Finance Research Letters 105 (2026) 110161

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

The use of AI in 10-K filings: An empirical analysis☆

Marcelo S. Perlin, Cristian R. Foguesatto *, Aliki K. Galanos, Felipe Affonso

Federal University of Rio Grande do Sul, Washington Luis 855, Porto Alegre RS, 90010-460, Rio Grande do Sul, Brazil

ARTICLE INFO

Keywords:
Corporate finance
10-K filings
Large language models (LLMs)
Artificial intelligence

ABSTRACT

With the rise on the use of AI in the corporate setting and the associated downside risk of bad disclose and overall credibility, we ask how much and which type of organizations rely on these tools when preparing their annual reports. Using a custom, publicly available, prediction model finetuned on FinBERT, we estimate the probability of AI-generated content and analyze the sentiment of three sections of the 10-K reports of Russell 1000 constituent companies from 2018 to 2024. Results indicate a relatively stable use of AI over the period, with distinct patterns across report sections, being the Business Description the section with the highest probability of AI adoption. Additionally, older, higher-risk (beta), and higher asset turnover firms appear more likely to adopt AI-assisted writing, in contrast to firms with higher accrual ratios and greater leverage. Moreover, there is a positive association between sentiment scores and AI detection within the Risk Factors section. These findings offer valuable insights for researchers who need to further examine AI-assisted writing in annual reports. They also carry important implications for practitioners such as investors and policy-makers who should pay closer attention to this emerging organizational practice.

1. Introduction

Financial reports are crucial for stakeholders to assess a company's financial health and profitability and to make informed decisions, such as investing, lending, or employment choices. As powerful tools for textual analysis with a human-like ability to generate text, Large Language Models (LLMs) have gained significant relevance in financial reporting (Huang et al., 2023; Yang et al., 2024). The reliance on these reports implies that firms should be wary of language selection, text preparation, length and readability (Hasan, 2020; Nguyen and Kimura, 2023). With ongoing technological advances, LLMs are expected to support professionals in generating reported information (Türegün, 2019; Csere et al., 2025) and possibly assist with writing it (De Villiers et al., 2024). Thus far, researchers have provided evidence on the effects of AI strategies and integration on different firm outcomes such as, corporate investment efficiency (Liu and Hu, 2025) and Environmental, Social and Governance (ESG) performance (Weng, 2025). Moreover, the disclosure of automated decision-making has also been examined (Bonsón et al., 2023). However, as far as we are concerned, there is no study on the use of AI to assist managers in writing financial reports.

Although the implications of such assistance can be advantageous with professionals having more time to focus on decision making over data collection and manual analysis (Türegün, 2019), it could also bias reports by compromising their transparency and accountability (De Villiers et al., 2024). The use of LLMs can increase the risk of artificial and inconsistent information, with a direct impact on the quality of financial reports, and also on the governance and transparency of the information.

☆ This article is part of a Special issue entitled: 'AI in Corporate Finance' published in Finance Research Letters.

* Corresponding author.
E-mail addresses: m.perlin@ufrgs.br (M.S. Perlin), cristian.foguesatto@ufrgs.br (C.R. Foguesatto), aliki.g.kara@gmail.com (A.K. Galanos), felipe-affonso@hotmail.com (F. Affonso).

https://doi.org/10.1016/j.frl.2026.110161
Received 27 February 2026; Received in revised form 29 April 2026; Accepted 9 May 2026
Available online 29 May 2026
1544-6123/

Therefore, we aim to examine whether organizations use AI-assisted writing in their annual reports, covering three main section of the 10-K: Risk Factors (RF), Management Discussion and Analysis (MD&A) and Business Description (BD). These sections are relevant because provide a set of key companies' information. In addition, we investigate which firms are more likely to do so. In doing so, we extend the existing literature on the use of AI in the organizational spectrum and further discuss the implications of its use. Our research is relevant to corporate finance literature by examining the use of AI-assisted writing in 10-K filings, a key channel through which firms communicate information to investors and capital markets. By identifying the presence of AI-generated content and its association with disclosure tone, the study provides new evidence on how emerging technologies, such as AI, may influence information transparency from financial reports.

2. Data and methods

We collect 10-K reports and financial/price data for all firms that are currently part of the Russel 1000 index, which covers approximately 90% of all traded volume in the American exchanges. The financial data covers the time period between 2015 and 2024. This time frame is relevant as it coincides with the widespread adoption of AI-based natural language processing tools in corporate reporting. For the 10-K reports, we use EdgarTools, a Python library for downloading company data directly from the official SEC API. The Russel 1000 composition and all financial data is obtained from EODHD, a commercial data vendor and a common source of information for academic research. Within each downloaded 10-K document, we save the text of three distinct sections that are later used for the analysis: Business Description (BD), Risk Factors (RF) (de Jong, 2025), and management discussion (MD) of results (Fengler and Phan, 2025).

Using the 10-K text data, we fine-tune a LLM model based on FinBERT (Huang et al., 2023) that outputs the probability of AI use in writing text.[1] Based on data from 2015 to 2020, we randomly gather a large dataset (3000 cases) of financial text from the MD (Management Discussion) between 2015 and 2020 (pre AI), rewrite it using three distinct AI models (Gemini, chatgpt and Claude), and use it in a train/test machine learning framework. The resulting model presented high accuracy (89%) and precision (92%), justifying its usage in the study. Details on how we executed such an estimation are found in Appendix. The model is also available to the public on Hugging Face. Additionally, we compute the sentiment of all 10-K sections using standard FinBERT (Huang et al., 2023), a large language model for extracting sentiment from financial texts.

One caveat of using FinBERT based models in analyzing large financial text is their maximum input tokens (512 for FinBert). We find that, for our dataset, this limit was significantly lower than the total number of tokens in the majority of the texts. To solve this issue, our estimation strategy was to use a sliding window and split large texts into smaller parts with an overlap of 10% of tokens for each split (Wang et al., 2019; Ding et al., 2020). In our case, we use 500 as the chunk size and 50 as overlap. We then aggregate the results of AI detection by taking an average of all chunk probabilities. For the sentiment, we first ignore all *neutral* sentiment cases and, each chunk of text, assign +1 for positive sentiments and −1 for negative. Finally, we average the results for a final sentiment score for each of the 10-K sections.

After cleaning for missing data, where we lost approximately 18% of total rows, our text and financial data consists of 3939 data points for 570 companies during the time period between 2018 and 2024. Table 1 presents the descriptive statistics for our final sample: Panel A details the company's financial and market characteristics, and Panel B provides statistics for the text analysis variables, including the AI detection scores and sentiment scores derived from the 10-K reports.

In Table 1, Panel A, we observe significant variability in performance, as indicated by the high standard deviation and wide range of Return on Equity (ROE). We can also see that the median IPO year is 1997, that is, most are well established companies. The average age for the firms in the sample is 23 years old. Moreover, it is observed that firms do not considerably differ in terms of size. However, there are significant gaps among companies in terms of leverage and asset turnover, as suggested by the means, minimum, and maximum values of these variables. The negative value of −6 at the minimum age is due to the fact that we calculated age based on IPO year. Some companies provided 10-K for the years previous to the their IPO. Which then resulted in negative values for the age variable.

Panel B of Table 1 shows that the *Business Description* section (BD Score) has the highest average probability of AI use at 0.178. This is substantially higher than the scores for the *Management Discussion of Results* (MD Score) at 0.158, and *Risk Factors* (RF Score) at 0.129. The sentiment scores also behave as expected by content: *Risk Factors* (RF Sentiment) are almost uniformly negative, with a mean of −0.843. In contrast, *Business Description* (BD Sentiment) is positive on average 0.301, while *Management Discussion* (MD Sentiment) is slightly negative at −0.170.

2.1. Models

We use econometric models to investigate the profile of companies that are mostly using AI in their written financial reports. For that, we use the probabilities of AI written text from our custom AI Text Detector as the dependent variable. Given its truncated nature ([0–1]), we estimate a Generalized Linear Model (GLM), with Logit link function and quasibinomial distribution to model the data, Eq. (1).

$$E(AI_{i,t}) = f(z_{i,t}) \tag{1}$$

[1] The initial version of this paper used Desklib, a general AI detector that scored highly in RAID Benchmark Leaderboard (access in 2026-01-15). However, the model performance significantly worse in financial text.

Table 1
Descriptive Statistics (2018–2024). All monetary values are reported in millions of USD.

Variable	N	Mean	Std Dev	Min	Max
Panel A: Company variables					
ROE	6301	0.206	5.755	−105.913	239.107
ROA	6301	0.055	0.413	−8.949	21.117
IPO year[a]	6098	1997	15	1919	2024
Age	6098	23.0	14.8	−6.0	105.0
Log size	6301	10.122	0.647	7.389	12.602
Leverage	6301	0.693	3.020	0.000	238.549
Beta	6156	1.019	0.474	−4.054	5.950
Asset turnover	6301	0.718	2.326	−0.069	95.086
Accrual ratio	6301	0.024	0.628	−1.897	44.745
Panel B: Text analysis variables					
BD Score	5405	0.178	0.095	0.009	0.959
RF Score	5465	0.129	0.063	0.011	0.845
MD Score	5483	0.158	0.085	0.007	0.849
BD Sentiment	6294	0.301	0.624	−1.000	1.000
RF Sentiment	6294	−0.843	0.352	−1.000	1.000
MD Sentiment	6294	−0.170	0.426	−1.000	1.000

Note:
[a] We report the median (and not mean) value for IPO year.

Table 2
Description and calculation of explanatory variables.

Variable	Calculation	Hypothesis (sign)
Panel A: Hypothesis variables		
ROE	Total Net Income divided by Total Equity	Positive
Age	Number of years since IPO date	Negative
Log size	Log of total assets	Negative
Risk (Beta)	Beta from market model, calculated using SP500 returns	Positive
Leverage	Total debt divided by total assets	Positive
Sentiment	Sentiment value of text, based on FinBert	Positive
Asset turnover	Total Revenue divided by Total Assets	Positive
Accrual ratio	Free cash flow minus net income, divided by total assets	Positive
Panel B: Control variables		
Sector	Sector of company	–
Year	Year of data	–

$$pAI_{i,t} = \alpha + \theta X_{i,t} \tag{2}$$

where pAI_i is the dependent variable, measuring the probability of AI use in text generation, α is the intercept, and the θ is a vector of parameters attached to $X_{i,t}$. As for the explanatory variables, we follow the work of Fengler and Phan (2025) and set our hypothesis and control variables as show in Table 2.

Following Table 2, we expect a positive relationship between **Return on Equity (ROE)** and AI usage, suggesting that more profitable firms are more likely to adopt AI-assisted writing (Shiyyab et al., 2023). High-performing companies typically possess greater financial resources, allowing them to invest in emerging technologies and absorb the costs associated with implementing new software systems.

As for **Age**, we believe in a negative relationship with AI usage, implying that younger firms are more prone to using AI than their older counterparts. We suspect that established corporations often have deeply entrenched manual reporting processes and bureaucratic layers that resist technological change. Conversely, younger companies can be viewed as more agile and innovative (Protogerou et al., 2017).

Both **Risk (Beta)** and **Leverage** carry a positive hypothesis. Companies with high leverage or market risk face intense scrutiny from creditors and investors, requiring extensive and careful disclosure of risk factors (Zhu et al., 2024). Likewise, we expect **Asset Turnover** to be positively associated with the use of AI in corporate reports. Firms that employ their assets more efficiently tend to place greater value on operational scalability and process optimization. Prior research shows that more efficient firms are better positioned to benefit from information technologies, as automation helps reduce administrative frictions and managerial overload by reallocating routine tasks away from senior management (Brynjolfsson et al., 2017).

The variables for **Sentiment** and **Accrual Ratio** are both hypothesized to be positive. For Accrual Ratio, higher accruals often indicate complex accounting estimates that require detailed technical explanations; firms may use LLMs to help generate this dense, technical text more quickly. Regarding Sentiment, the positive link suggests that AI tools may be used to polish the report of positive reports. That is, managers might use the technology's tendency to produce smoother, more professional, and generally more optimistic or neutral-positive corporate prose compared to raw human drafts (Huang et al., 2023).

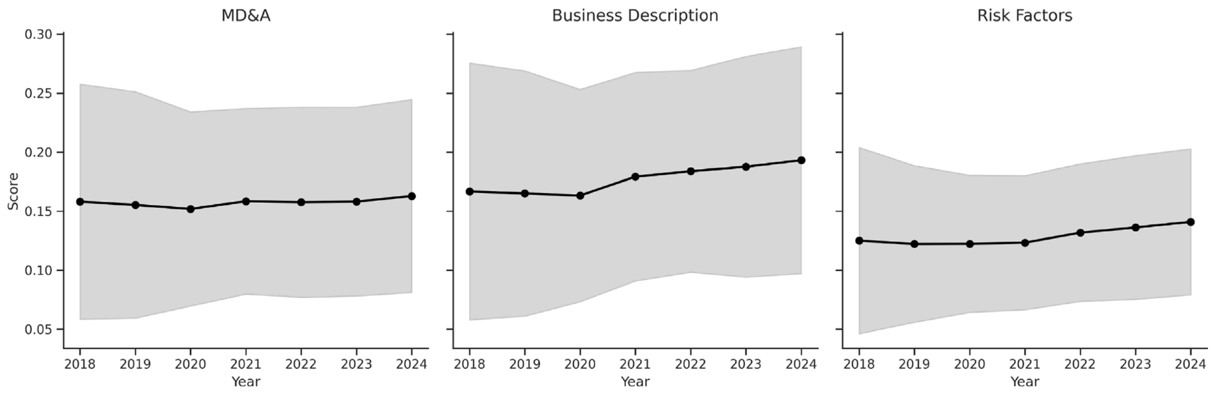

Fig. 1. Average AI score per 10-K text segment - per year.

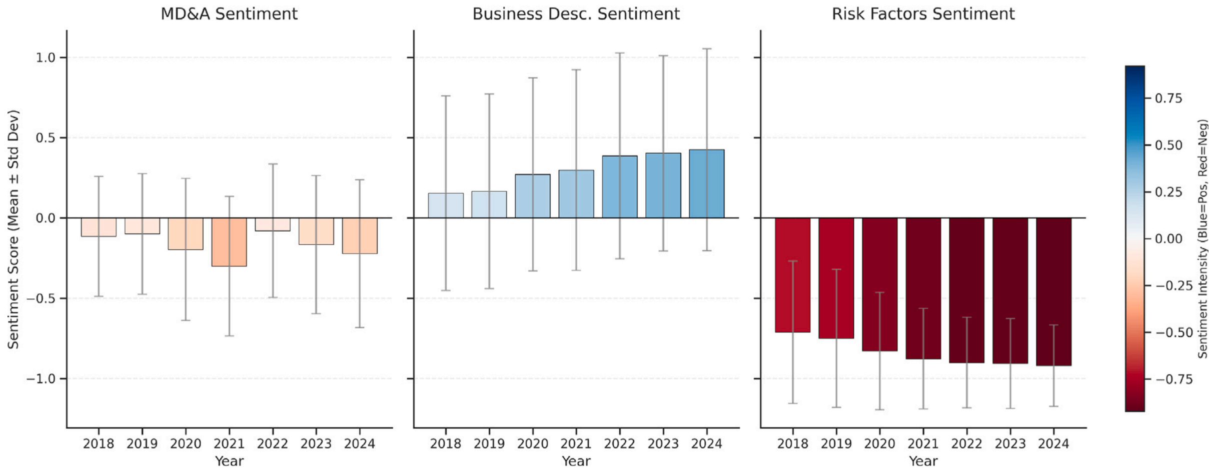

Fig. 2. Distribution of sentiment score per 10-K text segment.

3. Results

3.1. Descriptive statistics

Fig. 1 displays the average AI detection scores across the Management Discussion and Analysis (MD), Business Description (BD), and Risk Factors (RF) sections of 10-K reports from 2018 to 2024. The overall probability of AI use remained relatively steady during this period, with the Business Description section consistently showing the highest AI usage among the three. Although the averages are stable, a minor peak occurred across all sections in 2021. Additionally, the widening standard error bands in 2024 indicate increased variation in the MD and BD sections. This implies that while average adoption did not change significantly, certain firms increased their AI-assisted writing in these sections recently.

To formally check for a possible structural break of AI usage around 2022, which is the first year where AI became popular, we perform a simple Chow test for each of the sections of the 10-K documents. We use the averages probabilities of AI usage over the years as the target variable, with a simple intercept only model ($y_t = \alpha + \epsilon_t$). The results in Table 3 show that, even for a small sample of 7 observations (2018–2024) for each type of text, we still get some evidence of structural break, with p-values ranging from 2.29% (RF) to 15.17% (MD). This indicates that the data partially corroborates with a higher usage of AI after the ChatGPT release in 2022.

In Fig. 2 we present the average sentiment per document and year. First, we see that, as expected, the average RF sentiment is far lower than for the MD and BD. This is justified by the purpose of this sections, which is to alert potential investors about all risks of the company. Likewise, Business Description presents a more positive tone, which is also explained by its nature, to convince possible investors about the feasibility of the business, usually on a suggestive and positive tone.

Table 3
Results of structural break test (Chow).

Type text	AI prob Before (mean)	AI prob After (mean)	Statistic	p-value
Business Description (BD)	16.85%	18.78%	4.501	10.12%
Risk Factors (RF)	12.16%	13.56%	12.923	2.29%
Management Discussion (MD)	15.44%	15.86%	3.128	15.17%

Notes: This table reports the results of using a Chow test for structural break in the AI probabilities over the years with a simple intercept-only model. The data is aggregate by calculating the mean value over the years. We use 2022 as the break-point in the test.

Table 4
Estimation results for GLM model (Eq. (2)).

	Business Description	Risk Factors	Management Discussion
ROE	−0.027 (0.019)	−0.001 (0.016)	−0.004 (0.020)
Age	0.003 (0.001)***	−0.003 (0.001)***	0.002 (0.001)*
Log size	−0.077 (0.026)***	0.153 (0.021)***	0.024 (0.026)
Risk (Beta)	0.026 (0.030)	0.065 (0.025)**	0.014 (0.031)
Leverage	−0.024 (0.056)	0.008 (0.047)	−0.127 (0.058)**
Sentiment	−0.016 (0.019)	0.196 (0.089)**	−0.025 (0.027)
Asset turnover	0.018 (0.025)	0.056 (0.021)***	0.049 (0.026)*
Accrual ratio	−0.739 (0.203)***	−0.314 (0.170)*	−0.299 (0.211)
Year	0.038 (0.014)***	0.048 (0.012)***	0.023 (0.015)
Year effect	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes
Deviance	115.543	63.518	104.962
Num. obs.	2409	2409	2409

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

3.2. Econometric results

Table 4 presents the results from the estimation of Eq. (2) for each of the 10-K sections. We estimate the model using data only from 2022 to 2024, which is the period where AI use was first introduced. The financial performance of firms (ROE) is the only variable that shows no statistically significant relationship with AI use in any section, thus rejecting the hypothesis that more profitable firms are more likely to adopt AI-assisted writing. Firm age exhibits a significant positive association with AI use in the *Business Description* and *Management Discussion* sections, while showing a significant negative relationship in the *Risk Factors* section partially rejecting the hypothesis that younger firms are more prone to using AI in their filling writing. Additionally, firm size negatively predicts AI utilization in the *Business Description* but is a significant positive predictor for the *Risk Factors* section.

AI-assisted writing in the BD section is also negatively predicted by the accrual ratios of the firms, which are likewise negatively associated with the use of AI in the *Risk Factors* section, thereby rejecting the hypothesis that firms with higher accruals could rely on AI to assist in detailed technical explanations. Contrary to the accrual ratio, AI use in the RF section is positively associated with firms' risk and asset turnover, as predicted. Asset turnover also positively affects the use of AI in the *Management Discussion*, while leverage has the opposite effect, contrary to our expectations.

Additionally, we also find that the coefficient for *Sentiment* is a positive predictor of AI use. That is, reports with more positive sentiment scores correlates to a significantly higher probability of AI generation in the *Risk Factors* section. A possible explanation is that, in a section whose purpose is to alert potential investors to all company risks, when the news is good and the sentiment is positive, managers are more likely to use AI to produce more professional, optimistic, or neutral-positive corporate prose, as we formulated in our hypothesis referring to sentiment.

4. Discussion of results

Our empirical analysis reveals that, while the aggregate adoption of AI in annual reports has remained relatively stable between 2018 and 2024, there is a distinct heterogeneity in usage across different report sections. The Business Description (BD) section consistently exhibits the highest probability of AI-assisted writing. This section describes the company's main operations and, as such, the manual work could be optimized using AI. This finding align with the perception that AI enhances the efficiency of management activities, automation in management activities and optimizes management processes (Li, 2025). This result also highlights a critical boundary in current AI assistance: while efficiency is desirable, the potential for inaccuracies or generic disclosures in risk reporting (De Villiers et al., 2024) likely deters full automation in high-stakes areas where precise legal language is paramount such as MD&A and *Risk Factors*.

Regarding firm characteristics, while accrual ratio and leverage are negatively associated to the AI-assisted writing in some 10-K sections, asset turnover and risk have the opposite effect. Moreover, results on firm age and size are not conclusive considering the sometimes positive and sometimes negative relationships found depending on the 10-K section. For example, older firms use AI in the *Business Description* and *Management Discussion* sections while younger firms use AI in the *Risk Factors* section. Although contrary to our expectations, Yang and Yang (2025) found mature firms more susceptible to AI adoption. Regarding firm size, bigger firms use AI to assist them in the writing of *Risk Factors* and are less prone to use AI to write *Business Description*.

A interesting finding of this study is the positive association between sentiment scores and AI probability, particularly within the Risk Factors section. This implies that managers may be utilizing LLMs not merely for text generation, but as a strategic mechanism for impression management to "polish" the tone of disclosures. By leveraging the tendency of LLMs to produce neutral-to-positive professional prose (Huang et al., 2023), firms might benefit from AI-assisted writing in a section where "bad news" is common, whereas "good news" might seem weightless. This reinforces the perception that AI adoption can improve disclosure quality instead of compromising it (Yang and Zhou, 2025). However, we point out that, the correlation does not necessarily imply causality, and this should be taken into consideration in the interpretation of our results.

Collectively, these results portray a nuanced landscape of AI integration in financial reporting. While widespread, unchecked automation is not yet the norm, specific subsets of firms are leveraging these tools to reshape the narrative tone and efficiency of their filings. The strategic distinction in usage across report sections suggests that managers are deploying AI where automation can lead to efficiency while exercising caution where legal precision is critical. As noted by Yang et al. (2024), this evolution necessitates that auditors and regulators adapt their monitoring frameworks to effectively detect and evaluate machine-generated nuances in financial disclosures.

5. Conclusion

This study aimed to examine the prevalence and determinants of AI-assisted writing within the annual reports of Russell 1000 companies. We estimate and use a custom, FinBERT based, large language model to detect probabilities of AI-generated content and FinBERT to assess sentiment across three key sections: Management Discussion, Business Description, and Risk Factors.

We investigate the presence and determinants of AI-assisted writing in companies' annual reports. Using LLM-based detection and sentiment analysis, we provide evidence that AI usage is not uniformly distributed across disclosure sections. Furthermore, AI usage is systematically associated with firm characteristics. We conclude that companies are more likely to employ AI in narrative sections where standardization and presentation are more salient, such as the Business Description text in 10-Ks, while its use in more sensitive disclosures, such as Risk Factors, follows a different pattern. Moreover, firm fundamentals – including age, size, risk exposure, and operational efficiency – help explain cross-sectional variation in AI-assisted writing.

Another interesting piece of evidence in our paper is the positive association between AI detection and sentiment, particularly in risk-related disclosures. This finding suggests that AI may be used not only to improve readability but also to shape the tone of corporate communication. However, as discussed before, the used methodology cannot rule out the case of reverse causality. Such a gap motivates further studies with a more specific methodology to deal with the causality issue such as the use of an instrumental variable or a quasi-experimental research design.

Overall, our results show that AI is already embedded in corporate reporting practices in a non-random and economically meaningful way. This has direct implications for investors and regulators. Because AI-assisted writing can alter how corporate information is framed and interpreted, there is a risk of exacerbating information asymmetry and undermining market discipline. Furthermore, while AI might improve readability, it can also produce overly general disclosures that increase ambiguity and reduce the overall informativeness of financial reports. Consequently, policymakers and regulators must adapt their monitoring frameworks to effectively detect and evaluate these machine-generated nuances in financial disclosures.

CRediT authorship contribution statement

Marcelo S. Perlin: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cristian R. Foguesatto:** Writing – review & editing, Writing – original draft, Visualization. **Aliki K. Galanos:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation. **Felipe Affonso:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology.

Acknowledgment

Cristian R. Foguesatto gratefully acknowledges the financial support provided by FAPERGS (Grant No. 25/2551-0000831-0).

Appendix. Details on AI detection model estimation

This document outlines the methodological steps undertaken to retrieve, clean, and process the dataset used for training an AI text detection model on corporate financial documents. We use the base model of *yiyanghkust/finbert-pretrain* (Yang et al., 2020) to fine tune the model to our needs. The training data is built using three different and popular AI tools (Gemini, Chatgpt and Claude). We achieve an overall accuracy of approximately 89% in the testing data. This documents describes the details in the estimation of the model.

A.1. Data retrieval

We gather textual data for the *Management Discussion* (MD) section of 10-Ks of all companies part of the Russell 1000 index, between 2015 and 2020. We decided to use only the MD section, and not *Risk factors* (RF) or *Business description* (BD) as these two may not change much in between the years.[2] The MD section, on the other hand, describes the performance of past fiscal year and much richer in diversified text. Also, The choice of the time period, 2015 to 2020, is not accidental. In this time period AI models were not yet available to the public. We label this text data as *Human written*.

While we strongly suspect that text written in the time period between 2015 to 2020 are not written by AI, we cannot prove or accurately confirm it. Some forms of automated text generation system, such as, online proof readers and template-based-system were already available at this time. Therefore, it is expected some form of noise and contamination in the training data. Nonetheless, we confirm that, despite this problem, the model presented a acceptable performance in the out of sample test data.

One of the problems at this stage is that the underlying fine-tuned model (*yiyanghkust/finbert-pretrain*) process a maximum of 512 tokens, which is far lower than the size of the MD section of each 10-K. To solve this issue, we first split all texts into chunks of 500 tokens. We then randomly sampled 3000 text chunks across all companies and years, which are then used in the next stage. In the sampling process, we ignore empty chunks of text and those with less than 150 tokens.

A.2. Artificial text generation

To construct the "AI-generated" counterpart class, we used the randomly selected chunks of text from previous stage and provided it as context prompts to state-of-the-art Large Language Models. Specifically, we utilized Google's Gemini (*gemini-2.5-flash*), OpenAI's ChatGPT (*gpt-5.4-nano*), and Anthropic's Claude (*claude-haiku-4-5*). The LLMs were tasked with regenerating or rewriting the financial texts at a controlled temperature of 0.7, yielding comparable synthetic financial narratives. The same prompt was used in all LLMs:

```
# Context
You are an expert corporate lawyer and CFO. Your task is to
paraphrase a section from a 10-K filing for {{display_name}} (dated {{year}}).

# Instructions
1. Rewrite the text entirely, improving readability by enhancing
sentence structures and vocabulary.
2. Maintain the formal, legalistic, and dense tone of
a standard SEC filing.
3. Strictly preserve all original facts, financial figures,
and semantic meaning. Do not add external information or hallucinate data.
4. Ensure the output is approximately the same length as the input.

# Input Text
{{text}}

# Expected Output
Output ONLY a valid JSON object matching the exact structure below.
Do not include markdown formatting blocks (like ‘‘‘json).

{
 "new_text": "your rewritten text here"
}
```

A.3. Data cleaning and preparation

In order to have a quality training data, we use a simple text cleaning pipeline to enhance training stability and mitigate confounding noise. We used the following steps to clean the textual data:

(i) Whitespace standardization to replace multiple spacing characters with a single space.
(ii) Removal of typical conversion pagination artifacts (e.g., regex filtering for "Page X of Y").
(iii) A strict character boundary that filters out non-ASCII symbols, retaining only standard English alphanumeric syntax and a select subset of prevalent currency signs and accented Latin characters.

[2] We also performed a robustness check by calculating the performance of the prediction model in the MD and RF section of a different dataset than the one used in the estimation. Results showed a very comparable and high accuracy, indicating that the model performs well in other parts of the 10-K, besides the MD section.

Table A.5
Performance metrics of the fine-tuned FinBERT AI detector.

Metric	Value
Accuracy	89.17%
F1-Score	88.58%
Precision	92.65%
Recall	84.85%

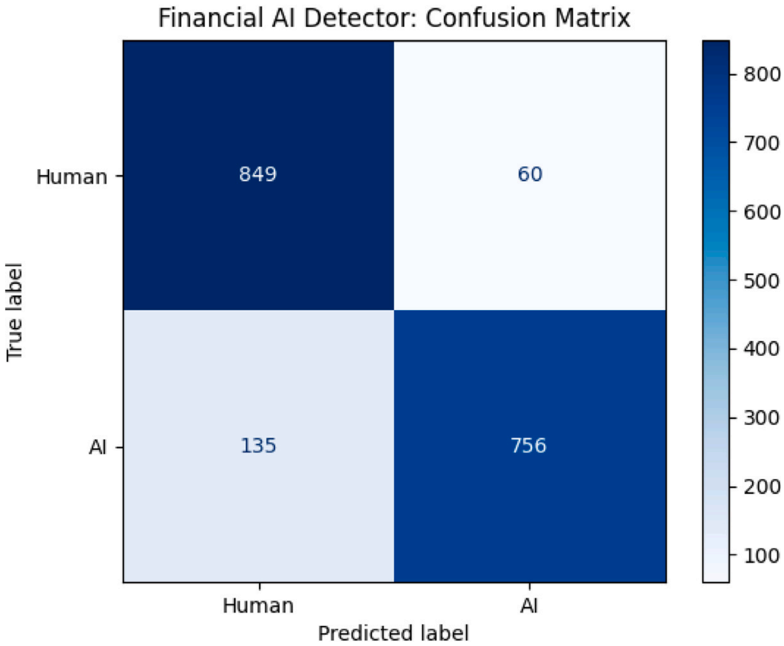

Fig. A.3. Confusion matrix evaluated on the test set.

After applying the sequence normalization, texts containing missing values (`NA`) were entirely dropped from the dataset. Subsequently, a length-based filter was enforced, discarding any textual sequence shorter than 10 characters to rule out uninformative fragments.

A.4. Sampling and dataset alignment

Once previous stage was complete, we have 3000 chunks of text from the raw, human written, 10-K files, and 3000 chunks of text written by the three AI models: Gemini, Chatgpt and Claude. In the final step of data preparation, the sampled dataset was divided: 70% of the observations were allocated to the training set for model estimation, while the remaining 30% were reserved as an out-of-sample testing set to validate the model's accuracy on unseen data.

A.5. Model estimation

We formulated the text detection problem as a binary sequence classification task. Our approach leverages a pre-trained language model, specifically FinBERT (`yiyanghkust/finbert-pretrain`) (Huang et al., 2023; Yang et al., 2020), which provides robust contextualized representations fine-tuned initially on a massive corpus of financial documents.

A.6. Model performance

The trained model was evaluated against the held-out 30% test dataset to gauge its generalization capability on unseen financial texts. The classifier yielded robust performance across multiple metrics, detailed in Table A.5.

The classification performance is further illustrated by the confusion matrix in Fig. A.3, which captures the true positive, true negative, false positive, and false negative predictions. The high precision underscores the model's reliability in flagging AI-generated text with minimal false accusations.

Data availability

Data will be made available on request.

References

Bonsón, E., Bednárová, M., Perea, D., 2023. Disclosures about algorithmic decision making in the corporate reports of western European companies. Int. J. Account. Inf. Syst. 48, 100596.
Brynjolfsson, E., Rock, D., Syverson, C., 2017. Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. Technical Report, National Bureau of Economic Research.
Csere, P., Estep, C., Griffith, E.E., MacKenzie, N.L., 2025. Financial executives' responses to the use of artificial intelligence in financial reporting and auditing. Curr. Issues Audit. 1–9.
de Jong, P., 2025. Do risk warnings predict capital raising? Evidence from 10-k disclosures and subsequent 8-k filings. Finance Res. Lett. 108000.
De Villiers, C., Dimes, R., Molinari, M., 2024. How will AI text generation and processing impact sustainability reporting? Critical analysis, a conceptual framework and avenues for future research. Sustain. Account. Manag. Policy J. 15 (1), 96–118.
Ding, M., Zhou, C., Yang, H., Tang, J., 2020. Cogltx: Applying bert to long texts. Adv. Neural Inf. Process. Syst. 33, 12792–12804.
Fengler, M.R., Phan, T.M., 2025. Unveiling themes in 10-k disclosures: A new topic modeling perspective. Int. Rev. Finance Anal. 103, 104121.
Hasan, M.M., 2020. Readability of narrative disclosures in 10-k reports: does managerial ability matter? Eur. Account. Rev. 29 (1), 147–168.
Huang, A.H., Wang, H., Yang, Y., 2023. FinBERT: A large language model for extracting information from financial text. Contemp. Account. Res. 40 (2), 806–841.
Li, J., 2025. Artificial intelligence innovation and financial report quality. Int. Rev. Econ. Finance 104832.
Liu, Y., Hu, R., 2025. Artificial intelligence, supply chain finance, and corporate investment efficiency. Finance Res. Lett. 107851.
Nguyen, P.T.T., Kimura, A., 2023. Disclosure characteristics of annual reports and information asymmetry: Evidence from foreign firms listed on the US stock exchange. Finance Res. Lett. 54, 103776.
Protogerou, A., Caloghirou, Y., Vonortas, N.S., 2017. Determinants of young firms' innovative performance: Empirical evidence from europe. Res. Policy 46 (7), 1312–1326. http://dx.doi.org/10.1016/j.respol.2017.05.011, URL: https://www.sciencedirect.com/science/article/pii/S004873331730094X.
Shiyyab, F.S., Alzoubi, A.B., Obidat, Q.M., Alshurafat, H., 2023. The impact of artificial intelligence disclosure on financial performance. Int. J. Finance Stud. 11 (3), http://dx.doi.org/10.3390/ijfs11030115, URL: https://www.mdpi.com/2227-7072/11/3/115.
Türegün, N., 2019. Impact of technology in financial reporting: The case of amazon go. J. Corp. Account. Finance 30 (3), 90–95.
Wang, Z., Ng, P., Ma, X., Nallapati, R., Xiang, B., 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. arXiv preprint arXiv:1908.08167.
Weng, M., 2025. Green innovation through artificial intelligence technology: Enhancing environmental, social, and governance performance. Finance Res. Lett. 75, 106921.
Yang, Y., Uy, M.C.S., Huang, A., 2020. Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.
Yang, G., Yang, X., 2025. AI adoption and ESG performance: Evidence from China. Int. Rev. Econ. Finance 104659.
Yang, X., Zang, S., Ren, Y., Peng, D., Wen, Z., 2024. Evaluating large language models on financial report summarization: An empirical study. arXiv preprint arXiv:2411.06852.
Yang, S., Zhou, N., 2025. Artificial intelligence and the quality of corporate accounting information disclosure. Finance Res. Lett. 108136.
Zhu, X., Li, J., Wang, Y., 2024. Are risk disclosures in financial reports informative? A text mining-based perspective. Humanit. Soc. Sci. Commun. 11 (1), 1–18.