LLMs and Systematic Measurement Error*

J. Anthony Cookson
Pennsylvania State University

Maryam Fathollahi
University of Oklahoma

William Grieser
Oklahoma State University

Eshwar Venugopal
University of Alabama

July 26, 2026

Abstract

Measures derived from Large Language Models (LLMs) are pervasive while researchers commonly treat such measures as observed data rather than generated regressors. For S&P 500 firms from 2005–2024, we show that LLM-derived measures contain non-classical measurement error driven by two distortions: omission and injection. Due to these distortions, LLM measures neglect meaningful terminology, shift meaning and reshape tone. Using the predictive content of omission and injection rates, we develop and apply a measurement error diagnostic to three settings: Item 1A risk disclosures, earnings calls, and direct-prompt measures. Measurement error is most severe for construct-style ratings of long disclosures, milder for the same constructs on shorter earnings-call texts, and substantially attenuated for direct-prompt measures, though the diagnostic reveals measurement-related contamination across all three designs.

Keywords: Large Language Models, Text Analysis, Measurement, Corporate Disclosure, Generated Regressors, Risk Factors, Natural Language Processing, LLM Diagnostics

JEL Codes: C18, G14, G30, C81, M41

*We are grateful for the computing resources provided by our universities, which were instrumental for pursuing this project. Cookson acknowledges financial support from the Smeal Chair endowment. We are also grateful for comments from Tania Babina, Andrew Chen, Lauren Cohen, Gregor Schubert, Tian Qiu, Lucy Wang, and seminar participants at the AI in Finance Conference, Oklahoma State University, the University of Alabama, the University of Oklahoma, and the University of Piraeus.

1 Introduction

Large language models (LLMs) have quickly become a standard tool for extracting quantitative measures from unstructured text. LLMs are adaptable to new problems and highly scalable to large tasks. They also often outperform traditional natural language processing approaches (e.g., Loughran and McDonald, 2011; Tetlock, 2007) in measuring sentiment, risk, readability, and narrative tone. These capabilities have led to a rapid incorporation of LLM-generated measures into empirical work.[1] While this progress is remarkable, scholars have predominantly treated these LLM measures as data rather than generated regressors, and the properties of the resulting measurement error remain largely unexamined.

In this paper, we examine LLM measurement error organized around three key facts. First, measurement error is built into the architecture of LLMs. LLMs operate by compressing input text before generating output. Indeed, we show LLMs allocate as much as 86% of their attention to the first 10% of a document and almost none to the middle 80%, regardless of the downstream task. Second, the resulting distortion is large. When we prompt four widely used LLMs to summarize a risk disclosure, they drop 62% to 81% of the source's most distinctive vocabulary. They also introduce novel terms that constitute 15% to 35% of the summary's most distinctive language. Altogether, 37% to 51% of the source's semantic content does not survive into the summary. Third, this summarization is not neutral. LLMs shift tone toward hedged language and negativity, and the novel terminology reflects what was salient during the model's training period. These properties are intrinsic to how LLMs compress text in measurement tasks, which is not specific to a given model or prompt. For this reason, these issues are likely to persist, even as LLMs advance.

We study these errors in Item 1A risk disclosures and earnings call transcripts. We formalize an LLM-generated measure as a noisy proxy for a latent construct, contaminated by a model-specific wedge that reflects how the model compresses input, allocates attention, and draws on its priors. Because the wedge is not directly observable, we measure it through two summary-based proxies. *Omission* is the share of the source's most distinctive vocabulary that the model drops, and *injection* is the share of the summary's most distinctive vocabulary that appears nowhere in

[1] LLM measurement applications are multifaceted. Just within corporate finance, LLMs have shed light on aspects of corporate disclosure, firm exposure to generative AI, earnings predictability, supply-chain risk, and ESG narratives, among many other topics (Kim et al., 2023; Eisfeldt et al., 2023; Lopez-Lira and Tang, 2023; Sikochi et al., 2024).

the source. Applying this framework to four widely used models (GPT-4o-mini, Grok-3-mini, Phi-4, Llama-3.3-70B) on finance corpora, we show that the wedge proxies enter downstream regressions significantly, in ways that signal non-classical measurement error rather than attenuation, and we develop a diagnostic that flags the severity of measurement error for inference.

LLM summaries exhibit significant omission and injection. Across the four models, LLM summaries omit 59% to 73% of the top-1,000 most distinctive source terms, with omission rates varying widely by risk category. Categories with repetitive, templated language — regulatory compliance, supply chain, litigation — show the highest omission, while categories with firm-specific financial content retain closer to half its discrimintive vocabulary. Injection is pervasive: novel terms absent from the source account for 15% to 35% of the summary's most discriminative vocabulary, and these injected terms are systematically evaluative and tonal rather than firm-specific. The same patterns replicate qualitatively in the earnings call corpus, suggesting that the measurement wedge is a property of LLM processing, not a feature of Item 1A disclosures. Further, we analyze the content of omitted and injected terms, finding that these vocabulary distortions translate directly into lost meaning. Using Doc2Vec to embed semantic context of omitted and injected terms, we find that between 35% and 38% of a source disclosure's semantic content is absent from its LLM summary, and the resulting tone is *systematically* more negative and hedged than the source text.

The wedge is heterogeneous across the 26 risk categories of Fathollahi et al. (2026). Omission is greatest in categories with templated, repetitive language. For example, GPT-4o-mini omits roughly two-thirds of the most distinctive terms in regulatory compliance, supply chain, competition & strategy, and litigation risk categories. Omission is smallest in categories with firm-specific operational and financial content (growth management, credit, liquidity, retaining roughly half).

Injected vocabulary differs systematically from human risk disclosures and follows patterns shaped by each model's training data. For Natural and Physical Environment risks, LLMs introduce terms such as *droughts*, *wildfires*, *flooding*, and *radioactive*, while human disclosures emphasize *contamination*, *cleanup*, and *temperature changes*. For Political and Social Environment risks, LLMs often add *Taiwan*, *wars*, and *Chinese*, whereas firms emphasize *nationalization*, *hostilities*, and *customs*. Similarly, models trained in 2023–2024 disproportionately weight Taiwan-related tensions and GDPR compliance. The injected terms reflect what was salient in news discourse during training, which may not match the compliance categories firms use in practice.

These textual patterns shape downstream LLM constructs pervasively. We prompt each LLM to score four constructs — severity, specificity, complexity, and boilerplate — from each Item 1A disclosure, and find that Omission and Injection jointly predict each construct with strong significance across all four LLMs. Because Omission and Injection are proxies for *how* the LLM selected its vocabulary rather than for what the source said, a random text-processing model would not produce systematic associations between them and the constructs. Their strong joint significance is therefore a first sign of non-classical error, and it motivates the formal diagnostic that follows.

To diagnose whether LLM errors are non-classical, we estimate three specifications for each construct–outcome pair: a *naive* regression of the outcome on the LLM construct; a *joint* regression that adds Omission and Injection as wedge proxies (with construct-coefficient sensitivity $|\Delta\%|$ relative to naive); and a *decomposition* that splits the construct into a wedge-predictable component (coefficient θ) and a wedge-orthogonal residual (coefficient λ). If we find that Omission and Injection explain corporate financial outcomes beyond the corresponding LLM measures, this is a strong indication of non-classical measurement error.[2]

We apply our diagnostic to 6 key corporate finance outcomes across 4 different LLMs. With three fixed effects variants, we obtain a grid of 288 cells (4 LLMs $\times$ 4 constructs $\times$ 6 outcomes $\times$ 3 fixed effects variants). Across these scenarios, the coefficient sensitivity is at least 10% in 69% of cells, and that sensitivity remains elevated at 63% when we restrict to cells where the naive regression is significant. The earnings-call corpus produces the same qualitative pattern with smaller magnitudes (with sensitivity above 10% in 39% of cells). These findings indicate that measurement error is quantitatively important for the typical case. In addition, there is also significant heterogeneity across LLMs, constructs, and specifications. Thus, even when the typical case is moderate bias as with earnings calls, we see some cases with significant bias.

We also show how contamination from measurement error depends on the design of the LLM-derived measure. Replicating Jha et al. (2024), we find that direct-prompt measures on focused texts — where the model answers a single outcome-specific question — are much less contaminated than construct-style ratings of long disclosures: median coefficient sensitivity falls from 20.9% for Item

[2]Under a control function interpretation, we also note that the decomposition specification doubles as a partial correction: the coefficient on the wedge-orthogonal residual (λ) estimates the construct-outcome relationship after stripping away wedge-correlated variation, leveraging the LLM's genuine signal while neutralizing its processing artifacts.

1A construct ratings to below 7% for the five direct predictors with significant naive coefficients. Even so, omission and injection enter significantly in the direct-prompt setting, confirming that the diagnostic is useful across prompt designs.

If used without adjustment, LLM-derived measures import the model's own assumptions into academic work, overstating effects in categories with heavy term injection and compressing meaningful differences across firms. Our findings are not an argument against using LLMs — their scale and adaptability are substantial. They are an argument for treating LLM outputs as generated regressors with a measurable wedge, which would suggest new practices and diagnostics. For example, researchers could report omission and injection alongside their main results, run the joint and decomposition specifications as a screen for non-classical error, and aggregate across models to filter out idiosyncratic wedge variation that no single model's proxies can detect.

Our findings contribute to four strands of research. The first strand uses LLMs as a productivity tool to scale textual analysis, treating outputs as reliable measures of firm fundamentals (Kim et al., 2023; Eisfeldt et al., 2023; Lopez-Lira and Tang, 2023; Jha et al., 2024; Bybee, 2024). Kim et al. (2023) use ChatGPT to summarize 10-K MD&A sections and treat the summaries as reliable extractions; Jha et al. (2024) construct firm-level LLM scores and validate them against CFO survey responses; Huang et al. (2023) build a finance-pretrained transformer (FinBERT) that outperforms dictionary methods on sentiment tasks. These papers demonstrate impressive explanatory power of LLMs while showcasing their adaptability to different settings, but none opens the black box to examine how the internal workings of LLMs contaminate inference. Our paper contributes by formalizing LLM outputs as generated regressors with a measurable wedge and providing diagnostics that screen for non-classical measurement error.

Our research also relates to a growing literature that critiques textual methods on measurement grounds, documenting problems with bag-of-words approaches, dictionary saturation, negation handling, and context insensitivity (Loughran and McDonald, 2011; Gentzkow et al., 2019; Cohen et al., 2020; Grimmer et al., 2022). Although LLMs overcome some of these limitations, our findings imply they introduce new biases that raise concern because they look like semantic understanding. Our measurement of omission connects to the NLP summarization literature, where Grusky et al. (2018) introduce extractive coverage as a metric for how much of a summary's vocabulary derives from the source, and See et al. (2017) and Narayan et al. (2018) document that

neural summarizers systematically discard source vocabulary. Our measurement of injection connects to the NLP faithfulness literature, where Maynez et al. (2020) formalize extrinsic hallucination as the introduction of content not supported by the source, and Ji et al. (2023) survey its pervasiveness across generation tasks. We extend this literature by showing these properties are severe in regulatory risk disclosures — a domain with a legalistic, hedged register that differs fundamentally from the news and Wikipedia corpora on which prior work focuses — and by translating the lexical-level findings into measurement-error consequences for finance regressions.

A third strand formalizes LLM outputs as generated regressors and develops inference tools for downstream regressions using AI-generated variables (Battaglia et al., 2024; Egami et al., 2023; Ludwig et al., 2025; Fong and Tyler, 2021). Battaglia et al. (2024) prove that regressions using AI-generated variables suffer from a bias that does not vanish in large samples, and propose bias-corrected confidence intervals using validation data. Egami et al. (2023) develop Design-based Supervised Learning (DSL), a doubly-robust estimator that combines many LLM labels with a small gold-standard sample. Ludwig et al. (2025) provide an applied econometric framework distinguishing prediction from estimation uses of LLMs. Our contribution to this literature is mechanistic: we identify omission and injection as specific text-processing pathologies that produce the error structure Battaglia et al. and Egami et al. seek to correct. Our joint and decomposition diagnostic complements these corrective estimators: researchers can run it cheaply as a screening tool before investing in validation data, and it flags when deeper corrections (DSL, validation samples) are warranted.

Fourth, our paper contributes to recent research identifying problems with LLMs in predictive contexts. After the initial wave of research into LLM return prediction properties, research was quick to point out forward-looking bias (Ludwig et al., 2024; Sarkar and Vafa, 2024; Lopez-Lira et al., 2025). Sarkar and Vafa (2024) formalize lookahead contamination and show that prompt-based mitigations fail. He et al. (2025) train chronologically consistent LLMs (ChronoBERT, ChronoGPT) pretrained only on text available at each historical timestamp, and Engelberg et al. (2025) propose "entity neutering" when chronologically-consistent models are unavailable. Other research checks whether conclusions continue to hold after the model's training cutoff (Jha et al., 2024; Lopez-Lira and Tang, 2024; Bybee, 2025; Bhagwat et al., 2025). Our findings complement this literature by providing a cross-sectional perspective on how LLM text processing contaminates inference. Distinct

from the temporal dimension, omission and injection produce non-classical measurement error that biases cross-sectional regressions regardless of training cutoff. Our research thus generalizes the forward-looking bias critique to a form of cross-sectional contamination, and by offering diagnostics and best practices, it provides a path forward for future research using these tools.

2 LLMs as Generated Regressors

Researchers often use LLM outputs as explanatory variables in regressions of the form

$$Y_i = \alpha + \beta \cdot L_{im} + \delta' X_i + \epsilon_i, \tag{1}$$

where L_{im} is an LLM-generated output, such as a score, classification, or extracted quantity. Standard practice treats L_{im} as if it were directly observed, reporting $\hat{\beta}$ with t-statistics computed as if the score were measured without error.[3] Implicitly, the goal of estimating Equation (1) is to estimate the relation between Y_i and a latent economic construct θ_i,

$$Y_i = \alpha + \beta\theta_i + \delta' X_i + u_i, \tag{2}$$

where L_{im} serves as a proxy for θ_i. Effectively, L_{im} is the output of a first-stage statistical model whose properties the researcher does not control or observe. In econometric terms, every LLM-derived measure is a generated regressor (Pagan, 1984; Murphy and Topel, 1985) in which the first stage is a high-dimensional black-box predictor. Standard generated regressor corrections assume the researcher either observes the first-stage estimator or can resample estimates cheaply. Neither condition holds for LLMs: the first stage conditions on a researcher-chosen prompt whose effect is not separately identified from the model, and resampling requires repeated API queries that are prohibitively costly at scale.

We formalize the LLM-generated regressor as a stochastic mapping from documents and prompts to a scalar output. Let T_i denote unit i's source text.[4] A model m with pretraining priors Ω_m,

[3] LLMs are also used to generate outcome variables, and the issues we develop apply in spirit to that case as well, though through a different mechanism. We focus on the case where the LLM output is used, directly or indirectly, as an explanatory variable, since it is the most common in current empirical work and the bias mechanism is most transparent.

[4] The unit of analysis is general; i may index firms, filings, transcripts, contracts, or any text-bearing entity.

applied to T_i with prompt P_i, produces

$$L_{im} = \theta_i + w_m(T_i, P_i, \Omega_m). \tag{3}$$

The term $w_m(\cdot) = L_{im} - \theta_i$ is the measurement wedge, the part of the output that reflects model processing rather than the construct θ_i. By construction, the wedge is not a channel for additional signal about θ_i; any covariation between $w_m(\cdot)$ and the outcome generates distortion. Combining the estimable Equation (1) with Equations (2) and (3) yields the disturbance $\epsilon_i = u_i - \beta w_m$. Thus, the wedge must be mean-independent of the structural disturbance u_i, the construct θ_i, and covariates X_i: $E[w_m \mid \theta_i, X_i, u_i] = 0$. Under this condition L_{im} is an unbiased noisy proxy for θ_i and the classical measurement-error case obtains: the wedge induces only attenuation, and standard errors-in-variables corrections recover β in expectation.

The classical case breaks down when the wedge is correlated with θ_i or u_i conditional on X_i. The wedge may be correlated with θ_i if the LLM misreads the construct in a way that depends on the level of θ_i, such as the model compressing a rating scale by pulling genuinely extreme disclosures toward the middle. This is systematic miscalibration rather than additional signal: it rescales β, and distorts the magnitude of the estimated relationship. A more serious issue arises when $Cov(w_m, u_i) \neq 0$, indicating that w_m affects Y_i through unobserved variables other than through θ_i.

Non-classical measurement error from both sources of correlation arises through the three arguments of $w_m(\cdot)$, the text, the prompt, and the model's priors. First, the wedge may vary with features of the source text T_i, such as document length, linguistic register, or boilerplate and topical content, which bias the estimate if these features capture multiple concepts that are correlated with θ_i or with Y_i for reasons other than θ_i. Second, the wedge may vary with the prompt P_i in ways the researcher does not realize, such as ordering effects, how the prompt defines the LLM-generated regressor, or implicit anchors in the prompt language. Third, it may reflect the model's priors Ω_m, which over- or under-weight certain content, so that disclosures whose register matches the pretraining distribution are processed differently from those whose register does not.

2.1 LLMs and Compression

All three channels of measurement error stem from a single feature of how language models process text: they compress information. The model transforms input text into a smaller output, and that reduction is selective. Processing a long disclosure into a single score or category forces the model to discard most of what the text contains, and what it keeps is chosen nonrandomly.

For example, consider a disclosure stating:

> "Our operations depend on a limited number of suppliers for certain key components, and we do not have long-term supply agreements with most of these suppliers. A transition to new suppliers could result in delays, quality issues, and increased costs that could materially affect our results of operations."

A human reader retains the specifics: limited suppliers, no long-term agreements, transition risks enumerated as delays, quality, and cost, with the impact qualified to operations. A model — in this case GPT4o-mini — summarizing the passage may return only:

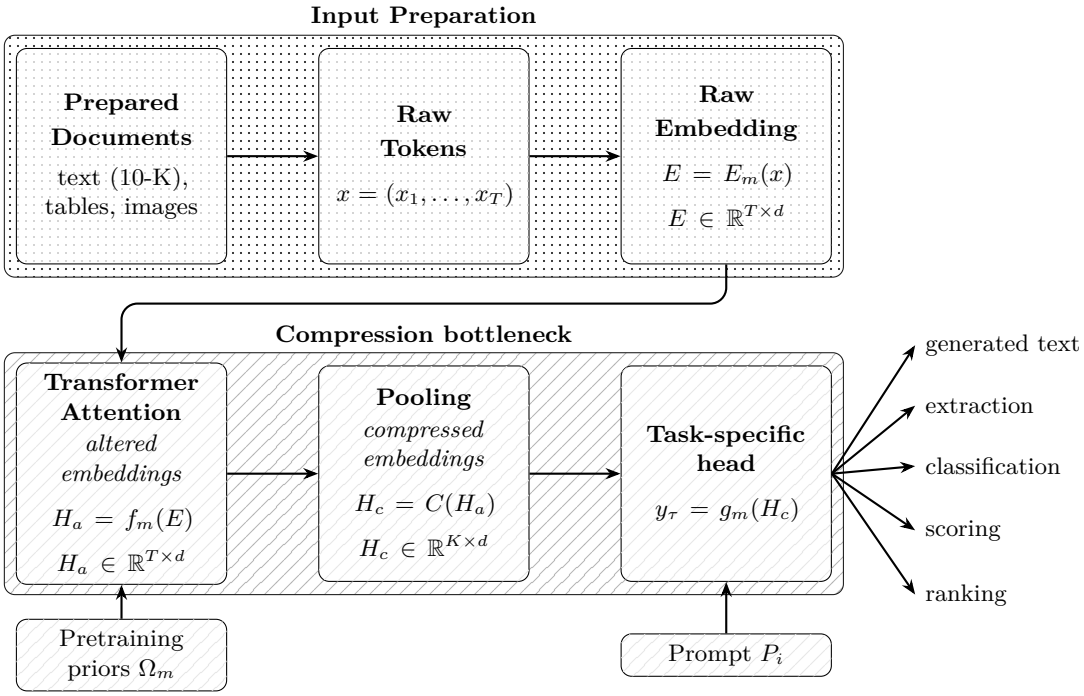
> "The company faces supply chain risks that could negatively impact financial performance."

The conditional structure, the absence of contracts, and the enumerated consequences are compressed away.

Exhibit 1 illustrates the compression process through the LLM-based measurement pipeline that transforms the prompt and source material into a compressed task output. A document of text, tables, and images is first converted into tokens $x = (x_1, \ldots, x_T)$, and then into embedded token representations $E = E_m(x) \in \mathbb{R}^{T \times d}$. The tokenization and embedding process does not discard any information and is best understood as representation rather than measurement.

Compression begins in the transformer layers, which build contextualized token representations $H_a = f_m(E) \in \mathbb{R}^{T \times d}$. Each token's representation becomes a weighted mixture of information from the others, so the model is, in effect, deciding which parts of a long document matter for representing any given part, and tokens that receive little attention have less opportunity to shape what follows. The same layers also fold in content the model learned during pretraining, so the representation reflects its priors as well as the source. Attention is therefore an informational compression stage

Exhibit 1: LLM compression as structured measurement error



and the entry point for model priors Ω_m and prompt-driven emphasis P_i, even though the size of the representation has not yet fallen.

Interacting with the representation stage, pooling imposes a more literal informational bottleneck. After attention has rewritten each token in light of the broader document, pooling discards much of the per-token detail and produces a smaller object, $H_c = C(H_a) \in \mathbb{R}^{K \times d}$, with $K \ll T$ and often $K = 1$. This is where the size drops, and it is the step that decides which of the rewritten content survives into the compact object the task-specific head can use. The head, $y_\tau = g_m(H_c)$, then maps that representation into the requested output L_m.

The central measurement problem is that the reported output is not produced from an exhaustive read of the full disclosure. It is produced from whatever survives attention, pooling, and the task-specific projection. As a result, LLM-generated measures carry structured rather than merely random measurement error whenever economically relevant content is de-emphasized, averaged away, or replaced by prompt- or model-driven judgment. Importantly, because every task reads the same compressed representation H_c, a distortion in H_c propagates to all outputs: the wedge that contaminates a score also shapes the category or summary the model would produce from the same input.

The compression *architecture* is set by the chosen LLM and prompt design, not a direct property of the source text (i.e., how token-level information is pooled and projected into the properties of output). However, the *effective* information loss can still vary with source text length, complexity, and how well the model's representation captures the language. A prompt may ask an LLM to return a summary, a coarse category label, an ordinal rating, or a single score on risk severity, all based on the same source text. These task-specific outputs sit on a spectrum from least to most compressed, which comes with a crucial tradeoff. Richer outputs give the LLM more room to extract signals and audit its output, but they also provide more room to editorialize, drift, or import content that the source text does not support. More compressed outputs reduce this discretion, but they force the model to collapse a multidimensional risk profile into a score or rating. The source text can be too nuanced for a single score to capture its content accurately and is often consistent with a range of values. The model thus settles on a point estimate by leaning heavily on its priors, which reflects its training as much as the filing content.

Exhibit 2: LLM Compression Grid: Signal vs. Distortion

	Intrinsic (from reading the source text)	**Extrinsic (from the model's parameters)**
Signal	The model condenses the disclosure into the latent construct efficiently, emphasizing what matters.	The model applies correct, temporally consistent outside knowledge, as an analyst or industry expert would at the time of disclosure.
	Example: *LLM reads mild single-supplier language as the serious concentration risk it is, which traditional NLP approaches miss.*	Example: *Knowing a copper miner faces persistently volatile commodity markets, it scores vague price-exposure language higher than the flat words suggest.*
Distortion	The model misinterprets the source text.	The model imports outside content that overrides the document.
	Example: *It scores boilerplate ("could materially and adversely affect") as real severity, or reads a hypothetical ("if a key customer were to leave") as a realized exposure.*	Example: *It scores a 2006 mortgage disclosure as catastrophic because its training data knows 2008 happened, information no reader had at the time.*

The grid in Exhibit 2 organizes the effects of LLM compression in the context of risk disclosures along two axes. The rows indicate whether the model generates a signal that moves a score toward the true latent economic construct or a distortion that moves it away by adding noise or systematic bias. The columns separate whether that signal or distortion is *intrinsic*, coming from reading the source text itself, or *extrinsic*, coming from extrapolating beyond what the document contains. Intrinsic signal is faithful extraction, where the model efficiently condenses relevant information from

the text. Intrinsic distortion is misreading the text, treating irrelevant information as substantive, or omitting pertinent content.

The extrinsic column holds the sharpest distinction because its two outputs can look alike. Extrinsic signal is faithful extrapolation: correct, temporally constrained outside knowledge. Scoring a copper miner's vague price-volatility language higher because commodity markets are persistently volatile uses stationary structural knowledge available when the disclosure was written, consistent with what a reasonable analyst or industry expert would do. Extrinsic distortion is fabrication, where the model introduces content that overrides what the document can support. Scoring a 2006 mortgage disclosure as catastrophic because the training data encodes 2008 uses a specific future event that no contemporaneous reader could have known. The first is legitimate context, the second is look-ahead leakage, and only a measure anchored to the disclosure's own vintage can tell them apart.

The grid in Exhibit 2 raises two questions. The first is what LLMs offer over traditional NLP methods such as bag-of-words, TF-IDF, or LDA. Those methods are easily vintage-restricted and, by construction, occupy only the intrinsic column. They cannot reach outside the text available to them, which mitigates fabrication at the expense of losing good extrinsic signal. Their intrinsic reading is also blunt, keying on surface words without resolving, for instance, the asset-management sense of "investment" from the corporate-finance one. The LLM's advantage is a rich, context-specific representation that situates the document's words against everything the model has seen. Although extrinsic reasoning can make the signal more precise, it opens the door to fabrication alongside context.

The second question is whether the four channels will evolve differently as models advance, and thus, whether the measurement concern fades with better AI. Both intrinsic and extrinsic distortion should recede as models improve and as more human-validated data enters training. But some distortion is imposed by compression itself, not by any limitation the model can outgrow. Condensing an eight-thousand-token Item 1A into a few sentences or a single severity score discards substantial information, and the quantity discarded is not tied to the model's skill. Irrespective of model capability, a summary this compressed cannot have a zero noise-to-signal ratio because most of the information is gone before any inference begins.

3 Empirical Analysis

3.1 Data

Our primary corpus consists of Item 1A risk factor disclosures extracted from annual 10-K filings of S&P 500 firms over 2005–2024, obtained from SEC EDGAR. We treat each individual risk factor disclosure as the unit of analysis, following standard practice in the empirical disclosure literature (Campbell et al., 2014; Kravet and Muslu, 2013; Hope et al., 2016). The resulting sample contains 889,223 unique risk factor disclosures spanning 1,675 firms. We lemmatize all tokens, remove standard stopwords, and retain only terms with a minimum corpus frequency of fifteen. We assign each disclosure to one of twenty-six risk categories (with 5 parent categories) using the taxonomy developed in Fathollahi et al. (2026) and presented in Appendix Table A.1. On average, Item 1A contains 11,575 words (median is 9,740), and approximately 34 distinct material risk disclosures, though length varies substantially across risk categories: operational and regulatory disclosures tend to be longer and more technically specific than financial risk disclosures.

Our second corpus consists of quarterly earnings call transcripts for S&P 500 firms over the same sample period, obtained from Refinitiv Workspace. Each transcript is structured into a prepared management section and an analyst question-and-answer session. Relative to Item 1A filings, earnings calls contain more timely language and, in the Q&A session, unscripted responses produced under real-time constraints. The earnings call corpus contains 97,776 transcripts spanning 1,563 unique firms, with an average of 9,781 words per transcript. As in the Item-1A corpus, we lemmatize all tokens, remove standard stopwords, and retain only terms with a minimum corpus frequency of fifteen. Together, the two corpora represent complementary modes of firm communication. Item 1A disclosures are legally scrutinized, drafted annually by counsel and management, and designed to satisfy regulatory requirements. Earnings call transcripts are produced quarterly and less scripted, especially in the Q&A session.

We extract company financial information from Compustat yearly files for 10-K Item 1A corpus and the quarterly file for the earnings call transcripts. Stock prices come from CRSP. Finally, we take the word lists from Loughran and McDonald (2011) to measure tone in raw corpora and their corresponding summaries. Table A.4 contains all variable definitions used in our analysis.

3.2 Measuring Omission and Injection from LLM-Generated Descriptions

Since θ_i is latent, the wedge w_{im} is not directly observable. We can, however, observe features of the LLM's processing through its output. We focus on the summarization task because it provides a direct source-to-output comparison that is informative about the model's priors: we can ask which source vocabulary the model retained, which it omitted, and which terms it introduced from outside the source. Summarization is the least compressed output (Exhibit 1) and serves as our laboratory. Because every task runs the same source text through a common compression process, summary omission and injection patterns reflect the model's priors and handling of the input, not summarization specifically, and carry over to scores and classifications.

For any source-summary pair, we partition the vocabulary into three disjoint sets. Retained terms appear in both the source and the summary; omitted terms appear only in the source; injected terms appear only in the summary. These are mechanical properties of LLM text processing that correspond to well-studied NLP concepts and map directly into the four cases illustrated in Exhibit 2. *Omission* is intrinsic: it improves signal when it sheds redundant or non-discriminating language, and it distorts signal when it selectively drops content relevant to θ_i or Y_i, or retains irrelevant context. *Injection* is primarily extrinsic, enhancing signal when injected terms supply source-consistent context drawn from the model's priors, while distorting the signal when those priors import associations or outcomes the source does not support.

To measure omission and injection rates for a given document, we compare the most discriminating vocabulary in the LLM summary against that in the original text, restricting attention to meaningful words so that vocabulary differences are not trivial. For each disclosure i and model m, let $V^{(n)}_{i,source}$ and $V^{(n)}_{im,summary}$ denote the top-n most discriminating terms in the source and summary, respectively, ranked by lift. The omission rate O_{im} is the share of the source's top-n discriminating terms absent from the summary's top-n:

$$O_{im} = 1 - \frac{|V^{(n)}_{i,\text{source}} \cap V^{(n)}_{im,\text{summary}}|}{|V^{(n)}_{i,\text{source}}|}. \tag{4}$$

The injection rate I_{im} is the share of the summary's top-n discriminating terms that appear nowhere in the source:

$$I_{im} = 1 - \frac{|V^{(n)}_{im,\text{summary}} \cap V_{i,\text{source}}|}{|V^{(n)}_{im,\text{summary}}|}. \tag{5}$$

Note the asymmetry in the denominators' reference sets. Omission compares source terms against the summary's top-n, while injection compares summary terms against the *full* source vocabulary. Thus, a term counts as injected only if it is genuinely absent from the source rather than merely outside its top-n terms. Appendix B details term ranking, the lift computation, and the full procedure. We set $n = 1000$ for our primary analysis and consider alternative depths as robustness.

LLM Compression at Document-level

Table 1 reports summary statistics for the Item 1A corpus in Panel A (four LLMs) and the earnings-call corpus in Panel B (two LLMs).[5] These statistics provide our initial exploration of vocabulary omission and injection via LLM summary and keyword generation tasks.

Panels A and B both show that LLM summaries contain considerably lower token counts relative to full disclosures and earnings transcripts, but the degree of compression varies strongly across models. Phi-4 and Llama-3.3-70B compress most aggressively, cutting tokens by roughly 92% (from 115–118M to 8.7–9.8M), followed by GPT-4o-mini at 86% (103M to 14.8M) and Grok-3-mini at 78% (126M to 27.2M). Panel B shows that GPT-4o-mini and Llama-3.3-70B transcript summaries reduce token counts by approximately 86% (139M to approximately 20M), which is comparable in magnitude to Item 1A compression rates.

Token counts indicate the total size of each source of text. The effect on vocabulary size (unique words represented) is more revealing. Panel A indicates that full risk disclosures contain 36,904–41,879 unique words. LLM summaries use a far more restricted vocabulary than human disclosures. Grok employs the most expansive list at 24,909 unique terms, while Llama-3.3-70B restricts the most aggressively, retaining only 14,965 unique terms. Concentration in vocabulary is the most telling empirical footprint of the pooling stage in Exhibit 1: as the representation collapses, common words absorb a larger share of the summary while distinctive, low-frequency terms drop out. In full disclosures, the top-100 words account for approximately 29% of all tokens; in LLM summaries, this rises to 34–41%. Top-500 concentration climbs from roughly 62% to 66–74%. LLMs appear to funnel diverse, firm-specific language into a narrower vocabulary where common words dominate. Panel B shows that the total earnings-call vocabulary of 132,539 unique words is roughly

[5] We restrict the earnings-call analysis to gpt-4o-mini and llama-3.3-70B, the two models for which summaries are available; their omission and injection patterns closely track those in the Item 1A sample, so little is lost by the restriction.

3.5 times that of the Item 1A disclosures, consistent with the largely unscripted Q&A section drawing on more varied language.

Keywords yield the greatest concentration. The top-100 keywords account for 43–60% of all keyword tokens, and the top-500 account for 77–89%. Llama-3.3-70B's keywords exhibit the strongest concentration with only 4,404 unique terms, and the top-100 and 500 terms accounting for 60% and 88% of tokens, respectively. Normalized entropy is stable across all text sources (0.71–0.76), as expected since it is relatively insensitive to changes in the low-frequency tail; the top-k concentration measures, which key on the top of the distribution, reveal the asymmetry directly. LLMs retain high-frequency vocabulary while omitting the long tail of distinctive terms with lower frequency.

Beyond omission and injection rates, the specific terms LLMs introduce show that summarization is not neutral. Table 2 reveals that injection in summaries can change the context of a disclosure. Under natural and physical environment risk, for example, where disclosures use terms like 'temperatures' or 'pollution,' the LLM summaries introduce vocabulary absent from the source: 'droughts,' 'wildfires,' 'toxic,' 'discharges.' This is injection in the sense of the extrinsic axis of Exhibit 2. But raw injection rates cannot indicate whether the injection adds signal or distortion. If the source describes climate exposure that reasonably entails wildfire risk, the added terms are faithful extrapolations; if the model is importing associations that the disclosure does not support, they are fabrications. Distinguishing the two requires reading whether the injected content is grounded in the source, which we explore in Section 3.4.

Injection creates a complementary vocabulary distortion. Table 2 catalogs the highest-frequency injected terms by risk category, showing that injection introduces general evaluative language — 'significant,' 'substantial,' 'primary' — rather than firm-specific disclosure content. This pattern is consistent with models imposing an editorial layer on the factual content, an interpretation we return to in Section 3.6.

Omission and Injection by Risk Factor

Figure 1 reports omission and injection at the category-vocabulary level[6] — comparing the discriminative vocabulary pooled across all disclosures within each risk category — whereas Table 3 reports

[6]Please see Figures A.1 and A.2 in the appendix for a more granular view of omission and injection rates across our full taxonomy of risk factors.

means at the disclosure level. Both levels show the same qualitative ranking across models.

At the category level, omission ranges from 57% to 74%, with considerable variation across models and risk types. At the model level, Llama-3.3-70B omits terms the most aggressively at 73.3% across all risk, followed by Phi-4 at 70.6%, GPT-4o-mini at 66.9% and Grok-3-mini at 58.8%, consistent with the document level patterns. Financial-risk disclosures have 1–2 percentage points more omission than other categories, while strategic risks have the lowest omission rates.

At the disclosure level, Table 3 indicates that omission is, on average, 67.8% while injection is 25.2%. Across models, omission rates range from 66.9% (GPT) to 73.1% (Llama-3.3-70B) while injection rates range from 17.3% (Phi) to 40.2% (GPT). These rates indicate that injection and omission are conceptually interdependent but can differ significantly across models and contexts.

To illustrate the degree to which the omitted terms and injected terms are central to the semantic content of the source text, we also report *Semantic Cosine Similarity* of the omitted and injected terms. After embedding via Doc2Vec (described in more detail in Section 3.4), we compute the cosine similarity of the omitted bag of words and injected bag of words to the original source's. Omitted terms are highly central to the source context, exhibiting a cosine similarity of 64% overall while injected terms are, on average, less relevant to the semantics of the source text. This pattern of dropping relevant terms while replacing them with less relevant ones foreshadows our analysis of fidelity loss in Section 3.4.

It is instructive which categories have high injection rates. For GPT-4o-mini, Regulatory and Business Model disclosures each show 44% injection—nearly half of its top discriminative terms for these categories are absent from the source. Phi-4's worst category is Competition Strategy (24%), meaning about one-quarter of its discriminative vocabulary is not in the original source. Llama-3.3-70B also struggles with Organizational Development (29%). Categories involving operational processes, physical systems, and regulatory language consistently elicit high injection across models. These domains require specialized vocabulary that firms describe idiosyncratically. Models fill the gaps with prototypical terminology from pretraining.

Turning to low-injection categories, Macroeconomic Environment exhibits low injection rates with Llama-3.3-70B injecting just 18%. Credit ranges from 14-35%, Financial Markets from 18-38%. These categories feature standardized financial and legal vocabulary that models extract without supplementation. The pattern is consistent: established disclosure conventions permit faithful

extraction, while operational or technical language corresponds to higher injection rates.

This heterogeneity varies across models in ways that matter for research design. A study using GPT-4o-mini to analyze Regulations would face 44.3% injection while switching to Llama-3.3-70B would reduce this to 23%. While GPT-4o-mini injects the most, the pattern switches if we also take Omission into consideration.

3.3 Attention Weights in Open Weight Models

The omission rates in Table 1 provide an input-output comparison of LLM compression. The table shows that source vocabulary drops out, but not why. As Exhibit 1 highlights, the mechanism for this is transformer attention when the model decides which input tokens shape the downstream representation. In this section, we provide direct evidence on the attention layer itself.

We measure model attention using the open weight models Llama-3.3-70B and Phi-4, which expose their internal states. For each Item 1A disclosure chunk, we feed the chunk under our scoring prompt, and capture the attention distribution at the token where the model emits each construct's continuous score, averaged across all layers and heads. We partition the input into first-10%, middle-80%, and last-10% regions and compute each region's per-token attention mass, its share of total attention divided by its share of tokens, so that 1.0× represents uniform allocation.

Figure 2 shows that both models attend to the beginning of the input far above its token share. Llama-3.3-70B and Phi-4, respectively, place 8.6× and 6.4× of the attention that a uniform allocation would give the first 10% of tokens. Yet each model dedicates only 0.1–0.3× of the attention to the middle 80% of the text. The attention distributions are nearly identical across the four LLM-generated constructs (severity, specificity, complexity, boilerplate), indicating that what the model reads is a property of how it ingests input, not of the semantics of the output construct. These loading patterns are unchanged by the length of the input text and by variations in the ordering of prompt instructions. Attention concentration is larger when focusing on an earlier cutoff for the first part of the document — 17× on the first 5% for Llama-3.3-70B and 12× for Phi-4.[7]

Collectively, these patterns provide evidence that attention is concentrated on a non-random subset of input tokens. Most of the input text has very little weight on LLM scoring. Moreover, consistent with distortion, this attention bottleneck follows token position, not instruction ordering,

[7]See Appendix Figure A.3.

input length, or natural location in the document where a construct is relevant. This also means the model reads the input the same way regardless of task type, so the omission patterns rooted in what it reads carry over from model summarization to all task types (e.g., scoring and classification), which supports our focus on summaries to study broad properties of LLM measurement error.

3.4 LLM Compression and Fidelity Loss

The omission and injection rates in Section 3.2 capture the degree of dropping and adding terms without measuring directly whether meaning changes. When a model paraphrases 'key customer' as 'important client,' the lexical measures record both an omission and an injection even though the economic content is unchanged. Conversely, if the model recasts 'key customer' as 'supply chain concentration,' this can change meaning and bias inference. Omission and injection rates do not distinguish between these cases of meaning-preserving paraphrase versus meaning-altering drift. These cases were illustrated in Exhibit 2. When omission and injection rates are particularly high, it can be a useful warning sign, but these rates do not directly measure lost meaning.

We capture this idea by adapting the computer science concept of *fidelity*, the degree to which a compression process reproduces the original signal without distortion. Reproducing a full disclosure from an LLM summary is infeasible. We therefore need to measure how much of the disclosure's meaning, rather than its exact vocabulary, survives summarization.

We measure *fidelity* as the cosine similarity between the Doc2Vec embedding of a source risk-factor disclosure and the embedding of its LLM summary, and we define *fidelity loss* as one minus fidelity. Doc2Vec maps a document to a fixed-dimensional vector whose geometry reflects semantic content, so two texts that convey similar meaning in different words receive similar embeddings. Specifically, we train a separate Doc2Vec model for each date t on the full Item 1A corpus until then, so the embedding reflects the vocabulary and usage of Item 1A disclosures and cannot encode information that postdates a disclosure at time t. Doing this ensures chronological consistency in the semantic structure.[8]

[8]This time-consistent vocabulary constraint matters because the resulting LLM measures are used to study future firm outcomes. An embedding trained on later text could let post-disclosure information leak into a measure of the disclosure itself. A modern transformer embedding offers considerably better semantic understanding, but it is pretrained on a corpus running past most of our sample period and therefore cannot provide the same time-consistent guarantee. The period restriction is robust to the training window, as a five-year rolling window and a fixed year embedding space produce the same patterns. See Figure IA.1 in the Internet Appendix.

Figure 3 reports results on fidelity and fidelity loss by LLM. Panel A shows total fidelity (blue) and fidelity loss (red) for each model, averaged across all disclosure-summary pairs for S&P 500 firms. Fidelity is highest for Grok-3-mini (0.63), followed by Phi-4 (0.57), Llama-3.3-70B (0.56), and GPT-4o-mini (0.49). Across models, 37 to 51 percent of the source's semantic content is absent from the summary. Beyond the raw omission and injection rates, the embedding measure quantifies how much *meaning* is lost.

In Panel B, we use the Doc2Vec semantic structure to decompose further this fidelity loss into omission, injection, and how each corresponds to lost meaning. Panel B decomposes both fidelity and fidelity loss by vocabulary source. For each disclosure-summary pair, we partition terms into retained (in both), omitted (source only), and injected (summary only), embed each group using Doc2Vec embeddings, and compare these embeddings to those in the full source disclosure. On the signal side (left, blue), the solid band is fidelity carried by retained vocabulary and the hashed band is fidelity carried by injected vocabulary that still tracks the source's themes. On the distortion side (right, red), the solid band is fidelity loss from omitted content the summary drops, and the checkered band is loss from injected content with no source basis. Retained vocabulary projects most directly onto the source, as expected. Injected vocabulary can correspond to both signal and distortion, and our decomposition allows us to see each channel. Roughly 0.14 to 0.18 of each unit bar is fidelity recovered through injected terms, consistent with the model substituting synonyms for omitted terms. On the distortion side, injection accounts for most of the fidelity loss, the empirical analogue of the fabrication cell in Exhibit 2.

3.5 Ensemble Agreement across LLMs and Fidelity Differences

The difference in fidelity across models is another important aspect of these results. For example, Grok-3-mini's fidelity is 0.63 while GPT-4o-mini's is 0.49. This 14-point gap matters because higher fidelity implies less noise.

When several noisy measures capture the same underlying construct, ensemble agreement offers a way to raise the fidelity of the LLM measure. To highlight the value of ensemble agreement, we consider the risk factor classification task of Fathollahi et al. (2026). For each model, we consider subsets of Item-1A disclosures in which the model's risk category is the same as 0, 1, 2 and 3 other LLMs. This allows us to relate LLM agreement in risk factor classification to each model's fidelity.

Employing ensemble agreement across LLMs can be an effective solution to reduce the wedge and enhance the fidelity of the LLM measure used. Figure 4 shows ensemble agreement is effective by plotting each model's average fidelity as a function of how many other models agree with its own risk category classification.[9] The baseline level of fidelity is normalized to zero in the zero-agreement case, so that the bars for 1, 2 and 3 correspond to changes in fidelity from that baseline. Within LLM, we find economically large increases in fidelity when we focus on cases where different LLMs agree on the same classification. Moving from no agreement to unanimous agreement among the other three models raises GPT-4o-mini's fidelity by 0.21 standard deviations—a meaningful shift given GPT-4o-mini's average fidelity of 0.49. All of the other models have a comparable increase in fidelity when insisting on ensemble agreement. The finding that consensus with other models predicts a model's own source-summary similarity is unique evidence that the LLM measures are capturing the same underlying idea, albeit with noise. Put differently, cross-model agreement functions like inter-rater reliability, and the fact it enhances fidelity provides an actionable solution to reduce noise in LLM measurement errors.[10]

3.6 LLM Semantic Drift

In the previous sections, we show there is substantial fidelity loss associated with omission and injection, and through the Doc2Vec decomposition, we also provide unique semantic evidence on the amount of lost meaning. In this section, we examine the notion of semantic *drift* as a complementary analysis. Specifically, we show how important characteristics of the source text are amplified or dampened through LLM summarization. An example of this is when LLMs rewrite hedged legal language ("may have a material adverse effect") as causal assertions ("leads to," "faces," "poses significant threats") absent from the original filing. Panel A of Figure 5 asks whether such rewrites

[9] Figure A.4 provides a complementary view. For each risk category, it plots the share of terms in one model's dictionary that appear in another model's dictionary. Agreement is highest in categories with standardized financial vocabulary — credit, liquidity, financial markets — and lowest in operational and regulatory categories, precisely where injection rates are highest.

[10] The ensemble of LLMs setting is useful for another reason—inference. For generated variables like LLM measures, conventional standard errors are known to be biased because they ignore the variance in the generative process. Re-prompting the model at every bootstrap iteration is prohibitive at scale. When multiple models have been used in concert (i.e., as in using ensemble agreement), the cross-model variance at the observation level gives an estimate of generator uncertainty at no additional inference cost. It can be carried into heteroskedasticity-robust standard errors or used as inverse weights in WLS. As another alternative, token-level log probabilities give a within-model analog to this cross-model variance. We are currently testing whether log-prob–based variance estimates align with the cross-model estimates used here.

are anecdotes or a systematic tonal shift across several key aspects of language, and Panel B asks whether the shift is stable over the sample period.

Panel A shows how summarization shifts tone across seven dimensions. Six are the Loughran-McDonald word lists from Loughran and McDonald (2011): positive, negative, uncertainty, litigious, strong modal, and weak modal, each measured as the share of a document's tokens that appear on the respective list. The seventh dimension is average per-token concreteness from Brysbaert et al. (2014). We compute each measure for every source disclosure and its summary, then standardize both using the cross-sectional mean and standard deviation of the source disclosures. Each bar reports the average summary-minus-source difference for one LLM on one dimension, in units of source standard deviations. A bar at $+1\sigma$ indicates that summaries shift the dimension one source standard deviation above their own source texts. The dashed line at zero marks no drift, and dimensions are ordered by the average shift across the four LLMs.

Three patterns emerge. First, every model raises the per-token density of negative, uncertainty, and weak-modal language above the source baseline, with Phi-4 producing the largest amplification of uncertainty and weak-modal tone. The amplification is a selection effect of compression rather than an injection of new tone. Summaries drop non-tonal words at a faster rate than they drop tonal words in these categories, so density rises even as absolute counts fall.[11] Second, strong-modal words become relatively less frequent while positive and litigious tone hold their source shares, so summarization does not amplify tone uniformly. It shifts documents toward hedged, uncertain language and away from committed language. Third, the cross-model ordering of tonal drift, largest for Phi-4 and smallest for Grok-3-mini, does not match the semantic-drift ordering from Section 3.4 (GPT > Llama > Phi-4 > Grok). Phi-4 preserves overall meaning relatively well while distorting tone the most, and GPT-4o-mini shows the reverse. Fidelity loss is therefore not a single quantity that varies across models. It has a magnitude and a direction, and models that lose similar amounts of meaning can push the surviving text toward different tonal profiles.

The concentration effect in Panel A yields a concrete prediction about the form of the measurement error. Because summaries drop non-tonal vocabulary faster than tonal vocabulary, per-token hedge and negativity density rise even though absolute counts fall. Any downstream measure that reads its input token-by-token and normalizes by length, which describes most tone and

[11] Absolute counts of tonal words fall in every summary relative to its source (Appendix Figure A.6).

sentiment indices, will therefore inherit a denser hedged-and-negative text than the source contained, a systematic shift unrelated to the firm's actual disclosure. Whether this inherited shift translates into construct bias for the four measures we study (severity, complexity, specificity, boilerplate) is the question we take up in Section 4.

Panel B turns attention to temporal stability of semantic drift, plotting three standardized (z-score) series per model over the 2006–2024 window, the document-level Doc2Vec cosine similarity from Section 3.4, the lexical omission rate, and the lexical injection rate, with 95% confidence bands on per-year means. Every series shows year-to-year variation, largest in 2006–2008.[12] The within-LLM co-movement among the three series is model-specific in sign, not merely magnitude, and does not collapse to a shared trend. For GPT-4o-mini, Grok-3-mini and Phi-4 source–summary fidelity co-moves negatively with both the omission and injection rates, so that years of heavier omission and injection coincide with lower fidelity, whereas Llama-3.3 is the exception, its fidelity moving in step with the two rates. [13]

4 Implications of LLM Measurement Error

In this section, we ask whether the LLM measurement wedge matters in practice for empirical research that uses LLM-generated regressors. Specifically, we employ the wedge proxies, omission O_{im} and injection I_{im}, as regressors that proxy for elements of the wedge term in Equation (3). The intuition of this test is to question whether these mechanical properties of summarization carry information about firm outcomes beyond the LLM score itself. If they do, this test is a strong indication of non-classical measurement error. We apply the diagnostic in two forms: to LLM-generated constructs, for both Item 1A disclosures and earnings calls, and to the direct-prompt predictors of Jha et al. (2024).

[12] Note that only 51 firms in the sample included an Item 1A section in their 10-Ks in 2005, the year in which the section was mandated by the SEC. The cosine similarity z-score for 2005 was about 4 and was visually compressing variation in the plots. So we drop the 2005 to better illustrate the time-series variation in the series.

[13] Figure A.5 in the appendix plots omission and injection rates separately by model over the same window. The omisison and injection rates trend upward for gpt-4o-mini and grok-3-mini whereas they are relatively flat for Llama3.3-70B. The omission rate of Phi4 is the outlier that falls by 3 points in 2009 and 2020. However, the rates for all models move within a small band over time.

4.1 Diagnostic Tests of Non-Classical Measurement Error

We begin by examining LLM measures constructed via a scoring task. Using a structured prompt (see Appendix Figure A.10), we elicit four LLM-generated constructs—severity, complexity, specificity, and boilerplate—from all four LLMs in our study. We elicit these four constructs because they are commonly used in the literature.[14]

For each LLM-derived regressor L_{im} (each LLM-construct combination) and a firm outcome $y_{i,t}$, we estimate three specifications that jointly shed light on measurement error.

The first is the naive regression of equation (1), which embodies standard empirical practice,

$$y_{i,t} = \beta_{\text{naive}}\, L_{im,t} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t}, \tag{6}$$

where $X_{i,t} = \{\text{Size, MTB, CashFlow, Tangibility, Age}\}$ are standard firm controls and τ_t are year fixed effects. The second specification additionally includes the wedge proxies (omission and injection rates) as controls,

$$y_{i,t} = \beta_{\text{joint}}\, L_{im,t} + \rho_O\, O_{im,t} + \rho_I\, I_{im,t} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t}. \tag{7}$$

When estimating this specification, we follow exactly the same specification choices as the naive regression, and we summarize how much these controls shift the estimated coefficient by reporting the coefficient sensitivity $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$.

The third specification decomposes the measure into a wedge-predictable part and a wedge-orthogonal residual. We first estimate a Tobit regression of L_{im} on the wedge proxies and controls, yielding the predictable component $\widehat{L_{im}}$ and the residual $\hat{e}$, then estimate

$$y_{i,t} = \theta\, \widehat{L_{im,t}} + \lambda\, \hat{e}_{i,t} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t}. \tag{8}$$

By the Frisch–Waugh theorem, λ equals β_{joint}. This formulation enables a direct test of whether

[14] *Severity* is studied by Campbell et al. (2014) and Kravet and Muslu (2013), who link risk-factor severity to stock-return volatility and cost of equity. *Specificity*, the degree to which disclosures provide firm-specific rather than generic risk information, was developed by Hope et al. (2016), who show that specific risk-factor disclosures improve analyst forecast accuracy. *Complexity* is related to readability and plain-English literature (Loughran and McDonald, 2014; Bonsall IV et al., 2017). *Boilerplate*, the share of disclosure that is formulaic rather than informative, was studied by Lang and Stice-Lawrence (2015) and Dyer et al. (2017).

the wedge-driven channel θ and the wedge-free channel λ relate to the outcome differently.

Measurement error is *classical* when the true LLM wedge w_m is uncorrelated with the true construct. Our test is informative because omission and injection, if they relate to the outcome, are not uncorrelated with the construct. As we highlighted in Exhibit 2, omission can bias inference by dropping meaningful terms. Similarly, injection imports context from the LLM's priors, which can bias inference on the extrinsic dimension. When the true LLM wedge w_m is systematically linked to the outcome via these channels, the error is *non-classical*. The diagnostic regressions we propose help to flag non-classical measurement error. For example, a large difference between the naive and joint coefficient estimates ($\beta_{\text{naive}} \neq \beta_{\text{joint}}$) rejects the classical benchmark, as does a nonzero wedge loading on omission and injection rates (ρ_O or ρ_I), or different weights on the wedge-predictable versus wedge-orthogonal components ($\theta \neq \lambda$).

4.2 Measurement Error in Corporate Finance LLM Measures

To implement this diagnostic in a representative corporate finance setting, we estimate the three specifications across four LLMs, six corporate finance outcomes, and four constructs (severity, complexity, boilerplate, specificity). Table 4 reports the diagnostics aggregated to the outcome level, separately for the Item 1A sample (four LLMs) and Table 5 reports the same for the earnings-call sample (two LLMs). For each outcome, the tables report the mean coefficient across LLMs and specifications for β_{naive}, β_{joint}, the wedge coefficients ρ_O and ρ_I, and the decomposition components θ and λ, together with the median coefficient sensitivity $|\Delta\%|$. Bracketed entries give the share of underlying specifications (LLM $\times$ fixed effects choice) significant at the 5% level, except the bracket under λ, which reports the share of specifications where θ and λ differ significantly.

Our diagnostic reveals non-classical measurement error for Item 1A disclosures in Table 4. Specifically, the wedge-predictable component and the wedge-free residual frequently exhibit statistically different relationships to the outcome ($\theta \neq \lambda$). This is the defining signature of non-classical error. Further, the omission and injection load directly on outcomes, with ρ_O significant in a majority of specifications for five of six outcomes. The median coefficient sensitivity $|\Delta\%|$ is large, from 12.7% on the R&D ratio to 40.3% on leverage. The constructs' own significance also diminishes once the wedge terms enter, most visibly for capital expenditure, where the share of significant specifications falls from 25% to 12.5%.

For earnings-call transcripts in Table 5, the same patterns appear, albeit smaller in magnitude. Median coefficient sensitivity $|\Delta\%|$ ranges from 6.8% on leverage to 19.3% on capital expenditure, and only capital expenditure exceeds 10%. The construct's naive significance is largely preserved under the joint specification, and β_{naive} and β_{joint} move little across outcomes. The omission and injection rate terms remain significant in several specifications, with ρ_O significant in all R&D specifications and ρ_I in all sales-growth specifications, but the pattern is sparser, and θ and λ differ significantly for only a minority of specifications on most outcomes. The construct-outcome relationship on transcripts does not split cleanly into wedge and residual channels the way it does on Item 1A. One interpretation, consistent with the tonal drift of Section 3.6, is that the passive, hedged legal prose of Item 1A gives the model more to rewrite than the already-active, causal language of earnings-call commentary.

4.3 Heterogeneity and Richer Wedge Proxies

For these main diagnostic results, there are two important caveats. First, coefficient sensitivity might be mechanically larger if the true relationship is small. Second, omission and injection merely capture the *potential for bias* without characterizing the full meaning loss and semantic drift we documented earlier. Meaningful shifts in the estimated coefficients from including omission and injection rates, thus, provide a useful warning sign about measurement error but not a full characterization of it. In this section, we empirically characterize each of these issues by presenting evidence on the full distribution of coefficient sensitivity, we also consider the distribution restricting to statistically significant cases, and we directly include semantic wedge terms that capture meaning loss and semantic drift.

4.3.1 Coefficient Sensitivity Distributions

The median coefficient sensitivity describes the typical bias a researcher can expect, but wide dispersion across specifications may flag greater risk in any single application. To characterize this risk, Figure 6 plots the full distribution of $|\Delta\%|$ across the 288 Item 1A specifications and 144 earnings-call specifications (LLM $\times$ construct $\times$ outcome $\times$ fixed effects variant). As we reported in the tabular output, for Item 1A, most coefficient sensitivities exceed 10% (indicated by the dashed line), while for earnings calls, fewer coefficient sensitivities exceed this threshold. However, the

heterogeneity in these estimates is meaningful. There is a nontrivial right tail still exceeds 20%, even in the earnings call corpus. Given that the empirical variation in coefficient sensitivities may reflect a true underlying risk that measurement error could have meaningful influence, this evidence highlights a distinct risk of using LLM measures.

To help alleviate the concern that high coefficient sensitivity is a mechanical feature of having small estimates, Panels (B) and (d) restrict to naive-significant specifications. The Item 1A sensitivity survives the restriction—the median remains above 10%—so the high coefficient sensitivity is not an artifact of small naive estimates.For earnings calls, a few naive-significant cells still show sizable magnitudes, though the picture is distinctly less severe than for Item 1A disclosures.

4.3.2 Content-Aware Wedge Proxies

In the above tests, the omission and injection rates count what fraction of the important vocabulary changed, not whether the change altered meaning. Next, we extend the previous analysis to also incorporate wedge terms that are "content aware" by leaning on the Doc2Vec embeddings from the earlier fidelity analysis. Specifically, for each disclosure and model, we collect the omitted discriminative terms into one set and the injected discriminative terms into another, embed each set, and measure each set's cosine similarity to the embedding of the full source document. The omitted-set's cosine ρ_{BO} asks whether the content the model dropped was thematically central to the source, and the injected-bag cosine ρ_{BI} asks whether the content the model added was grounded in the source's themes or imported from elsewhere. These proxies are potentially powerful because they distinguish a dropped synonym from a dropped material risk, and a grounded paraphrase from a fabricated association.[15]

To incorporate these new wedge terms, we add ρ_{BO} and ρ_{BI} alongside O_{im} and I_{im} in equations (7) and (8) and re-estimate our diagnostic specifications. Figure 7 presents the full distribution of coefficient sensitivities. For Item 1A disclosures in the first two panels, this distribution is similar to using just omission and injection. However, for the earnings calls, we see a distinct shift to the right in the coefficient sensitivities. Whereas the median coefficient sensitivity with omission and injection alone roughly doubles when also considering content-aware wedge proxies . Viewed through this

[15]Another advantage of using the Doc2Vec embeddings is that they are trained only on disclosures up to time t, avoiding lookahead bias.

lens, the richer proxies deepen our evidence of non-classical measurement error, rather than explain it away. Stepping beyond coefficient sensitivity, the aggregated per-outcome diagnostic tests tell a similar story. See these results summarized in the appendix Table A.2.

4.3.3 Control Function Interpretation as a Partial Correction

Whereas the content-agnostic Omission and Injection wedge terms map purely to the distortion potential of LLM-generated regressors, the content-aware versions provide an additional advantage. In the decomposition of equation (8), λ is the coefficient on the part of the score orthogonal to the wedge proxies, so it estimates the construct-outcome relationship after the wedge-correlated variation has been filtered out. Read this way, the diagnostic using the added terms is closer to a traditional control-function.

Two properties make the filter attractive as a potential correction beyond diagnosing non-classical measurement error. First, this correction leverages the LLM's signal rather than discarding it. The score L_{im} is retained, and only its wedge-correlated component is removed, so the coefficient is identified from variation from the genuine construct extracted from the LLM while stripping away variation due to omission, injection, and with the content-aware wedges, semantic distortion. Second, because the Doc2Vec embedding is trained only on pre-t text, this approach neutralizes the extrinsic tendency of LLMs to exhibit lookahead bias.

4.4 Application to Direct Prompting

The previous analysis shows construct-style LLM measures (severity, complexity, boilerplate, specificity) have non-trivial potential for bias. However, LLM-derived regressors often take a different form. For example, Jha et al. (2024) prompt an LLM with single-question, outcome-specific tasks and use the resulting directional scores as predictors of realized firm outcomes (e.g., *based on the earnings call, will the firm increase capital expenditure?*). This is the leading empirical example of what we call *direct-prompt* LLM measurement. Given the prevalence of each approach in empirical research, it is natural to ask whether this prompt form is subject to the same bias we found for construct-style measures.

To apply our diagnostic in this setting, we start by replicating Jha et al. (2024) on our S&P 500 earnings-call corpus (2006–2020) using `gpt-4o-mini` and `llama-3.3-70B`. Both LLM choice and

sample scope differ, so this first step is crucial to establish a correspondence between our settings.

Specifically, we extract the seven direct-prompt elicitations for the outcomes considered in Jha et al. (2024) (capital expenditure, cash holdings, dividend, leverage, operating cash flow, R&D, and sales-growth direction) and following their prompting. We then relate each direct-prompt measure at t to its matched outcome at $t+2$ as is done in their paper, predicting future capital expenditure with the capital-expenditure prompt, future cash holdings with the cash prompt, and so on. We use their controls $\{\text{Size}, q_{\text{tot}}, \text{CashFlow}, \text{Leverage}_t, y_{i,t}\}$, where q_{tot} is the Peters–Taylor Tobin's Q and $y_{i,t}$ is the date-t value of the dependent variable, and firm and year-quarter fixed effects.

In Table A.3, we replicate three of Jha et al. (2024)'s headline regressions, the ChatGPT Investment score forecasting capital expenditure at t+2 (their Table 3, col. 4), capital expenditure at t+4 (Table 4, col. 2), and Peters–Taylor-scaled R&D at t+2 (Table 5, col. 8), following their equation (2). The capital-expenditure result reproduces with a somewhat larger magnitude than the published estimate. At t+2, the standardized betas are 0.080 (`gpt-4o-mini`) and 0.084 (`llama-3.3-70B`), versus 0.055 to 0.079 in the original, and at t+4 they are 0.090 and 0.105, slightly above their 0.060 to 0.080 range. The R&D result is weaker, with standardized betas of 0.006 and 0.007 that fall below their 0.020 to 0.025 range and are insignificant ($t = 0.92$ and $t = 1.23$). Since Jha et al. (2024) themselves report a smaller effect for R&D than for capital expenditure, the relative weakness is consistent across their setting and ours, though the absolute magnitudes are not identical. Given the sample and LLM differences, this constitutes a tight replication of their main findings.

We then apply the same wedge diagnostic used on the construct measures, adding omission and injection to each direct-predictor regression alongside the predictor itself. For each firm-quarter and LLM, the source-side vocabulary is the discriminative-term set from the full earnings-call transcript. Because the direct-prompt task returns a score rather than a summary, we use the model's own rationale text as the summary-side object, pooling the rationales across the seven scores to obtain sufficient text. The omission rate O_{im} is the share of source-side discriminative terms absent from the rationales, and the injection rate I_{im} is the share of rationale discriminative terms absent from the transcript. This is the closest available analog to the construct wedge, and if anything it overstates the wedge for direct prompts, since short rationales mechanically inject vocabulary.

Table 6 reports the diagnostic on the seven direct predictors. The naive coefficient is significant for five of the seven outcomes, capital expenditure, cash, dividend, leverage, and sales growth,

and insignificant for operating cash flow and R&D. In the five significant outcomes the predictor coefficient shifts only modestly when the wedge proxies enter. Median $|\Delta\%|$ ranges from 0.2% on sales growth to 6.6% on dividend, with capital expenditure (2.3%), leverage (3.6%), and cash (4.1%) in between, and no naive-significant cell loses significance under the joint specification. The two outcomes where median $|\Delta\%|$ exceeds 10%, operating cash flow at 28.3% and R&D at 35.2%, are exactly the two where the naive coefficient is itself insignificant, so the large ratios reflect near-zero denominators rather than contamination.

The decomposition tells the same story. Across every cell, λ closely tracks β_{naive} and θ on the wedge-predictable component is small and statistically indistinguishable from zero. The wedge proxies do load on outcomes, with ρ_O individually significant in five of fourteen cells and ρ_I in two, but that loading does not flow through the predictor's own coefficient. Capital expenditure is the clearest case. ρ_O is significant in 100% of the underlying cells, so omission genuinely predicts future capital expenditure, yet the predictor coefficient shifts by only 2.3% between the naive and joint specifications. The wedge carries information about the outcome, but not information the direct predictor is already carrying, which is the signature of a predictor whose signal is orthogonal to the wedge.

Turning to heterogeneity, Figure A.7 shows that the low average sensitivity comes with a correspondingly low chance of a case in which measurement error meaningfully shifts the estimated relationship. Across matched predictor-outcome pairs, LLMs and fixed effects variants (year-quarter; year-quarter plus industry; year-quarter plus firm) we diagnose 42 scenarios. Panel (a) plots the full coefficient sensitivity $|\Delta\%|$ distribution and Panel (b) restricts to naive-significant cells. The naive-significant distribution concentrates below the 10% threshold with a median near 6%, confirming that the within-cell magnitudes in Table 6 are not an artifact of the firm plus year-quarter specification. However, it is notable that a few scenarios — even among the naive-significant cells — produce coefficient sensitivity greater than 20%. That is, the risk is lower in the case of direct prompting, but our proposed diagnostic applied to a particular or novel setting remains useful even when direct prompting.

Our analyses reveal a clear ordering of where measurement error is most severe, and in doing so demonstrate the diagnostic's value beyond our setting — across prompting designs, samples, and model technologies. For the corporate outcomes we study, contamination is largest for construct

ratings on long Item 1A disclosures, milder for the same constructs on shorter earnings-call transcripts, and negligible for single-question direct prompts. The ordering links compression severity to prompt abstraction in a natural fashion. The risk of non-classical bias rises with the length of the source text and the abstraction of the task, and it appears to be greatest exactly where LLMs are most useful—distilling long, unstructured disclosures into compact scores.

5 Conclusion

Large language models are increasingly used as measurement instruments in empirical research, but every LLM output transforms the input text in non-random ways. We formalize this concept by treating any LLM output as a generated regressor. In our framework, LLM measures are noisy proxies for latent constructs that are contaminated by a model-specific wedge. This wedge reflects how the model compresses input, draws on its priors, and allocates attention. Considering these inner workings of LLMs, we find that they can violate classical measurement-error assumptions, and thus, lead to biased estimates when employed as predictors of firm outcomes.

Beyond this conceptual point, we identify two key underlying properties of LLM summarization – omission of terms from the source and injection of terms not in the source – as drivers of how LLMs transform output in non-random ways, and we measure these properties by comparing source and summary vocabulary. Both omitted and injected vocabulary are non-random and reshape meaning, which we show in tests using Doc2Vec embedding of the source text. Further, we document that injection rates, omission rates, and lost content rates correlate with corporate finance outcomes above and beyond LLM measures.

These findings signal meaningful and non-classical measurement error, which exhibits substantial heterogeneity across different types of risk while being present across all four LLMs we analyze. Cross-LLM differences in injection and omission suggest a partial antidote – ensemble agreement, which mirrors cross-rater agreement in data collection tasks. When we condition on cases when LLMs agree with one another on a classification, the measurement wedges from omission and injection are substantially smaller.

Finally, we develop a control function diagnostic to evaluate non-classical measurement error, and we apply it to three corporate finance settings. Construct-level ratings in Item 1A disclosures,

those same ratings in earnings calls, and direct-prompt measures. These analyses indicate that accounting for measurement error can dramatically change the estimated coefficients, especially when LLMs are used to measure constructs like severity of a particular corporate finance risk.

We expect that our proposed diagnostic can be useful beyond our application, and that ensemble agreement can sharpen LLM empirical measurement. More generally, by shifting the focus from LLM outputs as data to LLM outputs as generated regressors with a measurable wedge, we expect future work to keep extracting useful signals that can better withstand scrutiny.

References

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2024). Inference for regression with variables generated by ai or machine learning. *arXiv preprint arXiv:2402.15585*.

Bhagwat, V., Cookson, J. A., and Chukwuma Dim, M. N. (2025). The market's mirror: Revealing investor disagreement with LLMs. *SSRN Working Paper*.

Bonsall IV, S. B., Leone, A. J., Miller, B. P., and Mergenthaler, K. (2017). A plain english measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Bybee, L. (2024). The ghost in the machine: Generating beliefs with large language models. *Working Paper*.

Bybee, L. (2025). LLM-generated beliefs and macroeconomic expectations. *Working Paper*.

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):396–455.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *Journal of Finance*, 75(3):1371–1415.

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):221–245.

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*.

Eisfeldt, A. L., Schubert, G., Zhang, M. B., and Taska, B. (2023). Generative AI and firm values. *SSRN Working Paper No. 4436627*.

Engelberg, J., Manela, A., Mullins, W., and Vulicevic, L. (2025). Entity neutering. *SSRN Working Paper 5182756*.

Fathollahi, M., Grieser, W., and Venugopal, B. (2026). Operational resilience: Managing unhedgeable risks. *Working Paper*.

Fong, C. and Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 708–719.

He, S., Lv, L., Manela, A., and Wu, J. (2025). Chronologically consistent large language models. *arXiv preprint arXiv:2502.21206*.

Hoberg, G. and Phillips, G. M. (2025). Scope, scale, and concentration: The 21st-century firm. *The Journal of Finance*, 80(1):415–466.

Hope, O.-K., Hu, D., and Lu, H. (2016). The benefits of specific risk-factor disclosures. *Review of Accounting Studies*, 21(4):1005–1045.

Huang, A. H., Wang, H., and Yang, Y. (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Jha, M., Qian, J., Weber, M., and Yang, B. (2024). Chatgpt and corporate policies. *NBER Working Paper 32161*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Kim, A., Muhn, M., and Nikolaev, V. (2023). Bloated disclosures: Can ChatGPT help investors process information? *Working Paper*.

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

Lang, M. and Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2-3):110–135.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Working Paper*.

Lopez-Lira, A. and Tang, Y. (2024). Large language models and return prediction. *Working Paper*.

Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Working Paper*.

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

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69(4):1643–1671.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2024). The unreasonable effectiveness of large language models in forecasting. *Working Paper*.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. *NBER Working Paper 33344*.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919.

Murphy, K. M. and Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 3(4):370–379.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of EMNLP*, pages 1797–1807.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1):221–247.

Sarkar, S. K. and Vafa, K. (2024). Lookahead bias in pretrained language models. *SSRN Working Paper 4754678*.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of ACL*, pages 1073–1083.

Sikochi, A. S., Cohen, L., and Malloy, C. (2024). LLM-assisted ESG disclosure credibility assessment. *Working Paper*.

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

Figure 1: **Omission and Injection Across Risk Factors**

In panels A and B, we plot the omission and injection rates for discriminatory terms identified in our S&P 500 sample for each risk category. Omission is defined as $O_{im} = 1 - \frac{|V_{i,\text{source}} \cap V_{im,\text{summary}}|}{V_{i,\text{source}}}$. Injection is defined as $I_{im} = 1 - \frac{|V_{im,\text{summary}} \cap V_{i,\text{source}}|}{V_{im,\text{summary}}}$. We define *source* and *summary* vocabularies as the count of unique discriminative terms used by companies in risk disclosures Item 1A that we identify and LLM summaries, respectively. Higher omission indicates greater vocabulary reduction, and higher injection indicates more novel terminology introduced by the LLM.

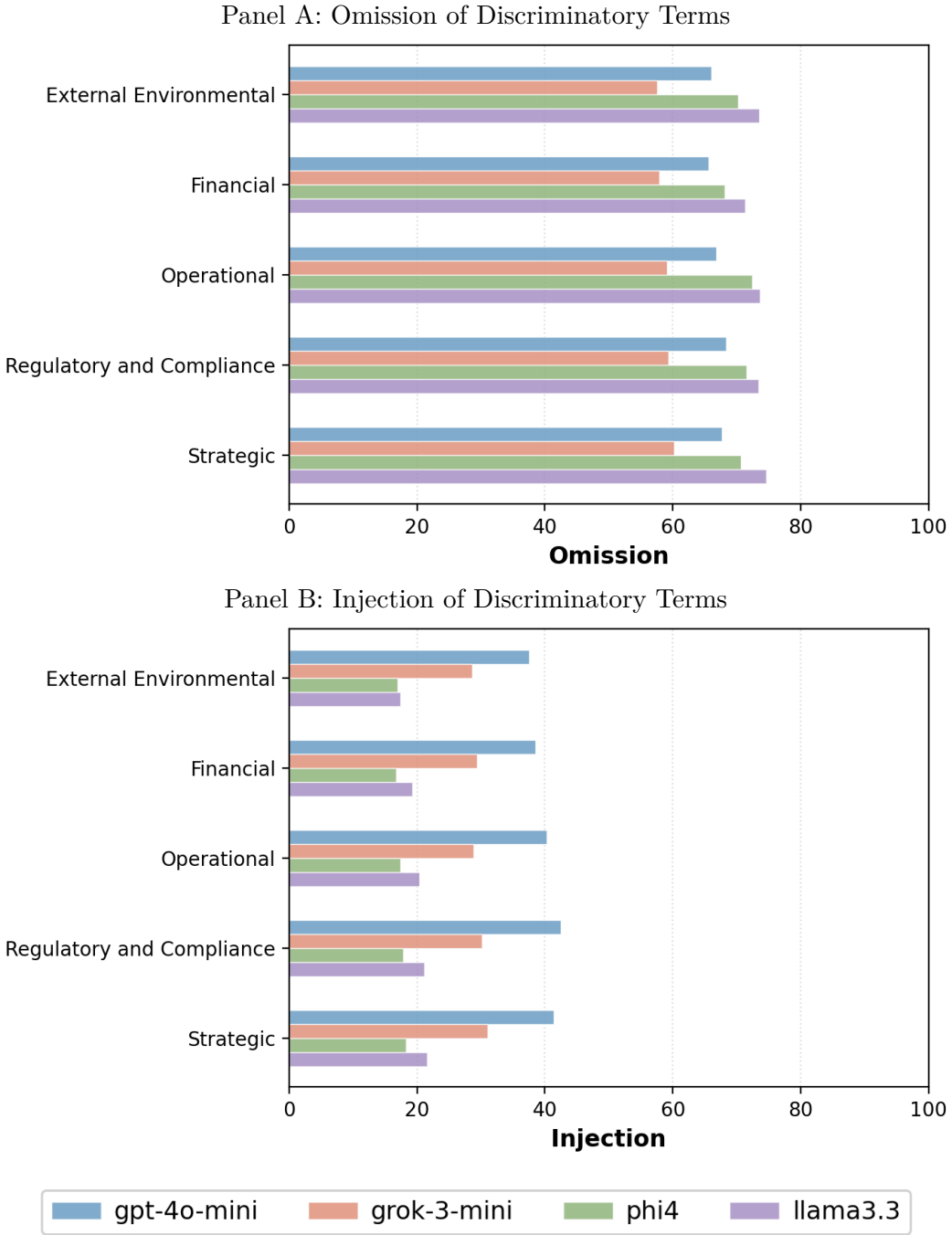


Figure 2: **Model Attention and LLM-generated Regressors.** In this figure, we plot the average per-token attention mass that Llama-3.3-70B and Phi-4 allocate to the first 10%, middle 80%, and last 10% of the input Item 1A when prompting for severity, specificity, complexity, and boilerplate (generated regressors). Per-token attention mass is the share of total attention in a position bin divided by the bin's share of input tokens; a value of 1.0 (dotted line) denotes uniform attention. Attention weights are extracted at the token where each model emits its continuous score for that construct, averaged across all transformer layers and attention heads. Means are taken across the 4,829 disclosure chunks (each up to 2,000 tokens) in our sample.

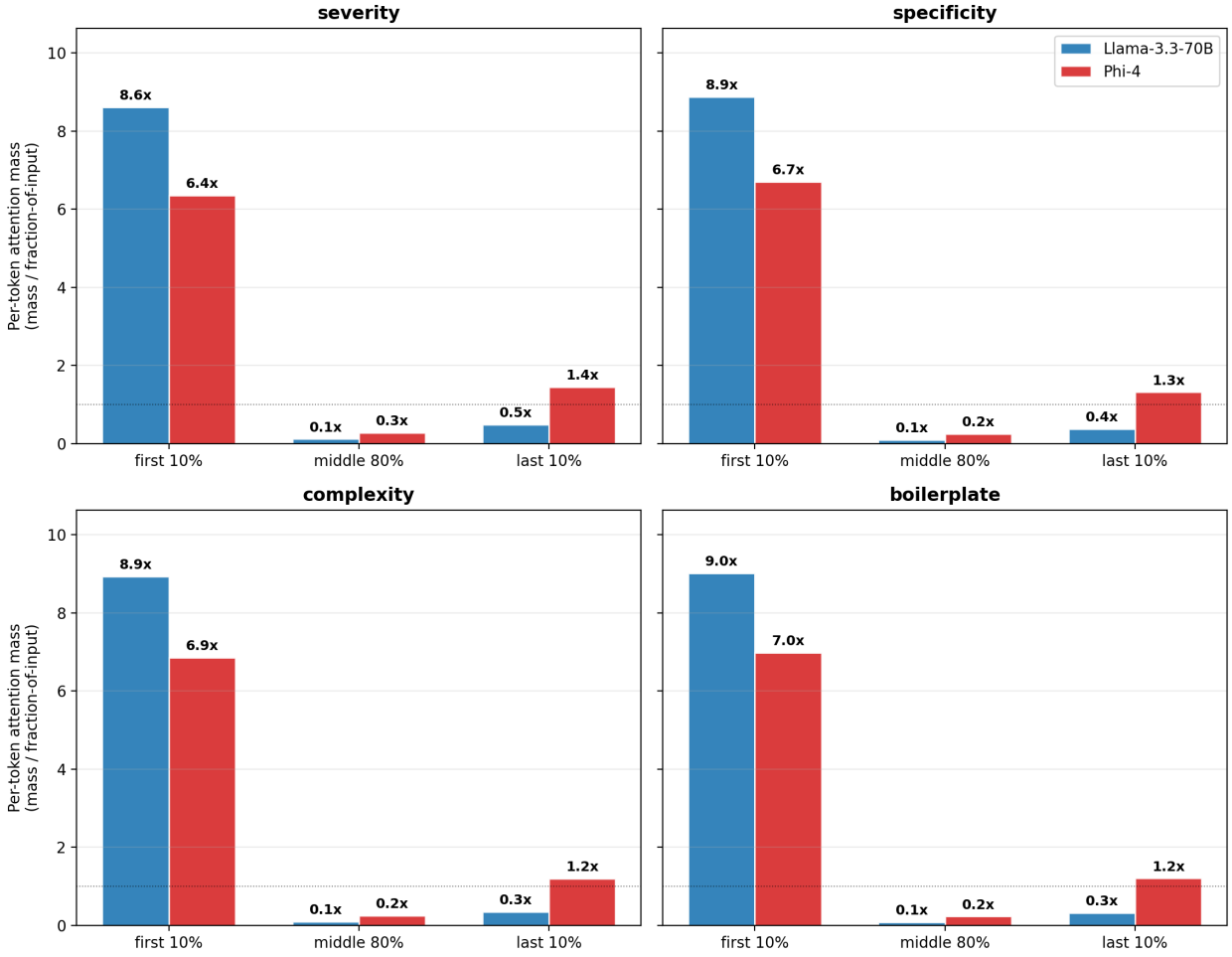
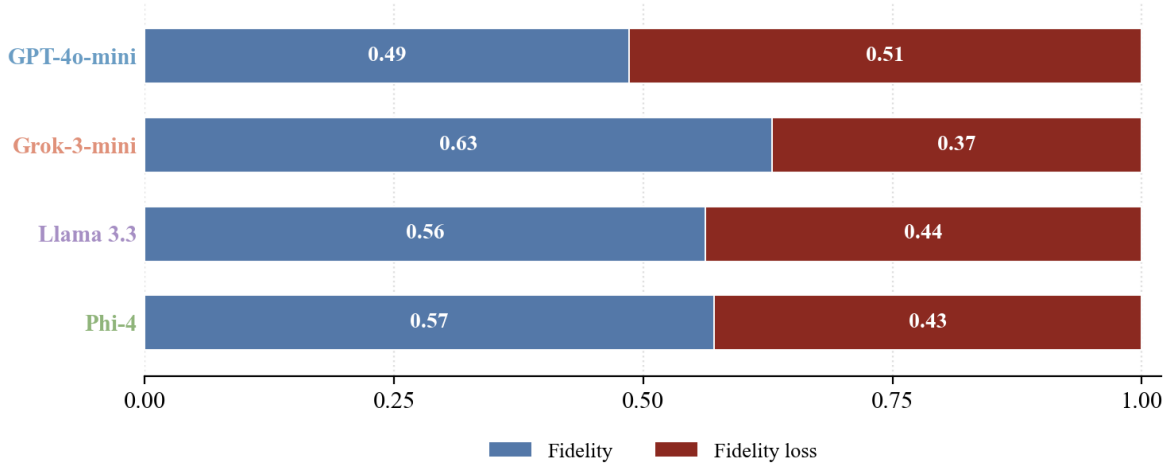


Figure 3: **Fidelity Lost During LLM Summarization**

This figure summarizes how much of an Item 1A risk disclosure survives in LLMs' risk factor summaries, and decomposes the signal and distortion. We measure Fidelity as the cosine similarity between the Doc2Vec embedding of an Item 1A disclosure (source) and the embedding of summaries of all risk factors disclosed in that disclosure pooled into a single document; fidelity loss is one minus fidelity. To keep the measure free of look-ahead bias, the Doc2Vec model for each period is trained only on source documents available up to time t, so it cannot encode information that postdates the disclosure. Panel A reports the average total fidelity (blue) and total fidelity loss (red) over 25,312 Item 1A–pooled LLM summaries for S&P 500 firms, separately for each of four LLMs (GPT-4o-mini, Grok-3-mini, Phi-4, and Llama-3.3-70B). Panel B presents the fidelity decomposition. For each disclosure-summary pair we split all words into three groups by where they appear: *retained terms* occur in both the source and the summary, *omitted terms* occur in the source but not the summary, and injected terms occur in the summary but not the source. We embed each group and compare it to the complete source disclosure. The signal side (left) is content the summary preserves: vocabulary retained from the source, plus injected vocabulary that still tracks the source's themes, which reflects benign aggregation rather than fabrication. The distortion side (right) is content the summary breaks: source vocabulary the summary fails to anchor (omission), plus injected vocabulary with no source basis (injection). The two signal components scale to total fidelity, the two distortion components to total fidelity loss, and the four sum to one.

Panel A: LLM Fidelity



Panel B: Fidelity Decomposition (Signal vs. Distortion)

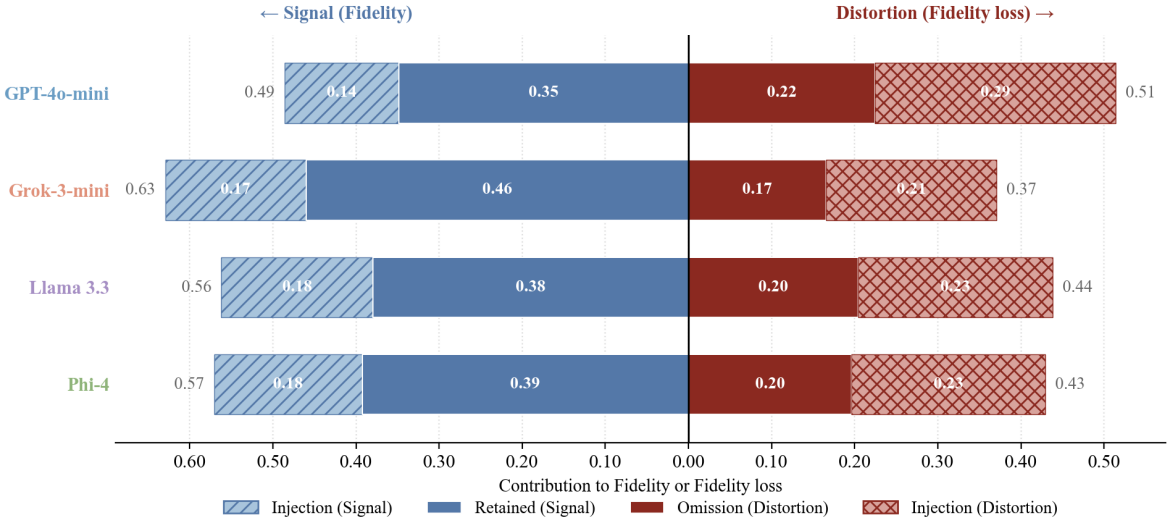


Figure 4: **LLM Ensemble Agreement of Risk Factor Categories and Embedding Similarity**

The figure plots the shift in mean Doc2Vec cosine similarity between an LLM's source-text disclosure and its corresponding summary text (the LLM's own description of the risk factor), expressed in standard-deviation units and measured relative to a bucket-0, grouped by the number of other LLMs that assigned the same primary subcategory to the underlying risk factor. The x-axis takes values in 0, 1, 2, 3 representing the number of other models that agree with the focal model's risk factor category classification. The y-axis is the cosine similarity shift in σ units (vs bucket 0): for each LLM, we subtract the bucket-0 mean cosine from each disclosure's cosine and divide by that LLM's standard deviation, then average within each bucket; bucket 0 sits at exactly zero by construction and bucket k reads as how many standard deviations of that LLM's cosine distribution the bucket-k mean sits above bucket 0. This anchors all four LLMs at zero in bucket 0 and isolates the agreement-driven shift on a comparable scale across models. Embeddings come from a single PV-DBOW Doc2Vec model trained on the union of source and summary documents (300-dimensional vectors, window size 15, minimum count 5, 40 training epochs; hyperparameters following Hoberg and Phillips (2025)).

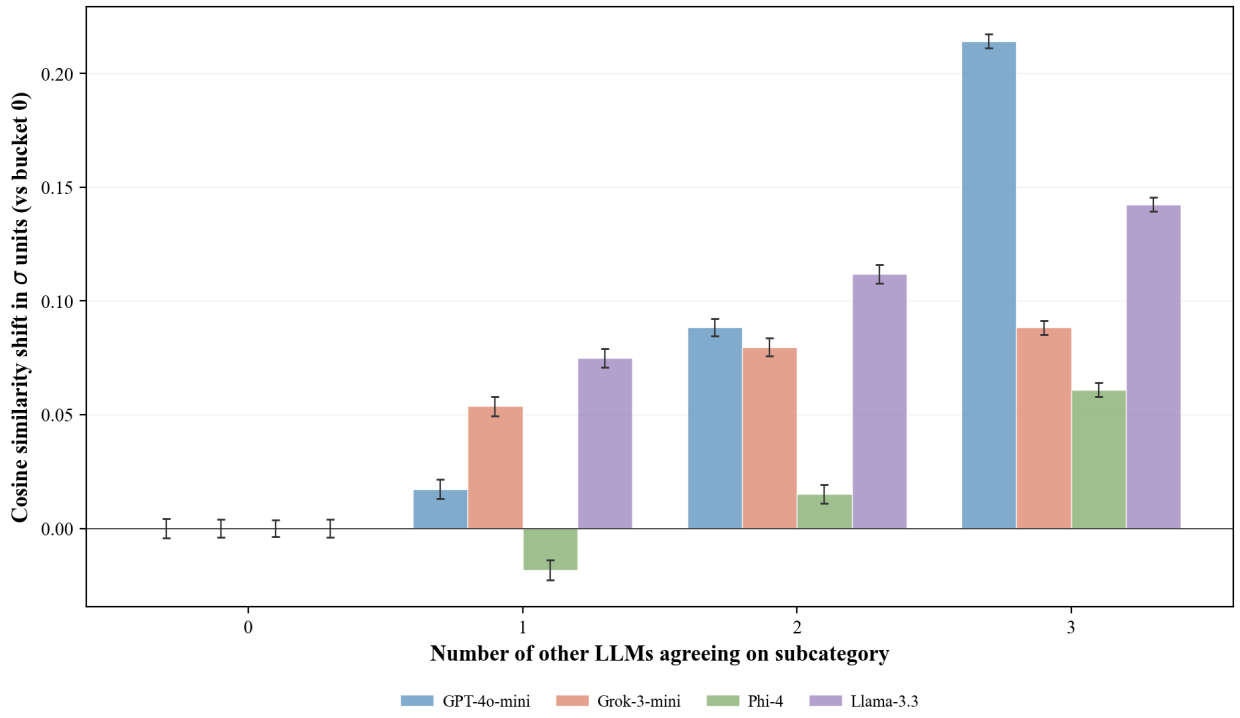
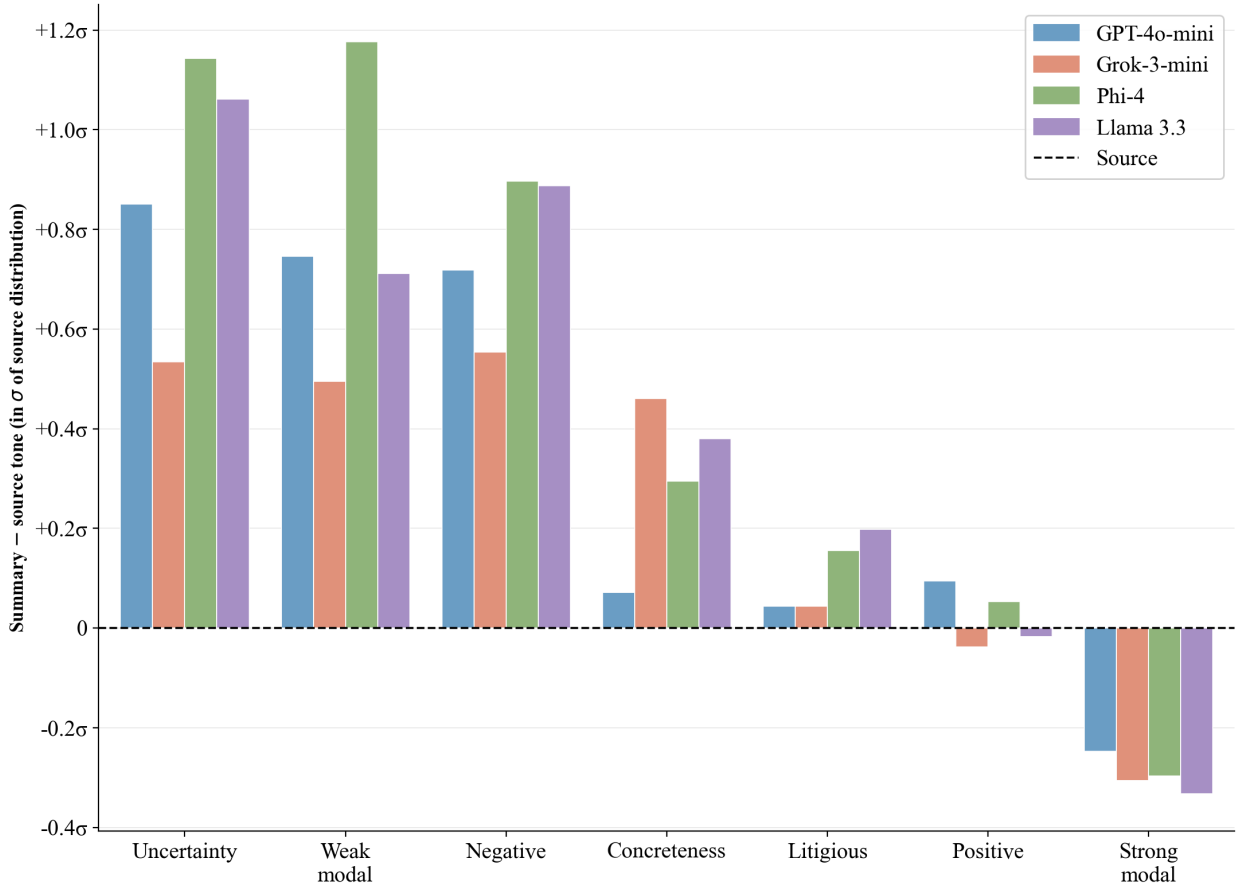


Figure 5: **LLM Summary Tone and Semantic Drift**

In panel A we plot, for each LLM, the average shift in tone from the source Item 1A risk disclosure to its LLM summary across a seven-dimensional tonal space: six Loughran-McDonald tone categories — positive, negative, uncertainty, litigious, and strong- and weak-modal — each measured as a share of the text's tokens, plus Brysbaert et al. (2014) concreteness (a per-token average). Each dimension is standardized against the distribution of source disclosures, so a bar of $+1\sigma$ means the summary's density on that dimension lies one source-standard-deviation above the typical source disclosure. Colored bars give the summary-minus-source difference for each LLM; the dashed black line at 0σ marks the source baseline, and the vertical axis is labeled in standard-deviation units. Dimensions are ordered by their average shift across the four LLMs. In panel B we plot three time series for each LLM: source–summary document-level Doc2Vec cosine similarity (black) computed over an expanding training corpus, Omission rate (red), and Injection rate (blue). The lexical omission and injection rates are computed over each LLM's discriminative-term vocabulary (top-1000 by lift per risk-factor subcategory). All three series are z-scored within (LLM × metric) so that the within-LLM year-to-year co-movement between metrics reads on a common scale. Shaded bands are 95% confidence intervals.

Panel A: Changes in tone: Summary−Source, by LLM and Dimension



Panel B: Changes in Semantic similarity, Omission, and Injection rates over time

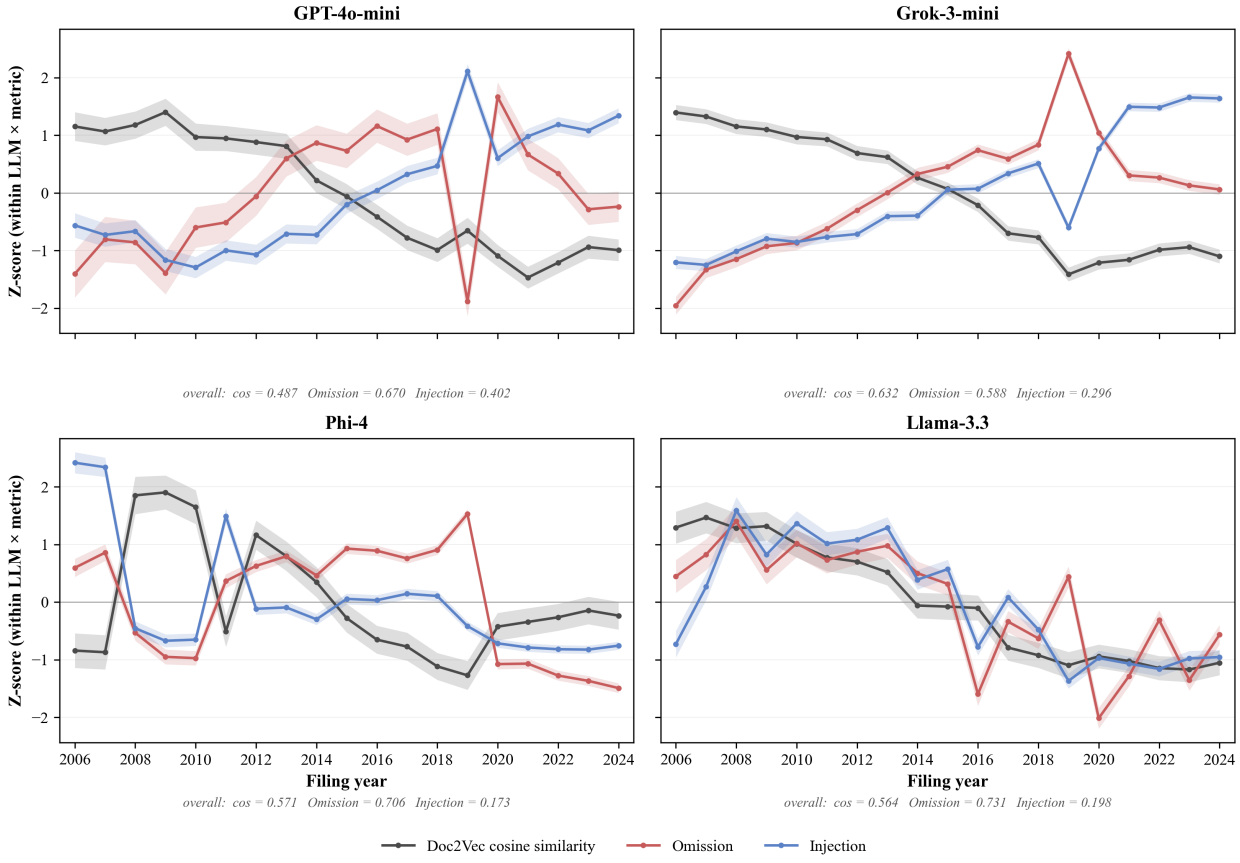
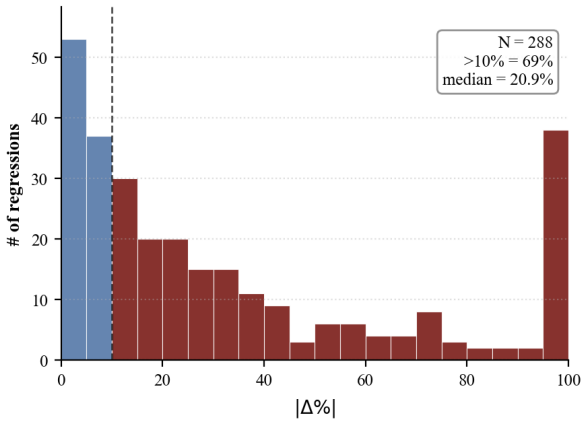


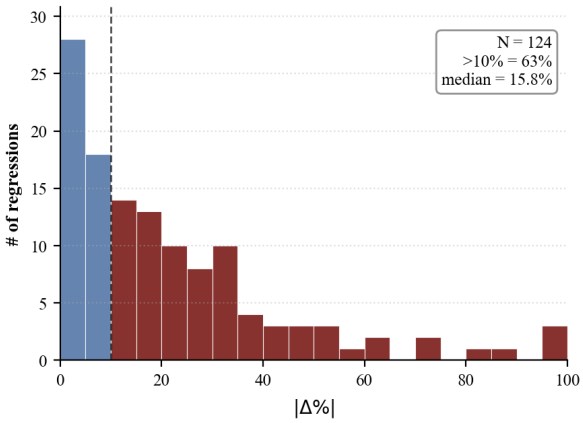
Figure 6: **Diagnostic Distributions**

This figure plots the difference between LLM-generated regressor coefficients in naive (Eq. 6) and joint (Eq. 7) regressions, summarized in Table 4. The joint regressions in these plots include the lexical wedges: Omission and Injection rates. Panels (A) and (B) report the results for Item 1A sample; panels (C) to (D) report the results from the earnings call sample. Coefficient sensitivity is $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$. The Item 1A panels are based on 4 LLM models (gpt-4o-mini, grok-3-mini, phi-4, llama-3.3-70B) $\times$ 4 constructs $\times$ 6 outcomes $\times$ 3 fixed effects variants (288 cells). The earnings call panels are based on 2 LLM models (gpt-4o-mini, llama-3.3-70B) $\times$ 4 constructs $\times$ 6 outcomes $\times$ 3 fixed effects variants (144 cells). The red dashed line marks $|\Delta\%| = 10\%$; blue bars sit below the threshold and red bars above. The right-hand summary box reports N, the share of cells above 10%, and the median. Panels (B) and (D) restrict the plots to regressions where the naive construct coefficient is significant at the 5% level.

Item 1A Risk Disclosures

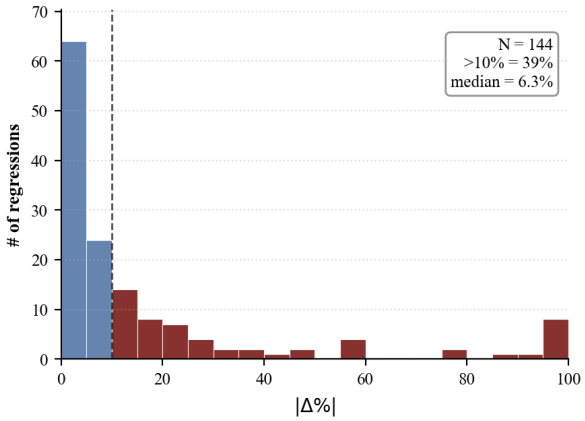


Panel A: All regressions.

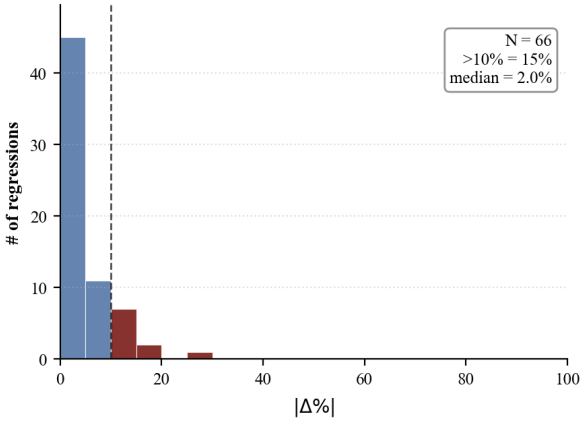


Panel B: Naive significant regressions

Earnings Call Transcripts



Panel C: All regressions.

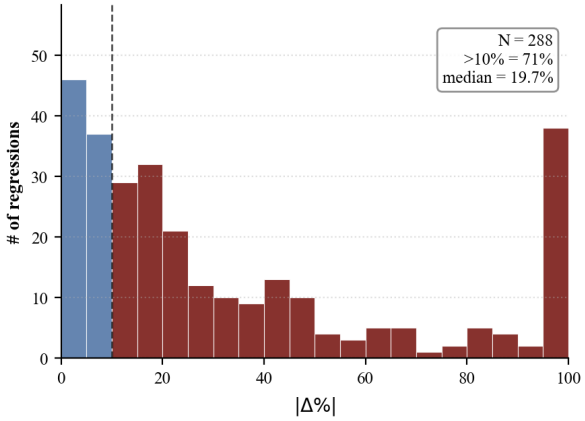


Panel D: Naive significant regressions

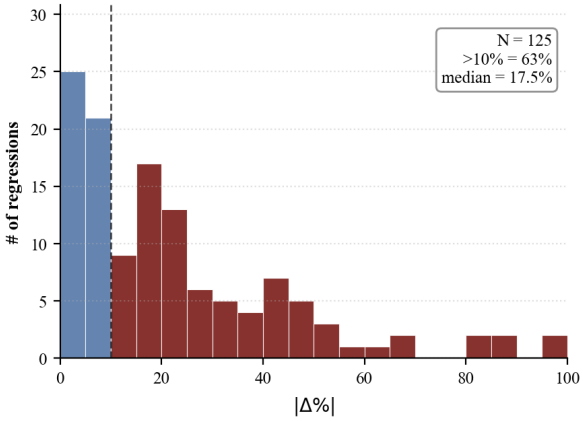
Figure 7: **Diagnostic Distributions with Semantic Wedges Included**

This figure plots the difference between LLM-generated regressor coefficients in naive (Eq. 6) and joint (Eq. 7) regressions, summarized in Table A.2. The joint regressions in these plots include the lexical and semantic wedges: Omission and Injection rates, and the cosine similarity of omission and injection bags. Panels (A) and (B) report the results for Item 1A sample; panels (C) to (D) report the results from the earnings call sample. Coefficient sensitivity is $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$. The Item 1A panels are based on 4 LLM models (gpt-4o-mini, grok-3-mini, phi-4, llama-3.3-70B) $\times$ 4 constructs $\times$ 6 outcomes $\times$ 3 fixed effects variants (288 cells). The earnings call panels are based on 2 LLM models (gpt-4o-mini, llama-3.3-70B) $\times$ 4 constructs $\times$ 6 outcomes $\times$ 3 fixed effects variants (144 cells). The red dashed line marks $|\Delta\%| = 10\%$; blue bars sit below the threshold and red bars above. The right-hand summary box reports N, the share of cells above 10%, and the median. Panels (B) and (D) restrict the plots to regressions where the naive construct coefficient is significant at the 5% level.

Item 1A Risk Disclosures

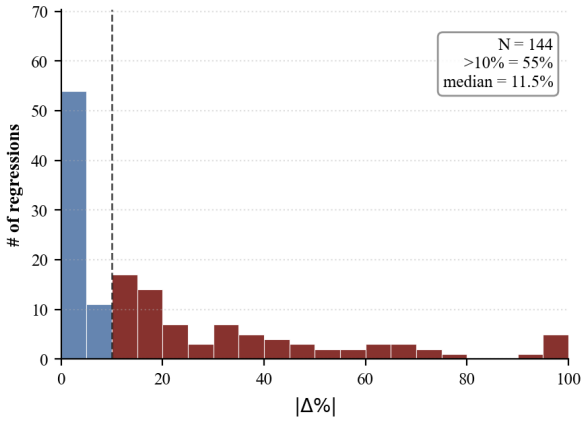


Panel A: All regressions.

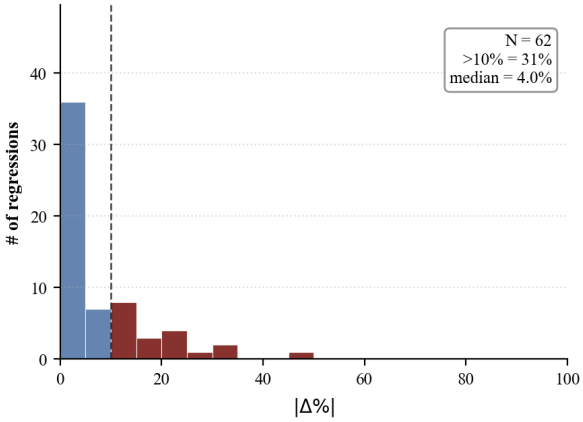


Panel B: Naive significant regressions.

Earnings Call Transcripts



Panel C: All regressions.



Panel D: Naive significant regressions

Table 1: **Text Summary Statistics**

Panel A of this table reports aggregate summary statistics for four text sources across four LLMs; these are based on Item 1A risk disclosures. *Sentences* contain individual raw sentences discussing risk factors. *Keywords* contain uni- or bi-grams that LLMs identify as salient terms for grouping sentences into company-specific risks. *LLM* contains AI-generated summaries of Item 1A risks. *Full Disclosure* contains the complete set of Item 1A sentences associated with each identified risk. Top-N *Concentration* shows the percentage of total word frequency accounted for by the N most common words. *Entropy* extends this concept to the full word distribution. It is the normalized Shannon entropy of word frequencies, on [0,1], capturing how evenly tokens are spread across the vocabulary: values near 1 indicate diverse vocabulary use, lower values indicate concentration in a core vocabulary. Panel B is based on earnings call transcripts and reports the metrics for raw earnings call text and the summaries generated by gpt-4o-mini and llama-3.3-70B.

Panel A: Item 1A risk disclosure text

			Concentration (%)			
	Total Words	Vocabulary Size	Top-100	Top-500	Top-1000	Entropy
		gpt-4o-mini				
Keywords	1,185,692	5,017	54.30	85.99	93.60	0.739
Sentences	48,867,009	34,865	27.70	60.76	75.29	0.724
LLM Summary	14,826,475	15,010	41.39	74.03	85.78	0.718
Full Disclosure	103,587,174	36,904	29.40	63.16	77.57	0.708
		grok-3-mini				
Keywords	1,983,044	7,604	43.37	77.79	88.60	0.758
Sentences	66,656,778	39,174	27.14	59.99	74.62	0.719
LLM Summary	27,228,390	24,909	34.19	66.06	79.03	0.722
Full Disclosure	126,502,779	41,879	28.81	62.21	76.73	0.704
		phi4				
Keywords	1,454,344	6,020	50.77	83.84	92.33	0.740
Sentences	62,632,936	37,944	27.32	60.19	74.84	0.721
LLM Summary	9,813,635	16,801	36.13	69.82	82.43	0.736
Full Disclosure	118,078,881	39,633	28.99	62.48	76.98	0.706
		llama3.3				
Keywords	1,132,059	4,404	60.07	88.99	95.21	0.721
Sentences	58,914,054	37,329	27.53	60.44	75.04	0.721
LLM Summary	8,755,130	14,965	38.73	72.12	84.39	0.732
Full Disclosure	115,353,982	39,006	29.13	62.70	77.17	0.707

Panel B: Earnings call transcripts

Earnings call	139,868,942	132,539	31.21	61.20	74.49	0.648
gpt-4o-mini summary	19,626,168	53,692	38.01	68.25	79.93	0.665
llama-3.3-70B summary	21,998,782	62,991	38.94	66.79	78.17	0.657

Table 2: **Sample Vocabulary Injections**

This table contrasts discriminative terms that appear exclusively in LLM-generated 1-shot summaries (injected) against terms in human-authored disclosure text for selected risk categories. LLM-injected terms appear in at least three of four models (GPT-4o-mini, Grok-3-mini, Phi-4, Llama-3.3-70B) but are absent from human text and reflect model priors.

Risk Factor	LLM-Injected Terms	Human Text Terms
Natural/Physical Environment	*droughts, wildfires, toxic, flooding, radioactive, spills, discharges*	*temperatures, contaminated, cleanup, winter, catastrophes, tornadoes, pollution*
Reputation & Communication	*criticism, confusion, unethical, complaints, sentiment, sustainability*	*enhancing, allegations, recalls, misconduct, regulators, investigations*
Political Environment	*taiwan, wars, tensions, border, threat, chinese, priorities*	*nationalization, violence, hostilities, customs, greenhouse, ukraine*
Regulatory Compliance	*gdpr, bribery, epa, labeling, fcc, inspections*	*debarment, dodd, imprisonment, hipaa, disabilities, occupational*

Table 3: **Summary Statistics: Omission, Injection, and LLM-generated Regressors**

This table reports summary statistics for the LLM wedges and the LLM-generated regressors. Panel A covers the S&P 500 Item 1A risk-factor disclosures (unit of observation: firm-year × LLM); Panel B covers the earnings-call transcripts (firm-quarter × LLM). We report the mean of each variable with the standard deviation in square brackets below. The *Lexical Rates* and *Semantic Cosine Similarity* are reported in percent (i.e., multiplied by 100), while the *LLM-generated Scores* are on their original 0-to-10 scale. The *Overall* column pools across LLMs; the remaining columns report each LLM separately. The *Lexical Rates* are wedges using the set of unique discriminative terms: *Omission* is the share of source terms dropped from the summary, and *Injection* is the share of summary terms absent from the source. The *Semantic Cosine Similarity* measures embed, with Doc2Vec, the bag of omitted (injected) terms and report its cosine similarity to the embedded source document, capturing how central the dropped (added) content is to the disclosure. The *LLM-generated Scores*—Severity, Boilerplate, Complexity, and Specificity—are the four construct scores each LLM assigns to a risk factor, averaged to the disclosure level. These scores are continuous scores generated by LLMs on a scale of 0 to 10.

Panel A: LLM wedges and generated-regressors from Item 1A Risk Disclosures

	Overall	gpt-4o-mini	grok-3-mini	phi4	llama-3.3-70B
Lexical Rates:					
Omission	67.84	66.96	58.83	70.62	73.10
	[19.94]	[16.68]	[16.03]	[22.09]	[19.89]
Injection	25.20	40.22	29.58	17.34	19.83
	[21.47]	[18.40]	[17.76]	[19.36]	[22.30]
Semantic Cosine Similarity:					
Omission	64.03	62.09	62.98	65.02	66.05
	[9.93]	[10.25]	[9.88]	[9.34]	[9.71]
Injection	22.18	19.67	21.89	22.93	24.29
	[7.07]	[6.74]	[7.99]	[6.32]	[6.29]
LLM-generated Regressors:					
Severity	7.106	7.255	6.776	7.152	7.244
	[0.897]	[0.459]	[1.146]	[0.877]	[0.820]
Boilerplate	7.240	7.460	7.293	7.203	7.085
	[1.236]	[0.438]	[1.180]	[1.248]	[1.566]
Complexity	5.744	5.687	5.762	5.590	5.933
	[1.061]	[0.466]	[1.273]	[1.137]	[1.031]
Specificity	4.595	4.674	4.994	4.333	4.475
	[1.164]	[0.604]	[1.506]	[0.878]	[1.277]

Panel B: LLM wedges and generated-regressors from Earnings Call Transcripts

	Overall	gpt-4o-mini	llama-3.3-70B
Lexical Rates:			
Omission	80.81	82.03	79.58
	[6.51]	[6.41]	[6.37]
Injection	23.35	27.69	19.02
	[12.18]	[10.84]	[11.90]
Semantic Cosine Similarity:			
Omission	88.62	89.00	88.24
	[2.31]	[2.21]	[2.35]
Injection	18.87	18.37	19.38
	[6.26]	[6.29]	[6.20]
LLM-generated Regressors:			
Severity	3.818	4.617	3.018
	[1.081]	[0.667]	[0.784]
Boilerplate	5.178	5.161	5.194
	[0.545]	[0.331]	[0.696]
Complexity	6.777	5.738	7.815
	[1.139]	[0.385]	[0.538]
Specificity	7.244	7.141	7.346
	[0.505]	[0.423]	[0.557]

Table 4: **Generated Regressor Diagnostics: Item 1A Risk Disclosures**

This table diagnoses whether the LLM-generated text constructs (severity, boilerplate, complexity, and specificity) are contaminated by LLM-processing characteristics, omission (O_{im}) and injection (I_{im}) rates, when used as regressors for *forward* firm outcomes $y_{i,t+1}$. β_{naive} is the mean coefficient on the construct in the naive baseline $y_{i,t+1} = \beta\,\text{Construct}_{i,t} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t+1}$ with firm controls $X_{i,t} = \{\text{Size}, \text{MTB}, \text{CashFlow}, \text{Tangibility}, \text{Age}\}$ and time fixed effects τ_t. The joint specification adds the wedges: $y_{i,t+1} = \beta\,\text{Construct}_{i,t} + \rho_O\,O_{im,t} + \rho_I\,I_{im,t} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t+1}$, yielding β_{joint}, ρ_O, ρ_I. $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$ measures coefficient sensitivity. The decomposition splits the construct, via a first-stage Tobit regression on the wedges and controls, into the wedge-predictable part $\widehat{\text{Construct}}$ (coefficient θ) and the orthogonal residual $\hat{e}$ (coefficient λ); under the maintained assumption that O_{im} and I_{im} are LLM-processing artifacts orthogonal to firm fundamentals, λ is the bias-corrected coefficient. This table summarizes regressions on S&P 500 Item 1A risk disclosures for all four LLMs. Rows were pooled across LLMs and constructs within each outcome. For each outcome variable, in the first row, we report the mean coefficients for β_{naive}, β_{joint}, wedge coefficient ρ, and the median for $|\Delta\%|$, θ, and λ. The bracketed entries beneath each cell report the share of underlying cells significant at the 5% level; except the bracketed entry under λ is the fraction of cells where $p(\theta = \lambda) < 0.05$. We include Year fixed effects in all regressions and standard errors are clustered at the firm level.

		Construct coeff.		Lexical Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	Med $\lvert\Delta\%\rvert$	θ	λ
NI/Assets	21,534	-0.0036	-0.0050	-0.0100	-0.0608	16.1%	0.0089	-0.0079
		[75.0%]	[68.8%]	[75.0%]	[100.0%]			[62.5%]
CapEx/Assets	21,534	0.0002	-0.0002	0.0153	-0.0014	40.1%	0.0053	0.0001
		[25.0%]	[12.5%]	[93.8%]	[0.0%]			[81.2%]
R&D ratio	21,534	0.0073	0.0078	0.0567	0.0656	12.7%	0.0178	0.0095
		[87.5%]	[87.5%]	[81.2%]	[100.0%]			[81.2%]
Leverage	20,390	0.0044	0.0005	0.1742	0.0716	40.3%	0.0565	0.0034
		[25.0%]	[25.0%]	[87.5%]	[43.8%]			[68.8%]
Sales growth	20,345	0.0055	0.0060	0.0755	0.1152	27.2%	-0.0045	0.0093
		[43.8%]	[68.8%]	[50.0%]	[100.0%]			[75.0%]
WW index	14,111	0.0021	0.0019	0.0526	0.0445	21.6%	0.0029	0.0050
		[81.2%]	[75.0%]	[62.5%]	[81.2%]			[56.2%]

Table 5: **Generated Regressor Diagnostics: Earnings Call Transcripts**

This table diagnoses whether the LLM-generated text constructs (severity, boilerplate, complexity, and specificity) are contaminated by LLM-processing characteristics, omission (O_{im}) and injection (I_{im}) rates, when used as regressors for *forward* firm outcomes $y_{i,t+1}$. β_{naive} is the mean coefficient on the construct in the naive baseline $y_{i,t+1} = \beta\,\text{Construct}_{i,t} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t+1}$ with firm controls $X_{i,t} = \{\text{Size}, \text{MTB}, \text{CashFlow}, \text{Tangibility}, \text{Age}\}$ and time fixed effects τ_t. The joint specification adds the wedges: $y_{i,t+1} = \beta\,\text{Construct}_{i,t} + \rho_O\,O_{im,t} + \rho_I\,I_{im,t} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t+1}$, yielding β_{joint}, ρ_O, ρ_I. $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$ measures coefficient sensitivity. The decomposition splits the construct, via a first-stage Tobit regression on the wedges and controls, into the wedge-predictable part $\widehat{\text{Construct}}$ (coefficient θ) and the orthogonal residual $\hat{e}$ (coefficient λ); under the maintained assumption that O_{im} and I_{im} are LLM-processing artifacts orthogonal to firm fundamentals, λ is the bias-corrected coefficient. This table summarizes regressions on S&P 500 quarterly earnings-call transcripts for gpt-4o-mini and Llama-3.3-70B only. Rows were pooled across LLMs and constructs within each outcome. For each outcome variable, in the first row, we report the mean coefficients for β_{naive}, β_{joint}, wedge coefficient ρ, and the median for $|\Delta\%|$, θ, and λ. The bracketed entries beneath each cell report the share of underlying cells significant at the 5% level; except the bracketed entry under λ is the fraction of cells where $p(\theta = \lambda) < 0.05$. We include Year-quarter fixed effects in all regressions and standard errors are clustered at the firm level.

		Construct coeff.		Lexical Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	Med $\lvert\Delta\%\rvert$	θ	λ
NI/Assets	18,868	−0.0008	−0.0009	+0.0079	−0.0128	8.5%	+0.0027	−0.0009
		[62.5%]	[75.0%]	[25.0%]	[25.0%]			[37.5%]
CapEx/Assets	18,866	+0.0000	+0.0001	+0.0031	+0.0086	19.3%	−0.0014	+0.0001
		[37.5%]	[37.5%]	[0.0%]	[50.0%]			[12.5%]
R&D ratio	18,868	+0.0007	+0.0008	+0.0435	−0.0044	9.4%	−0.0158	+0.0003
		[62.5%]	[62.5%]	[100.0%]	[0.0%]			[62.5%]
Leverage	18,868	+0.0085	+0.0080	−0.1248	−0.0737	6.8%	+0.0766	+0.0070
		[25.0%]	[25.0%]	[0.0%]	[0.0%]			[0.0%]
Sales growth	18,864	−0.0211	−0.0200	+0.1551	+0.2044	7.7%	−0.0754	−0.0149
		[62.5%]	[62.5%]	[37.5%]	[100.0%]			[62.5%]
WW index	18,868	+0.0008	+0.0008	+0.0324	−0.0076	8.2%	−0.0095	+0.0006
		[37.5%]	[12.5%]	[0.0%]	[0.0%]			[0.0%]

Table 6: **Generated Regressor Diagnostics: Jha et al. (2024) Replication**

This table diagnoses whether the LLM-generated direct predictors used by Jha et al. (2024) are contaminated by LLM-processing characteristics, omission (O_{im}) and injection (I_{im}) rates, when used to investigate firm outcomes. The predictors are the seven scores extracted in the Jha framework (capex, cash, div, leverage, ocf, rd, sales). Each row reports results for one firm outcome variable at time $t+2$ (per the paper's convention). β_{naive} is the mean coefficient on the predictor in the paper's baseline $y_{i,t+2} = \beta\,\text{Score}_{i,t} + \gamma' X_{i,t} + \alpha_i + \alpha_t + \varepsilon$ with $X_{i,t} = \{\text{Size}, q_{\text{tot}}, \text{CashFlow}, \text{Leverage}, y_{i,t}\}$ where q_{tot} is the Peters–Taylor Tobin's Q. The joint specification adds the omission and injection rates: $y_{i,t+2} = \beta\,\text{Score}_{i,t} + \rho_O\,O_{im,t} + \rho_I\,I_{im,t} + \gamma' X_{i,t} + \alpha_i + \alpha_t + \varepsilon$, yielding β_{joint}, ρ_O, ρ_I. $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$ measures coefficient sensitivity. The decomposition splits the predictor via a first-stage projection on the wedges and controls into the wedge-predictable component (θ) and the orthogonal residual (λ); λ is the bias-corrected coefficient under the maintained orthogonality assumption. Rows pool across LLMs and direct predictors within each outcome; bracketed entries beneath each cell report the share of underlying cells significant at the 5% level. All reported coefficients are scaled by 100 (percentage units) to match the convention of Jha et al. (2024). We use earnings call transcripts for during the period 2005 to 2020 for this analysis. All regressions include Firm and year-quarter fixed effects and standard errors are clustered at the firm level.

		Construct coeff.		Lexical Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	Med $\lvert\Delta\%\rvert$	θ	λ
CapEx/Assets	13,665	+0.4799 [100.0%]	+0.4911 [100.0%]	+1.0214 [100.0%]	+0.1195 [0.0%]	2.3%	−0.4089	+0.4911 [100.0%]
Cash Ratio	13,678	+3.2204 [100.0%]	+3.3877 [100.0%]	+5.1636 [50.0%]	−1.2582 [0.0%]	4.1%	−10.1426	+3.3877 [50.0%]
Div. Quarterly	13,638	+0.2411 [100.0%]	+0.2567 [100.0%]	+0.5238 [0.0%]	+0.0689 [0.0%]	6.6%	−0.3301	+0.2567 [0.0%]
Leverage	13,678	+3.3116 [100.0%]	+3.4080 [100.0%]	−7.6825 [50.0%]	+2.2970 [50.0%]	3.6%	−6.9571	+3.4080 [50.0%]
OpCF Ratio	13,672	+0.1110 [0.0%]	+0.1388 [0.0%]	+2.4553 [50.0%]	−0.2103 [0.0%]	28.3%	−1.7497	+0.1388 [0.0%]
R&D/Total Capital	13,678	+0.0503 [0.0%]	+0.0681 [0.0%]	+0.5763 [0.0%]	−0.3167 [50.0%]	35.2%	−0.2345	+0.0681 [0.0%]
Sales Growth	13,665	+16.7067 [100.0%]	+16.7168 [100.0%]	+20.1993 [0.0%]	−0.5186 [0.0%]	0.2%	+22.2239	+16.7168 [0.0%]

A Appendix: Additional Figures and Tables

Figure A.1: **Vocabulary Omission: Granular Risk Factor Categories**

In this figure, we plot the Omission rates for discriminatory terms identified for each of the 26 risk categories in our sample. Omission is defined as $(Vocab_{\text{Human}} - Vocab_{\text{LLM}} \cap Vocab_{\text{Human}})/Vocab_{\text{Human}}$. We define Human and LLM vocabularies as the count of unique discriminative terms used by companies in their Item 1A and LLM summaries, respectively.

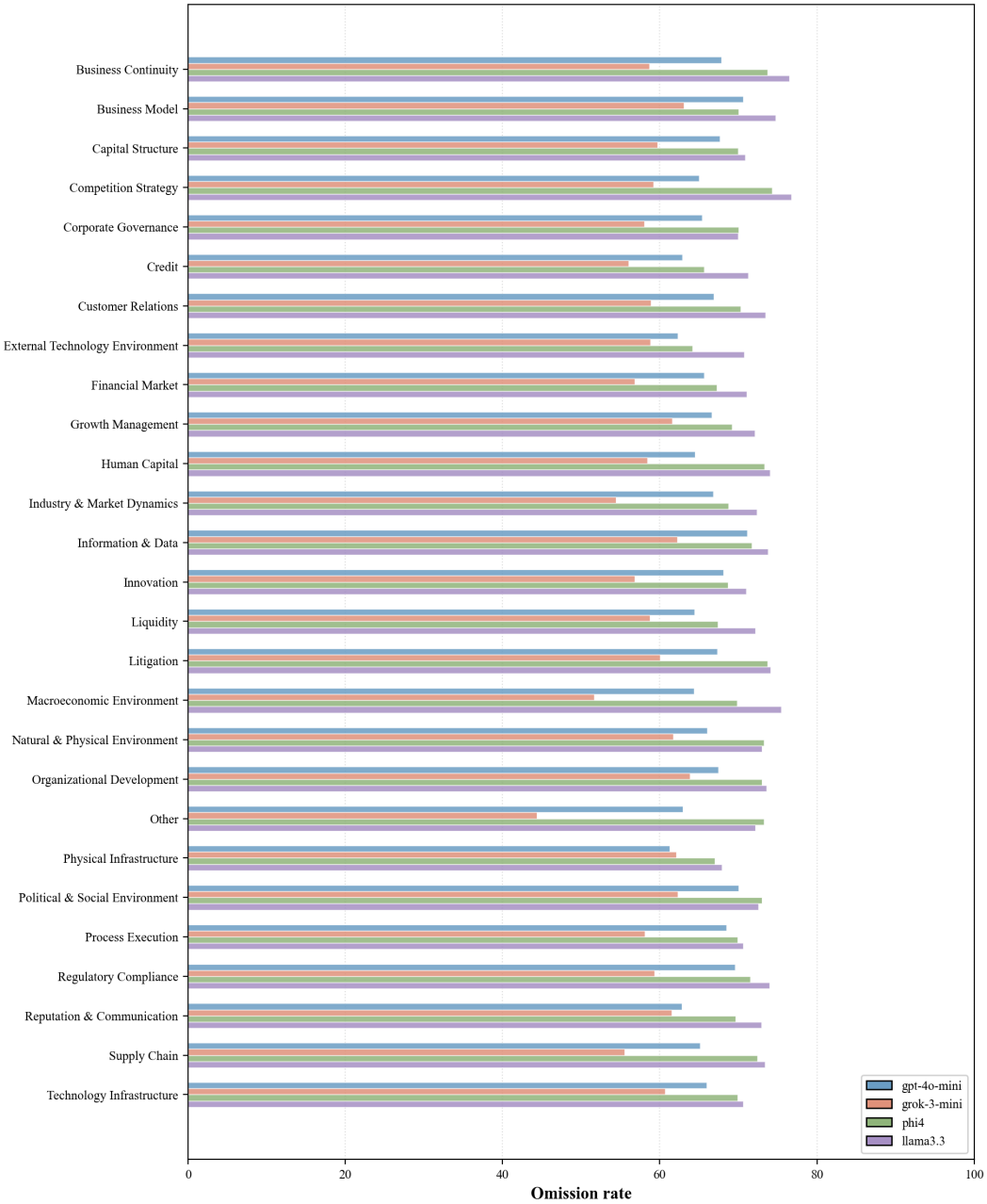


Figure A.2: **Injection of Vocabulary: Granular Risk Factor Categories**

In this figure, we plot the injection rates for discriminatory terms identified for each of the 26 risk categories in our sample. Injection is defined as $(Vocab_{\text{LLM}} - Vocab_{\text{LLM}} \cap Vocab_{\text{Human}})/Vocab_{\text{LLM}}$. We define Human and LLM vocabularies as the count of unique discriminative terms used by companies in their Item 1A and LLM summaries, respectively.

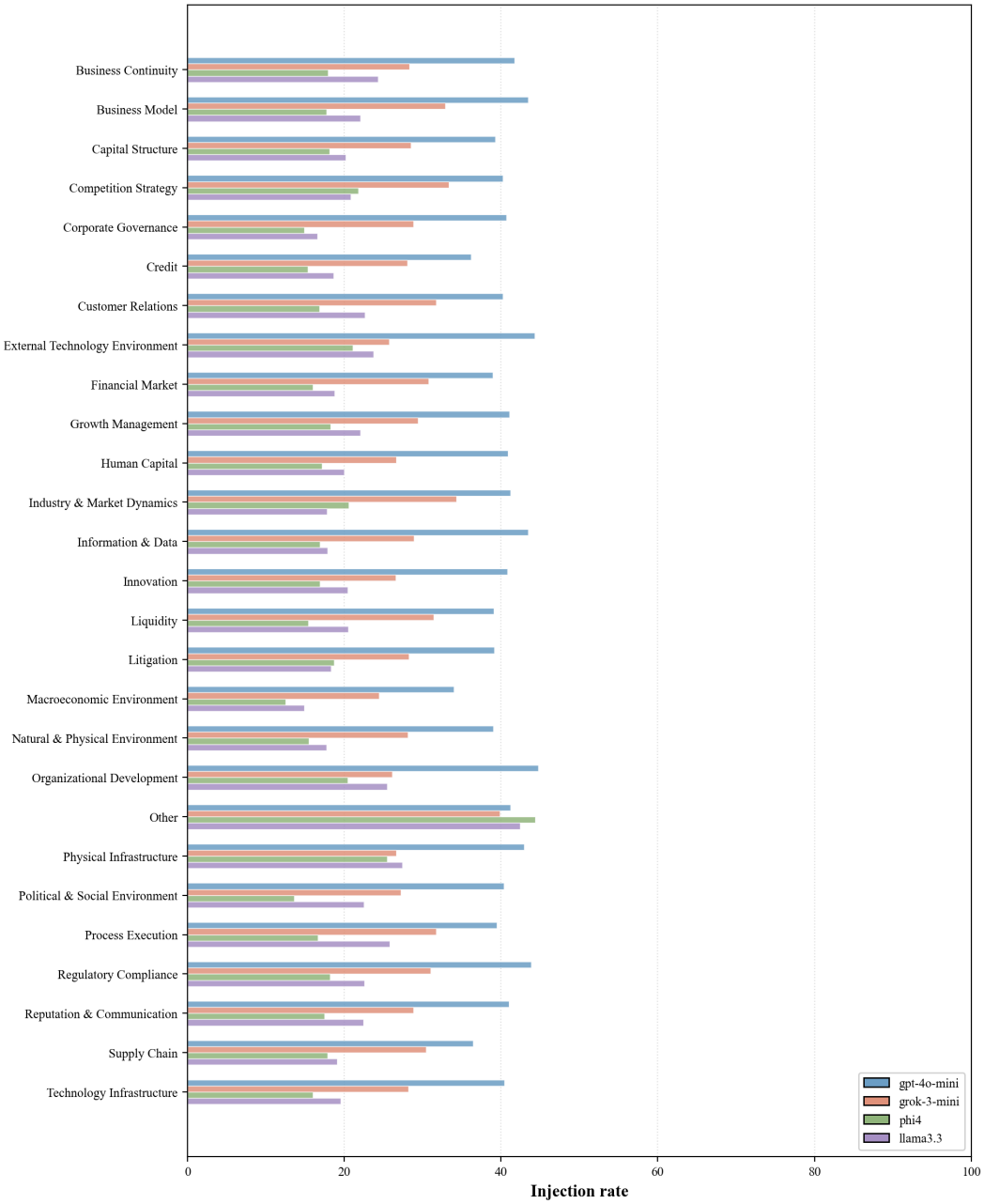
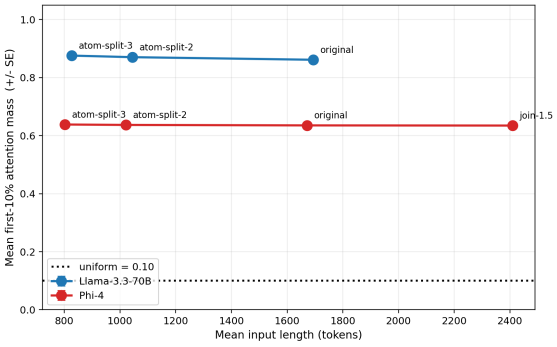
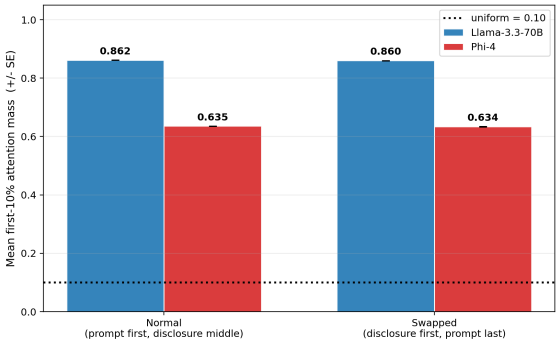


Figure A.3: **Robustness of the Attention Bottleneck**

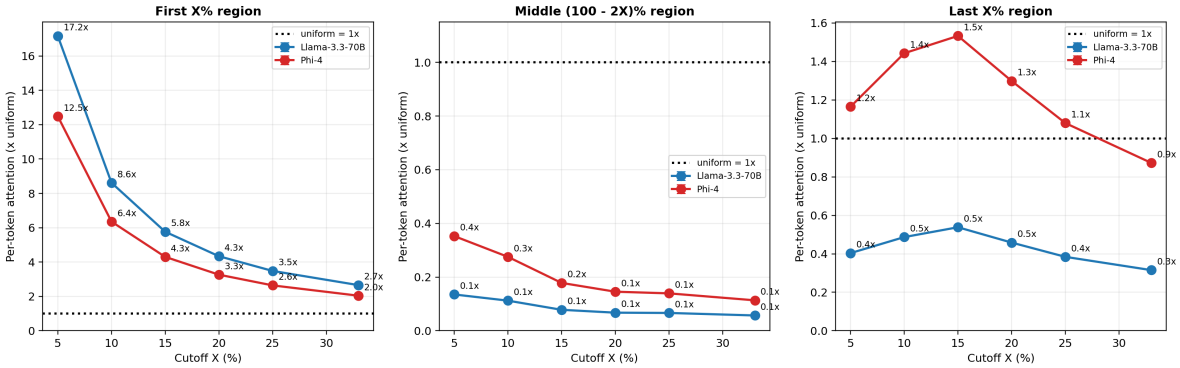
The figure assesses the robustness of the first-10% attention bottleneck reported in Figure 2 to three natural alternative explanations: input length, the position of the user's instructions within the prompt, and the choice of region boundary. All three panels use the score prompt with the severity construct on the same 500-chunk subsample of Item 1A disclosures, run on Llama-3.3-70B and Phi-4; markers, bars, and error bars are means and ± 1 standard error across the 500 chunks; dotted lines mark the values implied by uniform attention. Panel (a) plots mean first-10% attention mass against mean input length. Size variants are constructed from each base chunk's underlying sentence-level "atoms": atom-split-3 (one third of the original text, ≈ 800 tokens), atom-split-2 (one half, ≈ 1,000 tokens), original (≈ 1,700 tokens), and, for Phi-4 only, join-1.5 (adjacent same-firm chunks concatenated, ≈ 2,400 tokens; Llama-3.3-70B's 80-layer attention tensor exceeds 95 GB and does not fit in single-GPU memory). Panel (b) compares the first-10% attention mass under the standard score prompt (system message and user instructions first, disclosure body in the middle) against a swapped variant in which the disclosure body appears first and the user's instructions are appended at the end of the input. Panel (c) re-expresses the per-token attention measurement of Figure 2 for region boundaries ranging from 5% to 33% of the input. The input is divided into twenty 5% bins; for each cutoff X we report the per-token concentration multiplier (mean attention mass divided by the region's token share, so that 1× corresponds to uniform attention) in the first-X%, middle-$(100-2X)$%, and last-X% regions. Across all three panels, first-10% attention mass is essentially unchanged from its Figure 2 value, indicating that the bottleneck is not driven by input length, by the position of the user's instructions within the input, or by the choice of the 10% region boundary.



(a) Input Length Sensitivity



(b) Prompt-Position Swap



(c) Region Cutoff Sensitivity

Figure A.4: **LLM-generated Dictionary Terms Agreement by Risk Factor Category**

In this figure, we plot the pairwise agreement rates between LLM-generated dictionaries across 26 risk factor categories. For each category, we compute the share of terms in a base model's dictionary that appear in each comparison model's dictionary.

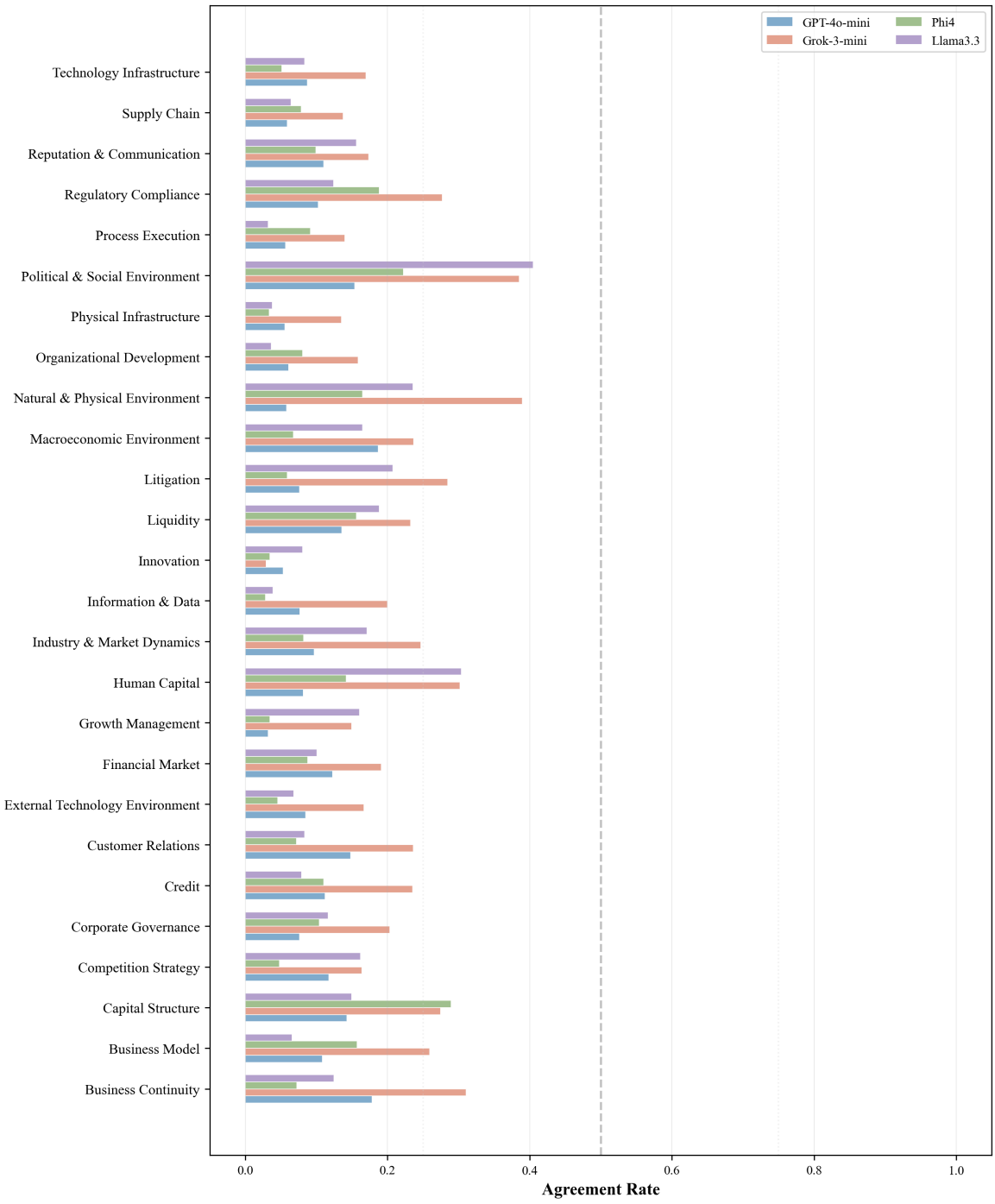
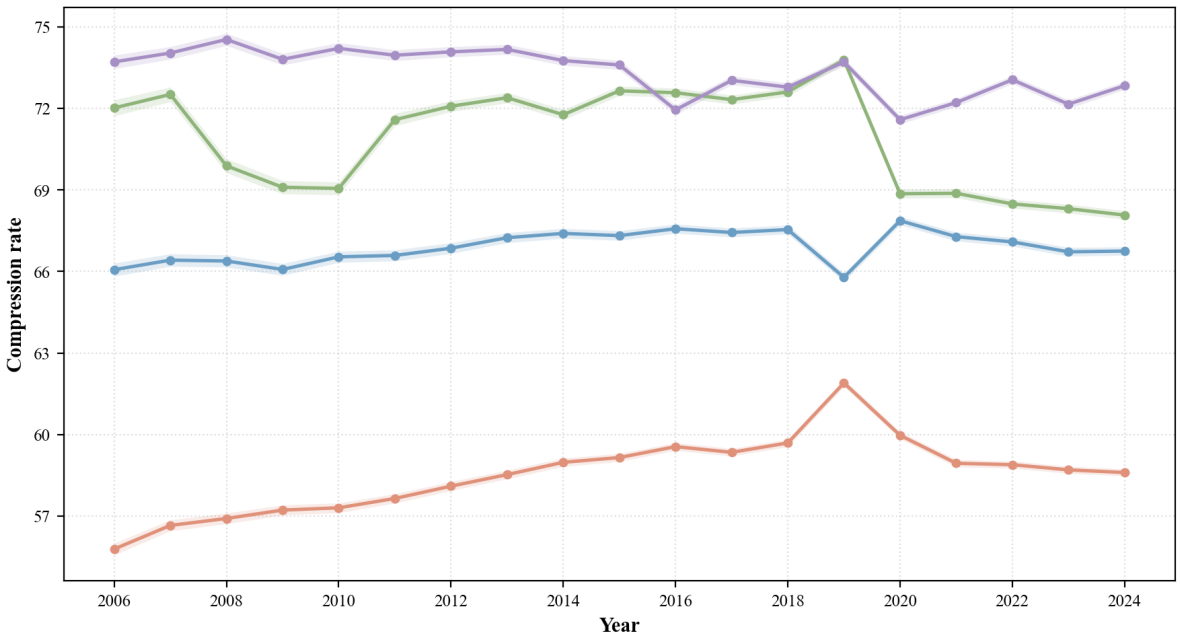


Figure A.5: **Omission and Injection Rates over the Sample Period**

In panels (a) and (b), we plot the Omission and Injection rates for discriminatory terms identified in our S&P 500 sample across all risk factors separately for each model. Omission is defined as $O_{im} = 1 - \frac{|V_{i,\text{source}} \cap V_{im,\text{summary}}|}{V_{i,\text{source}}}$. Injection is defined as $I_{im} = 1 - \frac{|V_{im,\text{summary}} \cap V_{i,\text{source}}|}{V_{im,\text{summary}}}$. We define Human and LLM vocabularies as the count of unique discriminative terms used by companies in their Item 1A and LLM summaries, respectively.

(a) Omission rate



(b) Injection rate

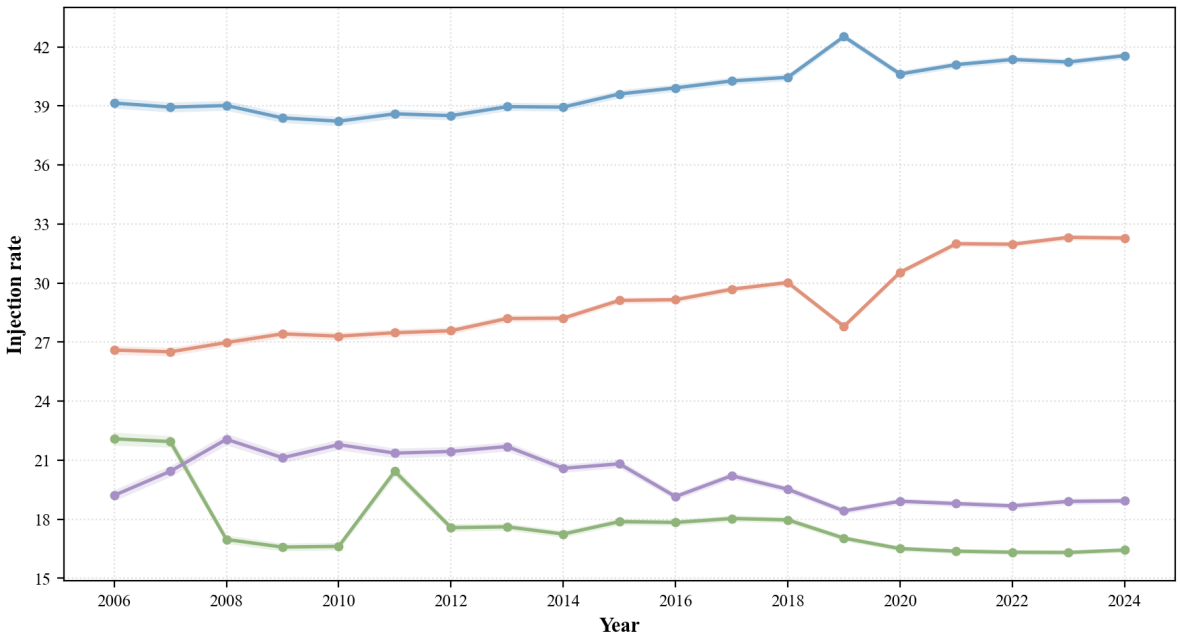




Figure A.6: **Tone Distribution**

This figure shows the average number of words in each of the seven Loughran-McDonald tone dimensions per disclosure, broken out by vocabulary object: Source (black, is the risk sentences from Item 1A), Compressed (red, words dropped by LLM), Retained (green, words retained by LLM), Injected (blue, words added by LLM), Summary (gray, is the LLM summary). Four panels, one per LLM.

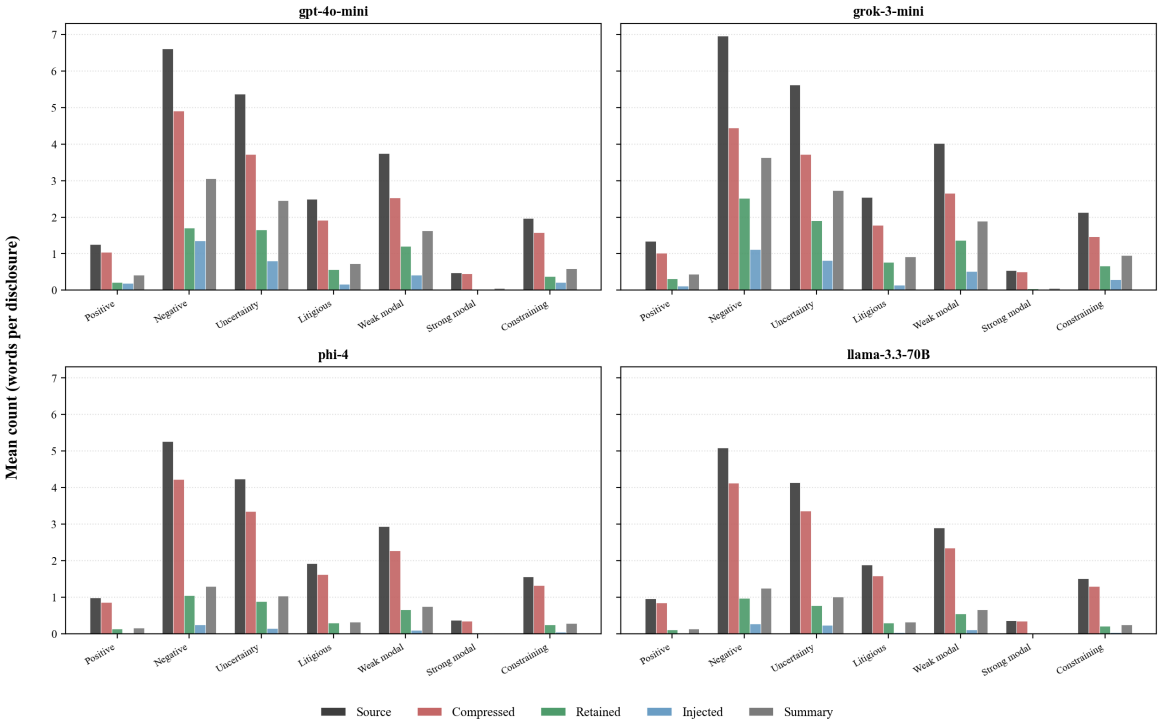
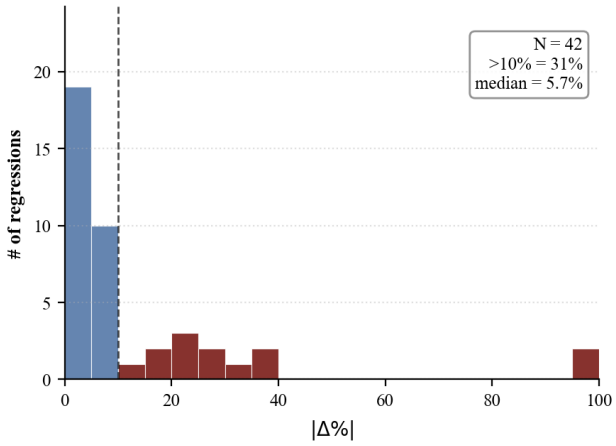
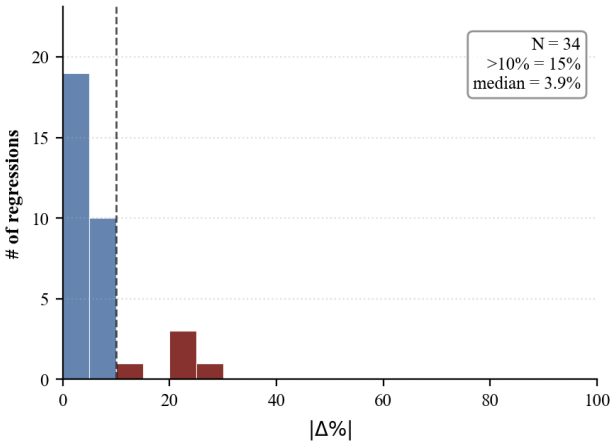


Figure A.7: **Diagnostic Distributions: Jha et al. (2024) Replication**

This figure plots the sensitivity of the Jha et al. (2024) direct-predictor regressions summarized in Table 6 to the inclusion of the LEX omission and injection wedges. Coefficient sensitivity is $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$. Cells pool 2 LLMs (gpt-4o-mini, Llama-3.3-70B) $\times$ 7 direct predictors matched to outcomes $\times$ 3 fixed effects variants (year-quarter; year-quarter + industry; year-quarter + firm) = 42 cells. The red dashed line marks $|\Delta\%| = 10\%$; the right-hand summary box reports N, the share of cells above 10%, and the median. Panel (b) restricts to cells where the naive predictor coefficient is significant at the 5% level.



(a) $|\Delta\%|$ distribution — all cells



(b) $|\Delta\%|$ distribution — naive-significant cells

Figure A.8: **LLM Prompt for Risk Factor Summaries from Item 1A Disclosure**

This figure reproduces the prompt issued to each large language model (GPT-4o-mini, Grok-3-mini, Llama-3.3-70B, and Phi-4) to produce several tasks ranging from summary descriptions to risk factor labels, applied to individual Item 1A risk-factor disclosures.

```
instructions = """
You are a corporate risk manager reviewing this item 1A section from
a company's 10-K disclosure. You do not know of future events beyond
the content in this document. Do not use any information, terminology,
or market context beyond the provided financial disclosure and
disclosure date. Analyze the disclosure and present the following:

1. Identify all risk factors and, if available, the associated
mitigation plans to address the identified risks **explicitly
discussed** by the company in the disclosure.
Do not infer or introduce unstated risks.

2. For each risk factor, present the analysis in JSON format
with the following structure:

{
 "risk_factors": [
 {
  "risk_factor": "Bi-gram Label",
  "risk_factor_syn": "Bi-gram synonym of Risk Factor label",
  "risk_severity": "High/Medium/Low/Not Specified",
  "risk_desc": "100-150 word description based ONLY on given text",
  "risk_sentences": "All exact sentences from text about this risk",
  "risk_keyword": "Keyword indicating this risk",
  "mitigation": "Bi-gram Label or 'Not Disclosed'",
  "mitigation_syn": "Bi-gram synonym OR 'Not Disclosed'",
  "mitigation_desc": "100-150 word description
  OR 'No plan discussed'",
  "mitigation_sentences": "All related sentences OR 'None'",
  "mitigation_keyword": "Keyword for mitigation OR 'None'",
  "bus_function": "One from: |'Research & Development'|'Operations'|
  'Sales & Marketing'|'Customer Experience'|'Financing'|
  'Capital Investment'|'Human Capital Management'|
  'Information Technology'|'Supply Chain Management'|
  'Strategy \& Planning'|'Risk & Compliance'"
 }
 ]
}
Important Notes:
- Explicit Information Only
- Bi-gram Labels (e.g., "Market Competition", "Supply Chain")
- Time-Appropriate Language
- Valid JSON format required
"""
```

Figure A.9: **LLM Prompt for Risk Factor Classification**

This figure reproduces the prompt issued to each large language model (GPT-4o-mini, Grok-3-mini, Llama-3.3-70B, and Phi-4) to classify risk factors identified by the prompt in Figure A.10. We ask the LLMs to classify a risk factor into a primary and a secondary risk factor category. We provide the LLMs with a list of 26 risk factor categories in our taxonomy, the raw sentences in Item 1A pertaining to the risk factor, the LLM-generated risk factor summary, risk factor bigrams, and key words identified by the prompt in Figure A.10.

```
prompt = """You are an insightful and precise Financial Economist. You will be provided with risk
factor information from Item 1A of 10-K filings.

For each risk factor, assign:
1. "Primary" Risk Factor Category - directly related to the central risk theme
2. "Secondary" Risk Factor Category - captures additional, less central risk dimensions

Choose ONLY from these 27 categories:
- External Technology Environment
- Industry and Market Dynamics
- Macroeconomic Environment
- Natural & Physical Environment
- Political & Social Environment
- Capital Structure
- Credit
- Liquidity
- Financial Market
- Business Continuity
- Human Capital
- Information & Data
- Organizational Development
- Physical Infrastructure
- Process Execution
- Supply Chain
- Technology Infrastructure
- Corporate Governance
- Litigation
- Regulatory Compliance
- Reputation & Communication
- Business Model
- Competition Strategy
- Customer Relations
- Growth Management
- Innovation
- Other

KEY RULES:
- Use EXACT category names from the list
- Choose specific categories over general ones
- Never create new categories
- If unsure between categories, pick the most relevant one from the list

Return a JSON array with one object per risk factor:
[
  {
    "primary_risk_categ": "Category Name",
    "secondary_risk_categ": "Category Name"
  }
]
Risk factors to classify:
"""
```

Figure A.10: **LLM Scoring Prompt for Risk-Disclosure Constructs**

This figure reproduces the scoring prompt issued to each large language model (GPT-4o-mini, Grok-3-mini, Llama-3.3-70B, and Phi-4) used to score Item 1A risk-factor disclosures along the four constructs studied in the paper. Panel A presents the task instructions, the closed lists of evaluable constructs and business functions, and the ordinal and continuous scoring rubrics. Panel B presents the required JSON output schema.

Panel A: Task Instructions, Business Functions, and Constructs

```
Evaluate each of the following {n} risk disclosures on four constructs.
For EACH disclosure, provide:

PART A: EVALUABILITY
Which constructs have CLEAR indicators? Select all that apply from:
Severity, Complexity, Specificity, Boilerplate

PART B: BUSINESS FUNCTIONS
Which business functions are explicitly affected? Select all that apply from:
Research & Development, Operations, Sales & Marketing, Customer Experience,
Financing, Capital Budgeting, Human Capital Management, Information Technology,
Supply Chain Management, Strategy & Planning, Risk & Compliance

PART C: SCORING
- Ordinal: 1-5 scale (integers only)
Severity:    1 = negligible       to 5 = catastrophic
Complexity:  1 = straightforward  to 5 = intractable
Specificity: 1 = generic          to 5 = unique
Boilerplate: 1 = entirely original to 5 = pure boilerplate
- Continuous: 0.0-10.0 scale (decimals allowed)

CONSTRUCTS:
1. SEVERITY -- Potential impact if risk materializes
Ordinal:    1=negligible, 2=minor, 3=moderate, 4=major, 5=catastrophic
Continuous: 0=no impact,  10=existential threat

2. COMPLEXITY -- Difficulty of assessing/interpreting this risk
Ordinal:    1=straightforward, 2=somewhat complex, 3=complex,
4=highly complex,  5=intractable
Continuous: 0=trivially simple, 10=impossible to evaluate

3. SPECIFICITY -- How unique is this risk to this firm
Ordinal:    1=generic/any firm, 2=industry-wide, 3=sector-specific,
4=firm-specific,    5=unique
Continuous: 0=universal risk, 10=completely unique to this firm

4. BOILERPLATE -- Degree of template/standard language used
Ordinal:    1=entirely original, 2=mostly original, 3=mixed,
4=mostly template,   5=pure boilerplate
Continuous: 0=fully novel, 10=entirely standard language
```

Panel B: Output Format and Required JSON Schema

```
CRITICAL INSTRUCTIONS:
- Return ONLY a valid JSON array with exactly {n} objects, one per
disclosure in order
- Each object must have: evaluable, functions, ordinal, continuous
- Use pipe-delimited strings for evaluable and functions
- Ordinal scores: integers 1-5. Continuous scores: decimals 0.0-10.0.

REQUIRED FORMAT (array of {n} objects):
[
{
        "evaluable": "Severity|Complexity",
        "functions": "Operations|Risk & Compliance",
        "ordinal":    {"Severity": 4,   "Complexity": 3,
                "Specificity": 2, "Boilerplate": 4},
        "continuous": {"Severity": 7.5, "Complexity": 6.0,
                "Specificity": 4.0, "Boilerplate": 8.0}
},
...
]

---
{disclosures}
```

Table A.1: Risk Factor Taxonomy: Detailed

Main Category	Sub-category	Key Elements
External Environmental	External Technology Environment	Technological disruption, Industry-wide digital transformation, Innovation pace, Tech adoption trends, Emerging technologies, Digital transformation
	Industry and Market Dynamics	Industry-wide consolidation and price trends, Shifts in Industry-wide supply and demand, New market entrants, Competitive intensity, Product market fluidity
	Macroeconomic Environment	Macroeconomic conditions, Economic growth/recession, Interest rates/inflation, Currency fluctuations, Trade conditions, Market liquidity
	Natural/Physical Environment	Natural disasters, Climate change impacts, Weather-related disruptions, Resource scarcity, Environmental degradation
	Political/Social Environment	Government stability, Regulatory changes, Social/demographic shifts, Geopolitical tensions, Public sentiment
Financial	Capital Structure	Leverage levels, Funding mix, Dividend policy, Investment capacity, Solvency, Tax Management, Bankruptcy
	Credit Risk	Counterparty default, Customer credit, Collateral value, Recovery rates, Borrower concentration
	Liquidity Risk	Funding access, Cash flow management, Asset liquidity, Working capital, Emergency funding, Sinking fund, Treasury operations, Capital market access
	Financial Market Risk	Interest rate exposure, Foreign exchange, Asset valuations, Commodity prices, Hedging risks, Value at Risk Analysis, Investment Portfolio management
Operational	Business Continuity	Disaster recovery, Crisis management, Backup systems, Emergency response, Business restoration, Continuity planning, Critical function maintenance
	Human Capital	Talent management, Labor relations, Skills development, Talent acquisition, Performance management, Employee engagement
	Information & Data Risks	Cybersecurity, Data privacy, Information protection, System access, Cloud security, Network security, Third-party data handling
	Organizational Development	Succession planning, Leadership development, Organizational culture
	Physical Infrastructure	Physical assets, Equipment reliability, Facility management, Workplace safety
	Process Execution	Quality control, Efficiency metrics, Capacity utilization, Process integration, Standard operating procedures, Operational efficiency, Transaction processing, Project delivery, Timeline management
	Supply Chain	Supplier management, Logistics network, Inventory control, Distribution systems, Supply chain resilience, Strategic supply chain partnerships, Supplier concentration, Supplier service quality
	Technology Infrastructure	System reliability, Company-specific tech capabilities, System architecture, IT strategy, Technology obsolescence, Digital capabilities, System integration, Enterprise systems
Regulatory and Compliance	Corporate Governance	Board oversight, Internal controls, Ethics/conduct, Decision rights, Accountability structure, Audit risk
	Litigation	Contract compliance, Legal proceedings, Dispute resolution, Liability exposure, Intellectual property
	Regulatory Compliance	Regulations, Compliance standards, Operating licenses, Reporting requirements, Regulatory changes, Environmental compliance, Labor regulations, Capital adequacy requirements
	Reputation & Communication	Media relations, Stakeholder management, Corporate social responsibility, Public relations, Social media risks, Corporate reputation, Community relations, Crisis communication
Strategic	Business Model	Value proposition, Revenue streams, Cost structure, Resource allocation, Business model disruption, Strategic agility, Adaptation capability, Development costs, Strategic initiatives
	Competition Strategy	Firm-level market share, Product differentiation, Firm-specific pricing power, Price-cost margin, Brand value, Product adoption, Competitive advantage, Strategic positioning
	Customer Relations	Customer satisfaction, Customer retention, Service quality, Digital engagement, Customer concentration, Customer trust, Client relations, Customer service
	Growth management	Organic growth, Market development, Market expansion, M&A execution, Post-merger integration, Strategic partnerships other than supply chain
	Innovation Risks	R&D effectiveness, Innovation pipeline, Development timing, Product testing, Patent development, IP protection, Research rights, Technical feasibility

Table A.2: **Generated Regressor Diagnostics: LLM Text Constructs with Bag-of-Words Wedges**

This table augments Table 4 with two bag-of-words wedges that summarize the source-anchored content of the LLM's vocabulary choices. For each (*document* × *LLM*), we form three pooled bags from the top-1,000 χ^2-discriminative terms, omitted, injected, and retained, embed each bag via the Doc2Vec model, and compute its cosine against the source-document embedding. We denote the cosine for the omitted bag as ρ_{BO} and for the injected bag as ρ_{BI}. The joint specification augments the naive baseline of Table 4 with the two lexical wedges ρ_O, ρ_I and these two bag-of-words wedges: $y_{i,t+1} = \beta\,\text{Construct}_{i,t} + \rho_O\,O_{im,t} + \rho_I\,I_{im,t} + \rho_{BO}\,\cos_{BO} + \rho_{BI}\,\cos_{BI} + \gamma' X_{i,t} + \tau_t + \varepsilon_{i,t+1}$, yielding β_{joint}, ρ_O, ρ_I, ρ_{BO}, ρ_{BI}. $|\Delta\%| = 100 \times |\beta_{\text{joint}} - \beta_{\text{naive}}|/|\beta_{\text{naive}}|$ measures coefficient sensitivity. The decomposition θ, λ follows Table 4, but with the wedge-predictable part now spanning the four wedges $\{\rho_O, \rho_I, \rho_{BO}, \rho_{BI}\}$. Panel A summarizes regressions on S&P 500 Item 1A risk disclosures for all four LLMs; Panel B summarizes regressions on S&P 500 quarterly earnings-call transcripts for gpt-4o-mini and Llama-3.3-70B only. Rows were pooled across LLMs and constructs within each outcome. For each outcome variable, in the first row, we report the mean coefficients for β_{naive}, β_{joint}, wedge coefficient ρ, and the median for $|\Delta\%|$, θ, and λ. The bracketed entries beneath each cell report the share of underlying cells significant at the 5% level; except the bracketed entry under λ is the fraction of cells where $p(\theta = \lambda) < 0.05$. Panel A regressions include Year FE and Panel B regressions include Year-quarter FE. Standard errors are CR1 firm-clustered. Raw per-LLM panel tables underlying these summaries are reported in Appendix B.

Panel A: Item 1A risk disclosures

		Construct coeff.		Lexical Wedges		BOW Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	ρ_{BO}	ρ_{BI}	Med $\lvert\Delta\%\rvert$	θ	λ
NI/Assets	21,342	-0.0076	-0.0091	-0.0536	-0.0378	0.0878	0.0073	17.0%	-0.0362	-0.0101
		[75.0%]	[75.0%]	[68.8%]	[100.0%]	[100.0%]	[18.8%]			[93.8%]
CapEx/Assets	21,342	0.0008	-0.0001	0.0180	-0.0008	0.0013	-0.0017	37.5%	0.0108	0.0002
		[50.0%]	[18.8%]	[81.2%]	[0.0%]	[0.0%]	[0.0%]			[81.2%]
R&D ratio	21,342	0.0144	0.0146	0.0857	0.0480	-0.0582	0.0041	13.3%	0.0440	0.0174
		[87.5%]	[87.5%]	[93.8%]	[100.0%]	[100.0%]	[68.8%]			[81.2%]
Leverage	20,209	0.0037	-0.0031	0.1608	0.1125	-0.0065	-0.1417	74.8%	0.0560	-0.0010
		[12.5%]	[37.5%]	[75.0%]	[75.0%]	[0.0%]	[100.0%]			[62.5%]
Sales growth	20,164	0.0120	0.0103	0.1473	0.0969	-0.0877	0.0147	18.7%	0.0570	0.0107
		[68.8%]	[50.0%]	[87.5%]	[100.0%]	[100.0%]	[0.0%]			[56.2%]
WW index	13,961	0.0038	0.0035	0.0344	0.0533	0.0194	-0.0136	15.6%	0.0115	0.0056
		[56.2%]	[56.2%]	[50.0%]	[75.0%]	[25.0%]	[25.0%]			[43.8%]

Panel B: Earnings call transcripts

		Construct coeff.		Lexical Wedges		BOW Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	ρ_{BO}	ρ_{BI}	Med $\lvert\Delta\%\rvert$	θ	λ
NI/Assets	18,196	−0.0008	−0.0008	+0.0106	−0.0161	−0.0170	−0.0020	10.8%	+0.0010	−0.0009
		[62.5%]	[75.0%]	[50.0%]	[75.0%]	[25.0%]	[0.0%]			[25.0%]
CapEx/Assets	18,194	+0.0000	+0.0001	+0.0031	+0.0086	+0.0035	+0.0000	25.3%	−0.0012	+0.0001
		[37.5%]	[37.5%]	[0.0%]	[50.0%]	[0.0%]	[0.0%]			[0.0%]
R&D ratio	18,196	+0.0007	+0.0007	+0.0389	+0.0108	+0.0462	+0.0171	13.4%	+0.0009	+0.0004
		[62.5%]	[62.5%]	[100.0%]	[37.5%]	[100.0%]	[100.0%]			[50.0%]
Leverage	18,196	+0.0087	+0.0086	−0.0324	−0.1991	−0.7458	−0.0986	16.6%	+0.0033	+0.0065
		[25.0%]	[25.0%]	[0.0%]	[0.0%]	[100.0%]	[12.5%]			[12.5%]
Sales growth	18,192	−0.0207	−0.0200	+0.1432	+0.2365	+0.1860	+0.0470	7.5%	−0.0311	−0.0141
		[62.5%]	[62.5%]	[37.5%]	[100.0%]	[0.0%]	[25.0%]			[25.0%]
WW index	18,196	+0.0007	+0.0007	+0.0164	+0.0126	+0.1404	+0.0168	28.8%	−0.0001	+0.0004
		[12.5%]	[12.5%]	[0.0%]	[0.0%]	[100.0%]	[37.5%]			[50.0%]

Table A.3: Jha et al. (2024) Primary Results Replication

This table replicates the three headline regressions of Jha et al. (2024): the firm-quarter ChatGPT Investment Score forecasts capital expenditure at t+2 (their Table 3 col. 4), capital expenditure at t+4 (Table 4 col. 2), and PT-scaled R&D at t+2 (Table 5 col. 8). The regression specification follows their equation (2): $y_{i,t+h} = \beta\,\text{ChatGPT Investment Score}_{i,t} + \beta_2\, q_{\text{tot},i,t} + \beta_3\, y_{i,t} + \gamma' X_{i,t} + \alpha_i + \alpha_t + \varepsilon_{i,t}$, with the control set $X_{i,t}$ = {Size, q_{tot} (Peters–Taylor Total q), CashFlow, Leverage}, lagged $y_{i,t}$, firm fixed effects α_i, and year-quarter fixed effects α_t. Standard errors are clustered on firm. The predictor is the −1 to +1 ChatGPT Investment Score (page 14 of Jha et al. (2024)), averaged across earnings-call chunks. Outcomes are in percent (CapEx/Assets and PT-scaled R&D as pp), matching their reporting convention. The standardized beta $\beta \cdot \text{sd}(X)/\text{sd}(Y)$ is the unit-free magnitude they report in their text. Jha et al. (2024) used `gpt-3.5-turbo`; the replication uses `gpt-4o-mini` and `Llama-3.3-70B`. Sample is S&P 500 firms with quarterly earnings-call transcripts for the period 2006–2020. * $p < .10$, ** $p < .05$, *** $p < .01$.

	gpt-4o-mini					Llama-3.3-70B				
Source	coef	std β	t	R^2	N	coef	std β	t	R^2	N
Panel A: $CapEx_{i,t+2}$ (Jha Table 3 col. 4)										
Jha et al. (2024)	+0.4770 * **	0.055–0.079	13.91	0.575	74,429	+0.4770 * **	0.055–0.079	13.91	0.575	74,429
Replication	+0.5160 * **	+0.0795 * **	11.05	0.807	13,665	+0.4438 * **	+0.0841 * **	10.76	0.807	13,665
Panel B: $CapEx_{i,t+4}$ (Jha Table 4 col. 2)										
Jha et al. (2024)	+0.4740 * **	0.060–0.080	13.99	0.543	73,598	+0.4740 * **	0.060–0.080	13.99	0.543	73,598
Replication	+0.5732 * **	+0.0902 * **	10.03	0.798	12,720	+0.5423 * **	+0.1051 * **	10.22	0.800	12,720
Panel C: $R\&D_{i,t+2}$ (PT-scaled) (Jha Table 5 col. 8)										
Jha et al. (2024)	+0.1350 * **	0.020–0.025	5.37	0.905	34,253	+0.1350 * **	0.020–0.025	5.37	0.905	34,253
Replication	+0.0459	+0.0060	0.92	0.782	13,678	+0.0547	+0.0066	1.23	0.782	13,678

Table A.4: **Variable Definitions**

Variable	**Definition / Formula**	**Source**
Vocabulary descriptives (Table 1; computed per text source × model)		
Total words	Total word count in the corpus	Authors' calculation
Vocabulary size	Number of distinct words appearing at least once in the corpus	Authors' calculation
Top-N concentration	Percentage of total words accounted for by the N most frequent words: *Top-N* $= \left(\sum_{w \in \text{top-}N} c(w)\right)/T \times 100\%$. where $c(w)$ is the count of word w. We report $N \in \{100, 500, 1000\}$.	Authors' calculation
Entropy (normalized Shannon)	$H_{\text{norm}} = \left(-\sum_w p_w \log_2 p_w\right) \Big/ \log_2 V$, $p_w = c(w)/T$. Bounded on $[0, 1]$; $H_{\text{norm}} = 1$ indicates a uniform word distribution (maximum lexical diversity for the given V), $H_{\text{norm}} \to 0$ means a small set of words dominates the corpus.	Authors' calculation
LLM-mechanism regressors		
Discriminative term	A word identified as informative for distinguishing one risk-factor category from the others. We rank candidate words within each (model, category, text source) cell by lift, retain the top 1,000 with *in category count* ≥ 15 and significant χ^2 statistics, and treat the resulting vocabulary as the discriminative term set for that category. The same procedure is applied to the Item 1A text and to each LLM's summary, producing $V_{i,\text{source}}$ and $V_{im,\text{summary}}$.	Authors' calculation
Omission rate (lexical)	$O_{im} = 1 - \frac{\lvert V_{i,\text{source}} \cap V_{im,\text{summary}} \rvert}{V_{i,\text{source}}}$. Share of human discriminative terms not present in the LLM's discriminative-term set.	Authors' calculation
Injection rate (lexical)	$I_{im} = 1 - \frac{\lvert V_{im,\text{summary}} \cap V_{i,\text{source}} \rvert}{V_{im,\text{summary}}}$. Share of LLM discriminative terms not present in the human discriminative-term set.	Authors' calculation
Omission rate (semantic)	Cosine similarity between the Doc2Vec embedding of the bag of omitted terms and the embedding of the source document; captures how central the dropped content is to the disclosure.	Authors' calculation
Injection rate (semantic)	Cosine similarity between the Doc2Vec embedding of the bag of injected terms and the embedding of the source document.	Authors' calculation
Outcome variables		
NI/Assets	Net income divided by total assets	Compustat
CapEx/Assets	Capital expenditure divided by total assets	Compustat
R&D ratio	R&D expense divided by total assets	Compustat
Sales growth	Growth rate of sales	Compustat
WW index	Whited–Wu (2006) financial-constraint index	Compustat
Cash ratio	Cash divided by total assets	Compustat
Dividends/Assets	Dividends divided by total assets	Compustat
Operating cash flow/Assets	Operating cash flow divided by total assets	Compustat
Firm-level controls (firm-year, year t)		
Leverage	Total debt divided by total assets	Compustat annual
Cash/Assets	Cash and short-term investments divided by total assets	Compustat annual
Market-to-Book	Market value of common equity divided by book value: $(csho \times prcc_f) \big/ ceq$	Compustat annual
Operating Income/Assets	Operating income before depreciation divided by total assets: $oibdp/at$	Compustat annual
Size	Natural logarithm of total assets	Compustat
CashFlow	Cash flow divided by total assets	Compustat
Tangibility	Net property, plant, and equipment divided by total assets	Compustat
Age	Time since the firm's first appearance in the sample	Compustat
Peters–Taylor Total q	Total q of Peters and Taylor (2017), Tobin's q augmented with intangible capital	Compustat

B Appendix: Discriminative Term Ranking, Lift, and the LLM Wedge Measures

This appendix describes how we identify a document's *distinctive* vocabulary and how the omission and injection rates are built from it. There are four steps in the construction: tokenize the text, score each term's distinctiveness by *lift*, keep the statistically significant top terms, and compare the source and LLM summary vocabularies using simple set differences.

Step 1: Tokenization

We reduce each source disclosure and each LLM summary to a *set* of word stems. We lowercase the text, keep alphabetic tokens of at least two characters, map each token to its dictionary base form (lemmatization, so that *supplies*, *supplied*, and *supplying* all become *supply*), and remove standard English stopwords (*the*, *and*, *of*) along with a short list of domain-generic words. Repeated words count once. Crucially, the source text and the summary are processed with the identical pipeline, so any later difference between them reflects content, not preprocessing.

Step 2: Distinctive terms by lift

Common words such as *company* or *risk* appear throughout the corpus and say little about what a particular disclosure is about. We therefore rank terms by how *concentrated* they are in a given risk category relative to the rest of the corpus. For a term w and risk category c, its *lift* is

$$\text{lift}(w, c) \;=\; \frac{\text{fraction of category-}c\text{ words that are }w}{\text{fraction of non-category-}c\text{ words that are }w}. \tag{9}$$

A lift of one means the term is equally common inside and outside the category; a lift of, say, thirty means it is thirty times more concentrated in that category. Lift is the empirical counterpart of the familiar TF–IDF idea: both reward terms that are frequent in the target context but rare elsewhere.

Example. In supply-chain disclosures, *supplier* might occur at roughly eight times the rate it does in the text of other categories, giving lift ≈ 8, whereas *company* occurs at about the same rate everywhere, giving lift ≈ 1. The word *supplier* is distinctive of the category; *company* is not.

Step 3: Significance and ranking

A high lift can arise by chance for a rare word. We therefore also score each candidate term with a chi-squared statistic that tests whether its concentration in the category is statistically significant rather than a small-sample artifact, and we require a minimum corpus frequency of 15 occurrences so that estimates are stable. Retaining the significant terms and ranking them by lift, we keep the top n per category. We set $n = 1{,}000$ in the primary analysis and keep $n \in \{200, 500, 2{,}000, 5{,}000\}$ for robustness checks. This produces, for each risk category, a distinctive *source dictionary* and a distinctive *summary dictionary*.

Step 4: Omission and injection

Consider a single disclosure assigned to category i. Let $V^n_{i,source}$ be the set of its source lemmas that fall in category i's source dictionary, and let $V^n_{im,summary}$ be the set of its summary lemmas that fall in category i's summary dictionary. We set n=1000 and focus on the top-1000 discriminative

source and summary terms for each category in our primary analyses. Omission and injection are set differences over these distinctive vocabularies:

$$\text{Omission}(O_{im}) \;=\; 1 - \frac{|V^{(n)}_{i,\text{source}} \cap V^{(n)}_{im,\text{summary}}|}{|V^{(n)}_{i,\text{source}}|}, \qquad \text{Injection}(I_{im}) \;=\; 1 - \frac{|V^{(n)}_{i,\text{source}} \cap V^{(n)}_{im,\text{summary}}|}{|V^{(n)}_{i,\text{source}}|}.$$

Omission is the share of the source's distinctive terms that the summary drops; injection is the share of the summary's distinctive terms that are not among the source's distinctive terms.

Example. Suppose a supply-chain disclosure's distinctive source terms are $V^{n}_{i,source}$={tariff, supplier, logistics, inventory, freight} and its summary's distinctive terms are $V^{n}_{im,summary}$={supplier, logistics, disruption, delay} for model (m)=gpt-4o-mini. The summary drops *tariff*, *inventory*, and *freight*, so Omission = 3/5 = 60%. It introduces *disruption* and *delay*, which are not among the source's distinctive terms, so Injection = 2/4 = 50% for gpt-4o-mini.

Earnings-call implementation

The earnings-call replication uses the same logic with three implementation differences suited to that corpus and to closely follow Jha et al. (2024). First, because transcripts are analyzed in chunks rather than pooled by risk category, distinctive terms are ranked *per chunk*: a term's lift in (9) compares its frequency in the chunk to its frequency across the whole earnings-call corpus. Second, we apply an explicit significance cutoff—a chi-squared value above 3.84 (the 5% level, one degree of freedom) and a minimum of fifteen corpus occurrences. Third, tokenization replaces each firm's name with a placeholder token so that company names do not crowd the distinctive lists. The omission formula is unchanged; the injection numerator counts a summary term as injected if it is absent from the chunk's *full* vocabulary, a slightly stricter reference set than the top-n used above.

LLMs and Systematic Measurement Error

Internet Appendix

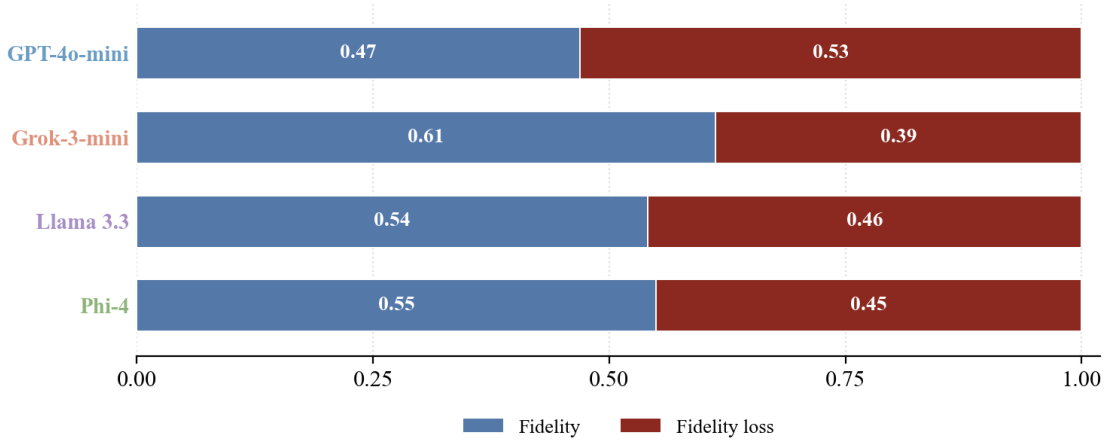
July 2026

In this Internet Appendix, we report additional supporting analyses for the main results in the paper. We begin by documenting the robustness of the fidelity measures to the expanding Doc2Vec training window in the paper by repeating the analysis with Doc2Vec models trained using documents from 2018 and a 5-year rolling window of documents. We also report the per-LLM base tables that underlie the pooled generated-regressor diagnostics in the main text.

Figure IA.1: **Fidelity Lost During LLM Summarization: Robustness Analysis**

In Panel A of this figure, we reproduce Figure 3 using a single Doc2Vec embedding space fixed at 2018 in place of the period-specific (expanding-window) models, holding everything else identical. The 2018 model is trained once on the S&P 500 Item 1A risk disclosures filed in 2018 and is then used to embed the source documents and LLM summaries. We report the average total fidelity (blue) and total fidelity loss (red) for each of the four LLMs (GPT-4o-mini, Grok-3-mini, Phi-4, and Llama-3.3-70B). In Panel B, we use Doc2Vec models estimated on a five-year rolling window instead of the expanding window. For each filing year Y the embedding is trained on the S&P 500 Item 1A disclosures from years $Y-4$ through Y, and the source documents, summaries, and discriminative-term bags are embedded with that year's model. The cross-model ordering and the signal–distortion split are unchanged, confirming that the fidelity result does not depend on the time-varying (expanding) embedding.

Panel A: Fidelity using 2018 embedding space



Panel B: Fidelity using 5-year rolling window embedding space

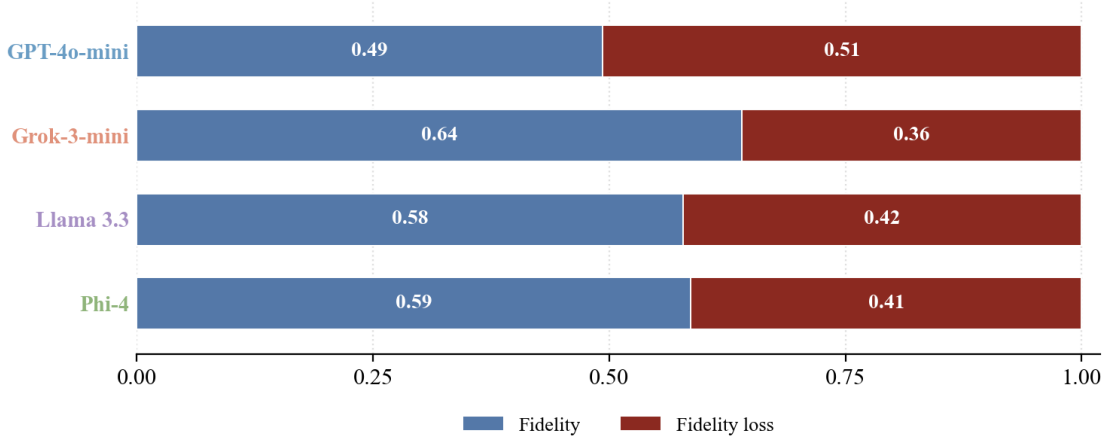


Table IA.1: **Generated Regressor Diagnostics by LLM: Item 1A Risk Disclosures**

This table reports the per-LLM regressions that Table 4 summarizes by pooling across the LLMs and the four constructs. Each panel is one LLM; within a panel the four blocks are the constructs (severity, boilerplate, complexity, and specificity) and the rows are the six forward firm outcomes. Columns report the naive and joint construct coefficients (β_{naive}, β_{joint}), the omission and injection wedge loadings (ρ_O, ρ_I), the coefficient sensitivity ($|\Delta\%|$), and the decomposition coefficients (θ, λ); see Table 4 for variable definitions and the full specification. Each cell is a single regression, with ***, **, and * denoting significance at the 1%, 5%, and 10% levels. Regressions include Year fixed effects and standard errors are clustered at the firm level.

Panel A: GPT-4o-mini

		Coefficient on construct		Lexical Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
				Severity				
NI/Assets	21,486	+0.0062	+0.0062	−0.0399 ∗ ∗∗	−0.0576 ∗ ∗∗	0.1%	+0.0210	+0.0062
CapEx/Assets	21,486	+0.0012	+0.0011	+0.0156 ∗ ∗∗	−0.0006	13.2%	+0.3351 ∗ ∗∗	+0.0011
R&D ratio	21,486	+0.0101 ∗ ∗∗	+0.0096 ∗ ∗∗	+0.1049 ∗ ∗∗	+0.0763 ∗ ∗∗	5.1%	+1.0712 ∗ ∗∗	+0.0096 ∗ ∗∗
Leverage	20,357	−0.0081	−0.0093	+0.2196 ∗ ∗∗	+0.1483 ∗ ∗∗	14.2%	+2.3359 ∗ ∗	−0.0093
Sales growth	20,312	+0.0099	+0.0093	+0.1278 ∗ ∗∗	+0.1085 ∗ ∗∗	5.5%	+1.1325∗	+0.0093
WW index	14,086	+0.0015	+0.0016	+0.0964 ∗ ∗∗	+0.0692 ∗ ∗∗	5.0%	−0.2066	+0.0016
				Boilerplate				
NI/Assets	21,486	+0.0042	+0.0042	−0.0395 ∗ ∗∗	−0.0577 ∗ ∗∗	0.2%	+0.0156	+0.0042
CapEx/Assets	21,486	−0.0008	−0.0006	+0.0157 ∗ ∗∗	−0.0006	23.8%	−0.3304 ∗ ∗∗	−0.0006
R&D ratio	21,486	−0.0046∗	−0.0040	+0.1051 ∗ ∗∗	+0.0762 ∗ ∗∗	13.6%	−1.1054 ∗ ∗∗	−0.0040
Leverage	20,357	+0.0013	+0.0028	+0.2194 ∗ ∗∗	+0.1484 ∗ ∗∗	115.1%	−2.5640 ∗ ∗∗	+0.0028
Sales growth	20,312	−0.0074	−0.0067	+0.1279 ∗ ∗∗	+0.1085 ∗ ∗∗	9.0%	−1.1605 ∗ ∗	−0.0067
WW index	14,086	−0.0152 ∗ ∗	−0.0145 ∗ ∗	+0.0958 ∗ ∗∗	+0.0693 ∗ ∗∗	4.8%	−1.1872 ∗ ∗∗	−0.0145 ∗ ∗
				Complexity				
NI/Assets	21,486	+0.0045	+0.0050	−0.0400 ∗ ∗∗	−0.0577 ∗ ∗∗	11.4%	−0.5128 ∗ ∗∗	+0.0050
CapEx/Assets	21,486	+0.0023	+0.0021	+0.0155 ∗ ∗∗	−0.0007	10.1%	+0.2410 ∗ ∗∗	+0.0021
R&D ratio	21,486	+0.0095 ∗ ∗∗	+0.0080 ∗ ∗∗	+0.1048 ∗ ∗∗	+0.0761 ∗ ∗∗	15.4%	+1.4868 ∗ ∗∗	+0.0080 ∗ ∗∗
Leverage	20,357	−0.0088	−0.0119	+0.2201 ∗ ∗∗	+0.1485 ∗ ∗∗	35.3%	+3.0531 ∗ ∗∗	−0.0119
Sales growth	20,312	+0.0080	+0.0062	+0.1277 ∗ ∗∗	+0.1083 ∗ ∗∗	22.4%	+1.7611 ∗ ∗∗	+0.0062
WW index	14,086	−0.0010	−0.0017	+0.0965 ∗ ∗∗	+0.0691 ∗ ∗∗	75.0%	+2.9764 ∗ ∗∗	−0.0017
				Specificity				
NI/Assets	21,486	−0.0042	−0.0042	−0.0393 ∗ ∗∗	−0.0578 ∗ ∗∗	0.2%	−0.0100	−0.0042
CapEx/Assets	21,486	−0.0005	−0.0007	+0.0157 ∗ ∗∗	−0.0007	44.7%	+0.1507 ∗ ∗∗	−0.0007
R&D ratio	21,486	+0.0011	+0.0004	+0.1053 ∗ ∗∗	+0.0762 ∗ ∗∗	67.6%	+0.5081 ∗ ∗∗	+0.0004
Leverage	20,357	+0.0001	−0.0016	+0.2194 ∗ ∗∗	+0.1484 ∗ ∗∗	1174.8%	+1.0479 ∗ ∗∗	−0.0016
Sales growth	20,312	+0.0052	+0.0044	+0.1277 ∗ ∗∗	+0.1085 ∗ ∗∗	14.9%	+0.4657∗	+0.0044
WW index	14,086	+0.0012	+0.0005	+0.0964 ∗ ∗∗	+0.0692 ∗ ∗∗	58.9%	+0.4217 ∗ ∗∗	+0.0005

Panel B: Grok-3-mini

Outcome	N	Coefficient on construct		Lexical Wedges		Sensitivity	Decomposition	
		β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
				Severity				
NI/Assets	21,526	$-0.0147***$	$-0.0143***$	$-0.0481***$	$-0.0734***$	3.0%	$-0.0171***$	$-0.0143***$
CapEx/Assets	21,526	$+0.0009$	-0.0003	$+0.0147***$	-0.0010	136.6%	$+0.0074***$	-0.0003
R&D ratio	21,526	$+0.0228***$	$+0.0214***$	$+0.0581***$	$+0.0697***$	6.1%	$+0.0305***$	$+0.0214***$
Leverage	20,390	$+0.0079$	-0.0065	$+0.2305***$	$+0.0830**$	183.0%	$+0.0961***$	-0.0065
Sales growth	20,345	$+0.0208***$	$+0.0149***$	$+0.1537***$	$+0.1413***$	28.3%	$+0.0566***$	$+0.0149***$
WW index	14,111	$+0.0116***$	$+0.0085***$	$+0.0786***$	$+0.0576***$	26.3%	$+0.0284***$	$+0.0085***$
				Boilerplate				
NI/Assets	21,526	$+0.0125***$	$+0.0094***$	$-0.0586***$	$-0.0648***$	24.9%	$+0.0457***$	$+0.0094***$
CapEx/Assets	21,526	$-0.0044***$	$-0.0038***$	$+0.0093*$	-0.0004	14.2%	$-0.0111***$	$-0.0038***$
R&D ratio	21,526	$-0.0314***$	$-0.0283***$	$+0.0559***$	$+0.0579***$	9.7%	$-0.0635***$	$-0.0283***$
Leverage	20,390	$+0.0064$	$+0.0218***$	$+0.2476***$	$+0.0858**$	243.1%	$-0.1688***$	$+0.0218***$
Sales growth	20,345	$-0.0320***$	$-0.0236***$	$+0.1466***$	$+0.1337***$	26.1%	$-0.1264***$	$-0.0236***$
WW index	14,111	$-0.0193***$	$-0.0155***$	$+0.0734***$	$+0.0534***$	20.0%	$-0.0672***$	$-0.0155***$
				Complexity				
NI/Assets	21,526	$-0.0124***$	$-0.0127***$	$-0.0425***$	$-0.0762***$	1.9%	$-0.0117***$	$-0.0127***$
CapEx/Assets	21,526	$+0.0025***$	$+0.0016$	$+0.0106**$	$+0.0008$	34.7%	$+0.0052***$	$+0.0016$
R&D ratio	21,526	$+0.0326***$	$+0.0361***$	$+0.0120$	$+0.0901***$	10.7%	$+0.0212***$	$+0.0361***$
Leverage	20,390	-0.0082	$-0.0293***$	$+0.2844***$	$+0.0596$	258.5%	$+0.0631***$	$-0.0293***$
Sales growth	20,345	$+0.0164***$	$+0.0104**$	$+0.1531***$	$+0.1421***$	36.5%	$+0.0365***$	$+0.0104**$
WW index	14,111	$+0.0112***$	$+0.0086***$	$+0.0735***$	$+0.0606***$	23.1%	$+0.0198***$	$+0.0086***$
				Specificity				
NI/Assets	21,526	$-0.0125***$	$-0.0104***$	$-0.0463***$	$-0.0650***$	17.4%	$-0.0257***$	$-0.0104***$
CapEx/Assets	21,526	$+0.0011$	$+0.0003$	$+0.0135***$	-0.0007	72.4%	$+0.0061***$	$+0.0003$
R&D ratio	21,526	$+0.0201***$	$+0.0176***$	$+0.0508***$	$+0.0572***$	12.6%	$+0.0354***$	$+0.0176***$
Leverage	20,390	$+0.0238***$	$+0.0130*$	$+0.1912***$	$+0.0872**$	45.2%	$+0.0932***$	$+0.0130*$
Sales growth	20,345	$+0.0226***$	$+0.0151***$	$+0.1413***$	$+0.1329***$	33.2%	$+0.0710***$	$+0.0151***$
WW index	14,111	$+0.0134***$	$+0.0100***$	$+0.0696***$	$+0.0535***$	25.6%	$+0.0365***$	$+0.0100***$

Panel C: Llama-3.3-70B

Outcome	N	Coefficient on construct		Lexical Wedges		Sensitivity	Decomposition	
		β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
				Severity				
NI/Assets	21,531	$-0.0035**$	$-0.0071***$	$+0.0356***$	$-0.0649***$	99.3%	$+0.0149***$	$-0.0071***$
CapEx/Assets	21,531	$+0.0010*$	$+0.0004$	$+0.0205***$	-0.0005	57.4%	$+0.0042***$	$+0.0004$
R&D ratio	21,531	$+0.0077***$	$+0.0093***$	$+0.0415***$	$+0.0732***$	21.8%	-0.0011	$+0.0093***$
Leverage	20,369	$+0.0105**$	$+0.0087$	$+0.0788*$	$+0.0097$	17.2%	$+0.0200*$	$+0.0087$
Sales growth	20,324	$+0.0037$	$+0.0078***$	$+0.0099$	$+0.1124***$	111.0%	$-0.0179**$	$+0.0078**$
WW index	14,095	$+0.0053***$	$+0.0068***$	$+0.0111$	$+0.0403***$	27.6%	-0.0014	$+0.0068***$
				Boilerplate				
NI/Assets	21,531	$+0.0049***$	$+0.0050***$	$+0.0161$	$-0.0523***$	2.5%	-0.0312	$+0.0050***$
CapEx/Assets	21,531	$-0.0011*$	-0.0010	$+0.0213***$	-0.0013	11.0%	$-0.0382***$	-0.0010
R&D ratio	21,531	$-0.0060***$	$-0.0056***$	$+0.0678***$	$+0.0564***$	6.2%	$-0.1148***$	$-0.0056***$
Leverage	20,369	$+0.0034$	$+0.0040$	$+0.1088**$	-0.0066	19.3%	$-0.1820**$	$+0.0040$
Sales growth	20,324	$-0.0099***$	$-0.0098***$	$+0.0291$	$+0.0988***$	1.0%	-0.0364	$-0.0098***$
WW index	14,095	$-0.0039**$	$-0.0039**$	$+0.0304**$	$+0.0258***$	0.3%	-0.0022	$-0.0039**$
				Complexity				
NI/Assets	21,531	$-0.0041***$	$-0.0088***$	$+0.0485***$	$-0.0679***$	112.8%	$+0.0112***$	$-0.0088***$
CapEx/Assets	21,531	$+0.0015***$	$+0.0008$	$+0.0187***$	$+0.0001$	48.6%	$+0.0040***$	$+0.0008$
R&D ratio	21,531	$+0.0142***$	$+0.0177***$	$+0.0000$	$+0.0881***$	24.8%	$+0.0026$	$+0.0177***$
Leverage	20,369	$+0.0020$	-0.0032	$+0.1191**$	-0.0120	255.5%	$+0.0193*$	-0.0032
Sales growth	20,324	$+0.0046$	$+0.0093***$	-0.0025	$+0.1149***$	102.3%	$-0.0111*$	$+0.0093**$
WW index	14,095	$+0.0043***$	$+0.0057***$	$+0.0111$	$+0.0372***$	30.8%	$+0.0000$	$+0.0057***$
				Specificity				
NI/Assets	21,531	$-0.0096***$	$-0.0111***$	$+0.0306**$	$-0.0581***$	15.5%	$+0.0260***$	$-0.0111***$
CapEx/Assets	21,531	$+0.0008$	$+0.0003$	$+0.0213***$	-0.0011	57.3%	$+0.0116***$	$+0.0003$
R&D ratio	21,531	$+0.0122***$	$+0.0121***$	$+0.0521***$	$+0.0628***$	1.1%	$+0.0154**$	$+0.0121***$
Leverage	20,369	$+0.0113*$	$+0.0095$	$+0.0918*$	-0.0010	16.3%	$+0.0562**$	$+0.0095$
Sales growth	20,324	$+0.0179***$	$+0.0194***$	$+0.0045$	$+0.1091***$	8.9%	-0.0205	$+0.0194***$
WW index	14,095	$+0.0085***$	$+0.0089***$	$+0.0192$	$+0.0326***$	5.5%	-0.0002	$+0.0089***$

Panel D: Phi-4

		Coefficient on construct		Lexical Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
				Severity				
NI/Assets	21,594	$-0.0049**$	$-0.0090**$	$+0.0128$	$-0.0665**$	84.3%	$+0.0128**$	$-0.0090**$
CapEx/Assets	21,594	$+0.0005$	-0.0002	$+0.0129**$	-0.0055	138.0%	$+0.0033*$	-0.0002
R&D ratio	21,594	$+0.0048**$	$+0.0063**$	$+0.0436**$	$+0.0474**$	31.7%	-0.0018	$+0.0063**$
Leverage	20,443	$+0.0059$	$+0.0070$	$+0.1262**$	$+0.0765*$	18.6%	$+0.0013$	$+0.0070$
Sales growth	20,398	$+0.0040$	$+0.0101**$	$+0.0161$	$+0.1139**$	151.5%	$-0.0216**$	$+0.0101**$
WW index	14,153	$+0.0058**$	$+0.0068**$	$+0.0197$	$+0.0250*$	16.9%	$+0.0023$	$+0.0068**$
				Boilerplate				
NI/Assets	21,594	$-0.0045*$	$-0.0037*$	$+0.0045$	$-0.0383**$	16.7%	$-0.0318**$	$-0.0037*$
CapEx/Assets	21,594	$-0.0043**$	$-0.0045**$	$+0.0150**$	-0.0012	5.2%	$+0.0038$	$-0.0045**$
R&D ratio	21,594	$-0.0113**$	$-0.0131**$	$+0.0575**$	$+0.0409**$	16.1%	$+0.0552**$	$-0.0131**$
Leverage	20,443	$+0.0151*$	$+0.0120*$	$+0.1278**$	$+0.0462$	20.7%	$+0.1252**$	$+0.0120*$
Sales growth	20,398	-0.0033	-0.0058	$+0.0307$	$+0.0902**$	74.2%	$+0.0838**$	-0.0058
WW index	14,153	$-0.0078**$	$-0.0088**$	$+0.0326*$	$+0.0130$	12.1%	$+0.0152$	$-0.0088**$
				Complexity				
NI/Assets	21,594	$-0.0113**$	$-0.0166**$	$+0.0343**$	$-0.0528**$	47.1%	$+0.0066$	$-0.0166**$
CapEx/Assets	21,594	$+0.0014$	-0.0001	$+0.0130*$	-0.0051	110.0%	$+0.0067**$	-0.0001
R&D ratio	21,594	$+0.0220**$	$+0.0225**$	$+0.0077$	$+0.0452**$	2.3%	$+0.0203**$	$+0.0225**$
Leverage	20,443	-0.0059	$-0.0242*$	$+0.1805**$	$+0.0403$	310.6%	$+0.0568**$	$-0.0242**$
Sales growth	20,398	$+0.0095*$	$+0.0117*$	$+0.0052$	$+0.0931**$	22.6%	$+0.0022$	$+0.0117*$
WW index	14,153	$+0.0064*$	$+0.0044$	$+0.0193$	$+0.0065$	30.4%	$+0.0129*$	$+0.0044$
				Specificity				
NI/Assets	21,594	$-0.0088**$	$-0.0121**$	$+0.0119$	$-0.0625**$	37.2%	$+0.0201**$	$-0.0121**$
CapEx/Assets	21,594	$+0.0013*$	$+0.0008$	$+0.0121*$	-0.0035	35.4%	$+0.0054*$	$+0.0008$
R&D ratio	21,594	$+0.0133**$	$+0.0150**$	$+0.0392**$	$+0.0558**$	12.8%	-0.0016	$+0.0150**$
Leverage	20,443	$+0.0143*$	$+0.0154*$	$+0.1224**$	$+0.0839*$	8.0%	$+0.0044$	$+0.0154*$
Sales growth	20,398	$+0.0176**$	$+0.0235**$	$+0.0095$	$+0.1268**$	33.6%	$-0.0333*$	$+0.0235**$
WW index	14,153	$+0.0115**$	$+0.0126**$	$+0.0180$	$+0.0292*$	10.0%	$+0.0036$	$+0.0126**$

Table IA.2: **Generated Regressor Diagnostics by LLM: Earnings Call Transcripts**

This table reports the per-LLM regressions that Table 5 summarizes by pooling across the LLMs and the four constructs. Each panel is one LLM; within a panel the four blocks are the constructs (severity, boilerplate, complexity, and specificity) and the rows are the six forward firm outcomes. Columns report the naive and joint construct coefficients (β_{naive}, β_{joint}), the omission and injection wedge loadings (ρ_O, ρ_I), the coefficient sensitivity ($|\Delta\%|$), and the decomposition coefficients (θ, λ); see Table 5 for variable definitions and the full specification. Each cell is a single regression, with ***, **, and * denoting significance at the 1%, 5%, and 10% levels. Regressions include Year-quarter fixed effects and standard errors are clustered at the firm level.

Panel A: GPT-4o-mini

		Coefficient on construct		Lexical Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
				Severity				
NI/Assets	18,868	−0.0012 ∗ ∗∗	−0.0012 ∗ ∗∗	−0.0000	−0.0086	0.3%	−0.0021	−0.0012 ∗ ∗∗
CapEx/Assets	18,866	−0.0006 ∗ ∗∗	−0.0006 ∗ ∗∗	+0.0027	+0.0109 ∗ ∗	0.2%	−0.0009	−0.0006 ∗ ∗∗
R&D ratio	18,868	−0.0006∗	−0.0005	+0.0474 ∗ ∗∗	+0.0021	17.6%	−0.0252 ∗ ∗∗	−0.0005
Leverage	18,868	+0.0268 ∗ ∗∗	+0.0265 ∗ ∗∗	−0.1387	−0.1219	1.0%	+0.0868	+0.0265 ∗ ∗∗
Sales growth	18,864	−0.0514 ∗ ∗∗	−0.0511 ∗ ∗∗	+0.1377∗	+0.2277 ∗ ∗∗	0.4%	−0.0997 ∗ ∗	−0.0511 ∗ ∗∗
WW index	18,868	+0.0003	+0.0004	+0.0455∗	−0.0041	41.9%	−0.0240∗	+0.0004
				Boilerplate				
NI/Assets	18,868	−0.0000	+0.0000	+0.0021	−0.0096	234.4%	−0.0022	+0.0000
CapEx/Assets	18,866	+0.0001	+0.0001	+0.0038	+0.0104 ∗ ∗	24.0%	−0.0013	+0.0001
R&D ratio	18,868	−0.0008∗	−0.0004	+0.0477 ∗ ∗∗	+0.0020	48.7%	−0.0296 ∗ ∗∗	−0.0004
Leverage	18,868	+0.0160∗	+0.0148∗	−0.1626	−0.1111	7.2%	+0.1044	+0.0148∗
Sales growth	18,864	−0.0198 ∗ ∗∗	−0.0184 ∗ ∗∗	+0.1993 ∗ ∗	+0.2003 ∗ ∗∗	6.7%	−0.1222 ∗ ∗	−0.0184 ∗ ∗∗
WW index	18,868	−0.0017	−0.0014	+0.0427∗	−0.0029	19.6%	−0.0280∗	−0.0014
				Complexity				
NI/Assets	18,868	−0.0016 ∗ ∗∗	−0.0017 ∗ ∗∗	+0.0016	−0.0108∗	4.3%	+0.0094	−0.0017 ∗ ∗∗
CapEx/Assets	18,866	+0.0007 ∗ ∗	+0.0008 ∗ ∗	+0.0040	+0.0111 ∗ ∗	12.6%	−0.0138 ∗ ∗	+0.0008 ∗ ∗
R&D ratio	18,868	+0.0027 ∗ ∗∗	+0.0029 ∗ ∗∗	+0.0492 ∗ ∗∗	+0.0039	7.8%	−0.0303 ∗ ∗	+0.0029 ∗ ∗∗
Leverage	18,868	−0.0026	−0.0041	−0.1865	−0.1047	55.1%	+0.2217	−0.0041
Sales growth	18,864	−0.0235 ∗ ∗∗	−0.0214 ∗ ∗∗	+0.2209 ∗ ∗∗	+0.1727 ∗ ∗∗	8.8%	−0.3431 ∗ ∗∗	−0.0214 ∗ ∗∗
WW index	18,868	+0.0050 ∗ ∗∗	+0.0051 ∗ ∗∗	+0.0465∗	−0.0000	3.5%	−0.0221	+0.0051 ∗ ∗∗
				Specificity				
NI/Assets	18,868	−0.0001	−0.0002	+0.0025	−0.0099	88.5%	+0.0031	−0.0002
CapEx/Assets	18,866	+0.0000	+0.0001	+0.0035	+0.0106 ∗ ∗	1675.4%	−0.0023	+0.0001
R&D ratio	18,868	+0.0017 ∗ ∗∗	+0.0015 ∗ ∗∗	+0.0448 ∗ ∗∗	+0.0048	11.1%	+0.0085 ∗ ∗	+0.0015 ∗ ∗∗
Leverage	18,868	+0.0027	+0.0029	−0.1920	−0.0959	8.0%	−0.0050	+0.0029
Sales growth	18,864	+0.0018	+0.0022	+0.2224 ∗ ∗∗	+0.1930 ∗ ∗∗	22.0%	−0.0124	+0.0022
WW index	18,868	+0.0020 ∗ ∗	+0.0018∗	+0.0408∗	−0.0002	10.5%	+0.0094	+0.0018∗

Panel B: Llama-3.3-70B

		Coefficient on construct		Lexical Wedges		Sensitivity	Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
				Severity				
NI/Assets	18,868	$-0.0020**$	$-0.0020**$	$+0.0117*$	-0.0123	1.2%	$-0.0145**$	$-0.0020**$
CapEx/Assets	18,866	$-0.0005**$	$-0.0005**$	$+0.0025$	$+0.0061$	0.1%	-0.0008	$-0.0005**$
R&D ratio	18,868	$-0.0011**$	$-0.0010**$	$+0.0406**$	-0.0161	6.2%	$-0.0370**$	$-0.0010**$
Leverage	18,868	$+0.0261**$	$+0.0261**$	-0.0652	-0.0389	0.3%	$+0.0664$	$+0.0261**$
Sales growth	18,864	$-0.0623**$	$-0.0623**$	$+0.0689$	$+0.2412**$	0.0%	-0.0511	$-0.0623**$
WW index	18,868	-0.0007	-0.0007	$+0.0216$	-0.0161	5.7%	-0.0219	-0.0007
				Boilerplate				
NI/Assets	18,868	$+0.0006**$	$+0.0006**$	$+0.0138*$	$-0.0134*$	1.0%	$+0.0125$	$+0.0006**$
CapEx/Assets	18,866	-0.0000	-0.0000	$+0.0030$	$+0.0058$	17.5%	-0.0129	-0.0000
R&D ratio	18,868	-0.0006	-0.0006	$+0.0416**$	$-0.0172*$	0.5%	-0.0064	-0.0006
Leverage	18,868	$+0.0111$	$+0.0111$	-0.0915	-0.0139	0.5%	$+0.1164$	$+0.0111$
Sales growth	18,864	$-0.0116**$	$-0.0114**$	$+0.1326*$	$+0.1895**$	2.1%	$-0.4692**$	$-0.0114**$
WW index	18,868	-0.0008	-0.0009	$+0.0223$	-0.0171	0.7%	$+0.0102$	-0.0009
				Complexity				
NI/Assets	18,868	$-0.0015**$	$-0.0017**$	$+0.0161**$	$-0.0206**$	12.7%	$+0.0034*$	$-0.0017**$
CapEx/Assets	18,866	$+0.0003$	$+0.0004$	$+0.0024$	$+0.0074$	21.2%	-0.0015	$+0.0004$
R&D ratio	18,868	$+0.0031**$	$+0.0031**$	$+0.0375**$	-0.0045	1.2%	$+0.0041$	$+0.0031**$
Leverage	18,868	-0.0082	-0.0087	-0.0802	-0.0548	6.5%	$+0.0053$	-0.0087
Sales growth	18,864	-0.0048	-0.0031	$+0.1373*$	$+0.1831**$	36.2%	$-0.0494**$	-0.0031
WW index	18,868	$+0.0010$	$+0.0009$	$+0.0211$	-0.0131	12.3%	$+0.0041$	$+0.0009$
				Specificity				
NI/Assets	18,868	-0.0004	$-0.0006**$	$+0.0151**$	$-0.0175**$	56.2%	$+0.0023*$	$-0.0006**$
CapEx/Assets	18,866	$+0.0000$	$+0.0001$	$+0.0028$	$+0.0063$	491.8%	-0.0009	$+0.0001$
R&D ratio	18,868	$+0.0012**$	$+0.0011**$	$+0.0393**$	-0.0103	11.1%	$+0.0029*$	$+0.0011**$
Leverage	18,868	-0.0041	-0.0047	-0.0819	-0.0481	14.5%	$+0.0031$	-0.0047
Sales growth	18,864	$+0.0025$	$+0.0053*$	$+0.1218$	$+0.2274**$	115.9%	$-0.0321**$	$+0.0053*$
WW index	18,868	$+0.0016**$	$+0.0015*$	$+0.0190$	-0.0074	6.0%	$+0.0028$	$+0.0015*$

Table IA.3: **Generated Regressor Diagnostics by LLM: Jha et al. (2024) Replication**

This table reports the per-LLM regressions that Table 6 summarizes by pooling across the LLMs and the four constructs. Here the LLM-derived regressor is the direct-prompt predictor of Jha et al. (2024) rather than a text construct. Each panel is one LLM; within a panel the four blocks are the constructs (severity, boilerplate, complexity, and specificity) and the rows are the six forward firm outcomes. Columns report the naive and joint construct coefficients (β_{naive}, β_{joint}), the omission and injection wedge loadings (ρ_O, ρ_I), the coefficient sensitivity ($|\Delta\%|$), and the decomposition coefficients (θ, λ); see Table 6 for variable definitions and the full specification. Each cell is a single regression, with ***, **, and * denoting significance at the 1%, 5%, and 10% levels. Regressions include Year-quarter fixed effects and standard errors are clustered at the firm level.

Panel A: GPT-4o-mini

			Coefficient on predictor		Lexical Wedges		Sensitivity	Decomposition	
Predictor	Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
capex	CapEx/Assets	13,665	$+0.5160^{**}$	$+0.5330^{**}$	$+1.4311^{**}$	$+0.0122$	3.3%	-0.4262	$+0.5330^{**}$
rd	R&D/Total Capital	13,678	$+0.0459$	$+0.0620$	$+0.7178^{*}$	-0.4168^{**}	34.9%	-0.3122	$+0.0620$
ocf	OpCF Ratio	13,672	$+0.0759$	$+0.1052$	$+4.1560^{**}$	-0.4374	38.5%	-3.2191^{*}	$+0.1052$
sale	Sales Growth	13,665	$+17.3514^{**}$	$+17.3246^{**}$	$+25.0817^{*}$	-1.0089	0.2%	$+37.5672$	$+17.3246^{**}$
lev	Leverage	13,678	$+2.2686^{**}$	$+2.3993^{**}$	-9.8387^{**}	$+2.4923^{**}$	5.8%	-5.2456^{**}	$+2.3993^{**}$
div	Div. Quarterly	13,638	$+0.2102^{**}$	$+0.2255^{**}$	$+0.6093^{*}$	$+0.0749$	7.3%	-0.1763	$+0.2255^{**}$
cash	Cash Ratio	13,678	$+4.3740^{**}$	$+4.7153^{**}$	$+10.8208^{**}$	-1.8807	7.8%	-22.9156^{**}	$+4.7153^{**}$

Panel B: Llama-3.3-70B

			Coefficient on predictor		Lexical Wedges		Sensitivity	Decomposition	
Predictor	Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	$\lvert\Delta\%\rvert$	θ	λ
capex	CapEx/Assets	13,665	$+0.4438^{**}$	$+0.4492^{**}$	$+0.6118^{**}$	$+0.2268$	1.2%	-0.3916	$+0.4492^{**}$
rd	R&D/Total Capital	13,678	$+0.0547$	$+0.0741$	$+0.4349$	-0.2166	35.5%	-0.1568	$+0.0741$
ocf	OpCF Ratio	13,672	$+0.1460$	$+0.1725$	$+0.7546$	$+0.0169$	18.1%	-0.2802	$+0.1725$
sale	Sales Growth	13,665	$+16.0620^{**}$	$+16.1090^{**}$	$+15.3169$	-0.0283	0.3%	$+6.8806$	$+16.1090^{**}$
lev	Leverage	13,678	$+4.3547^{**}$	$+4.4168^{**}$	-5.5263^{*}	$+2.1018$	1.4%	-8.6686	$+4.4168^{**}$
div	Div. Quarterly	13,638	$+0.2720^{**}$	$+0.2880^{**}$	$+0.4384^{*}$	$+0.0628$	5.9%	-0.4838	$+0.2880^{**}$
cash	Cash Ratio	13,678	$+2.0668^{**}$	$+2.0601^{**}$	-0.4937	-0.6358	0.3%	$+2.6304$	$+2.0601^{**}$

Table IA.4: **Generated Regressor Diagnostics by LLM with Bag-of-Words Wedges: Item 1A Risk Disclosures**

This table reports the per-LLM regressions that Panel A of Table A.2 summarizes by pooling across the LLMs and the four constructs. The joint specification augments the omission and injection wedges with the bag-of-words wedges ρ_{BO}, ρ_{BI}, as in Table A.2. Each panel is one LLM; within a panel the four blocks are the constructs (severity, boilerplate, complexity, and specificity) and the rows are the six forward firm outcomes. Columns report the naive and joint construct coefficients (β_{naive}, β_{joint}), the omission and injection wedge loadings (ρ_O, ρ_I), the coefficient sensitivity ($|\Delta\%|$), and the decomposition coefficients (θ, λ); see Table A.2 for variable definitions and the full specification. Each cell is a single regression, with ***, **, and * denoting significance at the 1%, 5%, and 10% levels. Regressions include Year fixed effects and standard errors are clustered at the firm level.

Panel A: GPT-4o-mini

		Construct coeff.		Lexical wedges		BOW Wedges			Decomposition	
Outcome	N	b_{naive}	b_{joint}	ρ_O	ρ_I	ρ_{BO}	ρ_{BI}	$\lvert\Delta\%\rvert$	θ	λ
Panel A: Severity										
NI / Assets	21,479	+0.006	+0.007*	−0.084***	−0.024**	+0.090***	−0.009	8.8%	−0.790***	+0.007
CapEx / Assets	21,479	+0.001	+0.001	+0.014**	+0.000	+0.002	−0.006	11.6%	+0.207**	+0.001
R&D ratio	21,479	+0.010***	+0.009***	+0.138***	+0.054***	−0.055***	+0.026***	9.3%	+1.400***	+0.009***
Leverage	20,350	−0.008	−0.009	+0.208***	+0.152***	−0.019	−0.093**	11.9%	+1.434*	−0.009
Sales growth	20,305	+0.010	+0.009	+0.187***	+0.075***	−0.101***	+0.019	11.8%	+1.805***	+0.009
WW index	14,080	+0.002	+0.002	+0.092***	+0.073***	+0.017	+0.017	11.1%	−0.340	+0.002
Panel B: Boilerplate										
NI / Assets	21,479	+0.004	+0.005	−0.083***	−0.024**	+0.090***	−0.009	12.6%	−1.150***	+0.005
CapEx / Assets	21,479	−0.001	−0.001	+0.014**	+0.000	+0.002	−0.006	19.2%	−0.362***	−0.001
R&D ratio	21,479	−0.004*	−0.004*	+0.138***	+0.054***	−0.055***	+0.026***	2.9%	−0.305	−0.004*
Leverage	20,350	+0.002	+0.003	+0.208***	+0.152***	−0.019	−0.093**	44.3%	−2.117*	+0.003
Sales growth	20,305	−0.007	−0.007	+0.188***	+0.074***	−0.101***	+0.019	0.9%	+0.149	−0.007
WW index	14,080	−0.015**	−0.014**	+0.092***	+0.073***	+0.016	+0.017	3.6%	−1.017***	−0.014**
Panel C: Complexity										
NI / Assets	21,479	+0.004	+0.005	−0.084***	−0.024**	+0.090***	−0.009	15.0%	−0.580***	+0.005
CapEx / Assets	21,479	+0.002	+0.002	+0.014**	+0.000	+0.002	−0.006	11.1%	+0.228***	+0.002
R&D ratio	21,479	+0.010***	+0.008***	+0.138***	+0.054***	−0.055***	+0.026***	17.1%	+1.439***	+0.008***
Leverage	20,350	−0.008	−0.012	+0.208***	+0.152***	−0.019	−0.093**	40.7%	+2.891***	−0.012
Sales growth	20,305	+0.008	+0.006	+0.187***	+0.074***	−0.101***	+0.019	27.5%	+1.808***	+0.006
WW index	14,080	−0.001	−0.002	+0.093***	+0.073***	+0.017	+0.017	98.9%	+2.100***	−0.002
Panel D: Specificity										
NI / Assets	21,479	−0.004	−0.005	−0.083***	−0.024**	+0.090***	−0.009	13.6%	+0.394***	−0.005
CapEx / Assets	21,479	−0.000	−0.001	+0.014**	+0.000	+0.002	−0.006	45.5%	+0.147***	−0.001
R&D ratio	21,479	+0.001	+0.001	+0.138***	+0.054***	−0.055***	+0.026***	31.2%	+0.233**	+0.001
Leverage	20,350	−0.001	−0.002	+0.208***	+0.152***	−0.019	−0.093**	222.4%	+0.822**	−0.002
Sales growth	20,305	+0.005	+0.005	+0.188***	+0.075***	−0.101***	+0.019	1.6%	+0.052	+0.005
WW index	14,080	+0.001	+0.000	+0.092***	+0.073***	+0.017	+0.017	64.8%	+0.451***	+0.000

Panel B: Grok-3-mini

		Construct coeff.		Lexical wedges		BOW Wedges			Decomposition	
Outcome	N	b_{naive}	b_{joint}	ρ_O	ρ_I	ρ_{om}	ρ_{inj}	$\lvert\Delta\%\rvert$	θ	λ
Panel A: Severity										
NI / Assets	21,492	$-0.017***$	$-0.014***$	$-0.099***$	$-0.042***$	$+0.078***$	$+0.004$	18.7%	$-0.037***$	$-0.014***$
CapEx / Assets	21,492	$+0.001$	-0.000	$+0.013**$	$+0.001$	$+0.002$	-0.003	100.5%	$+0.008***$	-0.000
R&D ratio	21,492	$+0.026***$	$+0.023***$	$+0.088***$	$+0.041***$	$-0.038***$	$+0.023**$	12.0%	$+0.046***$	$+0.023***$
Leverage	20,366	$+0.008$	-0.006	$+0.224***$	$+0.135***$	-0.023	$-0.122***$	175.6%	$+0.099***$	-0.006
Sales growth	20,321	$+0.022***$	$+0.013**$	$+0.212***$	$+0.103***$	$-0.078***$	$+0.015$	41.5%	$+0.079***$	$+0.013**$
WW index	14,090	$+0.012***$	$+0.009***$	$+0.055***$	$+0.078***$	$+0.031**$	-0.014	20.2%	$+0.027***$	$+0.009***$
Panel B: Boilerplate										
NI / Assets	21,492	$+0.013***$	$+0.007***$	$-0.112***$	$-0.034***$	$+0.081***$	$+0.003$	46.8%	$+0.070***$	$+0.007***$
CapEx / Assets	21,492	$-0.005***$	$-0.004***$	$+0.006$	$+0.003$	$+0.005$	-0.004	9.9%	$-0.009***$	$-0.004***$
R&D ratio	21,492	$-0.032***$	$-0.028***$	$+0.084***$	$+0.032***$	$-0.032**$	$+0.023**$	14.0%	$-0.075***$	$-0.028***$
Leverage	20,366	$+0.007$	$+0.023***$	$+0.250***$	$+0.132***$	-0.034	$-0.121***$	213.3%	$-0.143***$	$+0.023***$
Sales growth	20,321	$-0.033***$	$-0.022***$	$+0.200***$	$+0.100***$	$-0.072**$	$+0.015$	33.9%	$-0.140***$	$-0.022***$
WW index	14,090	$-0.019***$	$-0.016***$	$+0.047***$	$+0.075***$	$+0.036**$	-0.014	15.1%	$-0.054***$	$-0.016***$
Panel C: Complexity										
NI / Assets	21,492	$-0.013***$	$-0.011***$	$-0.094***$	$-0.046***$	$+0.078***$	$+0.005$	15.5%	$-0.020***$	$-0.011***$
CapEx / Assets	21,492	$+0.003***$	$+0.002*$	$+0.008$	$+0.004$	$+0.003$	-0.004	29.4%	$+0.005***$	$+0.002*$
R&D ratio	21,492	$+0.035***$	$+0.037***$	$+0.033***$	$+0.071***$	$-0.027**$	$+0.017*$	7.0%	$+0.027***$	$+0.037***$
Leverage	20,366	-0.009	$-0.031***$	$+0.292***$	$+0.101**$	-0.039	$-0.115***$	233.4%	$+0.060***$	$-0.031***$
Sales growth	20,321	$+0.017***$	$+0.008*$	$+0.213***$	$+0.104***$	$-0.080***$	$+0.015$	50.9%	$+0.045***$	$+0.008$
WW index	14,090	$+0.011***$	$+0.009***$	$+0.048***$	$+0.082***$	$+0.032**$	-0.015	16.1%	$+0.017***$	$+0.009***$
Panel D: Specificity										
NI / Assets	21,492	$-0.014***$	$-0.009***$	$-0.099***$	$-0.034***$	$+0.077***$	$+0.003$	34.9%	$-0.040***$	$-0.009***$
CapEx / Assets	21,492	$+0.001*$	$+0.000$	$+0.012**$	$+0.001$	$+0.003$	-0.003	61.0%	$+0.006***$	$+0.000$
R&D ratio	21,492	$+0.021***$	$+0.017***$	$+0.080***$	$+0.028***$	$-0.034***$	$+0.026**$	18.1%	$+0.043***$	$+0.017***$
Leverage	20,366	$+0.025***$	$+0.014**$	$+0.173***$	$+0.144***$	-0.006	$-0.123***$	43.7%	$+0.089***$	$+0.014**$
Sales growth	20,321	$+0.024***$	$+0.013***$	$+0.197***$	$+0.097***$	$-0.072**$	$+0.017$	43.2%	$+0.082***$	$+0.013***$
WW index	14,090	$+0.014***$	$+0.011***$	$+0.043**$	$+0.074***$	$+0.036***$	-0.013	21.0%	$+0.033***$	$+0.011***$

Panel C: Llama-3.3-70B

		Construct coeff.		Lexical wedges		BOW Wedges			Decomposition	
Outcome	N	b_{naive}	b_{joint}	ρ_O	ρ_I	ρ_{om}	ρ_{inj}	$\lvert\Delta\%\rvert$	θ	λ
Panel A: Severity										
NI / Assets	21,144	$-0.013***$	$-0.025***$	$+0.022$	$-0.049***$	$+0.099***$	$+0.012$	89.3%	$+0.010**$	$-0.025***$
CapEx / Assets	21,144	$+0.005***$	$+0.003$	$+0.019***$	$+0.001$	$+0.001$	-0.003	46.7%	$+0.009***$	$+0.003$
R&D ratio	21,144	$+0.034***$	$+0.039***$	$+0.016$	$+0.067***$	$-0.076***$	$-0.028***$	17.2%	$+0.022***$	$+0.039***$
Leverage	20,004	$+0.010$	-0.000	$+0.037$	$+0.033$	$+0.037$	$-0.198***$	100.7%	$+0.030*$	-0.000
Sales growth	19,959	$+0.026***$	$+0.035***$	$+0.025$	$+0.102***$	$-0.095***$	$+0.013$	33.5%	$+0.009$	$+0.035***$
WW index	13,793	$+0.013***$	$+0.016***$	-0.034	$+0.046***$	$+0.017$	$-0.033**$	27.8%	$+0.006$	$+0.016***$
Panel B: Boilerplate										
NI / Assets	21,144	$+0.006***$	$+0.006***$	$-0.051***$	$-0.033***$	$+0.103***$	$+0.010$	12.2%	$-0.035**$	$+0.006***$
CapEx / Assets	21,144	$-0.001**$	$-0.001**$	$+0.027***$	-0.002	$+0.000$	-0.003	4.2%	-0.005	$-0.001**$
R&D ratio	21,144	$-0.007***$	$-0.007***$	$+0.131***$	$+0.043***$	$-0.081***$	$-0.025**$	0.1%	-0.007	$-0.007***$
Leverage	20,004	$+0.009$	$+0.010$	$+0.035$	$+0.039$	$+0.043$	$-0.199***$	14.0%	-0.064	$+0.010$
Sales growth	19,959	$-0.013***$	$-0.013***$	$+0.127***$	$+0.077***$	$-0.103***$	$+0.017$	1.7%	-0.026	$-0.013***$
WW index	13,793	-0.002	-0.002	$+0.011$	$+0.036***$	$+0.017$	$-0.031**$	28.2%	$-0.050***$	-0.002
Panel C: Complexity										
NI / Assets	21,144	$-0.008***$	$-0.016***$	$+0.020$	$-0.049***$	$+0.093***$	$+0.014$	89.2%	$+0.003$	$-0.016***$
CapEx / Assets	21,144	$+0.004***$	$+0.002$	$+0.018*$	$+0.001$	$+0.002$	-0.003	46.3%	$+0.006***$	$+0.002$
R&D ratio	21,144	$+0.034***$	$+0.045***$	$-0.069***$	$+0.082***$	$-0.058***$	$-0.034***$	31.2%	$+0.019***$	$+0.045***$
Leverage	20,004	$-0.014*$	$-0.036***$	$+0.196***$	$+0.006$	$+0.022$	$-0.190***$	155.0%	$+0.018$	$-0.036***$
Sales growth	19,959	$+0.015***$	$+0.017**$	$+0.048$	$+0.098***$	$-0.088***$	$+0.012$	16.9%	$+0.011$	$+0.017**$
WW index	13,793	$+0.005$	$+0.006$	-0.014	$+0.042***$	$+0.020$	$-0.032**$	19.1%	$+0.004$	$+0.006$
Panel D: Specificity										
NI / Assets	21,144	$-0.016***$	$-0.015***$	$-0.028*$	$-0.033***$	$+0.096***$	$+0.012$	6.8%	$-0.038***$	$-0.015***$
CapEx / Assets	21,144	$+0.002$	$+0.001$	$+0.025***$	-0.001	$+0.001$	-0.003	46.9%	$+0.017***$	$+0.001$
R&D ratio	21,144	$+0.021***$	$+0.017***$	$+0.104***$	$+0.043***$	$-0.072***$	$-0.026**$	17.5%	$+0.095***$	$+0.017***$
Leverage	20,004	$+0.008$	$+0.006$	$+0.028$	$+0.031$	$+0.039$	$-0.199***$	24.9%	$+0.052$	$+0.006$
Sales growth	19,959	$+0.034***$	$+0.031***$	$+0.079**$	$+0.077***$	$-0.088***$	$+0.013$	9.9%	$+0.106***$	$+0.031***$
WW index	13,793	$+0.011***$	$+0.010***$	-0.002	$+0.035***$	$+0.019$	$-0.032**$	7.0%	$+0.031**$	$+0.010***$

Panel D: Phi-4

		Construct coeff.		Lexical wedges		BOW Wedges			Decomposition	
Outcome	N	b_{naive}	b_{joint}	ρ_O	ρ_I	ρ_{om}	ρ_{inj}	$\lvert\Delta\%\rvert$	θ	λ
Panel A: Severity										
NI / Assets	21,252	$-0.019**$	$-0.024***$	-0.020	$-0.050***$	$+0.088***$	$+0.022**$	25.0%	$+0.021*$	$-0.024***$
CapEx / Assets	21,252	$+0.001$	-0.002	$+0.026***$	-0.005	$+0.000$	$+0.005$	270.1%	$+0.021***$	-0.002
R&D ratio	21,252	$+0.022***$	$+0.019***$	$+0.092***$	$+0.041***$	$-0.073***$	-0.005	12.6%	$+0.045***$	$+0.019***$
Leverage	20,115	-0.006	-0.016	$+0.133**$	$+0.140***$	-0.020	$-0.157***$	170.9%	$+0.078*$	-0.016
Sales growth	20,070	$+0.016**$	$+0.013$	$+0.130***$	$+0.129***$	$-0.081***$	$+0.009$	19.2%	$+0.041$	$+0.013$
WW index	13,882	$+0.007$	$+0.006$	$+0.005$	$+0.026$	$+0.010$	-0.026	13.5%	$+0.015$	$+0.006$
Panel B: Boilerplate										
NI / Assets	21,252	$-0.007***$	$-0.006**$	$-0.039***$	$-0.046***$	$+0.084***$	$+0.023**$	18.5%	$-0.039***$	$-0.006**$
CapEx / Assets	21,252	$-0.006***$	$-0.007***$	$+0.031***$	-0.008	-0.003	$+0.006$	14.6%	$+0.016***$	$-0.007***$
R&D ratio	21,252	$-0.016***$	$-0.020***$	$+0.133***$	$+0.024**$	$-0.082***$	-0.004	23.0%	$+0.077***$	$-0.020***$
Leverage	20,115	$+0.028***$	$+0.028***$	$+0.086*$	$+0.161***$	-0.006	$-0.158***$	0.7%	$+0.033$	$+0.028***$
Sales growth	20,070	-0.003	-0.006	$+0.150***$	$+0.121***$	$-0.084***$	$+0.009$	79.0%	$+0.063**$	-0.006
WW index	13,882	$-0.008**$	$-0.008**$	$+0.020$	$+0.020$	$+0.006$	-0.026	4.0%	-0.000	$-0.008**$
Panel C: Complexity										
NI / Assets	21,252	$-0.018***$	$-0.027***$	$+0.007$	$-0.056***$	$+0.086***$	$+0.022**$	52.4%	$+0.008$	$-0.027***$
CapEx / Assets	21,252	$+0.003**$	-0.000	$+0.025***$	-0.005	$+0.000$	$+0.005$	109.6%	$+0.012***$	-0.000
R&D ratio	21,252	$+0.036***$	$+0.039***$	$+0.038***$	$+0.055***$	$-0.070***$	-0.005	6.6%	$+0.030***$	$+0.039***$
Leverage	20,115	-0.010	$-0.031**$	$+0.175***$	$+0.129***$	-0.022	$-0.157***$	198.1%	$+0.045*$	$-0.031**$
Sales growth	20,070	$+0.019**$	$+0.016*$	$+0.113***$	$+0.133***$	$-0.080***$	$+0.009$	18.2%	$+0.028*$	$+0.016*$
WW index	13,882	$+0.006$	$+0.005$	$+0.002$	$+0.027$	$+0.010$	-0.026	13.7%	$+0.008$	$+0.005$
Panel D: Specificity										
NI / Assets	21,252	$-0.025***$	$-0.024***$	$-0.029**$	$-0.036***$	$+0.085***$	$+0.020*$	2.4%	$-0.041**$	$-0.024***$
CapEx / Assets	21,252	$+0.003**$	$+0.002$	$+0.022***$	-0.005	$+0.000$	$+0.006$	28.8%	$+0.028***$	$+0.002$
R&D ratio	21,252	$+0.040***$	$+0.037***$	$+0.087***$	$+0.025**$	$-0.070***$	-0.002	8.3%	$+0.126***$	$+0.037***$
Leverage	20,115	$+0.018$	$+0.009$	$+0.110**$	$+0.142***$	-0.020	$-0.155***$	48.9%	$+0.262***$	$+0.009$
Sales growth	20,070	$+0.053***$	$+0.048***$	$+0.111***$	$+0.112***$	$-0.078***$	$+0.013$	9.5%	$+0.190***$	$+0.048***$
WW index	13,882	$+0.024***$	$+0.023***$	-0.001	$+0.019$	$+0.009$	-0.024	3.3%	$+0.049$	$+0.023***$

Table IA.5: **Generated Regressor Diagnostics by LLM with Bag-of-Words Wedges: Earnings Call Transcripts**

This table reports the per-LLM regressions that Panel B of Table A.2 summarizes by pooling across the LLMs and the four constructs. The joint specification augments the omission and injection wedges with the bag-of-words wedges ρ_{BO}, ρ_{BI}, as in Table A.2. Each panel is one LLM; within a panel the four blocks are the constructs (severity, boilerplate, complexity, and specificity) and the rows are the six forward firm outcomes. Columns report the naive and joint construct coefficients (β_{naive}, β_{joint}), the omission and injection wedge loadings (ρ_O, ρ_I), the coefficient sensitivity ($|\Delta\%|$), and the decomposition coefficients (θ, λ); see Table A.2 for variable definitions and the full specification. Each cell is a single regression, with ***, **, and * denoting significance at the 1%, 5%, and 10% levels. Regressions include Year-quarter fixed effects and standard errors are clustered at the firm level.

Panel A: GPT-4o-mini

		Construct coeff.		Lexical wedges		BOW Wedges			Decomposition	
Outcome	N	β_{naive}	β_{joint}	ρ_O	ρ_I	ρ_{BO}	ρ_{BI}	$\lvert\Delta\%\rvert$	θ	λ
					Severity					
NI/Assets	18,196	−0.0012 * **	−0.0012 * **	+0.0032	−0.0111*	−0.0181 * *	−0.0012	0.0%	−0.0011	−0.0012 * **
CapEx/Assets	18,194	−0.0006 * **	−0.0006 * **	+0.0025	+0.0112 * *	+0.0045	+0.0001	0.7%	−0.0009	−0.0006 * **
R&D ratio	18,196	−0.0007*	−0.0006*	+0.0405 * **	+0.0192 * *	+0.0475 * **	+0.0181 * **	3.7%	−0.0027	−0.0006*
Leverage	18,196	+0.0269 * **	+0.0270 * **	−0.0172	−0.2884	−0.7428 * **	−0.1371 * *	0.4%	+0.0182	+0.0270 * **
Sales growth	18,192	−0.0512 * **	−0.0515 * **	+0.1277	+0.3041 * **	+0.1573	+0.1026 * **	0.7%	−0.0227	−0.0515 * **
WW index	18,196	+0.0001	+0.0001	+0.0235	+0.0229	+0.1408 * **	+0.0222 * *	122.8%	−0.0057	+0.0001
					Boilerplate					
NI/Assets	18,196	−0.0000	−0.0000	+0.0050	−0.0127 * *	−0.0169*	−0.0023	10.1%	+0.0002	−0.0000
CapEx/Assets	18,194	+0.0001	+0.0001	+0.0036	+0.0103 * *	+0.0052	−0.0004	30.6%	−0.0018	+0.0001
R&D ratio	18,196	−0.0008*	−0.0002	+0.0412 * **	+0.0184*	+0.0481 * **	+0.0174 * **	72.9%	−0.0363 * **	−0.0002
Leverage	18,196	+0.0164*	+0.0137	−0.0376	−0.2578	−0.7696 * **	−0.1080	16.6%	+0.1865	+0.0137
Sales growth	18,192	−0.0196 * **	−0.0174 * **	+0.1802 * *	+0.2415 * **	+0.2097	+0.0490	10.9%	−0.1534 * **	−0.0174 * **
WW index	18,196	−0.0017	−0.0010	+0.0217	+0.0235	+0.1405 * **	+0.0221 * *	39.6%	−0.0430 * **	−0.0010
					Complexity					
NI/Assets	18,196	−0.0016 * **	−0.0016 * **	+0.0042	−0.0132 * *	−0.0146*	−0.0016	1.0%	−0.0024	−0.0016 * **
CapEx/Assets	18,194	+0.0007 * *	+0.0008 * *	+0.0039	+0.0106 * *	+0.0040	−0.0008	11.6%	−0.0038	+0.0008 * *
R&D ratio	18,196	+0.0026 * **	+0.0025 * **	+0.0430 * **	+0.0190 * *	+0.0447 * **	+0.0163 * **	6.5%	+0.0118*	+0.0025 * **
Leverage	18,196	−0.0025	+0.0009	−0.0587	−0.2511	−0.7735 * **	−0.1113	135.4%	−0.1815*	+0.0009
Sales growth	18,192	−0.0231 * **	−0.0229 * **	+0.1944 * *	+0.2267 * **	+0.2454	+0.0634	0.8%	−0.0334	−0.0229 * **
WW index	18,196	+0.0047 * **	+0.0040 * **	+0.0257	+0.0242	+0.1351 * **	+0.0205*	14.6%	+0.0410 * **	+0.0040 * **
					Specificity					
NI/Assets	18,196	−0.0001	−0.0002	+0.0055	−0.0131 * *	−0.0168*	−0.0023	62.2%	+0.0020	−0.0002
CapEx/Assets	18,194	+0.0000	+0.0001	+0.0033	+0.0105 * *	+0.0051	−0.0004	757.4%	−0.0019	+0.0001
R&D ratio	18,196	+0.0017 * **	+0.0014 * **	+0.0385 * **	+0.0207 * *	+0.0478 * **	+0.0171 * **	20.2%	+0.0130 * **	+0.0014 * **
Leverage	18,196	+0.0023	+0.0039	−0.0679	−0.2446	−0.7735 * **	−0.1122	68.5%	−0.0497	+0.0039
Sales growth	18,192	+0.0034	+0.0032	+0.2006 * *	+0.2388 * **	+0.2122	+0.0518	7.3%	+0.0117	+0.0032
WW index	18,196	+0.0018*	+0.0013	+0.0203	+0.0254	+0.1403 * **	+0.0219 * *	26.9%	+0.0181 * *	+0.0013

Panel B: Llama-3.3-70B

Outcome	N	Construct coeff.		Lexical wedges		BOW Wedges			Decomposition	
		β_{naive}	β_{joint}	ρ_O	ρ_I	ρ_{BO}	ρ_{BI}	$\lvert\Delta\%\rvert$	θ	λ
					Severity					
NI/Assets	18,196	$-0.0020**$	$-0.0020**$	$+0.0147**$	$-0.0154*$	$-0.0207**$	-0.0011	0.3%	-0.0026	$-0.0020**$
CapEx/Assets	18,194	$-0.0005**$	$-0.0005**$	$+0.0027$	$+0.0065$	$+0.0018$	$+0.0008$	0.4%	-0.0003	$-0.0005**$
R&D ratio	18,196	$-0.0012**$	$-0.0011**$	$+0.0377**$	-0.0004	$+0.0447**$	$+0.0183**$	1.9%	-0.0033	$-0.0011**$
Leverage	18,196	$+0.0262**$	$+0.0258**$	-0.0062	-0.1551	$-0.7029**$	-0.1008	1.6%	$+0.0660$	$+0.0258**$
Sales growth	18,192	$-0.0624**$	$-0.0627**$	$+0.0706$	$+0.2841**$	$+0.0956$	$+0.0738**$	0.6%	-0.0288	$-0.0627**$
WW index	18,196	-0.0010	-0.0008	$+0.0102$	-0.0002	$+0.1409**$	$+0.0129$	15.6%	$-0.0163**$	-0.0008
					Boilerplate					
NI/Assets	18,196	$+0.0007**$	$+0.0006**$	$+0.0163**$	$-0.0176**$	$-0.0153*$	-0.0027	11.5%	$+0.0053**$	$+0.0006**$
CapEx/Assets	18,194	-0.0000	-0.0000	$+0.0031$	$+0.0058$	$+0.0025$	$+0.0003$	65.1%	-0.0011	-0.0000
R&D ratio	18,196	-0.0006	-0.0004	$+0.0388**$	-0.0023	$+0.0453**$	$+0.0171**$	40.5%	$-0.0161**$	-0.0004
Leverage	18,196	$+0.0123$	$+0.0092$	-0.0310	-0.1101	$-0.7139**$	-0.0727	25.1%	$+0.2039**$	$+0.0092$
Sales growth	18,192	$-0.0117**$	$-0.0108**$	$+0.1282*$	$+0.1857**$	$+0.1638$	$+0.0098$	7.6%	$-0.0671**$	$-0.0108**$
WW index	18,196	-0.0008	-0.0003	$+0.0110$	-0.0017	$+0.1414**$	$+0.0120$	68.6%	$-0.0371**$	-0.0003
					Complexity					
NI/Assets	18,196	$-0.0015**$	$-0.0017**$	$+0.0185**$	$-0.0242**$	$-0.0168*$	-0.0023	11.5%	$+0.0029$	$-0.0017**$
CapEx/Assets	18,194	$+0.0003$	$+0.0004$	$+0.0026$	$+0.0072$	$+0.0024$	$+0.0002$	19.9%	-0.0013	$+0.0004$
R&D ratio	18,196	$+0.0031**$	$+0.0029**$	$+0.0350**$	$+0.0087$	$+0.0456**$	$+0.0162**$	4.7%	$+0.0067**$	$+0.0029**$
Leverage	18,196	-0.0083	-0.0079	-0.0190	-0.1477	$-0.7443**$	-0.0732	4.5%	-0.0177	-0.0079
Sales growth	18,192	-0.0042	-0.0029	$+0.1294*$	$+0.1855**$	$+0.2038$	$+0.0151$	31.5%	$-0.0381**$	-0.0029
WW index	18,196	$+0.0010$	$+0.0007$	$+0.0101$	$+0.0010$	$+0.1421**$	$+0.0118$	30.6%	$+0.0085$	$+0.0007$
					Specificity					
NI/Assets	18,196	-0.0004	$-0.0006**$	$+0.0177**$	$-0.0216**$	$-0.0171*$	-0.0025	49.2%	$+0.0019$	$-0.0006**$
CapEx/Assets	18,194	$+0.0000$	$+0.0001$	$+0.0029$	$+0.0063$	$+0.0025$	$+0.0003$	2079.3%	-0.0008	$+0.0001$
R&D ratio	18,196	$+0.0012**$	$+0.0009**$	$+0.0368**$	$+0.0032$	$+0.0463**$	$+0.0166**$	25.0%	$+0.0046**$	$+0.0009**$
Leverage	18,196	-0.0041	-0.0034	-0.0216	-0.1380	$-0.7456**$	-0.0738	16.6%	-0.0117	-0.0034
Sales growth	18,192	$+0.0029$	$+0.0053*$	$+0.1145$	$+0.2254**$	$+0.2005$	$+0.0103$	79.2%	$-0.0231*$	$+0.0053*$
WW index	18,196	$+0.0016*$	$+0.0012$	$+0.0083$	$+0.0055$	$+0.1418**$	$+0.0112$	21.8%	$+0.0055$	$+0.0012$