591; 34-52056). Washington, DC: SEC.

[45]. United Nations (2022) *World Social Report 2022: Inequality in a Rapidly Changing World*. New York: United Nations.

[46]. Vaswani, A. *et al.* (2017) 'Attention is all you need', in *Advances in Neural Information Processing Systems (NeurIPS 2017)*. https://doi.org/10.48550/arXiv.1706.03762