Large Language Model-Based Analysis of Corporate Risk Narratives for Investment Decision-Making

Bohaoran Zou
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
bohaoranmail@binghamton.edu

Rowan L. Burton
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
rowan1989@oregonstate.edu

Abstract

The increasing availability of unstructured corporate disclosures has created both opportunities and challenges for investment decision-making. This paper presents a system-level examination of large language model-based analysis of corporate risk narratives, focusing on architectural trade-offs, governance mechanisms, deployment infrastructure, robustness, fairness, and policy implications. Traditional textual analysis methods rely on lexicons and shallow classifiers, but recent advances in transfer learning and generative language models enable more context-sensitive interpretation of risk factors, management discussion, and forward-looking statements. The paper argues that effective deployment requires more than predictive accuracy: it demands careful corpus construction, continuous monitoring, explainable decision support, and alignment with regulatory expectations. It discusses how these models can be integrated into investment workflows without displacing human judgment, and it identifies structural risks including distributional drift, semantic manipulation, opacity, and systemic homogeneity. Through cross-domain comparisons with credit risk, regulatory supervision, and medical decision support, the paper highlights the need for institutional governance, auditability, and sustainability. The discussion extends to emerging policy questions such as disclosure integrity, model assurance, and market efficiency. The paper concludes that large language models should be treated as socio-technical infrastructural components whose value depends on organizational oversight, incentive alignment, and the capacity to detect anomalies in narrative structures rather than on isolated performance benchmarks.

Keywords

large language models; corporate risk narratives; investment decision-making; financial text analysis; governance; explainable artificial intelligence; systemic risk.

1. Introduction

Corporate risk narratives have long been recognized as important information sources for investment decision-making, yet the extraction of reliable signals from unstructured text remains difficult. Early research demonstrated that word choice in mandatory filings conveys economically meaningful information beyond quantitative accounting variables [1]. In particular, forward-looking statements and management tone have been shown to predict volatility, earnings surprises, and market reactions [2]. Media sentiment also influences short-term returns and can distort price discovery [3]. These findings established a foundation for

treating text as data, but early methods relied on dictionaries and regression-based classifiers that often struggled with contextual nuance, negation, and domain-specific language.

The evolution from static lexicons toward machine learning has enabled more flexible representations of textual content. Word weighting approaches and supervised models improved the measurement of sentiment in analyst reports and corporate communications [4]. Further advances in language modeling introduced contextual embeddings that capture word meaning as a function of surrounding discourse [7, 8]. The subsequent emergence of large language models has shifted the focus from feature engineering to transfer learning and instruction following [9, 10]. In financial applications, domain-adapted language models have refined sentiment classification and risk disclosure analysis [11]. Deep neural networks have also been used to support decision-making from financial disclosures through transfer learning with limited labeled data [12]. Text-as-data frameworks provide rigorous foundations for causal inference and measurement, but they also caution that machine learning models may reproduce existing biases or conflate sentiment with information content [13]. Systemic risk research has further shown that firm-level disclosures can signal interconnected vulnerabilities, although converting narrative signals into systemic measures remains an open challenge [14]. Recent work on semantic anomaly detection in risk factor disclosures has begun to focus not only on word frequency but also on structural deviation from normal narrative patterns, using explainable artificial intelligence to highlight anomalous clause relationships in 10-K filings [15].

This paper examines large language model-based analysis of corporate risk narratives from a system-level perspective. It considers the structural trade-offs in architecture, data governance, infrastructure, interpretability, and policy. The objective is not to propose a single model or algorithm but to analyze how these systems operate within institutional and regulatory environments. Such a perspective is necessary because investment decisions are not isolated classification tasks; they are embedded in complex workflows involving legal review, risk management, portfolio construction, and fiduciary obligations. Moreover, the widespread adoption of language models may alter the very narrative landscape they are designed to parse, creating feedback loops that require continuous monitoring and adaptation.

2. Background and Related Work

The academic literature on financial text analysis has progressed from manual coding to large-scale computational methods. Loughran and McDonald showed that general-purpose dictionaries misclassify many words in financial filings, motivating the development of domain-specific lexicons [1]. Li applied a naive Bayesian approach to forward-looking statements and found that tone predicts future earnings and liquidity [2]. Tetlock demonstrated that media pessimism forecast downward pressure on stock prices, followed by reversion, which suggested that linguistic measures capture investor sentiment rather than only fundamental information [3]. These studies illustrate a recurring theme: text contains multiple overlapping signals, including factual disclosure, managerial impression, and market sentiment.

Subsequent research improved measurement by treating documents as networks and by exploiting industry peers. Hoberg and Phillips constructed text-based industry classifications and showed that product descriptions reveal competitive dynamics absent from standard industry codes [6]. Kogan and colleagues used regression models on annual reports to predict risk, emphasizing the importance of large training corpora and the challenge of rare events [5]. Jegadeesh and Wu applied word power measures to assess the economic weight of terms in

financial documents [4]. These contributions established that textual analysis can complement traditional risk models, but they also revealed limitations related to interpretability, data bias, and sensitivity to document structure.

The introduction of transformer architectures enabled contextual representation at scale [8]. Pre-trained bidirectional models such as BERT have been adapted for financial text through additional domain pretraining, resulting in improved performance on sentiment and risk classification tasks [7, 11]. Generative models can now perform few-shot extraction, summarization, and question answering over lengthy disclosures [9, 10]. In financial settings, deep neural networks have been used to support decision-making from corporate filings [12]. Text-as-data frameworks provide rigorous foundations for measurement but caution that machine learning models may reproduce existing biases [13]. Systemic risk research has shown that narrative indicators may complement market-based measures of interconnectedness [14]. More recently, semantic anomaly detection approaches have directed attention toward structural deviations in risk factor disclosures, using explainable artificial intelligence to identify unusual clause relationships rather than isolated term frequencies [15].

3. Architectural Considerations for Language Model Systems

A central design choice in large language model-based analysis is the trade-off between general-purpose models, domain-adapted models, and hybrid architectures. General-purpose models benefit from broad world knowledge, but they may lack sensitivity to legal and accounting terminology. Domain adaptation through continued pretraining can align representations with financial discourse, yet it introduces costs related to data curation, compute, and validation. Fine-tuning for specific tasks such as risk factor classification can improve accuracy on narrow benchmarks, but it may reduce robustness when disclosures evolve. Hybrid architectures that combine a frozen language model with lightweight task-specific heads offer scalability and modularity, while multi-stage systems can separate retrieval, extraction, scoring, and explanation.

Another architectural consideration is the granularity of analysis. Corporate risk narratives appear in multiple sections, including risk factors, management discussion and analysis, legal proceedings, and forward-looking statements. Each section has distinct textual conventions and regulatory constraints. A system designed only for sentiment classification may fail to capture conditional statements, hedging, temporal dependencies, or cross-references. Therefore, document-level and clause-level representations must be jointly considered. Long-context transformer models allow more complete document representations, but they incur computational costs and may dilute local anomalies. Hierarchical models can first encode sentences or clauses and then aggregate them, preserving both local and global structure. The choice of architecture also affects inference latency, which matters for real-time investment workflows but less for quarterly report screening.

A third architectural issue concerns the encoding of uncertainty. Investment decisions require not only point predictions but also calibrated confidence and acknowledgment of ambiguity. Large language models can generate fluent outputs that appear authoritative even when the underlying evidence is weak. This property is particularly dangerous in high-stakes financial contexts. Architectural mechanisms such as retrieval-augmented generation, constrained decoding, and explicit evidential scoring can help anchor outputs to source text. However, these mechanisms introduce additional components and failure modes. Retrieval may surface irrelevant passages, and constrained decoding may reject legitimate alternative interpretations. System designers must therefore balance expressiveness, control, and auditability.

4. Corpus Construction and Data Governance

The quality of a language model system depends heavily on the construction and maintenance of its training and evaluation corpora. Corporate disclosures are heterogeneous across jurisdictions, industries, and time periods. A corpus that overrepresents large U.S. technology firms may not transfer well to small-cap financial institutions or non-U.S. issuers. In addition, regulatory changes alter disclosure formats and language conventions. For instance, the adoption of plain English rules and the expansion of risk factor requirements have changed the structure of annual reports over time. A governance framework must therefore address collection, versioning, cleaning, and deduplication while preserving source provenance.

Data governance also involves access control and confidentiality. Some corporate narratives may contain material non-public information when processed before public release, creating legal risks if systems are used inappropriately. Even after public release, the use of disclosures for trading may be subject to ethical and regulatory constraints. Institutions must implement role-based access, audit logs, and lineage tracking to ensure that model outputs can be traced to specific documents and versions. Without such infrastructure, a language model system may become an opaque component in a compliance chain, making it difficult to demonstrate due diligence to regulators.

Another governance challenge is the management of labels. Supervised training often requires human annotations of risk categories, sentiment, or materiality. These labels are expensive, noisy, and subject to disagreement. Different annotators may interpret the same sentence differently depending on their professional background. Crowdworkers may lack financial expertise, while expert analysts may introduce strategic biases. Active learning and weak supervision can reduce labeling costs, but they also introduce new forms of sample selection bias. A robust governance framework must document label definitions, annotator qualifications, inter-annotator agreement, and dispute resolution procedures. These practices are increasingly important as regulators and auditors demand explanations for algorithmic investment tools.

5. Deployment Infrastructure and System Integration

Deploying large language model-based analysis in an investment organization requires more than a model endpoint. The system must be integrated with data ingestion pipelines, document repositories, market data systems, and portfolio management interfaces. Latency, throughput, and availability requirements differ by use case. A pre-trade compliance check may require sub-second response, while a quarterly risk review can tolerate batch processing. The infrastructure must also support model versioning, rollback, and shadow evaluation. When a new model version is deployed, its outputs should be compared with the previous version on historical data before it influences live decisions.

Operational sustainability depends on monitoring distributional drift. Corporate risk narratives change over time due to economic cycles, regulatory reforms, and changes in issuer behavior. A model trained on pre-pandemic disclosures may misinterpret language introduced during supply chain disruptions or geopolitical crises. Continuous monitoring should include statistical drift detectors, human review of high-impact outputs, and periodic retraining schedules. However, frequent retraining can introduce instability if not governed carefully. Model registries, feature stores, and pipeline orchestration tools can provide the necessary control, but they also add operational complexity.

System integration also raises questions about interoperability. Investment firms often use multiple data vendors, analytics platforms, and compliance systems. Language model outputs must be represented in standardized formats that can be consumed by downstream risk models and decision support interfaces. The design of these formats affects whether outputs are treated as evidence, signals, or recommendations. Clear distinction between raw extracted text, scored features, explanations, and generated summaries is essential to avoid conflation of fact and interpretation. Interfaces should present model outputs with confidence indicators, source citations, and comparison baselines so that human analysts can interrogate them effectively.

6. Robustness, Fairness, and Interpretability

Robustness in corporate risk narrative analysis means that model outputs remain reliable under adversarial and naturally occurring perturbations. Companies may intentionally alter the wording of risk disclosures to obscure negative information while remaining technically compliant. Language models can be sensitive to small paraphrases, section reordering, or the insertion of boilerplate text. If a system overweights lexical cues, issuers may learn to manipulate those cues, reducing signal quality. Robustness therefore requires adversarial evaluation, stress testing, and detection of semantic anomalies that diverge from expected structural patterns [15]. This perspective aligns with broader concerns about the vulnerability of large language models to distribution shift and spurious correlations [16]. Foundation models, in particular, introduce new failure modes because their capabilities emerge from massive pretraining corpora that are difficult to audit comprehensively [17].

Fairness is particularly complex in investment settings because it may refer to different normative standards. A model should not systematically disadvantage particular issuers, industries, or investor groups due to biased training data. For example, if risk narratives from smaller companies use different vocabulary than those from large firms, a model may overestimate risk for small issuers simply because their language patterns are underrepresented. Similarly, models trained primarily on U.S. securities may misread cross-border disclosures. Fairness audits should examine subgroup performance, error rate disparities, and distributional impacts. However, fairness cannot be reduced solely to statistical parity because some differences in risk language are economically meaningful. Governance mechanisms must distinguish between legitimate variation and unjustified discrimination.

Interpretability and explainability are necessary for institutional trust. Investment professionals must be able to understand why a system flags a particular risk narrative as concerning. Techniques such as local explanation models can highlight influential words or clauses, but these explanations may be unstable or misleading when applied to deep transformers [19]. Model cards and structured documentation provide a higher-level mechanism for communicating intended use, limitations, and evaluation results [18]. Explainable artificial intelligence is particularly relevant for regulatory disclosure analysis because anomalous text may indicate emerging risks that lack established categories. A system that merely outputs a risk score without explaining the underlying semantic deviation is unlikely to satisfy fiduciary or regulatory expectations.

7. Investment Decision Integration and Human Oversight

Language model outputs should be treated as inputs to a broader investment decision process rather than as autonomous recommendations. Effective integration requires a clear division of

labor between algorithmic screening and human judgment. Models can scale the reading of thousands of documents, identify unusual disclosures, and generate structured comparisons, while analysts can assess context, verify sources, and evaluate strategic implications. To support this division, system outputs should be embedded in interactive interfaces that allow exploration of source passages, historical comparisons, and alternative scenarios. Decision records should capture which model outputs were used, how they were interpreted, and what human actions followed.

The introduction of language models into investment workflows changes organizational roles and accountability structures. Compliance teams must review not only the final investment decision but also the tools used to inform it. Portfolio managers must understand model limitations and resist automation bias, in which operators overtrust algorithmic outputs. Training programs and certification processes can help build appropriate competencies. At the same time, organizations should avoid excessive reliance on explanation as a substitute for validation. A plausible explanation may accompany an incorrect prediction, and the presence of an explanation can falsely increase user confidence. Oversight mechanisms should therefore include independent validation, red teaming, and post-trade review.

Human oversight is also needed to address strategic feedback loops. If many market participants use similar language models to interpret disclosures, their reactions may become correlated, potentially amplifying volatility and reducing the diversity of market views. This systemic homogeneity is a concern for financial stability because concentrated model behavior can create crowded trades and liquidity shocks [14]. To mitigate this risk, institutions can maintain ensembles of diverse models, incorporate non-textual data, and impose limits on the speed and size of automated responses to narrative signals. The governance challenge is to balance informational efficiency with market resilience.

8. Policy, Regulatory, and Sustainability Implications

The use of large language models in investment decision-making intersects with securities regulation, data protection, and artificial intelligence governance. Regulators may require firms to demonstrate that algorithmic tools are accurate, non-deceptive, and consistent with fair dealing obligations. In the United States, the Securities and Exchange Commission has expressed concern about the use of predictive data analytics in broker-dealer and investment adviser interactions. In Europe, the Artificial Intelligence Act and the General Data Protection Regulation impose transparency and risk management requirements on high-risk systems. Although corporate disclosures are generally public, the processing of personal data within them and the use of models for consequential financial decisions may trigger additional obligations.

A key policy question is how to ensure disclosure integrity when language models can be used by issuers to draft narratives that exploit model weaknesses. If risk factor text is optimized to appear benign to automated readers while remaining technically compliant, the information environment may degrade. Regulators may need to update disclosure standards to require substantively meaningful risk communication, not just compliance with formal language. Similarly, audit firms may need new methods to assess whether management narratives have been generated or influenced by language models. Ethical auditing practices developed for other algorithmic domains provide useful lessons for constructing accountability mechanisms around financial language systems [20].

Sustainability considerations include environmental costs, institutional knowledge preservation, and long-term maintenance. Training and serving large models consume significant computational resources, and frequent retraining exacerbates energy use. Organizations should evaluate whether smaller domain-adapted models or distilled architectures can achieve acceptable performance with lower environmental impact. In addition, the sustainability of these systems depends on maintaining internal expertise. If a small number of vendors control the underlying models, investment firms may face lock-in and loss of analytical independence. Open-model ecosystems, structured evaluation benchmarks, and shared auditing practices can support a more sustainable and competitive landscape. At the same time, open models raise their own governance issues because their failure modes may be less documented than those of closed commercial systems.

9. Conclusion

This paper has examined large language model-based analysis of corporate risk narratives as a socio-technical system rather than an isolated modeling task. The effective use of these tools depends on architectural choices that preserve semantic context, corpus governance that addresses bias and versioning, deployment infrastructure that supports monitoring and integration, robustness against adversarial disclosure practices, fairness across issuers and markets, and interpretability that enables meaningful oversight. Investment decisions should remain subject to human judgment, and institutional processes must be designed to prevent automation bias and correlated model behavior. Policy frameworks must evolve to address disclosure integrity, transparency, and sustainability in an environment where language models are used by both issuers and investors. Future research should develop standardized evaluation protocols, examine the feedback effects of machine-readable disclosure optimization, and investigate how explainable anomaly detection can support regulators in identifying emerging narrative risks. By attending to these structural dimensions, the financial system can harness the analytical power of large language models while preserving the trust, accountability, and diversity of interpretation that underpin well-functioning markets.

References

1. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65.

2. Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of Accounting Research, 48(5), 1049–1102.

3. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168.

4. Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. Journal of Financial Economics, 110(3), 712–729.

5. Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference, 366–374.

6. Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5), 1423–1465.

7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.

8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems 30, 5998–6008.

9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901.

10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35, 27730–27744.

11. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

12. Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems, 104, 38–48.

13. Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574.

14. Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. (2017). Measuring systemic risk. Review of Financial Studies, 30(1), 2–47.

15. Sun, F., He, S., Wang, R., Ke, L., Shen, H., & Liao, Q. (2026). Modeling Structural Deviation in 10-K Risk Factors: A Semantic Anomaly Detection and Explainable AI Approach. Risks, 14(4), 87.

16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

17. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A. S., Creel, K., Davis, J., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.

19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

20. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 145–151.