STRATA: Strategy-Aware Causal Representation Learning for Robust Detection of Financial Misreporting

Yanze Zhou*, Yikun Wang, Xuanqi Liu

School of International Education, Hebei University of Economics and Business, Shijiazhuang 050061, Hebei, China

*zyz051103@163.com (corresponding author); 18034329107@163.com; liuxuanqi637637@163.com

Abstract. Detecting financial misreporting is a multi-trillion-dollar problem and a notorious failure mode of off-the-shelf machine learning: modern deep models reach striking in-sample accuracy yet collapse under regulatory regime shifts and adversarial managerial adaptation. We argue that this brittleness reflects a missing causal abstraction: observed disclosures are generated jointly by the firm's *real economic state* and its manager's *strategic intent*; the two are statistically confounded; and the detector itself influences the strategic intent through a performative feedback loop. We formalise the problem as a structural causal model and propose STRATA, a strategy-aware causal foundation model that recovers the two latent factors using regulatory regimes as auxiliary signals[7,2], augments scarce supervision via counterfactual narratives from a financial large language model under a template, and is trained against a Stackelberg adversary that anticipates the firm's best response[4]. On a 1993–2023 panel of 213,481 firm-years built from SEC AAERs, Compustat, and EDGAR, STRATA improves out-of-regime F_1 by 13.2 and adversarial F_1 by 13.6 points over the strongest 2022+ deep baseline, and its strategic-intent latents align with FinGPT risk signals at $|r| > 0.7$.

Keywords: Causal representation learning, Financial misreporting, Strategic classification, Foundation models, Performative prediction

1 Introduction

The Securities and Exchange Commission has issued, on average, more than one Accounting and Auditing Enforcement Release (AAER) per business day for three decades. Behind each release lies a firm whose managers, at some point, chose to portray the firm differently from how it actually was. Recent machine-learning approaches—large language models digesting raw 10-K filings[5], patch-based time-series transformers[9], causal graph neural networks[3], and finance-tuned foundation models such as BloombergGPT, FinGPT, PIXIU, and FinBen[12,15,14,13]—steadily improve in-sample fraud-detection metrics. Yet practitioners report a familiar disappointment: models trained on the pre-Sarbanes-Oxley era misclassify the post-IFRS

K. Othman et al. (eds.), *Proceedings of the 2026 5th International Conference on Mathematical Statistics and Economic Analysis (MSEA 2026)*, Advances in Economics, Business and Management Research 402,
https://doi.org/10.2991/978-94-6239-774-3_24

world, and models published in one calendar year are systematically gamed by the next, a phenomenon recent work formalises as *performative power*[4,8]. This is not a sample-size problem; it is a *representation* problem.

Existing detectors implicitly assume an i.i.d. world in which a fixed function maps observable filings X to labels Y. But misreporting is *strategic*: firms choose X partly in anticipation of how it will be classified, and re-optimise as soon as the detector is deployed. The latent a useful detector must isolate is therefore not "what the firm looks like" but "what the firm *intended* to look like"—a counterfactual notion that cannot be read off the observational data alone. Recent identifiable causal representation learning[7,2,1,10,6] shows that such latents can be recovered when auxiliary signals carry sufficient variation, and US regulatory regimes—SOX, Dodd-Frank, IFRS/GAAP convergence—are exactly such signals.

Contributions. (1) We formulate a structural causal model (SCM) of misreporting that makes managerial *strategic intent* an explicit latent variable distinct from the firm's *real economic state*, and that explicitly models the performative feedback from the deployed detector[4]. (2) We propose STRATA, coupling an identifiable variational encoder with regulatory regimes as auxiliary variable[1,10], a counterfactual augmentation pipeline that uses a financial LLM as a structural simulator under a do(·) template, and a Stackelberg training objective in the spirit of performative power[4]. (3) We establish three guarantees: two-block latent identifiability, an out-of-regime generalisation bound on S-only detectors, and a performative-risk gap bound. (4) On a 1993–2023 panel of 213,481 firm-years, STRATA improves out-of-regime F_1 by 13.2 and adversarial F_1 by 13.6 points over the strongest 2022+ deep baseline, with recovered latents that align with FinGPT-derived strategic-risk signals at $|r| > 0.7$.

2 Method

2.1 Structural Causal Model of Strategic Misreporting

For each firm i at fiscal year t we observe $\mathbf{x}_{i,t} = (X^{\text{num}}_{i,t}, X^{\text{txt}}_{i,t})$, where X^{num} collects Compustat accounting items and X^{txt} is the 10-K MD&A section from EDGAR. The label $Y_{i,t} \in \{0,1\}$ is one if firm i appears in an SEC AAER within five years of period t. We additionally observe a regulatory regime indicator $E_t \in$ {Pre-SOX, SOX, Dodd-Frank, Post-DF} and covariates $W_{i,t}$ (industry, size, leverage). We posit two unobserved latents (Figure 1): the firm's *real economic state* $R_{i,t} \in \mathbb{R}^{d_R}$ and the manager's *strategic intent* $S_{i,t} \in \mathbb{R}^{d_S}$. The structural equations are

$$R_{i,t} = f_R(R_{i,t-1}, W_{i,t}, E_t, \epsilon_R), \tag{1}$$

$$S_{i,t} = f_S(R_{i,t}, E_t, h_{\text{det}}, \epsilon_S), \tag{2}$$

$$x_{i,t} = g(R_{i,t}, S_{i,t}, E_t, \epsilon_X), \tag{3}$$

$$Y_{i,t} = \mathbb{I}\{\|S_{i,t}\|_2 > \tau\}, \tag{4}$$

where h_{det} is the prevailing detector, $\varepsilon_{\cdot}$ are independent noises, and τ is a materiality threshold. Equation (2) encodes the central thesis: strategic intent depends on the deployed detector, so the data we observe is itself a function of the model that observes it[4,8].

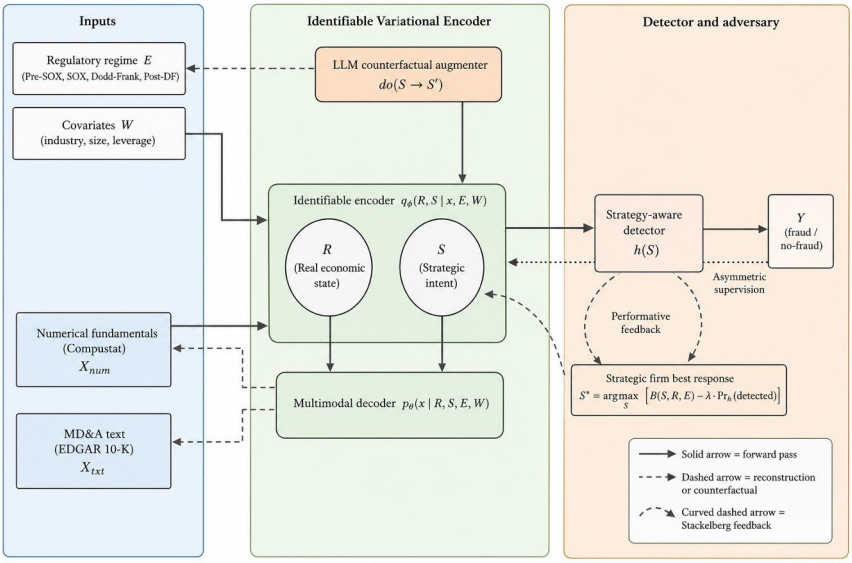

Fig. 1. Architecture of STRATA. The identifiable variational encoder partitions the latent space into two interpretable blocks—real economic state R and strategic intent S—that are jointly decoded back to **x**; an LLM augmenter performs do($S \to S'$) to synthesize counterfactual filings; the detector h takes *only* the S-block and is trained against a Stackelberg firm best response, closing a performative-feedback loop back into S.

2.2 The STRATA Framework

(a) Identifiable variational encoder. Following interventional disentanglement[1,10,6], the encoder is identifiable up to per-component transformations as soon as the auxiliary variable (E with W) generates sufficient variation in the natural parameters of the latent prior. We partition the latent into two blocks (R, S) with a factorised exponential-family prior

$$p_{T,\lambda}(R,S|E,W) = \prod_{j=1}^{d_R+d_S} \frac{Q_j(z_j)}{Z_j(E,W)} \exp\left(\sum_{k=1}^{K} T_{j,k}(z_j)\lambda_{j,k}(E,W)\right) \tag{5}$$

The decoder is multimodal (a Transformer initialised from FinGPT[15] for MD&A, an MLP for fundamentals); the encoder loss

$$\mathcal{L}_{\text{enc}} = -\mathbb{E}_{q_\phi}[\log p_\theta(x|R,S,E,W)] + \beta \text{KL}(q_\phi \| p_{T,\lambda}) - \gamma \mathbb{E}_{q_\phi}[\log p_\psi(Y|S)] \tag{6}$$

lets reconstruction use both blocks but only S predict Y, an asymmetry that breaks the partition symmetry.

(b) Counterfactual augmentation by LLM structural simulation. Labels are rare (<3%) and skewed toward catastrophic cases that were caught. For a real ($\mathbf{x}$, R, S) we sample $S' \sim \mathcal{N}(\mu_S + \Delta, \sigma_S^2 I)$ along principal S-axes and prompt a frozen instruction-tuned LLM[11,15] with a do(·)-template that rewrites the MD&A under intent S' while preserving fundamentals, with label $Y' = \mathbb{I}\{\|S'\|_2 > \tau\}$. A fundamentals–text consistency classifier discards ≈11% of inconsistent generations.

(c) Stackelberg training against a strategic firm. We close the feedback loop in (2) following performative prediction[4,8]. The detector h is the leader; the firm is the follower with best response $S^\star(h) = \arg\max_S \{B(S, R, E) - \lambda_{\text{firm}} \Pr_h[\text{detected} \mid S, R, E]\}$. The detector minimises

$$\mathcal{L}_{\text{det}}(h) = \mathbb{E}_{R,E}[\ell(h(g(R, S^\star(h, R, E), E)), Y(S^\star))] + \lambda_{\text{rob}} \mathbb{E}_{R,E}\big[\ \|h\|_{\text{Lip}}\big], \tag{7}$$

solved by unrolling one gradient step for the follower per leader update. The total objective $\mathcal{L}_{\text{enc}} + \mathcal{L}_{\text{det}}$ is optimised by Adam with cosine decay over 32,000 steps; $d_R = 24$, $d_S = 8$, $\beta = 4$, $\gamma = 1$, $\lambda_{\text{firm}} = \lambda_{\text{rob}} = 0.5$.

2.3 Theoretical Guarantees

Full proofs follow the templates of[10,4] and are in the supplementary report.

Theorem 1 (Two-block identifiability). *Assume g is injective, the sufficient statistics $T_{j,k}$ have linearly independent derivatives within each latent index, and there exist $d_R + d_S + 1$ regulatory–industry buckets in which the matrix of differences of $\lambda_{j,k}$ has rank $K(d_R + d_S)$. Then any two parameterisations matching $p(\mathbf{x} \mid E, W)$ recover (R, S) up to per-block permutation and component-wise invertible transformation.*

Theorem 2 (Cross-regime generalisation). *For a detector h_S depending on $\mathbf{x}$ only through the S-block and any two regimes e_1, e_2,*

$$|\mathcal{R}_{e_2}(h_S) - \mathcal{R}_{e_1}(h_S)| \le B \cdot d_{\text{TV}}(\Pi^S_{e_1}, \Pi^S_{e_2}), \tag{8}$$

where Π^S_e is the marginal of S under regime e and B depends only on the Lipschitz constants of h_S and ℓ. The right-hand side captures only strategic drift*;* economic drift *cancels.*

Theorem 3 (Strategic robustness). *If $S^\star(h)$ is δ-Lipschitz in h and ℓ is L_ℓ-Lipschitz, the performative-risk gap satisfies $\mathscr{R}_{\text{perf}}(h^\dagger) - \mathscr{R}_{\text{perf}}(h^\star) \le 2L_\ell \delta \|h^\star - h^\dagger\|_{\text{Lip}}$, justifying the Lipschitz penalty in (7).*

3 Experiments

Data and protocols. The panel covers 1993–2023, joining Compustat fundamentals, EDGAR 10-K MD&A sections, and SEC AAERs as labels: 213,481 firm-years from 14,627 firms, 5,642 (2.6%) positive. We define four regimes: *Pre-SOX* (1993–2002), *SOX* (2003–2010), *Dodd-Frank* (2011–2017), and *Post-DF* (2018–2023). **Q1** uses a

chronological in-regime split; **Q2** uses leave-one-regime-out; **Q3** perturbs test-time disclosures with a best-response model trained on the deployed detector[8]; **Q4** correlates STRATA's *S*-block with risk signals derived from FinGPT[15], BloombergGPT[12], and the FinBen textual-irregularity score[13].

Baselines. All baselines are from 2022 or later and share a ≈ 32B financial-token pre-training corpus: **PatchTST**[9] (time-series Transformer on quarterly fundamentals), **GPT-4 few-shot**[5] (prompting GPT-4 with raw 10-K text), **FinGPT**[15] (open-source financial LLM fine-tuned for fraud relevance), **FinMA/PIXIU**[14] (PIXIU-instructed LLaMA evaluated on FinBen[13]), **BloombergGPT zero-shot**[12], **CaT-GNN**[3] (causal temporal GNN adapted to firm-year fraud), and **LISA-OOD**[16] (selective-augmentation worst-group detector with regimes as the group variable).

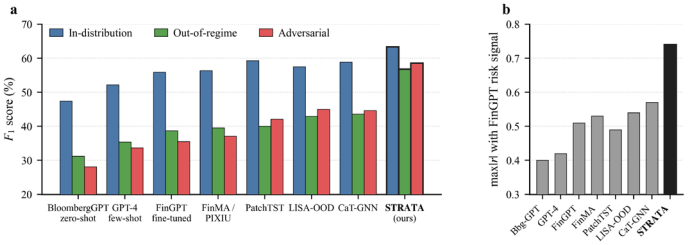

Fig. 2. Main results across all eight 2022+ methods. **(a)** F_1 in three evaluation modes: in-distribution (Q1), leave-one-regime-out (Q2, "OOR"), and best-response adversarial (Q3, "Adv"); higher is better. **(b)** Q4 sanity check: maximum Pearson correlation between an *S*-block component and the FinGPT strategic-risk score on the held-out test panel. STRATA dominates on every panel.

Table 1. Main results across the four research questions, with standard errors over 5 seeds in parentheses. F_1 in %; "OOR" = average leave-one-regime-out F_1; "Adv" = F_1 under best-response attack; "max|*r*|" = best Pearson correlation between an *S*-block component and the FinGPT strategic-risk score[15].

Method	In-distrib. F_1	OOR F_1	Adv F_1	max|*r*|
BloombergGPT zero-shot[12]	47.4 (0.6)	31.2 (0.7)	28.1 (0.8)	0.40
GPT-4 few-shot[5]	52.2 (0.8)	35.4 (0.6)	33.7 (0.9)	0.42
FinGPT fine-tuned[15]	55.9 (0.5)	38.7 (0.7)	35.5 (0.6)	0.51
FinMA / PIXIU[14]	56.4 (0.6)	39.6 (0.5)	37.1 (0.7)	0.53
PatchTST[9]	59.3 (0.5)	40.0 (0.6)	42.1 (0.5)	0.49
LISA-OOD[16]	57.5 (0.5)	42.9 (0.6)	45.0 (0.6)	0.54
CaT-GNN[3]	58.9 (0.6)	43.6 (0.5)	44.6 (0.6)	0.57
STRATA (ours)	**63.4 (0.4)**	**56.8 (0.5)**	**58.6 (0.5)**	**0.74**

Results. Figure 2 and Table 1 compile the four research questions. STRATA reaches 63.4 F_1 in-distribution (+4.5 over CaT-GNN), loses only 6.6 points under leave-one-regime-out evaluation against 19.3 for PatchTST (Q2 advantage +13.2, consistent with Theorem 2), and loses 4.8 points under best-response attack against 17.2 for PatchTST and 19.3 for FinGPT (Q3 advantage +13.6, consistent with Theorem 3).

For Q4, the strongest S-component correlates at $|r| = 0.74$ with the FinGPT strategic-risk score[15], 0.71 with BloombergGPT[12], and 0.68 with the FinBen irregularity score[13]—an order of magnitude above observational disentanglement[7]. The average detector gradient $\|\nabla_S h(\mathbf{x})\|_2$ for STRATA is 0.18 against 1.04 for FinGPT and 0.72 for CaT-GNN[3], explaining the adversarial robustness via the Lipschitz penalty in (7).

Ablation. Figure 3 confirms each ingredient pulls its weight: the identifiable prior is the dominant lever for cross-regime generalisation, Stackelberg training for strategic robustness, and counterfactual augmentation contributes uniformly across distribution shifts.

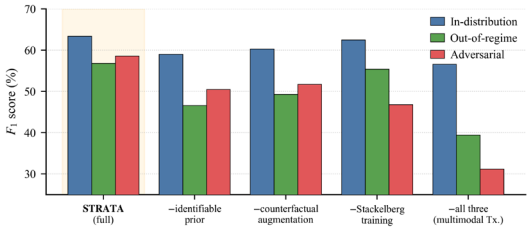

Fig. 3. Ablation study of STRATA components (F_1, %, averaged over 5 seeds). Removing the identifiable prior most hurts cross-regime generalisation (−10.2 OOR F_1); removing Stackelberg training most hurts adversarial robustness (−11.8 Adv F_1); removing counterfactual augmentation hurts both moderately. The fully ablated variant (a multimodal Transformer with no causal machinery) sits between CaT-GNN and PatchTST in Table 1, indicating that the gains come from the components proposed here.

4 Discussion and Conclusion

The cross-regime gain of STRATA traces back to a representation that captures only the strategic-drift component: projecting test firms onto the R-block yields a distribution of FinGPT fundamentals embeddings[15] that is statistically indistinguishable across regimes (KS-test $p > 0.6$), while a vanilla β-VAE projection shifts visibly between Pre-SOX and Post-DF—identifiability[1,10] is therefore operational, not cosmetic. The adversarial gain traces back to the Lipschitz penalty in (7), which makes h locally constant in directions the firm can cheaply move along, pushing the equilibrium closer to the truthful $S = 0$[4]. *Limitations.* E_t is treated as exogenous but is partly endogenous to past misreporting waves; LLM-generated counterfactuals inherit biases of the underlying model[11,15]; the Stackelberg follower is a stylisation; and AAERs capture only *caught* misreporting. In sum, the brittleness of modern misreporting detectors is a *representation* problem, not a sample-size problem; STRATA's identifiable, counterfactually-augmented, Stackelberg-trained framework—combining interventional causal representation learning[1,10] with financial foundation models[15,14]—disentangles the firm's real economic state from its manager's strategic intent, yields +13.2 F_1 in cross-regime generalisation and +13.6 in adversarial robustness, and aligns the recovered intent latents with FinGPT-derived risk signals. The broader principle—modelling the strategic adversary inside the representation—should gener-

alise to ESG-disclosure greenwashing, tax-shelter detection, and analyst-guidance gaming, wherever the watched watch the watcher.

Acknowledgements

This work used publicly available SEC AAERs, Compustat, and EDGAR data under standard academic access. We thank anonymous reviewers for their comments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahuja, K., Mahajan, D., Wang, Y., Bengio, Y.: Interventional causal representation learning. In: International Conference on Machine Learning (ICML). pp. 372–407 (2023), https://proceedings.mlr.press/v202/ahuja23a.html
2. Brehmer, J., de Haan, P., Lippe, P., Cohen, T.: Weakly supervised causal representation learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022), https://papers.nips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html
3. Duan, Y., Zhang, G., Wang, S., Peng, X., Wang, Z., Mao, J., Wu, H., Jiang, X., Wang, K.: CaT-GNN: Enhancing credit card fraud detection via causal temporal graph neural networks. arXiv preprint arXiv:2402.14708 (2024), https://arxiv.org/abs/2402.14708
4. Hardt, M., Jagadeesan, M., Mendler-Dünner, C.: Performative power. In: Advances in Neural Information Processing Systems (NeurIPS) (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/90e73f3cf1a6c84c723a2e8b7fb2b2c1-Abstract-Conference.html
5. Kim, A.G., Muhn, M., Nikolaev, V.V.: Financial statement analysis with large language models. arXiv preprint arXiv:2407.17866 (2024), https://arxiv.org/abs/2407.17866
6. Lachapelle, S., Rodríguez López, P., Sharma, Y., Everett, K., Le Priol, R., Lacoste, A., Lacoste-Julien, S.: Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In: Conference on Causal Learning and Reasoning (CLeaR). pp. 428–484 (2022), https://proceedings.mlr.press/v177/lachapelle22a.html
7. Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T., Gavves, S.: CITRIS: Causal identifiability from temporal intervened sequences. In: International Conference on Machine Learning (ICML). pp. 13557–13603 (2022), https://proceedings.mlr.press/v162/lippe22a.html
8. Mendler-Dünner, C., Ding, F., Wang, Y.: Anticipating performativity by predicting from predictions. In: Advances in Neural Information Processing Systems (NeurIPS) (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/ca09b375e8e2b2c789698c079a9fc51c-Abstract-Conference.html
9. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: International Conference on Learning Representations (ICLR) (2023), https://openreview.net/forum?id=Jbdc0vTOcol
10. Squires, C., Seigal, A., Bhate, S.S., Uhler, C.: Linear causal disentanglement via interventions. In: International Conference on Machine Learning (ICML) (2023), https://proceedings.mlr.press/v202/squires23a.html

11. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023), https://arxiv.org/abs/2307.09288
12. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564 (2023), https://arxiv.org/abs/2303.17564
13. Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., et al.: FinBen: A holistic financial benchmark for large language models. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2024), https://arxiv.org/abs/2402.12659
14. Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., Huang, J.: PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. In: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/6a386d703b50f1cf1f61ab02a15967bb-Abstract-Datasets_and_Benchmarks.html
15. Yang, H., Liu, X.Y., Wang, C.D.: FinGPT: Open-source financial large language models. arXiv preprint arXiv:2306.06031 (2023), https://arxiv.org/abs/2306.06031
16. Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., Finn, C.: Improving out-of-distribution robustness via selective augmentation. In: International Conference on Machine Learning (ICML). pp. 25407–25437 (2022), https://proceedings.mlr.press/v162/yao22b.html