Better AI, Worse Disclosures?

The Unintended Consequences of NLP on Financial Reporting

Pietro Ramella*

University of Chicago Booth

Preliminary draft

October 27, 2025

This paper studies how the growing use of algorithms to read and produce corporate disclosure can ultimately degrade disclosure quality. I develop a communication model in which an investor with limited information-processing capabilities relies on an algorithm to parse a manager's report. Two opposing forces emerge. Higher algorithmic accuracy enables the investor to extract information more precisely, but it also increases the incentive for reverse-engineering the algorithm, prompting managers to tailor their language strategically. The interaction of these forces yields a unimodal relationship: communication improves with accuracy up to a threshold, beyond which excessive tailoring reduces informativeness. Using a panel of U.S. conference-call transcripts (2007-2024), I examine the market response to textual surprise. The empirical evidence aligns with my model's prediction: with the advent of Large Language Models and Generative AI informativeness of corporate disclosure declined. My results highlight a previously overlooked feedback loop: when algorithms are employed by both investors and managers, "better" technology can ultimately harm disclosure quality.

*I thank Li Azinovic-Yang, Matthias Breuer, Anna Costello, Alex Frankel, Christian Leuz, Bradford Levy, Charles McClure, Michael Minnis, Delphine Samuels, Haresh Sapra, Douglas Skinner, Chris Stewart, Rimmy Tomy, Oktay Urcan, Luigi Zingales, the LBS Transatlantic Doctoral Conference participants and the lab class participants at Chicago Booth for their helpful comments and suggestions. All errors are my own. This is a preliminary draft, so please do not circulate.

1 Introduction

Over the past decade, algorithms have migrated from hedge-fund back offices to mainstream asset management, sell-side research, and even retail trading platforms (Bartram, Branke, and Motahari (2020), BlackRock (2024), and Reuters (2024))[1]. Machine learning models now outperform humans in stock return prediction and portfolio selection (Cao, Jiang, Wang, et al. (2024), Rossi and Utkus (2020), and Van Binsbergen, Han, and Lopez-Lira (2023)); sentiment classification algorithms have achieved human-level accuracy and are employed to generate profitable strategies (Ke, Kelly, and Xiu (2019) and Loughran and McDonald (2020)); and advanced detection models have become leading tools for uncovering financial fraud (Bao et al. (2020) and Bertomeu et al. (2021)). As a result, algorithmic trading is now estimated to accounts for more than 60% of U.S. equity volume.[2]

A large empirical literature documents this algorithmic supremacy, with the implicit premise of monotonicity: as algorithms become more accurate, capital markets should become better informed and more efficient. This view, however, overlooks an important equilibrium consequence. As investors increasingly rely on algorithms to interpret corporate disclosures, managers have growing incentives to tailor their language and exploit algorithmic predictability. In practice, this adaptation, or *tailoring*, involves choosing words and phrases to take advantage of known patterns in algorithmic classification. Tailoring means making lexical choices to influence an algorithm's output. For example:

"Smartphone revenue for the quarter was \$77.5 million on shipments of 330,000 units, a decline with respect to our November quarter revenue of \$171.0 million and shipments of 556,000 units."

A simple word-dictionary algorithm would classify this sentence as negative solely due to "decline". A minor rewrite, however, can shift the classification without altering the facts:

"Smartphone revenue for the quarter was \$77.5 million on shipments of 330,000 units versus smartphone revenue of \$171.0 million and shipments of 556,000 units in our November quarter."

[1] According to an AI talent report by Evident, 40% of hiring within the banking industry from October 2022 to March 2023 were for AI-related job functions.

[2] Goldman Sachs Global Marco Research

Here, replacing "decline" with "versus" renders the sentence neutral under the same algorithm, even though the underlying message is unchanged.

More sophisticated models such as ChatGPT or Claude are not as easily misled by a single word substitution. Yet these tools can themselves be used strategically. Rather than simply avoiding negative words, managers can use LLMs to reframe disclosures more positively. The NIRI 2023 Fall Report documents investor relations departments prompting models with requests such as:

"What specific recommended language changes would you have for the neutral sentiment expressed in pages 2 through 5 in order to make it more positive? Identify the original neutral language and then your corresponding recommendation."

Applied to the earlier example, prompting ChatGPT yields:

"Smartphone revenue for the quarter reached $77.5 million on shipments of 330,000 units, following a strong November quarter with $171.0 million in revenue and 556,000 units shipped."

Factually equivalent, this version would likely be classified as positive even by a Large Language Model (Can Turetken and Leippold (2024)). Thus, the same tools that improve the consumption of disclosure also enable its strategic production.

This example illustrates the mechanism at the heart of this paper: algorithmic advances boost investors' ability to process complex text but also create new incentives for manipulation. The very models designed to enhance precision become targets of strategic tailoring. Whether disclosure informativeness rises or falls with better technology depends on which force dominates: efficiency gains from improved processing, or losses from increased managerial obfuscation. This ambiguity motivates the central question of the paper: does technological progress ultimately strengthen or weaken the informational role of corporate disclosure? And more specifically, has the advent of Large Language Models (LLMs) and generative AI (GenAI) improved or worsen corporate communication?

To formalize this tradeoff, I develop a simple model of disclosure under limited investor attention. In the model, a manager communicates with an investor who relies on an algorithm to extract information from the report. The model captures a unique disclosure environment in which managers can write statements that are factually true yet strategically phrased to mislead an algorithm, though not a human reader. This requires bridging two strands of theory: disclosure models, where messages must be partially truthful, and cheap-talk models, where messages are unconstrained.

In the model, managers decide whether to invest in becoming tech-savvy, or write a tech-naive report. The tech-savvy manager buys her way into cheap talk communication, while the tech-naive managers are limited to truthful communication. Since managers that tailor to algorithm are mimicking good news, when investor use a more accurate algorithm, negative news are more likely to be truthful, and therefore entails lower prices. The contrary is not true for positive news. In fact, since better technology should make good report less likely to be misinterpreted, it widens the gap between the price assigned to good and bad news, making tailoring more appealing. In equilibrium, greater algorithmic accuracy improves the investor's ability to process text, but it also makes the algorithm itself a target of strategic tailoring. Informativeness therefore follows a non-linear path: it initially rises with algorithmic accuracy, but beyond a threshold, excess tailoring by managers erodes the gains from more precise classification.

This mechanism is particularly salient in the setting of accounting and narrative disclosure, where language is inherently malleable. Recent evidence shows that managers adapt to machine readers by embedding machine-readable tags (e.g., XBRL), avoiding accounting treatments that trigger fraud alerts, and rephrasing negative content to misdirect sentiment models (Allee, DeAngelis, and Moon Jr (2018), Cao, Jiang, Yang, et al. (2023), and Cao, Liang, and Moon (2023)). Modern disclosure is written not only for human investors, but increasingly for algorithms, making it a natural laboratory for studying the consequences of algorithmic advancements.

The model yields a sharp prediction. Informativeness should decline following the introduction of large language models (LLMs), as their human-level sophistication amplifies incentives for strategic reframing. By contrast, earlier technologies such as dictionary-based sentiment models should exhibit the opposite pattern: informativeness increases since the technology is less influential in price formation, and therefore the incentive to tailor for the technology is not as strong. Therefore, the model predicts different outcomes depending on the level of technological sophistication, providing both a primary test (LLMs) and a natural validation exercise (dictionary-based technology).

I examine this prediction in the setting of earnings conference calls, which include both a narrative, prepared disclosure and real-time Q&A portion. My primary focus is the release of GPT-3-class models in November 2022, which represents a salient, plausibly exogenous shock to the accuracy of natural language processing (NLP). Technological progress in this domain is cumulative and directional: successive innovations tend to deliver performance improvements.[3] I therefore treat the diffusion of LLMs as a technological shock that materially improves the accuracy of machine reading.

A first empirical challenge is measuring textual disclosure informativeness. To do that, I introduce the absolute *textual response coefficient* (TRC), which quantifies the sensitivity of absolute returns to textual surprise, in the spirit of the earnings response coefficient (ERC). Because textual expectations are not directly observable, I construct a measure of standardized textual surprise (SUT) as the component of textual dissimilarity between a firm's current call and its prior four calls, adjusted for length and contemporaneous events such as management guidance, 8K issuance and macro news. The idea mirrors the familiar ERC setting: just as earnings surprise captures unexpected performance and the earnings response coefficient (ERC) measures how markets react to it, textual surprise captures unexpected disclosure content and the textual response coefficient (TRC) measures the market's reaction. In this way, I can directly trace how the informativeness of disclosures changes around the ChatGPT shock.

My model prediction relies on the assumption of managerial tailoring for the algorithm. I validate this channel by constructing a *tailorability index* measuring the extent to which disclosures are phrased to evade the prevailing state-of-the-art algorithm. For Large Language Models (LLMs), the index leverages their limitations in numerical and comparative reasoning (Can Turetken and Leippold (2024)), calculated as the fraction of sentences containing neutral comparative statements (e.g., "compared to," "relative to") but lacking explicit directional negatives (e.g., "decline," "decrease"). Importantly, I exploit the structure of earnings calls: prepared speeches, written in advance, are more suitable for tailoring, while the Q&A segment, driven by analysts, provides a natural control group.[4] This within firm-year design enables separation of strategic adaptation from baseline linguistic drift. I find that tailorability increases in prepared speeches

[3] Frankel, Jennings, and Lee (2022) find that machine-learning approaches outperform dictionary-based routines in sentiment detection.

[4] Bushee, Gow, and Taylor (2018) provide a similar argument for managerial obfuscation.

after the release of ChatGPT-3.

Having validated the tailoring channel, I test the model's core prediction. I find that following the release of ChatGPT, the TRC declines. The simultaneous increase in tailorability and decline in textual informativeness suggest that the success of LLMs in parsing text creates new incentives for obfuscation, and that improved technology can reduce the informativeness of corporate communication.

I provide two sets of robustness tests. First, since tailoring involves an iterated process of writing and assessing the algorithmic evaluation to find the best version to communicate the same information, its effect should be concentrated in prepared remarks, that is drafted days in advance, rather than the Q&A, which is analyst-driven. I exploit the internal structure of conference calls, and re-estimate the TRC separately for these two components. I compare changes in market response to presenter speech before and after the release of ChatGPT-3 and use a "control" the market response to the Q&A section.[5] Consistent with the first test, I find that textual informativeness decreases for the presenter speech relative to the Q&A portion of the call after ChatGPT-3 is released.

I provide a second identification strategy to confer validity to the ChatGPT results by leveraging the functional form[6] of informativeness arising from my model. The theory predicts opposite effects at different levels of algorithmic accuracy: when accuracy is low, new tools should increase disclosure informativeness. To capture this setting, I turn to the earlier and less sophisticated Loughran and McDonald (2011) (LM) dictionary. My TRC test shows that informativeness rises after the release of the LM algorithm, reflecting that the tool enhanced investors' ability to parse tone without being sophisticated enough to trigger extensive managerial tailoring.

My theory predicts that the uni-modal pattern of informativeness is to be found exclusively in positive news. The informativeness of negative news is always increasing in algorithmic accuracy. I split conference calls based on the sign of the surprise. I find that after the introduction of ChatGPT the TRC increases for transcripts of calls following negative earnings surprise, but decreases for transcripts of calls following positive surprises.

[5] A key caveat is that the Stable Unit Treatment Value Assumption (SUTVA) likely do not hold here. The market reaction to the prepared remarks and the Q&A is jointly determined, so I cannot cleanly separate their effects.

[6] The functional form of a given relation can work as identification strategy, since omitted variables or endogeneity are unlikely to produce the observed shape (Samuels, Taylor, and Verrecchia (2021))

Taken together, these findings suggest that technological innovation, while improving investors' ability to process information, simultaneously strengthens managerial incentives to tailor disclosures. The net effect of these countervailing forces implies that advances in NLP can inadvertently reduce disclosure informativeness.

This paper contributes to several strands of literature. First, it extends the literature on obfuscation in financial reporting Li (2008) to a modern setting in which algorithms rather than humans increasingly interpret disclosures. I show that while earlier work emphasized managerial attempts to obscure meaning for human readers, today's obfuscation strategies reflect an adaptation to the inferential logic of NLP tools. This complements recent work on artificial intelligence and disclosure (Blankespoor, deHaan, and Li (2024) and Cao, Jiang, Yang, et al. (2023)) by emphasizing how algorithm design reshapes the strategic environment in which managers operate.

Second, it contributes to the disclosure theory by introducing a simple model in which informativeness displays a non-monotonic (unimodal) relationship with technological advancement, not due to noise or attention, but due to changes in processing cost that the receiver faces. This offers a novel rationale relative to prior work of Fang, Huang, and Wang (2017) and Samuels, Taylor, and Verrecchia (2021) that show uni-modal pattern in informativeness. My model shares the result of Frankel and Kartik (2019), that manipulation is more pronounced when payoffs for positive messages are higher. While they have an exogenous stakes parameter that determines the payoffs for positive messages, my model has processing noise determine the delta between the payoff for positive and negative news, and therefore determining the incentives to manipulate.

Third, I build on the classical ERC framework, and propose new measures to proxy for informativeness of the textual disclosure. I propose a novel methodology to measure the textual surprise (SUT) and estimate the absolute textual response coefficient (TRC), which may be useful in other applications.

Finally, the paper speaks to the literature on earnings conference calls (Allee and DeAngelis (2015), Frankel, Johnson, and Skinner (1999), Matsumoto, Pronk, and Roelofsen (2011), and Tasker (1997) by quantifying the relative informativeness of prepared remarks and Q&A segments and tracing how these dynamics evolve over time. The remainder of the paper is organized as follows. Section 2 presents the theoretical model and develops testable predictions. Section 3 describes the data and the construction of empirical measures. Section 4 presents the main results. Section 5 discusses extensions and robustness. Section 6 concludes.

2 Hypothesis Development

To formalize this tradeoff, I develop a simple model of disclosure in which investors cannot perfectly process all the information they receive. Instead, they rely on an algorithm that extracts signals from managers' reports. The algorithm is valuable because it allows investors to process large volumes of disclosure at low cost, but it is not flawless: it sometimes misclassifies reports, with the frequency of mistakes depending on the accuracy of the technology. In the model, managers also have access to the technology and can exploit it by tailoring their reports to the algorithm. In practice, this means rewriting a report until it conveys the same underlying facts but is interpreted more favorably by the algorithm, even if a human reader would not see any difference.[7] This implies that a manager who has invested in learning the technology, can keep rewriting until she finds a version the algorithm classifies favorably.[8] Appendix B contains the full model derivation.

2.1 The Model

My model builds on the disclosure framework of Rappoport (2020) and connects several strands of the information economics literature. First, following Krishna and Morgan (2004) and Blume, Board, and Kawamura (2007), it shows how stochastic message reception can affect communication. Second, it extends the literature on disclosure under agency frictions (Verrecchia 2001) by introducing algorithmic parsing as a new channel through which managers transmit and potentially manipulate information. Third, similar to Frankel and Kartik (2019), Samuels, Taylor, and Verrecchia (2021), and Fang, Huang, and Wang (2017), the model links informational stakes to manipulation incentives: as algorithms become more precise, the payoff to tailoring increases, amplifying strategic distortion.

[7] For example, "Sampo Bank's market share of lending was 13.6%, down from 14.4% in Q1 2008" is factually equivalent to "Compared with 14.4% in Q1 2008, Sampo Bank's current lending market share is 13.6%." Yet, a financial-specific generative AI algorithm, FinGPT, classifies the first as negative and the second as neutral.

[8] The model assumes that the manager can perfectly anticipate sentiment. In practice, manager and investor may not share the same technology, or the algorithm may have a stochastic element in its classification. In that case, a more natural assumption is that tailoring increases the probability of a positive signal above that of a non-tailored report, but not to certainty. This modification does not change the qualitative results.

2.1.1 Setup

A manager (M) communicates with an investor (I) who relies on an algorithm to process a corporate report. Nature draws three independent variables: the firm's fundamental news $i \in \{1, -1\}$ with equal probability, the algorithm's processing outcome $j \in \{1, -1\}$ with prior $\Pr(j = 1) = 1 - \kappa$ and $\Pr(j = -1) = \kappa$, and a private tailoring cost $c \sim U[0, 2]$.[9] The parameter κ captures the algorithm's error rate: higher κ corresponds to lower accuracy.

After observing (i, c), the manager decides whether to invest in learning how the algorithm works at cost c. If she invests, she can craft a report that guarantees the interpretation $m = 1$. If she does not invest, she reports truthfully, submitting i to the investor. In this case, the algorithm processes the report with outcome j, and the investor observes $m = i \times j$. When $j = -1$, which occurs with probability κ, the truthful report is misclassified.

Not all managers are equally willing or able to engage in this kind of tailoring. I capture this heterogeneity through the cost of tailoring c, which is assumed to be uniformly distributed between zero and 2.[10] For some managers the cost may be small, reflecting an ability to easily learn and deploy the technology. For others it may be high, either because their news is so negative that "positification" is infeasible, or because of reputational and legal risks.[11]

Managers will choose the costly tailoring action only if it yields a higher expected payoff than reporting truthfully. A tailored report is designed with knowledge of the algorithm, so its outcome is anticipated and guaranteed to be classified as good news. A truthful report, by contrast, is subject to the algorithm's processing step: the accuracy of the technology determines the likelihood that it is classified correctly or incorrectly. Among managers facing different costs but observing the same news, there will be a pivotal manager who is indifferent between tailoring and not tailoring. Any manager with the same news but a lower cost will tailor, while any manager with a higher cost will not. This creates an equilibrium with distinct pools for each type of news. While it is intuitive that managers with bad news will want to tailor, the choice is not limited to them.

[9] The maximum cost is assumed high enough to deter managers with bad news and the highest costs from tailoring.

[10] The maximum cost is assumed large enough that a manager facing the maximum cost and negative news is never willing to tailor. Moreover, uniformity is chosen for simplicity; the exact distribution affects the shape of the cutoff but not the existence of cutoffs or the qualitative results.

[11] The cost can also represent differences in awareness of the tailoring option itself, with managers ordered by their likelihood of being aware, and awareness depending on how profitable tailoring is. For example, an intermediary may sell the tailoring service, and its market share will be proportional to the benefit of tailoring.

Managers with good news may also tailor, since tailoring guarantees a positive classification while truth-telling leaves a risk of being misclassified as bad. As expected, however, the fraction of managers willing to tailor is always higher among those observing negative news than among those observing positive news.

The investor observes m and sets a price p. The manager seeks to maximize $U^M(p) = p$, while the investor minimizes quadratic loss $U^I(p,i) = -(p-i)^2$. Because the tailoring cost c is private information, investors form conjectures about how many managers with good or bad news will choose to tailor, and they set prices for good and bad reports accordingly.

2.1.2 Equilibrium

The price of a good report must reflect that it can be submitted by four types of managers: good-news managers who tailor, good-news managers who do not tailor and are correctly classified as positive, bad-news managers who tailor, and bad-news managers who do not tailor but are misclassified as positive. By contrast, a negative report can only come from managers who do not tailor. This pool[12] is made up of truthful bad-news managers correctly classified by the algorithm and truthful good-news managers misclassified as bad.

In equilibrium, investors' conjectures are confirmed: managers' choices validate their expectations. The fixed point problem delivers a unique equilibrium with two prices, one for the good report and one for the bad report, and with two threshold costs that determine which managers tailor. There exists a unique perfect Bayesian equilibrium in pure strategies characterized by threshold costs $c_1^*(\kappa)$ for good-news managers and $c_{-1}^*(\kappa)$ for bad-news managers, and prices $p^*(1)$ for positive reports and $p^*(-1)$ for negative reports. One cutoff applies to bad-news managers, below which they prefer to tailor and appear as good. The other (smaller) cutoff applies to good-news managers, below which they also find it worthwhile to tailor: $c_{-1}^*(\kappa) > c_1^*(\kappa)$.

The equilibrium masses of each type are: good-news managers who tailor (mass $\frac{1}{2} \cdot \frac{c_1^*}{2}$), good-news managers who report truthfully and are correctly classified (mass $\frac{1}{2} \cdot (1 - \frac{c_1^*}{2}) \cdot (1-\kappa)$), bad-news managers who tailor (mass $\frac{1}{2} \cdot \frac{c_{-1}^*}{2}$), bad-news managers who report truthfully but are misclassified as positive (mass $\frac{1}{2} \cdot (1 - \frac{c_{-1}^*}{2}) \cdot \kappa$), bad-news managers who report truthfully and are

[12] This is not a fully separating equilibrium, since both types can appear in the good-report pool, but rather a cutoff equilibrium with distinct pools defined by tailoring choices.

correctly classified (mass $\frac{1}{2} \cdot (1 - \frac{c^*_{-1}}{2}) \cdot (1 - \kappa)$), and good-news managers who report truthfully but are misclassified as negative (mass $\frac{1}{2} \cdot (1 - \frac{c^*_1}{2}) \cdot \kappa$).

Equilibrium prices satisfy rational expectations: given the tailoring thresholds, investors price each report based on its average quality using Bayes' rule, and given these prices, managers optimally choose whether to tailor by comparing the expected benefit to the cost c. The resulting fixed-point system yields unique values for c^*_1, c^*_{-1}, $p^*(1)$, and $p^*(-1)$.

2.1.3 Comparative Statics

Once optimal prices and tailoring behavior are defined, I can calculate the amount of information lost due to two forces: skepticism induced by tailoring and misclassification caused by the algorithm's inaccuracy. I measure information loss as the expected squared difference between the true value of the news and the price assigned to the report, weighted by the joint probability of the state and the report.[13] For good reports, this measure captures two effects: (i) bad-news managers who truthfully report but are misclassified as good by the algorithm, and (ii) bad-news managers who tailor and are priced as good reports. For bad reports, the measure captures good-news managers who truthfully report but are misclassified as bad by the algorithm. In other words, information loss measures how far prices deviate from the truth, on average. Informativeness is the expected payoff to the investor, which equals the negative of information loss.

Tailoring aside, when the technology improves the algorithm becomes more accurate, and reports are less likely to be misclassified. Clearer signals widen the gap between the price of a good report and the price of a bad report, which in turn makes tailoring more profitable. Importantly, the effect of accuracy on prices is not symmetric across good and bad reports.

Because bad reports are sent only by managers who do not tailor, their price reflects a mix of two types: bad-news managers who report truthfully and are correctly classified, and good-news managers who report truthfully but are misclassified as bad. As accuracy improves, misclassification becomes less frequent. The pool of non-tailoring managers therefore shifts toward a larger share of true bad news and a smaller share of misclassified good news. The price for negative reports, $p^*(-1)$, increases monotonically with accuracy (i.e., decreases monotonically in κ). Since no tailoring occurs in the negative pool, this compositional improvement translates directly into

[13] Squared loss is chosen for tractability; any convex measure of distance would generate the same qualitative results.

higher informativeness for negative reports.

The dynamics are different, however, for good reports. Good reports are sent by all managers who tailor, as well as by truthful managers whose good (bad) news is correctly (incorrectly) classified by the algorithm. As the technology becomes more accurate, the composition of truthful reporters shifts in the same way as for bad reports: the pool of good reports contains a higher fraction of true positives and a lower fraction of misclassified negatives.

The effect of accuracy on tailoring, however, changes this dynamic. When the technology is not very accurate, the signal extracted by the algorithm does not move prices much, so tailoring is not very profitable. In this region, a marginal increase in accuracy raises the price of good reports, just as it improves the informativeness of bad reports. By contrast, when the technology is already very accurate, the algorithm's signal plays a central role in pricing. Further improvements then encourage more tailoring, which offsets the gains from better classification of truthful managers. In this region, higher accuracy can actually lower the price of good reports.

Taken together, these forces generate a hump-shaped relation between the price of a good report and the accuracy of the algorithm. The price for good reports, $p^*(1)$, exhibits an inverted-U relationship with accuracy: it increases in accuracy (decreases in κ) at low accuracy levels, and decreases in accuracy (increases in κ) at high accuracy levels. Overall informativeness inherits this hump-shaped pattern, first rising then falling with accuracy.

2.2 Empirical Predictions

The model's comparative statics generate testable predictions about how disclosure informativeness responds to improvements in algorithmic technology. I organize these predictions around two mechanisms: tailoring behavior and informativeness dynamics.

2.2.1 Validating the Tailoring Mechanism

The model's logic rests on the assumption that managers actively tailor their reports to exploit algorithmic weaknesses. Before testing the main predictions, I validate this channel empirically:

Prediction 2.2.1 (H0)**.** Following the release of an algorithm with sentiment detection capabilities, textual disclosure will shift toward language that exploits the algorithm's weaknesses to obtain more positive inferences.

2.2.2 Main Predictions: Inverted-U Informativeness

Having established that tailoring occurs, I turn to the model's central implications for disclosure informativeness. The model predicts that report informativeness follows a hump-shaped path as algorithm accuracy increases. At low accuracy levels, better algorithms improve informativeness by reducing misclassification. At high accuracy levels, better algorithms reduce informativeness by intensifying strategic tailoring. Since I focus on the introduction of large language models (LLMs), a high-accuracy technology, my first main prediction concerns the downward-sloping portion of the inverted-U:

Prediction 2.2.2 (H1)**.** The informativeness of textual corporate communication decreased following the release of LLM technology.

The same theoretical result suggests the opposite pattern for less accurate technologies. When sentiment dictionaries such as the Loughran and McDonald (2011) lexicon were first introduced, algorithmic accuracy was relatively low, placing firms on the upward-sloping portion of the inverted-U. This motivates my second prediction:

Prediction 2.2.3 (H2)**.** The informativeness of textual corporate communication increased following the release of the Loughran and McDonald (2011) sentiment dictionary.

2.2.3 News-Type Asymmetry

The model's comparative statics also reveal an asymmetry by news type. The informativeness of negative reports always increases with accuracy, while the informativeness of positive reports can decrease at high accuracy levels. This asymmetry stems from the fact that negative reports are sent only by non-tailoring managers, whose pool becomes purer as accuracy rises. Positive reports, by contrast, attract tailoring managers, whose presence degrades informativeness when accuracy is high. This leads to my third prediction:

> **Prediction 2.2.4** (H3)**.** The informativeness of textual corporate communication decreased for positive news and increased for negative news following the release of LLM technology.

Finally, because the model builds on the assumption that investors face processing constraints, the predicted dynamics should be most pronounced in settings where algorithmic parsing is valuable, namely, text-rich disclosures that are costly to process manually. Having developed the theoretical foundation, I now turn to the empirical analysis to test whether the data exhibit these predicted patterns.

3 Data

3.1 Sample

I start with the universe of earnings conference call transcripts from Capital IQ, totaling 231,687 unique transcripts for firms listed on the NYSE and NASDAQ. For every conference call, I collect the latest version of the transcript.[14] Conference calls structure is typically a Prepared (or Presenter) Speech led by executives, followed by a Q&A where security analysts get to ask questions. Capital IQ tags each conference portion (presenter speech, question, answer, or operator) and speaker (exec, analysts, operator). I drop calls without both a tagged Presenter Speech and a Q&A section. I require for every company to host at least four conference calls in the previous year and a half. This drops the sample to 187,557.

I collect from I/B/E/S "street" measures of earnings and analysts estimates. I compute analysts' expectations as the median of latest individual analysts forecasts issued within the 90 days prior to the earnings announcement date (EAD). If management issues guidance within such window, I restrict the median estimates to only forecasts between the latest guidance and the EAD. Following Payne and Thomas 2003, I use unadjusted values, and apply CRSP adjustment factor to put both the forecast and the actual values on comparable per-share basis, and avoid look-ahead bias[15] while four digits precision in computing the median estimate[16]. I find I/B/E/S data from

[14] Note that the transcript creation date and time do not coincide with the event date and time. I use creation timestamps to retrieve the most recent transcripts, which should coincide with the most accurate version.

[15] Stock splits can be problematic when comparing EPS at different dates (i.e. at the announcement date vs the estimate date). Adjusted values in I/B/E/S use recent adjustment factor.

[16] Pre-compute median estimates in I/B/E/S summary stats are rounded at decimal level.

187,557 unique transcripts.

Following Anilowski, Feng, and Skinner (2007), I also collect management guidance that precedes the earnings announcement and conference call. I/B/E/S stores guidance for fourteen measures (capex, dividend-per-share, EBITDA, EBITDA-per-share, funds from operations-per-share, fully reported earnings-per-share, gross margin, net income, operating profits, pre-tax income, ROA, ROE, and sales). Guidance can be a point estimate or range guidance, and differ on forecast period (quarter or yearly). For every event, I compute total guidance, fraction of range guidance, whether EPS guidance is provided, what fraction of guidance is EPS (or sales) guidance, issued between the event date and the previous 90 days. I set the guidance metrics to zero if I can find a company in I/B/E/S but there is no guidance data in the Guidance Detail History database, and to missing otherwise. The dataset size drops to 174,288 observations with I/B/E/S EPS information, but a smaller fraction (55%) has also guidance information.

Next, I construct Cumulative Abnormal returns using CRSP daily stock data. The daily abnormal return is the difference between daily stock return and the CRSP market return over the same period. I compute Cumulative Abnormal Return metrics, $CAR_{EA}(-1,1)$, as the sum of three daily abnormal returns centered at the earnings announcement date; and $CAR_{CC}(-1,1)$ centered at the conference call date.[17] I center each window on the first trading day: that is, if the event is after hour, $t = 0$ is the day after the event, while if the event is before or during hours, date $t = 0$ is the day of the event.

I collect standard ERC controls to improve its estimate. I estimate Market Beta, and compute Size, Market-to-Book, Leverage, and Earnings Persistence. The latter is the coefficient of the basic earnings per share excluding extraordinary items of its lagged value, on at least three years of data, and up to ten.

Lastly, I follow Gipper, Leuz, and Maffett (2020), and truncate the continuous financial variables in the response coefficient regressions at 1 and 99%. Since ERC estimation suffers from the skeweness of unexpected earnings, I drop observations with earnings surprises below -1 and greater than 1, and observations with price two days before EAD lower than 1.

Table 1 illustrates data coverage, and Table 2 the summary statistics for the main variables.

[17] I also compute $CAR_{event}(0,2)$ for robustness.

3.2 Textual Metrics

In my empirical analysis I use textual metrics to validated the tailoring channel, and to measure informativeness of the transcript of the earnings conference calls. I describe them in this subsection.

3.2.1 Tailorability Indexes

I construct a transcript-level proxy that capture lexical choices explicitly designed to exploit weaknesses in Large Language Models (LLMs).

LLM–based tailoring. Transformer models such as GPT-3.5 have been shown to have contextual "understanding" of language, but have been shown to be less adept to numerical reasoning (Can Turetken and Leippold (2024) and Leippold (2023)). In each sentence of the transcript, I look for (i) comparators terms (e.g., "relative to", "compared with") and (ii) directional negatives (e.g., "decrease", "decline"). I list comparators terms and directional negatives in Panel B of Table 3.

Denote by numerical_{it} the number of sentences containing a comparator *without* a directional negative, and by comparator_{it} the number of sentences containing any comparator. The *LLM–tailorability index* is

$$\text{Tailorability}_{it}^{\text{LLM}} = \frac{\text{numerical}_{it}}{\text{comparator}_{it}},$$

High values indicate that managers rewrite potentially negative comparisons in linguistically neutral terms that LLMs underweight.

Dictionary–based tailoring. Cao, Jiang, Yang, et al. (2023) show positification in annual and quarterly reports catered to Loughran and McDonald (2011) dictionary, which is an older, less accurate algorithm, that was commonly used in the finance literature before the introduction of LLMs. They show that financial negative words (listed in the dictionary) become more infrequent after the release of the dictionary, while negative words in general language do not show this pattern. I build an index that tries to capture substitutability between listed and unlisted words in earnings conference calls. I start from the list of 2,303 negative words according to the LM vocabulary, excluding the tokens question(s). I call this set of tokens $\mathcal{N}_{\text{LM}}$. For each word, I collect all possible synonyms from WordNet, a large lexical database of English developed by Princeton University. In order to restrict the list of synonyms to words that also have a negative connotation, I keep only synonyms deemed negative by the Harvard dictionary. This yields a vocabulary of 1,438 synonymy with negative connotation, that I called $\mathcal{S}_{\mathrm{LM}}$.

For each transcript I count

$$\text{flagged}_{it} = \#\{\text{tokens} \in \mathcal{N}_{\mathrm{LM}}\}, \qquad \text{unflagged}_{it} = \#\{\text{tokens} \in \mathcal{S}_{\mathrm{LM}} \setminus \mathcal{N}_{\mathrm{LM}}\}.$$

The *LM–tailorability index* is the share of negative language that evades the dictionary:

$$\text{Tailorability}_{it}^{\mathrm{LM}} = \frac{\text{unflagged}_{it}}{\text{flagged}_{it} + \text{unflagged}_{it}}.$$

Large values indicate that managers substitute canonical LM words with semantically equivalent but unindexed synonyms.

To both indexes I add an infinitesimal term, 10^{-9}, to the denominator, to avoids division by zero when no match is found. Each index is computed at the transcript level for for the entire call and for different segments (presenter speech, Q&A, and questions only).

3.2.2 Textual Surprise (SUT)

I propose to measure textual informativeness with an absolute Textual Response Coefficient (TRC), analogous to the Earnings Response Coefficient (ERC), which measures how strongly markets react to textual surprise. As with the ERC, the challenge lies in identifying information set to construct the surprise. I define the information set as comprising the transcript of the four prior earnings calls; the amount, characteristics and timing of Form 8-K filings, and any management guidance issued during the quarter leading up to the current earnings announcement; and macro variables for the month preceding the call.

To quantify the novelty of information disclosed during the current earnings call relative to recent disclosures, I construct a firm-specific measure that captures linguistic changes between the transcript of the current call and those of the preceding four calls. I then regress the purely the textual dissimilairty on characteristics of the disclosure process to get a measure of textual surprise.

The construction proceeds in two steps. First, to measure pure textual dissimilarity between the earnings call and firm's history of conference calls, I follow the approach of Brown and Tucker (2011). The authors compute changes in 10-K language relative to the prior year; I adapt it to the higher-frequency and dialogic nature of conference calls. For each transcript, I compute the cosine similarity between the term-frequency–inverse-document-frequency (TF-IDF) representation of the current call and a benchmark composed of the four most recent transcripts from the same firm. The TF-IDF representation counts words in a given document, but discount each term by the count in the entire corpus.[18] The TF-IDF vectorizer is trained on a 20,000-transcript subsample to ensure representative weighting of common terms in conference calls.

As in Brown and Tucker (2011), the dissimilarity score is 1 minus the similarity score. Because longer transcripts naturally tend to exhibit greater linguistic dissimilarity, I follow Brown and Tucker (2011) methodology and project each score onto a fifth-degree polynomial of transcript length and use the residual as the length-adjusted abnormality metric. This metric is set up to capture the novelty of the current transcript with respect to the prior four, by computing how much variation the transcript "adds" to the vector space of past communication.

Conference calls are not the only disclosure events. In order to avoid attributing information disclosed via different channel, or already known to the market, in the second step, I regress the length-adjusted dissimilarity metric (i.e. the residual from the previous step) on a rich set of controls. I include: the number and nature (whether it is EPS guidance) of guidance events, the proximity in days to most recent 8-K filings, the complexity (measured with the Flesch Kincaidand index) of preceding regulatory filings in the same quarter, and prevailing macroeconomic conditions (interest rates, inflation, and industrial production).

I then use the residual from this regression, pass it through the Min-Max transformation to get values in [0,1], and consider this final measure the textual surprise, denoted *SUT*. This measure is meant to capture variation in language that cannot be attributed to other observable information events. I also construct an indicator variables for above-median *SUT*, *HSUT*. This binary trans-

[18] Prior to comparison, each transcript is pre-processed by lowercasing, stripping numeric and non-alphabetic characters, and stemming via the Porter algorithm. I do not have to drop stopwords, as is common among simple word counts approaches, because the TF-IDF transformation will already weight them down to zero.

formation serves two purposes. First, it reflects the notion, central to the model, that algorithmic tools are especially valuable in settings with high processing costs, such as unusually complex or novel disclosures. Second, it mitigates the influence of outliers and measurement noise inherent in high-dimensional textual similarity metrics. The resulting indicator, *HSUT*, flags transcripts where textual news is unusually high relative to a firm's recent history.

3.2.3 TRC

I validate my measure with a classical abnormal return test. Since conference calls have been found to be relevant informational events (Frankel, Johnson, and Skinner (1999) and Matsumoto, Pronk, and Roelofsen (2011)), if the measure I construct capturs relevant information, I expect absolute cumulative abnormal returns (ACAR) around the earnings conference calls to be positively associated with it. I use absolute cumulative abnormal returns since my SUT measure cannot distinguish positive from negative news. An important caveat is that my measure is additive in reported news, irrespective of their sign. This can hinder the relation with ACAR. As a purely illustrative example, assume that call i reports three positive news, and a call j reports two negative and one positive news, the SUT could be very close in value for call i and j, but ACAR for i may be larger than j.

Since textual information is to be consumed jointly with the accounting numbers, I also interact the absolute earning surprise, with the textual surprise. Then, the *TRC* has two components, a pure response to textual surprise, and an interaction of the textual and earnings surprises. Columns (1)-(3) of Table 4 shows the results without interacting earnings and textual surprise, while Columns (4)-(6) reports the estimates of the model:

$$ACAR(-1,1) = \beta_0 + \beta_1 Abs\ UE + \beta_2 SUT + \beta_3 Abs\ UE \times SUT + \varepsilon \tag{3.1}$$

I find that higher textual news is associated with higher absolute returns, but the effect is partially reversed with large surprises.

Note that while the interaction term is negative, it doesn't reverse the positive relation between the magnitude of the SUT and CAR; even at the 95 percentile of $Abs\ UE$ value, 0.023, the implied absolute TRC is $0.009 > 0$. Similar to ERCs being larger for small earnings surprise, the TRC is non linear in *Abs UE*. Textual surprise is more positively correlated with market returns when earnings surprise is low, which can suggests substitutability between the two signals.

The coefficients are stable to the addition of classical ERC controls designed to improve the ERC estimate: earnings loss, earnings persistence, size, market-to-book, leverage, and market beta. I also add controls specific to the TRC that are meant to capture other information event that can make the information embedded in the call redundant or less surprising: amount and type of guidance, length of the transcript, and distance in days from the earnings announcement date, as well as industry and quarter-year fixed effects.

4 Results

I start by showing empirical evidence of the tailoring channel 2.2.1, then move to the main prediction 2.2.2.

4.1 Tailoring for the algorithm

When algorithms are used to evaluate disclosures, managers have incentive to understand how they work, so they can write with the algorithm in mind, i.e. tailoring to the algorithm. Sentiment-detection algorithms are among the most widely used NLP techniques in finance. Although they can identify tone with high levels of accuracy, they are not perfect. The high-dimensional nature of language means that two differently worded sentences, despite carrying the same information, can still receive different sentiment scores from the very same model.

To examine whether firms strategically tailor their disclosure to get more favorable algorithmic outcomes, I inspect words distribution before and after the release dates of two algorithms, the Loughran and McDonald (2011) word dictionary (LM) and ChatGPT. I leverage the structure of earnings conference calls to obtain a cleaner estimate of tailorability patterns. Tailoring is a complex exercise and, therefore, should be most evident in the prepared remarks rather than in the more spontaneous Q&A segment. Because analysts do not face the same reporting incentives as managers, I treat the word distribution in analysts' Q&A questions as a control group, while the presenters' prepared speech serves as the treated group.

4.1.1 Tailoring for LM

I compute the LM-tailorability measure describe Section 3.2.1, on conference calls transcripts for the two years preceding and following the publications of LM on JF (2011). I then compare the frequency of words that elude the word-dictionary sentiment detection algorithm in the presenter speech and the analysts' questions of the Q&A section, running the following regression:

$$Tailorability_{it} = \alpha_0 + \alpha_1 Post_t + \alpha_2 Section_{it} + \alpha_3 Post_t \times Section_{it} + \varepsilon_{it} \tag{4.1}$$

The dependent variable, *Tailorabilty*, is a transcript-section level measure of frequency of words that elude the algorithm. The average $LM - tailorability$ is 0.7 in the presenter speech, and 0.78 in the analysts' questions of the Q&A section.

Post is an indicator taking value 1 if the call is held after January 6th, 2011, the date of the publication of the LM paper on the Journal of Finance[19]. *Section* is also an indicator variable, taking value 1 is the measure pertains the presenter speech, and zero if comes from the questions of the analysts.

In my empirical analysis, the Q&A portion of the call acts as control, and the presenter speech as the treated group. I can therefore apply quarter-year and firm fixed effects. I report the results in Table 5. The coefficient of interest, $\hat{\alpha}_3$, is 0.027 and remains positive and statistically significant both with and without fixed effects. This estimate implies that following the publication of its word list in the Journal of Finance, the usage of their unlisted synonyms in the prepared speech has increased by roughly 4 percent relative to the Q&A. This is a first piece of evidence for strategic tailoring. The usage of the synonym that escape negative tone detection increases increases after the dictionary is made public, and the effect is concentrated among the presenter speech, which is drafted in advanced, rather than the interactive Q&A portion.

[19] I follow Cao, Jiang, Yang, et al. (2023) and use the publication date, rather than the first posting date, January 23rd, 2009, as the word-index didn't gained popularity until after its publication)

4.1.2 Tailoring for LLMs

LLMs are robust to simple synonyms swap, but have limited numerical reasoning capacity. I use the tailorability catered for LLMs limitations explained in Section 3.2.1. I look for sentences that leverage the numerical reasoning limitation of LLMs, by calculating how many times comparative terms, like "with respect to", "relative to", "versus" appear in the transcripts without any negative connotation words like "decrease", "loss", "decline", etc. (the full list of terms is in Panel B of Table 3)

The average *Tailorability* score is 0.82 for the presenter speech and 0.72 for analysts' questions. I re-estimate specification 4.1 using the release date of ChatGPT-3 (30 November 2022) as the cut-off and restrict the sample to the two-year window on either side of that date. Table 3 shows that, tailorability rises within Presenter Speeches relative to analysts' questions in the two years following the release of ChatGPT-3. Specifically, I find an additional 1 percent increase in the share of sentences worded in a way that large language models are more likely to misinterpret.

4.2 Loss of Informativeness with the advent of LLMs

My model predicts that, as parsing technology becomes highly accurate, incentives to tailor disclosure to please the algorithm are heightened, and net, overall informativeness of the message decreases. Table 3 supports the tailoring channel behind this prediction by showing that tailorability of disclosures increased following the release of ChatGPT-3 class models. To empirically test the prediction on overall informativeness, I compare the Textual Response Coefficient (TRC) of conference calls before and after the ChatGPT-3 release date.

Since the model builds on a rational inattention assumption, the predicted dynamics should be found among reports subject to higher processing costs, and therefore the effects it to be expected among textual-rich disclosures. For this reason, I re-encode SUT into high versus low textual surprise, and compute an indicator for *above median textual surprise, (HSUT)*.[20]

I inspect changes in the absolute response coefficient of the binary transformation to textual

[20] The $HSUT$, unsurprisingly, behaves like to SUT: it is positively correlated with abnormal cumulative returns, its interaction with the length of the transcript is positively associated with absolute CARs, and shows substitutability with other informational events like, number of 8K, total guidance, presence of EPS guidance.

response before and after the release of ChatGPT-3. I run the following specification

$$\begin{aligned} ACAR(-1,1) = & \alpha_0 + \alpha_1 HSUT + \alpha_2 Abs\ UE + \alpha_3 HSUT \times Abs\ UE + \alpha_4 Post + \\ & \alpha_5 HSUT \times Post + \alpha_6 Abs\ UE \times Post + \alpha_7 HSUT \times Abs\ UE \times Post + \\ & \Gamma X + \varepsilon \end{aligned} \tag{4.2}$$

The dependent variable is the absolute value of the cumulative abnormal returns on the window $[-1, 1]$ centered on the trading day the call is held on, or the first trading day that follows the call, if held after-hours. I use controls that are standard in the ERC literature to better estimate the coefficient: *Earnings Loss*, *Size*, *Earnings Persistence*, *Market-to-Book*, and *Market Beta*. These metrics are deeply related to the ERC construct (Collins and Kothari (1989)). Since earnings surprises are mostly positive, and the coefficient estimated on positive surprises is typically higher than the one estimated on negative surprises, thus *Earnings Loss* captures the non-linearity in the ERC regression. Second, *Size*, is a proxy for the informational environment, while *Market-to-Book* and *Market Beta* proxy growth opportunities. Finally, *Earnings Persistence* that captures the stickiness of earnings, which is associated with a higher ERC.

Additionally, I control for the the length of the transcript (*Length*), since longer transcripts may be harder to process. I add three controls that measure other informational events that can the make price already reflect information discussed in the call. I include the number of 8K issued in the quarter prior the earnings announcement, the number of guidance, and whether it was EPS guidance. Finally, I control for the timespan between the call and other informational events. Specifically, I control for the distance in days between the earnings announcement and the conference call[21], and the distance in days between the call and the last 8K issued that quarter.

My coefficients of interest are $\hat{\alpha}_5$, $\hat{\alpha}_7$. Table 6 report the estimated parameters. The coefficient $\hat{\alpha}_5$ is positive, but insignificant, while the coefficient for the interaction term in the post period is negative ($\hat{\alpha}_7 < 0$) and significant a 1%.[22] The estimate is robust to the inclusion of my two sets of controls and industry or firm fixed effects. The estimates are qualitatively and quantitatively

[21] Almost 80% of the observation of the calls in my sample are held the same day of the earnings announcement, and over 99% within one day. The results are robust to restrict to earnings conference call held the same day of the earnings announcement.

[22] I also substitute the indicator *Post* with an indicator for the years 2020, 2021, 2023, 2024. I find that the coefficient of the interaction term in increasingly negative in the two years following the release of ChatGPT-2 models.

robust if I substitute $Abs\ UE$ with the decile of the absolute earnings surprise, $DAbs\ UE$.

In the period of examination, between November 2020 and November 2024, the mean absolute cumulative abnormal return, $ACAR$, is 0.066, and the mean absolute earnings surprise, $Abs\ UE$, is 0.008. Then, my estimates suggest that in the two years following the release of ChatGPT-3 class models, for the mean absolute surprise, the market response to textual-rich disclosures decreases by 11 to 17 basis points, corresponding to a 1.74 to 2.62% decrease in absolute cumulative abnormal returns.

The pre-post analysis of textual informativeness in 6, jointly with the tailorability result of 3, suggests that the textual information content of conference call might have decreased after LLMs introduction.

5 Robustness

I provide two additional set of test. First, I leverage the structure of the conference call and compare response to presenter speech versus Q&A portion, with the idea that tailoring will be more concentrated into the prepared portion,[23] so that the Q&A acts a "control". The issue with this test is that the two portions of the call are jointly consumed, possibly violating the Stable Unit Treatment Value Assumption (SUTVA).

Second, use my theory to guide identification vie functional form(Samuels, Taylor, and Verrecchia (2021)). The model predicts that for a less accurate algorithm, the predictions are flipped. I inspect changes in *HSUT* before and after the publication of the Loughran and McDonald (2011) word dictionary (LM), an earlier algorithm to detect financial tone. The LM is a simpler technology, a word dictionary, and therefore much less accurate. Frankel, Jennings, and Lee (2022) and Leippold (2023) both find that LM is outperformed in the tone detection task by machine learning techniques in financial texts. I find empirical evidence supporting this second prediction.

Finally, the model predicts that information loss should only appear in among positive news. Negative news is instead more informative when the algorithm improves. I inspect TRC changes separately for positive and negative earnings surprises, and find again that my empirical analysis support my theory.

[23]Bushee, Gow, and Taylor (2018) make a similar argument for obfuscation.

5.1 Robustness: Presenter Speech vs Q&A

The nature of technological progress is that algorithmic performance will improve with each successive innovation. I therefore treat the ChatGPT release as as an exogenous shock to the information environment. One obstacle in testing whether informativeness has ultimately decrease, is that increasing incentives to tailor apply to all firms. It is therefore difficult to create a "clean" control group of firms that were unaffected by the new technological environment. However, because presenter speech is prepared in advanced, vetted by lawyers, and is more suited to be vetted for algorithmic evaluation, while analysts don't share the agency incentives of managers, they should have no incentives to tailor their questions to the algorithms.

Estimating response coefficients separately for the presenter speech and the Q&A resembles the work of Lipe (1986) on understanding the information contained in the components of earnings. One caveat is that the response to the two components of the textual surprise of the call are jointly determined, and therefore violate the Stable Unit Treatment Value Assumption. With this caveat in mind, I can use the Q&A as "control" group and presenter speech as "treated" group. Each call is then split into its two component: presenter speech and Q&A, and the *HSUT* is separately computed for the two components.

I inspect the TRC of the components of the call in Appendix **??**. I find that the TRC for presenter speech is lower than the TRC for Q&A when interacted with the decile of absolute earnings surprise $DAbs\ UE$. The relation is less strong if I use absolute earnings surprises. Matsumoto, Pronk, and Roelofsen (2011) finds that the Q&A is more informative than the presenter speech, therefore, when running TRC-components regressions, I use deciles of absolute earnings surprises $(DAbs\ UE)$, rather than absolute earnings surprise.

I run the following specification

$$\begin{aligned} ACAR(-1,1) = {} & \alpha_0 + \alpha_1 HSUT + \alpha_2 DAbs\ UE + \alpha_3 HSUT \times DAbs\ UE + \alpha_4 Section + \\ & \alpha_5 HSUT \times Section + \alpha_6 DAbs\ UE \times Section + \alpha_7 HSUT \times DAbs\ UE \times Section + \\ & \alpha_8 Post + \alpha_9 HSUT \times Post + \alpha_{10} DAbs\ UE \times Post + \alpha_{11} HSUT \times DAbs\ UE \times Post + \\ & \alpha_{12} Section \times Post + \alpha_{13} HSUT \times Section \times Post + \alpha_{14} DAbs\ UE \times Section \times Post + \end{aligned} \tag{5.1}$$

$$\alpha_{15} HSUT \times DAbs\ UE \times Section \times Post + \Gamma X + \varepsilon$$

The coefficients of interests are $\hat{\alpha}_{13}$ and $\hat{\alpha}_{15}$. Table 7 reports the results of equation 5.1. In Column (3) $\hat{\alpha}_{13} = 0.0077$, and $\hat{\alpha}_{15} = -0.0014$. Then market response to textual-rich earnings conference calls increases for low earnings surprises ($DAbs\ UE \leq 5$), and decreases for larger earnings surprises ($DAbs\ UE \geq 6$). The median $DAbs\ UE$ is 6, therefore the market response to the textual-rich presenter speech, decreases by 7 basis points, with respect to the market response to the textual-rich Q&A, that is a 1.06% decrease in absolute cumulative abnormal returns.

Decomposing textual news into the component attributed to the prepared speech and to the Q&A portion of the call can help understand whether the advent of LLM technology has improved or worsen corporate communication, if we assume that response to the two sections has a common trend in absence of the technological shock. For the Q&A to be a valid control group, I have to examine the validity of the parallel trends assumption. ChatGPT-3 is released in late November 2022, so I consider 2023 the first year of treatment, and 2022 the baseline. Figure 1, Panel A and Panel B show that TRC-differences by year. Panel A looks at the coefficient for $HSUT$ (base, α_{13}), and Panel B at the coefficient for $DAbs\ UE \times HSUT$ (interaction, α_{15}). Both base and interaction coefficients are insignificant in the pre-period, which is consistent with the parallel trend assumption. In the post, the coefficient for the interaction term becomes negative and significant, by 2024, two years after the release of ChatGPT-3. Since $DAbs\ UE$ takes integer values from 1 to 10, the market response to $HSUT$ decreases, for all levels of absolute earnings surprise. Thus, the market reaction to the textual surprise in the presenter speech decreases relative to the reaction to the textual surprise in the Q&A.

There is a caveat for this analysis. It could be that the response to the textual surprise of one section affects the response to the other. Such spillover could bias my analysis. For this reason, I provide a second set of tests that relies on the functional form of the model to identify the effect of the emergence of LLMs and Generative AI on corporate disclosure.

5.2 Robustness: LM dictionary

In my second battery of robustness tests, I rely on my theory for identification. Prediction 2.2.3 states that when the baseline technology has low accuracy, improvements in accuracy are associated with higher informativeness. I propose that the introduction of the Loughran and McDonald (2011) dictionary fits the assumption behind this prediction. This dictionary based-algorithm is noisy proxy for financial sentiment, but has been, and still is, commonly used in the finance literature (Cao, Jiang, Yang, et al. (2023)).

In subsection 4.1.1 I validated the tailoring channel for LM-algorithm, so I can move directly to the test of prediction 2.2.3. I replicate the design of equation 4.2, but center around the publication of the publication date of the Loughran and McDonald (2011) paper. Table 8 reports the results. The coefficient $Post\ LM = 1 \times Abs\ UE \times HSUT$ is positive, therefore the market response to textual-rich earnings conference calls increases after the LM dictionary is released. The mean absolute cumulative abnormal return for the period January 2009 to January 2013 is 0.0561, and the mean $Abs\ UE$ is 0.0060. Thus, the market response to textual-rich earnings conference calls increases absolute CAR by 15 basis points, a 2.61% increase with respect to the mean return.

5.3 Robustness: Informativeness increases for negative news and decreases for positive news

I present one last set of results that leverage the implications of my model. Prediction 2.2.4 states that the informativeness of negative messages always increases in algorithmic accuracy, while informativeness of positive messages decreases with high the algorithm accuracy. Recall that my SUT measure is agnostic about the news. To investigate my last prediction, I let the earnings surprise determine the sign of the news, run the following regression:

$$\begin{aligned} CAR(-1,1) = \alpha_0 + \alpha_1 SUT + \alpha_2 UE + \alpha_3 SUT \times UE + \alpha_4 Post + \\ \alpha_5 SUT \times Post + \alpha_6 UE \times Post + \alpha_7 SUT \times UE \times Post + \Gamma X + \varepsilon, \end{aligned} \tag{5.2}$$

where I split the sample into positive (or zero) and negative earnings surprises UE. Since the model dynamics rely on rational inattention, I restrict my attention to transcripts of above median length. Table 9 reports the results for both subsets of negative and positive earnings surprises.

Estimates in columns (1)-(3) indicates that after the introduction of ChatGPT, for the median negative surprise, -0.0098, an increase of 1% in SUT, decreases the cumulative abnormal returns by an additional 3 basis points, or 1.13% decrease in cumulative abnormal return with respect to the mean CAR of -0.0299. Consistently with my theory, the informativeness of the textual news disclosed in the negative message increases with algorithmic accuracy.

My estimates suggest that the opposite is true for the subsample of positive news. Estimates in columns (4)-(6) indicates that after the introduction of ChatGPT, for the median positive surprise, 0.0044, an increase of 1% in SUT, decreases the cumulative abnormal returns by 2 basis points, or 2.66% decrease in cumulative abnormal return with respect to the mean CAR of -0.0299. Again, consistent with the model, the informativeness of the textual news disclosed in the positive message decreases with algorithmic accuracy.

6 Conclusion

This paper studies how the growing reliance on algorithms to parse corporate disclosure can have unintended consequences for disclosure quality. I develop a simple model in which investors rely on algorithms to process managers' reports. While greater algorithmic accuracy improves the investor's ability to extract information, it also raises the payoff to reverse-engineering the algorithm, intensifying incentives for managers to tailor their language. The interaction of these forces generates a hump-shaped pattern in disclosure informativeness.

Using a large panel of U.S. earnings conference calls from 2007 to 2024, I document evidence consistent with this prediction. Following the release of ChatGPT-3, informativeness of textual disclosure declined as shown by a decrease of 1.74 to 2.62% in absolute cumulative abnormal returns for textual-rich disclosures. I corroborate this result, and show that the market response to the textual information embedded in the prepared speech as declined after the release of Large Language Models, when compared to the same trend for the market response to textual information embedded in the Q&A portion of the call. Moreover, I leverage the prediction of my model and find that consistent with my theory, after the publication of the Loughran–McDonald (2011) dictionary, disclosures became more informative, as reflected in a stronger market response to textual surprise. Finally, also show that the information loss of the textual disclosure is concentrated among positive news, and that as theory predicts, informativeness of negative news has improved.

The findings highlight a feedback loop in modern disclosure: as algorithms become more sophisticated, they enhance information processing but simultaneously strengthen the incentives for managers to game their classifications. The net effect is non-monotonic. In equilibrium, better algorithms do not always lead to better disclosures.

These results contribute to the literature on disclosure and obfuscation by showing that modern tailoring strategies are aimed not at human readers but at machine interpreters. They also extend the economics of disclosure theory, offering a novel mechanism through which technological progress can diminish informativeness. Finally, the paper introduces new empirical measures (tailorability indexes and the absolute textual response coefficient) that may be useful in future research on the interplay between algorithmic readers and corporate disclosure.

References

Allee, Kristian D and Matthew D DeAngelis (2015). "The structure of voluntary disclosure narratives: Evidence from tone dispersion". In: *Journal of Accounting Research* 53.2, pp. 241–274.

Allee, Kristian D, Matthew D DeAngelis, and James R Moon Jr (2018). "Disclosure "scriptability"". In: *Journal of Accounting Research* 56.2, pp. 363–430.

Anilowski, Carol, Mei Feng, and Douglas J Skinner (2007). "Does earnings guidance affect market returns? The nature and information content of aggregate earnings guidance". In: *Journal of accounting and Economics* 44.1-2, pp. 36–63.

Bao, Yang et al. (2020). "Detecting accounting fraud in publicly traded US firms using a machine learning approach". In: *Journal of Accounting Research* 58.1, pp. 199–235.

Bartram, Söhnke M, Jürgen Branke, and Mehrshad Motahari (2020). *Artificial intelligence in asset management*. CFA Institute Research Foundation.

Bertomeu, Jeremy et al. (2021). "Using machine learning to detect misstatements". In: *Review of Accounting Studies* 26.2, pp. 468–519.

BlackRock (2024). “How AI is transforming investing”. In: *Systematic Investing*. July 29, 2024. https://www.blackrock.com/us/individual/insights/ai-investing. BlackRock.

Blankespoor, Elizabeth, Ed deHaan, and Qianqian Li (2024). “Generative AI in Financial Reporting”. In: *Financial Reporting (October 12, 2024)*.

Blume, Andreas, Oliver J Board, and Kohei Kawamura (2007). “Noisy talk”. In: *Theoretical Economics* 2.4, pp. 395–440.

Brown, Stephen V and Jennifer Wu Tucker (2011). “Large-sample evidence on firms’ year-over-year MD&A modifications”. In: *Journal of Accounting Research* 49.2, pp. 309–346.

Bushee, Brian J, Ian D Gow, and Daniel J Taylor (2018). “Linguistic complexity in firm disclosures: Obfuscation or information?” In: *Journal of Accounting Research* 56.1, pp. 85–121.

Can Turetken, Aysun and Markus Leippold (2024). “Battle of Transformers: Adversarial Attacks on Financial Sentiment Models”. In: *Swiss Finance Institute Research Paper* 23-59.

Cao, Sean, Wei Jiang, Junbo Wang, et al. (2024). “From man vs. machine to man+ machine: The art and AI of stock analyses”. In: *Journal of Financial Economics* 160, p. 103910.

Cao, Sean, Wei Jiang, Baozhong Yang, et al. (2023). “How to talk when a machine is listening: Corporate disclosure in the age of AI”. In: *The Review of Financial Studies* 36.9, pp. 3603–3642.

Cao, Sean, Ying Liang, and Jason Youngseok Moon (2023). “Machine Readership and Financial Reporting Decisions”. In: *Available at SSRN 4650113*.

Collins, Daniel W and Sagar P Kothari (1989). “An analysis of intertemporal and cross-sectional determinants of earnings response coefficients”. In: *Journal of accounting and economics* 11.2-3, pp. 143–181.

Fang, Vivian W, Allen H Huang, and Wenyu Wang (2017). “Imperfect accounting and reporting bias”. In: *Journal of Accounting Research* 55.4, pp. 919–962.

Frankel, Alex and Navin Kartik (2019). “Muddled information”. In: *Journal of Political Economy* 127.4, pp. 1739–1776.

Frankel, Richard, Jared Jennings, and Joshua Lee (2022). “Disclosure sentiment: Machine learning vs. dictionary methods”. In: *Management Science* 68.7, pp. 5514–5532.

Frankel, Richard, Marilyn Johnson, and Douglas J Skinner (1999). “An empirical examination of conference calls as a voluntary disclosure medium”. In: *Journal of Accounting Research* 37.1, pp. 133–150.

Gipper, Brandon, Christian Leuz, and Mark Maffett (2020). “Public oversight and reporting credibility: Evidence from the PCAOB audit inspection regime”. In: *The Review of Financial Studies* 33.10, pp. 4532–4579.

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu (2019). *Predicting returns with text data*. Tech. rep. National Bureau of Economic Research.

Krishna, Vijay and John Morgan (2004). “The art of conversation: eliciting information from experts through multi-stage communication”. In: *Journal of Economic theory* 117.2, pp. 147–179.

Leippold, Markus (2023). “Sentiment spin: Attacking financial sentiment with GPT-3”. In: *Finance Research Letters* 55, p. 103957.

Li, Feng (2008). “Annual report readability, current earnings, and earnings persistence”. In: *Journal of Accounting and economics* 45.2-3, pp. 221–247.

Lipe, Robert C (1986). “The information contained in the components of earnings”. In: *Journal of Accounting Research*, pp. 37–64.

Loughran, Tim and Bill McDonald (2011). “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks”. In: *The Journal of finance* 66.1, pp. 35–65.

— (2020). “Textual analysis in finance”. In: *Annual Review of Financial Economics* 12.1, pp. 357–375.

Matsumoto, D, M Pronk, and E Roelofsen (2011). “What makes conference calls useful”. In: *The information content of managers’ presentations and*.

Payne, Jeff L and Wayne B Thomas (2003). “The implications of using stock-split adjusted I/B/E/S data in empirical research”. In: *The Accounting Review* 78.4, pp. 1049–1067.

Rappoport, Daniel (2020). “Evidence and skepticism in verifiable disclosure games”. In: *Available at SSRN 2978288*.

Reuters (2024). “Trading app Robinhood agrees deal with Pluto Capital to tap into AI”. In: *Reuters Technology*. July 1, 2024. https://www.reuters.com/technology/artificial-intelligence/trading-app-robinhood-agrees-deal-with-pluto-capital-tap-into-ai-2024-07-01/. Reuters.

Rossi, Alberto G and Stephen P Utkus (2020). “Who benefits from robo-advising? Evidence from machine learning”. In: *Evidence from Machine Learning (March 10, 2020)*.

Samuels, Delphine, Daniel J Taylor, and Robert E Verrecchia (2021). “The economics of misreporting and the role of public scrutiny”. In: *Journal of Accounting and Economics* 71.1, p. 101340.

Tasker, Sarah C (1997). “Voluntary disclosure as a response to low accounting quality: evidence from quarterly conference call usage”. In: *Available at SSRN 2949*.

Van Binsbergen, Jules H, Xiao Han, and Alejandro Lopez-Lira (2023). “Man versus machine learning: The term structure of earnings expectations and conditional biases”. In: *The Review of financial studies* 36.6, pp. 2361–2396.

Verrecchia, Robert E (2001). “Essays on disclosure”. In: *Journal of accounting and economics* 32.1-3, pp. 97–180.

Tables

Table 1: **Sample Construction and Coverage**
This table presents the sample selection procedure for the full sample of conference call transcripts obtained from Capital IQ. The sample is constructed using the universe of earnings call transcripts from the Capital IQ between January 1, 2004 to December 31, 2024. In order to construct the main measure, *SUT*, I require data regarding informational events, like management guidance type and amount, characteristics of previous disclosure, and macro variable. I collect such information from I/B/E/S, CRSP, WRDS SEC Analytics and FRED St. Louis. In the last step of the sample construction, I perform truncation at 1-99% of relevant continuous variables and 2.5-97.5% of earning surprise, as is common practice for ERC studies.

Data Source / Filter	**Observations**	**Coverage**
Capital IQ transcripts	231,687	100%
with Brown–Tucker measures	187,557	81%
with Presenter Speech and Q&A tagged correctly:	187,557	81%
with key I/B/E/S metrics	174,288	75%
with key CRSP metrics	173,386	75%
with key Compustat metrics	154,907	67%
with guidance metrics	104,816	45%
with 8-K metrics	96,827	42%
with readability metrics	93,207	40%
with truncation at 1-99% and clipping of UE and $P(-2)$	77,157	33%

Table 2: **Descriptive Statistics of Key Variables**
This table presents the distribution of key variables used in my analysis. All variables are defined in Appendix A. Observations with stock price lower than $1, unexpected earnings lower than -1 and greater than 1, conference call of zero characters, and distance in days between the earnings announcement and the earnings call greater than 90 days are dropped. Cumulative Abnormal Returns, Size, Market-to-Book, Leverage, Market Beta, Earnings Persistence, and Brown Tucker dissimilarity measures are truncated at 1 and 99%.

Variable	**N**	**Mean**	**SD**	**Min**	**Q1**	**Median**	**Q3**	**Max**
ACAR(-1,1)	77,157	0.06	0.05	0.00	0.02	0.04	0.08	0.36
Absolute UE	77,157	0.0066	0.0240	0.0000	0.0006	0.0017	0.0048	0.9342
SUT	77,157	0.35	0.09	0.00	0.29	0.33	0.39	1.00
Loss	77,157	0.26	0.44	0.00	0.00	0.00	1.00	1.00
Size	77,157	7.90	1.60	2.48	6.78	7.87	9.01	12.29
Persistence	77,157	0.34	0.48	-2.54	0.03	0.32	0.63	5.92
MTB	77,157	3.43	4.82	-33.23	1.44	2.32	3.99	66.01
Leverage	77,157	1.77	2.94	-22.50	0.65	1.21	2.17	39.32
Beta	77,157	1.13	0.42	0.03	0.85	1.10	1.39	3.13
Dist. Conf.	77,157	0.40	3.61	0.00	0.00	0.00	0.00	90.00
Len. Comp.	77,157	10.51	0.31	8.56	10.32	10.57	10.73	11.93
Tot. Guidance	77,157	18.65	22.78	1.00	5.00	12.00	24.00	506.00
EPS Guidance	77,157	0.54	0.50	0.00	0.00	1.00	1.00	1.00

Table 3: **Tailorability ChatGPT**

Panel A reports regression estimates of the LLM–tailorability index for U.S. earnings-call transcripts, exploiting the weaknesses documented in Can Turetken and Leippold (2024) and Leippold (2023). For transcript i in quarter t I parse each sentence and detect (i) *comparator expressions* such as *relative to* or *versus*, and (ii) *directional negatives* such as *decrease*, *decline*, *drop*. Let numerical_{it} denote the number of sentences that contain at least one comparator but *no* directional negative, and let comparator_{it} be the total number of sentences that contain a comparator. The LLM–tailorability index is $\text{Tailorability}^{\text{LLM}}_{it} = \frac{\text{numerical}_{it}}{\text{comparator}_{it}+10^{-9}}$ where a small constant 10^{-9} avoids division-by-zero. Higher values imply that managers reframe potentially unfavourable comparisons in linguistically neutral terms that large language models tend to under-weight. I compute the index separately for the prepared remarks (*Presenter Speech* = 1) and the Q&A section (*Presenter Speech* = 0). The dummy Post ChatGPT equals 1 for calls held on or after 30 November 2022 (the public release date of GPT-3.5/ChatGPT) and 0 otherwise. The sample comprises U.S. conference calls from 30 November 2020 to 30 November 2024. Standard errors are in parentheses.

Panel A. Tailorability for LLMs

	LLM Tailorability	LLM Tailorability
Post ChatGPT=1	-0.035***	0.014
	(0.003)	(0.015)
Post ChatGPT=1 × Presenter Speech=1	0.006	0.008**
	(0.004)	(0.004)
CompanyFE	NO	YES
QYearFE	NO	YES
Observations	124474	124434
Adjusted R-squared	0.019	0.121

Panel B. Dictionaries Used to Construct the Index

Comparator expressions	**Directional-negative stems**
relative to; compared with; compared to; versus; vs; in comparison with; in comparison to; instead of; as opposed to; year over year; quarter over quarter; with respect to; on a year-over-year basis; against; in contrast to; over the prior year; over the previous quarter; on a quarterly basis; on an annual basis; from last year; from the prior period; from the previous year; from the same period last year; on a comparable basis; relative change; change from prior; change from previous; comparison with last year; comparison to last quarter	decreas; declin; drop; fall; deteriorat; worsen; reduc; shrink; weaken; compress; dip; plung; tumbl; sink; sunk; slip; lower; diminish; recede; collaps; decelerat

Table 4: **TRC Coefficient - Validation**

This table presents the baseline regression to validate the TRC measure. The ERC Controls are Size, Loss, Market-to-Book, Leverage, Earnings Persistence, and Market Beta. TRC-specific controls are Distance in days from the Earnings Announcement, Transcript Length, Total Guidance, and an Indicator of EPS Guidance issued throughout the Quarter. SUT is the residuals from regressing the Brown Tucker textual difference metric on informational events: guidance (total, if EPS guidance, if range guidance), other filings (readability, sentiment, timing), and macro variables (industrial and consumer indexes, treasury and bond spreads). The residual are then normalized with Min-Max transformation to ensure scores are between 0 and 1.

	(1) Abs CAR(-1,1)	(2) Abs CAR(-1,1)	(3) Abs CAR(-1,1)	(4) Abs CAR(-1,1)	(5) Abs CAR(-1,1)	(6) Abs CAR(-1,1)
Abs UE	0.310***	0.186***	0.134***	0.603***	0.397***	0.320***
	(0.010)	(0.010)	(0.010)	(0.042)	(0.041)	(0.040)
SUT	0.028***	0.021***	0.018***	0.032***	0.025***	0.021***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Abs UE $\times$ SUT				-0.790***	-0.567***	-0.500***
				(0.111)	(0.106)	(0.104)
ERC Controls	NO	YES	YES	NO	YES	YES
TRC Controls	NO	YES	YES	NO	YES	YES
FirmFE	NO	NO	YES	NO	NO	YES
QYearFE	NO	NO	YES	NO	NO	YES
Observations	75629	75629	74947	75629	75629	74947
Adjusted R-squared	0.014	0.089	0.149	0.015	0.090	0.150

Table 5: **Tailorability LM**

This table presents the results of the *Tailorability* validation test. The dependent variable captures the share of negative language that evades the Loughran McDonald tone dictionary. It is constructed as $\text{Tailorability}_{it}^{\text{LM}} = \frac{\text{unflagged}_{it}}{\text{flagged}_{it}+\text{unflagged}_{it}}$ where, $\text{flagged}_{it} = \#\{\text{tokens} \in \mathcal{LM} \text{ dictionary}\}$, and $\text{unflagged}_{it} = \#\{\text{tokens} \in \mathcal{LM} \text{ synonyms} \setminus \mathcal{LM} \text{ dictionary}\}$. The index is separately computed for the presenter speech (Section =1) and the Q&A portion of the call (Section = 0). $PostLM$ takes value equal to one if the call is held after January 6 2011, the day the paper is published on the Journal of Finance, and zero otherwise. The sample contains tailorability and pre- post-indicator for conference calls held on January 6 2009 to January 6 2013. Robust standard errors in parentheses.

Panel A. Tailorability for LM

	Tailorability LM	Tailorability LM
Post LM=1	0.018***	0.000
	(0.004)	(.)
Post LM=1 × Section=1	0.027***	0.027***
	(0.005)	(0.004)
Company FE	NO	YES
Quarter-Year FE	NO	YES
Observations	9692	9689
Adjusted R-squared	0.125	0.357

Panel B. Algorithm for Constructing the LM Synonym Set

1. **Input** Original LM negative lexicon $\mathcal{L}$ (2,303 lemmas); WordNet database WN; Harvard–IV negative lexicon $\mathcal{H}$.
2. **WordNet expansion** For each word $w \in \mathcal{L}$ retrieve every synset in WN (noun, verb, adjective, adverb) and collect the lemmatised lemmas ℓ.
3. **Polarity screen** Keep ℓ only if $\ell \notin \mathcal{L}$ *and* $\ell \in \mathcal{H}$, ensuring the candidate synonym carries negative sentiment.
4. **Inflectional closure** For every retained lemma generate all inflected forms using `lemminflect`.
5. **Deduplication** Remove duplicates and any token already present in $\mathcal{L}$.
6. **Output** Extended synonym set $\mathcal{S}$ with $|\mathcal{S}| = 1{,}438$ additional negative tokens; this is the set $\mathcal{LM}$ synonyms used in the tailorability measure.

Table 6: **ChatGPT Event Study**
This table estimates equation (4.2) around the introduction of ChatGPT-3 class models on 30 November 2022. The dependent variable is $ACAR(-1,1)$, the absolute cumulative abnormal stock return from one trading day before to one trading day after the earnings call (or the first trading day thereafter if the call occurs after hours). *Abs UE* is the absolute earnings surprise, $HSUT$ is an indicator equal to 1 when textual surprise (SUT) is above the median for the year, and *Post ChatGPT* equals 1 for calls on or after 30 November 2022. Column (1) contains no additional controls; Column (2) adds the controls: Size, Loss, Market-to-Book, Earnings Persistence, and Market Beta are classical controls to better estimate the ERC coefficient, while Transcript Length, Total Guidance for the quarter, an Indicator whether EPS Guidance issued throughout the Quarter, Distance in days from the Earnings Announcement, Number of 8Ks issued throughout the quarter, and Distance in Days from the most recent 8K issuance are TRC-specific and discussed in 4. Columns (3) and (4) add Industry and Firm fixed effects, respectively. Robust standard errors are in parentheses.

	(1)	(2)	(3)	(4)
	ACAR(-1,1)	ACAR(-1,1)	ACAR(-1,1)	ACAR(-1,1)
Abs UE	0.342***	0.195***	0.181***	0.159***
	(0.049)	(0.045)	(0.045)	(0.049)
HSUT	0.006***	0.003**	0.003***	0.002
	(0.001)	(0.001)	(0.001)	(0.001)
Abs UE × HSUT	-0.059	-0.040	-0.047	-0.032
	(0.052)	(0.051)	(0.051)	(0.056)
Post GPT=1	0.006***	0.007***	0.008***	0.008***
	(0.001)	(0.001)	(0.001)	(0.001)
Post GPT=1 × Abs UE	0.092	0.042	0.052	0.067
	(0.067)	(0.062)	(0.061)	(0.066)
Post GPT=1 × HSUT	0.003*	0.002	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)
Post GPT=1 × Abs UE × HSUT	-0.209***	-0.137*	-0.139*	-0.144*
	(0.072)	(0.072)	(0.071)	(0.076)
Controls	NO	YES	YES	YES
IndustryFE	NO	NO	YES	NO
FirmFE	NO	NO	NO	YES
Observations	31865	19339	19046	19197
Adjusted R-squared	0.020	0.107	0.138	0.207

Table 7: **ChatGPT Event Study: Presenter Speech vs Q&A**
This table reports the estimates of equation 5.1 for the sample period 30 November 2020 to 30 November 2024. The dependent variable is $ACAR(-1,1)$, the absolute cumulative abnormal stock return from one trading day before to one trading day after the earnings call. $Abs\ UE$ is the absolute earnings surprise; $HSUT$ equals 1 when textual surprise (SUT) is above the median textual surprise; *Section* equals 1 for prepared remarks and 0 for analyst questions; and *Post ChatGPT* equals 1 for calls on or after 30 November 2022. Column (1) contains no additional controls; Column (2) adds the controls: Size, Loss, Market-to-Book, Earnings Persistence, and Market Beta are classical controls to better estimate the ERC coefficient, while Transcript Length, Total Guidance for the quarter, an Indicator whether EPS Guidance issued throughout the Quarter, Distance in days from the Earnings Announcement, Number of 8Ks issued throughout the quarter, and Distance in Days from the most recent 8K issuance are TRC-specific and discussed in 4. Column (3) adds Industry and Quarter-Year fixed effects. Robust standard errors are in parentheses.

	(1) ACAR(-1,1)	(2) ACAR(-1,1)	(3) ACAR(-1,1)
DAbs UE	0.004***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)
HSUT	0.005**	0.004*	0.007***
	(0.002)	(0.002)	(0.002)
DAbs UE $\times$ HSUT	-0.000	-0.001	-0.001***
	(0.000)	(0.000)	(0.000)
Section=1 $\times$ Post ChatGPT=1	-0.004	-0.004	-0.004
	(0.004)	(0.003)	(0.003)
Section=1 $\times$ Post ChatGPT=1 $\times$ DAbs UE	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
Section=1 $\times$ Post ChatGPT=1 $\times$ HSUT	0.005	0.007	0.008*
	(0.004)	(0.005)	(0.005)
Section=1 $\times$ Post ChatGPT=1 $\times$ DAbs UE $\times$ HSUT	-0.001	-0.001	-0.001*
	(0.001)	(0.001)	(0.001)
Controls	NO	YES	YES
QYearFE	NO	NO	YES
IndustryFE	NO	NO	YES
Observations	63730	38678	38092
Adjusted R-squared	0.040	0.113	0.159

Table 8: **LM Event Study**

This table estimates equation (4.2) around the publication of Loughran and McDonald (2011) paper on 6 January 2011. The dependent variable is $ACAR(-1,1)$, the absolute cumulative abnormal stock return from one trading day before to one trading day after the earnings call (or the first trading day thereafter if the call occurs after hours). $Abs\ UE$ is the absolute earnings surprise, $HSUT$ is an indicator equal to 1 when textual surprise (SUT) is above the median for the year, and *Post LM* equals 1 for calls on or after 6 January 2011. The sample includes transcripts of calls held between ± 2 years of the publication date. Column (1) contains no additional controls; Column (2) adds the controls: Size, Loss, Market-to-Book, Earnings Persistence, and Market Beta are classical controls to better estimate the ERC coefficient, while Transcript Length, Total Guidance for the quarter, an Indicator whether EPS Guidance issued throughout the Quarter, Distance in days from the Earnings Announcement, Number of 8Ks issued throughout the quarter, and Distance in Days from the most recent 8K issuance are TRC-specific and discussed in 4. Columns (3) and (4) add Industry and Firm fixed effects, respectively. Robust standard errors are in parentheses.

	(1)	(2)	(3)	(4)
	ACAR(-1,1)	ACAR(-1,1)	ACAR(-1,1)	ACAR(-1,1)
Abs UE	0.857***	0.649***	0.623***	0.623***
	(0.066)	(0.063)	(0.063)	(0.063)
HSUT	0.003**	0.001	0.002	0.002
	(0.001)	(0.001)	(0.001)	(0.001)
Abs UE × HSUT	-0.660***	-0.504***	-0.489***	-0.489***
	(0.068)	(0.070)	(0.069)	(0.069)
Post LM=1	-0.006***	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Post LM=1 × Abs UE	-0.272***	-0.256***	-0.243***	-0.243***
	(0.089)	(0.085)	(0.084)	(0.084)
Post LM=1 × HSUT	-0.000	0.001	0.002	0.002
	(0.001)	(0.001)	(0.001)	(0.001)
Post LM=1 × Abs UE × HSUT	0.294***	0.261***	0.244***	0.244***
	(0.094)	(0.095)	(0.094)	(0.094)
Controls	NO	YES	YES	YES
IndustryFE	NO	NO	YES	NO
CompanyFE	NO	NO	NO	YES
Observations	24651	19869	19690	19690
Adjusted R-squared	0.022	0.098	0.121	0.121

Table 9: **Informativeness dynamics for positive and negative news**
This table estimates equation (5.2) around the introduction of ChatGPT-3 class models on 30 November 2022. The dependent variable is $CAR(-1,1)$, the cumulative abnormal stock return from one trading day before to one trading day after the earnings call (or the first trading day thereafter if the call occurs after hours). UE is the earnings surprise, SUT is the textual surprise, and *Post ChatGPT* equals 1 for calls on or after 30 November 2022. Columns (1)-(3) report the result for equation (5.2) on the subsample of *negative* earnings surprises $(UE < 0)$; while columns (4)-(6) estimate the same specification on non-negative earnings surprises $(UE \geq 0)$. Columns (1) and (4) contains no additional controls; columns (2) and (5) adds the controls: Size, Market-to-Book, Earnings Persistence, and Market Beta are classical controls to better estimate the ERC coefficient, while Transcript Length, Total Guidance for the quarter, an Indicator whether EPS Guidance issued throughout the Quarter, Distance in days from the Earnings Announcement, Number of 8Ks issued throughout the quarter, and Distance in Days from the most recent 8K issuance are TRC-specific and discussed in 4. Columns (3) and (6) add Industry fixed effects, respectively. Robust standard errors are in parentheses.

	(1)	(2)	(3)	(4)	(5)	(6)
	CAR(-1,1)	CAR(-1,1)	CAR(-1,1)	CAR(-1,1)	CAR(-1,1)	CAR(-1,1)
UE	0.247	0.270	0.548	0.083	0.014	0.069
	(0.454)	(0.455)	(0.457)	(0.675)	(0.675)	(0.679)
SUT	-0.009	-0.009	-0.034	-0.039**	-0.040**	-0.040**
	(0.031)	(0.031)	(0.032)	(0.018)	(0.018)	(0.019)
UE × SUT	-0.138	-0.203	-0.811	1.202	1.308	1.078
	(1.065)	(1.065)	(1.069)	(2.010)	(2.012)	(2.021)
Post ChatGPT=1 × UE	-1.048*	-1.103*	-1.369**	2.120**	2.083**	1.995**
	(0.567)	(0.567)	(0.570)	(0.955)	(0.957)	(0.962)
Post ChatGPT=1 × SUT	-0.030	-0.022	-0.013	0.028	0.031	0.034
	(0.044)	(0.044)	(0.045)	(0.025)	(0.025)	(0.025)
Post ChatGPT=1 × UE × SUT	2.705*	2.772**	3.484**	-5.544**	-5.592**	-5.388**
	(1.403)	(1.403)	(1.411)	(2.621)	(2.624)	(2.638)
Controls	NO	YES	YES	NO	YES	NO
IndustryFE	NO	NO	YES	NO	NO	NO
Observations	2367	2367	2333	7526	7526	7448
Adjusted R-squared	0.006	0.013	0.034	0.005	0.007	0.010

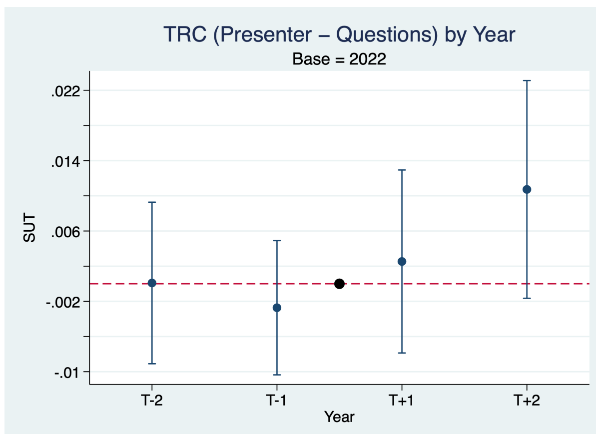
Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

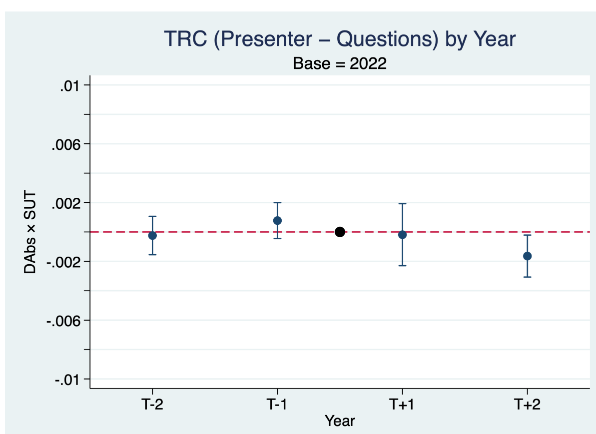
Figures

Figure 1: **ChatGPT Event Study: Presenter Speech vs Q&A - Dynamics**
The figure presents simultaneously estimated trends in textual response coefficients of equation 5.1 for the sample period 01 January 2020 to 31 December 2024. The dependent variable is $ACAR(-1,1)$, the absolute cumulative abnormal stock return from one trading day before to one trading day after the earnings call. $Abs\ UE$ is the absolute earnings surprise; $HSUT$ equals 1 when textual surprise (SUT) is above the median textual surprise; *Section* equals 1 for prepared remarks and 0 for analyst questions; $Abs\ UE$, $HSUT$, and *Section* are interacted with a saturated model of year-indicators for 2020, 2021, 2023, 2024, so that the baseline is 2022, the year before the introduction of ChatGPT. Additional controls are Size, Loss, Market-to-Book, Earnings Persistence, and Market Beta, Transcript Length, Total Guidance for the quarter, an Indicator whether EPS Guidance issued throughout the Quarter, Distance in days from the Earnings Announcement, Number of 8Ks issued throughout the quarter, and Distance in Days from the most recent 8K issuance. Fixed effects for Industry are also included. Robust standard errors are in parentheses.

Panel A: Coefficient $HSUT$



Panel B: Coefficient of the interaction $DAbs\ UE \times HSUT$



A Variables Definition

Table 10: Appendix A: Variable Definitions

Variable Name	**Description**	**Source**
Text Similarity & Transcript Measures. Each metric in evaluated on five transcript subsets: entire call, presenter speech, Q&A, Q, A)		
Textual Dissimilarity	Base dissimilarity is computed as 1 minus the cosine similarity between the vectorized current transcript's processed text and the vectorized document resulting from concatenating the prior four transcripts of the same firm. Textual vectors are computed using TF-IDF vectorization. Textual Dissimilarity is the residual of the base dissimilarity regressed on the five-degree polynomial of the transcript length.	Capital IQ, Brown and Tucker (2011)
SUT	The residual from regressing *Textual Dissimilarity* on a set of characteristics of the information distribution process. *Management guidance* in previous 90 days: both number of estimates, which fraction is range rather than point guidance, and which fraction is EPS guidance. *8-K issuance*: distance in days between the most recent 8-K and the conference call, the number of 8-K issued in the previous 90 days, readability (Flesch Kincaid grade level) and sentiment (frequency of LM negative words) of the most recent filing. *Macro variables:* industrial product index (IP), consumer price index, crude oil future, three month treasury, spread between 10 years and 3 month treasury bill, spread between BAA and AAA. The residual is them passed through the min-max transformation to get values between 0 and 1.	Capital IQ, Brown and Tucker (2011)
HSUT	Indicator: 1 if $SUT > median(SUT)$, 0 otherwise.	Computed
Length	Logarithm of character count of the processed transcript text in each subset, after tokenization and stemming. Proxy for transcript length by section.	Computed

Variable Name	Description	Source
Earnings Forecasts & Guidance		
Abs UE	Absolute value of the unexpected earnings (UE) . UE are computed as in Gipper, Leuz, and Maffett (2020) as the actual EPS (unadjusted, quarterly) as reported by IBES minus the median estimate divided by the stock price two days before the earnings announcement.	IBES, CRSP
LOSS	Indicator: 1 if any $UE < 0$, 0 otherwise.	IBES
MEDEST	Median analyst EPS forecast. For each analysts, retain the closest forecast within 90 days before earnings announcement.	IBES
STDEV	Standard deviation of analyst EPS forecasts in window; proxy for uncertainty/dispersion.	IBES
MEDESTG	Analyst median estimate, but filtering to only those forecasts issued after the latest management guidance within prior 90 days (if any).	IBES, TR IBES Guidance
GUIDANCE	Indicator: 1 if any prior management guidance exists in the 90-day window, 0 otherwise.	IBES, TR IBES Guidance
LATEST GUID	Timestamp of most recent management earnings guidance prior to the announcement, within 90 days.	TR IBES Guidance
TOT GUIDANCE	number of management guidance datapoints in the 90-day window, 0 otherwise.	IBES, TR IBES Guidance
EPS GUIDANCE	Indicator: 1 if any prior management guidance is EPS guidance, 0 otherwise.	IBES, TR IBES Guidance

Variable Name	Description	Source
NUM ANALYSTS	Number of unique analysts issuing forecasts 3 to 360 days before earnings announcement.	IBES
Market-Based Variables (Event Windows)		
$CAR(-1,1)_{event}$	Cumulative abnormal return (CAR), sum of abnormal returns from day -1 to +1 surrounding the conference call (cc) or earnings announcement (ea). Benchmark: S&P 500 daily returns.	CRSP + S&P500
$AbnVOL(-1,3)_{event}$	Abnormal trading volume, sum over days -1, 0, +1, +2, +3, normalized by previous 41-day baseline; centered at conference call date (cc) and earnings announcement date (ea).	CRSP
$P(-2)_{event}$	Stock price two trading days prior to conference call or earnings announcement.	CRSP (WRDS)
$iVOL_{event}$	Idiosyncratic volatility, estimated two trading days before event using rolling one-year Fama-French regression.	CRSP + FF Factors
$BETA_{event}$	Beta with respect to market excess return, estimated over rolling one-year regressions up to the event date.	CRSP + FF Factors
$timeliness_{event}$	Timeliness/price efficiency measured as average negative absolute log-price change versus last pre-event price, over days -82 to -2. Higher values indicate more informative price discovery.	CRSP
Firm Characteristics		

Variable Name	Description	Source
SIZE	log(Market Value), measured as the natural logarithm of the market value at fiscal year-end prior to the event. Proxy for firm size.	Compustat
MTB	Market-to-book ratio, defined as market capitalization divided by book equity at the fiscal year-end prior to the event.	Compustat
LEVERAGE	Financial leverage ratio, computed as total debt (long-term plus short-term) divided by total assets, at fiscal year-end prior to the event.	Compustat
EARNINGS PERSISTENCE	Earnings persistence, measured as the coefficient on lagged earnings per share excluding extraordinary items (epspx) from a firm-level regression of current epspx on past epspx. Higher values indicate greater earnings persistence.	Compustat
DIST EC EA	Calendar days between the earnings conference call and earnings announcement.	Computed

B The Model

This appendix presents a one-period communication game between a *sender* and a *receiver*. Communication is noisy: the report submitted by the sender does not necessarily coincide with the message processed by the receiver. Mismatches arise from (i) manual processing frictions when no technology is used and (ii) algorithmic misclassification when technology is employed.

The two players are a manager (*Manager*, M, she/her) and an investor (*Investor*, I, he/him). The manager, who holds proprietary information, drafts a report r and transmits it to the investor; the investor reads the report and set up the price for the manager's firm. While M wishes to maximize investment in her firm, I wants to set up a price that matches the fundamental value of the business.

The investor relies on an algorithm to parse the report. The algorithm is imperfect and may produce an incorrect signal. The manager does not know whether the algorithm will classify her report correctly. She can, however, pay a private cost to learn how the algorithm operates; doing so enables her to tailor the report so that she can choice precisely r. If she does not, she will report truthfully (i) and the report observed by the investor will be $r = i \times j$. The cost of acquiring this knowledge is private information to M; the investor observes only its distribution.

B.1 Timing

1. **Nature.** Nature draws a state $x = (i, j, c) \in X = \{(i, j, c) : (i, j) \in \{-1, 1\}^2,\ c \sim U[0, 2]\}$:
   - $i \in \{-1, 1\}$ represents fundamental news.
   - $j \in \{-1, 1\}$ captures limited processing capacity (or algorithmic inaccuracy). When M truthfully reports i, the message observed by I is $m = i \times j$. Therefore if $j = 1$, the report is processed correctly, and if $j = -1$, it is misinterpreted. Although it may seem unnatural to assume the processing outcome j is determined before the state is observed by M, this timing is without loss of generality. In fact, since M never observes j, the setup is equivalent to a more natural interpretation where j is realized only after the manager discloses the report.
   - $c \sim U[0, 2]$ is the private cost of learning the algorithm. The maximum cost is assumed to be high enough to deter managers with bad news and the highest possible costs from

engaging in algorithmic tailoring.

The prior μ satisfies

$$\mu(i=1) = \mu(i=-1) = \tfrac{1}{2}, \quad \mu(j=i \mid i) = 1-\kappa, \quad \mu(j \neq i \mid i) = \kappa,\ \kappa \in [0, 1/2]$$

and c is independent of (i, j).

2. **Manager Information.** The manager observes (i, c) but not j. Although j is drawn before she moves, the timing is inconsequential because j remains unobservable to her.

3. **Tailoring decision.** The manager decides whether to learn the algorithm at cost c. Let $T = 1$ if she tailors and $T = 0$ otherwise.

4. **Report.** If $T = 0$, the manager write a truthful, but tech-unaware report: $r = i \times j$ which is subject processing noise, and therefore can be interpreted correctly or not depending on j. If instead, $T = 1$, a tech-savvy manager can write a report that will be interpreted as she prefers: $r \in \{-1, 1\}$.

5. **Pricing.** The investor observes the message m and chooses a price $p \in [-1, 1]$. Payoffs are

$$U^M = p, \qquad U^I = -(p-i)^2.$$

B.2 Equilibrium

The solution concept is *perfect Bayesian equilibrium* (PBE) in pure strategies.

Proposition B.2.1. There exists a unique PBE in pure strategies:

$$T(i,c) = \mathbf{1}_{c \leq c_i^*}, \quad r^*(i, c \leq c_i^*) = 1,$$
$$r^*(i, c > c_i^*) = i \times j, \quad p(r)^* = \mathbb{E}[i|r], \text{ for } i \in \{-1, 1\}$$

Proof. Note that message depends on the second entry of the report, and such entry can differ from the state only when the technology is acquired. So the firm's price depend on the fraction of manager' types that want to acquire the technology.

Since, the value of the firm with good news is higher than the firm with bad news, I start conjecturing that technologically savvy managers will write a report $r = 1$. Tech-naive managers will stick with their truthful report that κ times gets misclassified.

$$
\begin{aligned}
&\text{If } i = 1: \\
&\qquad \mathbb{E}[p|T=0] = (1-\kappa)p(1) + \kappa p(-1) \\
&\qquad \mathbb{E}[p|T=1] = p(1) - c \\
&\implies T = 1 \text{ if: } c < \kappa\left[p(1) - p(-1)\right] = c_1
\end{aligned}
$$

$$
\begin{aligned}
&\text{If } i = -1: \\
&\qquad \mathbb{E}[p|T=0] = \kappa p(1) + (1-\kappa)p(-1) \\
&\qquad \mathbb{E}[p|T=1] = p(1) - c \\
&\implies T = 1 \text{ if: } c < (1-\kappa)\left[p(1) - p(-1)\right] = c_{-1}
\end{aligned}
$$

Note that for $\kappa < \frac{1}{2}$, the threshold cost for the manager's type $i = 1$ is higher than for types with $i = -1$. Since truthful reporting yields lower payoffs in expectation for the bad news holders, the incentives to acquire the technological knowledge are higher.

Since the investor cannot observe c, he will work in expectation. The fraction of managers, holding good (bad) news, willing to invest in learning how to tailor are $F(c1) = c_1/2$ $(F(c1) = c_{-1}/2)$. Then, *Investor* prices in a Bayesian fashion:

$$
\begin{aligned}
p(1) &= \frac{\left[F(c_1) + (1-\kappa)(1-F(c_1))\right] + v(-1)\left[F(c_{-1}) + \kappa(1-F(c_{-1}))\right]}{F(c_1) + (1-\kappa)(1-F(c_1)) + F(c_{-1}) + \kappa(1-F(c_{-1}))} \\
&= \frac{2 - 4\kappa + \left[p(1) - p(-1)\right](2\kappa - 1)}{2 + \left[p(1) - p(-1)\right](2\kappa^2 - \kappa + 1)} \qquad \text{(B.1)} \\
p(-1) &= \frac{\kappa(1-F(c_1)) + v(-1)(1-\kappa)(1-F(c_{-1}))}{\kappa(1-F(c_1)) + (1-\kappa)(1-F(c_{-1}))} \\
&= -\frac{2 - 4\kappa + \left[p(1) - p(-1)\right](2\kappa - 1)}{2 - \left[p(1) - p(-1)\right](2\kappa^2 - 2\kappa + 1)} \qquad \text{(B.2)}
\end{aligned}
$$

Then,

$$\begin{aligned}
p(1) - p(-1) &= \frac{[p(1) - p(-1)](2k-1) + 2 - 4\kappa}{4 - [p(1) - p(-1)]^2(2\kappa^2 - 2\kappa + 1)} \\
\text{Call } x \; = p(1) - p(-1): \quad x &= \frac{x(2k-1) + 2 - 4\kappa}{4 - x^2(2\kappa^2 - 2\kappa + 1)^2}
\end{aligned}$$

I find $p(1) - p(-1) = x(\kappa)$ as the solution of a fixed point problem.

$$A(x(k), k) = -x(k)^3(2\kappa^2 - 2\kappa + 1)^2 - x(k)(8 - 8\kappa) + (8 - 16\kappa) \tag{B.3}$$

Note that $p(1), p(-1)$ being Bayesian prices can take values at most in $[-1, 1]$, and therefore their distance cannot be any greater than 2. Also, without tailoring the distance of $p(1) - p(-1) \geq 0$. It cannot be sustained an equilibrium where $p(1) < p(-1)$, if that was the case, all bad type would be able to send $m = -1$ at a cheaper costs that the respective good type. But a pool of mostly bad types would result in lower Bayesian prices than the pool of remaining types (mostly good types): $(a - 1) < p(1)$. Then $p(1) - p(-1) \geq 0$.

In equilibrium $A(x(k), k) = 0$. As is standard in the literature, I prove that $\exists! x(k): \; A(x(k), k) = 0$ inspecting at the extreme values that x can take, $x = 0, \; x = 2$, and show monotonicity. Since $A(x, k)$ is polynomial of degree three, it is continuous and twice differentiable: for the intermediate value theorem, exists a unique $x(k)$ such that $A(x(k), k) = 0$.

In fact, for $x \in \left(0, \frac{1}{2}\right)$:

$$\begin{aligned}
A(0, \kappa) &= 8 - 16\kappa > 0 \\
A(2, \kappa) &= -8(2\kappa^2 - 2\kappa + 1)^2 - 8 < 0 \\
\frac{\partial A(x, k)}{\partial x} &= -3x^2(2\kappa^2 - 2\kappa + 1)^2 - (8 - 8\kappa) < 0
\end{aligned}$$

Then, since $A(x, k)$ is continuous on the interval, for the intermediate value theorem: $A(0, \kappa) > 0, \; A(2, \kappa) < 0, \; \frac{\partial A(x,k)}{\partial x} < 0 \implies \exists! x(k): \; A(x(k), k) = 0$.

Since $x(k)$ is unique fo a given κ, there exists a unique pair of valid thresholds $c_1^*(\kappa), \; c_{-1}^*(\kappa)$ and prices $a^*(1; \kappa)$, and $a^*(-1; \kappa)$.

B.2.1 Uniqueness

To complete the proof of uniqueness, it's still left to show that $p(-1) > p(1)$ cannot be sustained in equilibrium.

First, the thresholds for c_1 and c_{-1} become:

$$c_1 = (1-\kappa)[p(-1) - p(1)]$$
$$c_1 = \kappa[p(-1) - p(1)]$$

And therefore Bayesian prices are:

$$\begin{aligned} p(1) &= \frac{(1-\kappa)(1-F(c_1)) - \kappa(1-F(c_{-1}))}{(1-\kappa)(1-F(c_1)) + \kappa(1-F(c_{-1}))} \\ &= \frac{2(1-2\kappa) + \kappa(c_1 + c_{-1}) - c_1}{2 + \kappa(c_1 - c_{-1}) - c_1} && \text{(B.4)} \\ p(-1) &= \frac{F(c_1) + \kappa(1-F(c_1)) - (1-\kappa)(1-F(c_{-1})) + F(c_{-1})}{F(c_1) + \kappa(1-F(c_1)) + (1-\kappa)(1-F(c_{-1})) - F(c_{-1})} \\ &= -\frac{2(1-2\kappa) + \kappa(c_1 + c_{-1}) - c_1}{2 - \kappa(c_1 - c_{-1}) + c_1} && \text{(B.5)} \\ \implies p(-1) - p(1) &= -\frac{4[2(1-2\kappa) + \kappa(c_1 + c_{-1}) - c_1]}{4 - [\kappa(c_1 - c_{-1}) - c_1]^2} && \text{(B.6)} \\ &= -\frac{4[2(1-2\kappa) + \kappa x - (1-\kappa)x]}{4 - [\kappa(1-2\kappa)x - (1-\kappa)x]^2} && \text{(B.7)} \end{aligned}$$

Call $x = p(-1) - p(1) > 0$:

$$B(x, \kappa) = -x^3(2\kappa^2 - 2\kappa + 1)^2 + 8\kappa x + 8(1-2\kappa)$$

When assuming $p(1) > p(-1)$, $A(x, \kappa)$ satisfied the intermediate value theorem. For $B(x, \kappa)$ instead I get:

$$B(0, \kappa) = 8(1-2\kappa) > 0 \ \text{ for } \kappa \in \left(0, \frac{1}{2}\right)$$

And monotonically decreasing in x:

$$\frac{dB(x,\kappa)}{dx} = -3x^2(2\kappa^2 - 2\kappa + 1)^2 + 8\kappa$$

$$\begin{aligned}\text{From } B(x,\kappa) = 0 : x^2 &= \frac{8(1-2\kappa)+8\kappa x}{x(2\kappa^2-2\kappa+1)^2}\\ \implies \frac{dB(x,\kappa)}{dx} &= -\frac{24(1-2\kappa)-16\kappa x}{x} < 0\end{aligned}$$

Finally, for the largest possible value of $x = 2$:

$$\begin{aligned}B(2,\kappa) &= -8(2\kappa^2-2\kappa+1)^2+8\\ &= 8[1-(2\kappa^2-2\kappa+1)^2] > 0 \text{ for } \kappa \in \left(0, \frac{1}{2}\right)\\ &\text{Since } 2\kappa(2\kappa-1) < 0 \text{ for } \kappa \in \left(0, \frac{1}{2}\right)\end{aligned}$$

Then, $B(0,\kappa) > B(2,\kappa) > 0$, and $\frac{dB(x,\kappa}{dx} < 0$ implies that cannot exist x^* such that $B(x^*,\kappa) = 0$, which concludes the proof.

B.2.2 Comparative statics

Note that $\frac{\partial x(\kappa)}{\partial \kappa} < 0$. The implicit function theorem shows that:

$$\begin{aligned}&\frac{\partial A}{\partial \kappa} = 0\\ \iff& -3x(\kappa)^2\frac{\partial x(\kappa)}{\partial \kappa}(2\kappa^2-2\kappa+1)^2 - \frac{\partial x(\kappa)}{\partial \kappa}(8-8\kappa) - 2x^3(2\kappa^2-2\kappa+1)(4\kappa-2) + 8x(\kappa) - 16 = 0\\ \iff& \frac{\partial x(\kappa)}{\partial \kappa} = \frac{(8x(\kappa)-16) - 2x(\kappa)^3(2\kappa^2-2\kappa+1)(4\kappa-2)}{3x(\kappa)^2(2\kappa^2-2\kappa+1)^2+(8-8\kappa)} \qquad \text{(B.8)}\end{aligned}$$

The denominator is always positive, so the sign of the derivative is determined by the nominator. Note that I can rewrite $A(x(\kappa),\kappa) = 0$ as:

$$x(\kappa)^3(2\kappa^2-2\kappa+1) = \frac{(8-16\kappa)-x(\kappa)(8-8\kappa)}{(2\kappa^2-2\kappa+1)} \qquad \text{(B.9)}$$

I plug B.9 into B.8, and get that

$$\begin{aligned}\frac{\partial x(\kappa)}{\partial \kappa} > 0 &\iff (8x(\kappa)-16)(2\kappa^2-2\kappa+1) - (8\kappa-4)[8-16\kappa - x(\kappa)(8-8]\kappa) > 0\\ &\iff x(\kappa) > \frac{16(2\kappa^2-2\kappa+1)+(8\kappa-4)(8-16\kappa)}{8(2\kappa^2-2\kappa+1)+(8\kappa-4)(8-8\kappa)}\end{aligned}$$

$$\iff x > \frac{16}{8}\frac{-6\kappa^2+6\kappa-1}{-6\kappa^2+10\kappa-3} = 2h(\kappa)$$

Now, recall that $x \in [0,2]$, then if $h(\kappa) > 1$, I have $\frac{\partial x(\kappa)}{\partial \kappa} < 0$.

$$\frac{-6\kappa^2+6\kappa-1}{-6\kappa^2+10\kappa-3} > 1 \iff \kappa < 1/2$$

Which is always satisfied on $\kappa \in (0,1/2)$.

Finally, I can move to the comparative statics for of $p(1)$ and $p(-1)$.

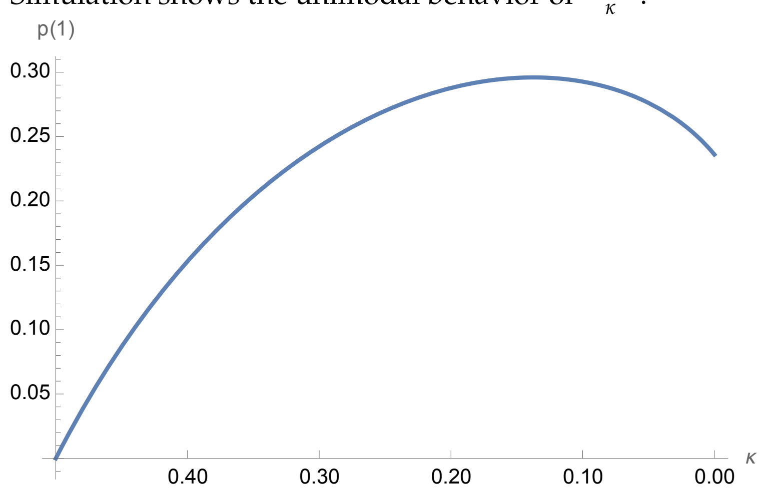
$$\frac{\partial p(-1)}{\kappa} = \frac{[4-\frac{\partial x(\kappa)}{\partial \kappa}(2\kappa-1)-2x(\kappa)][2-x(\kappa)(2\kappa^2-2\kappa+1)]+[-2+4\kappa-x(\kappa)(2\kappa-1)][\frac{\partial x(\kappa)}{\partial \kappa}(2\kappa-1)+x(\kappa)(4\kappa-2)]}{[2-x(\kappa)(2\kappa^2-2\kappa+1)]^2}$$

$$\propto \underbrace{\frac{\partial x(\kappa)}{\partial \kappa}}_{<0}\underbrace{[2\kappa^2-3\kappa+1]}_{<0}\kappa + \underbrace{x(\kappa)}_{>0}\underbrace{[2\kappa^2-2\kappa-1]}_{<0}\kappa + \underbrace{x(\kappa)^2}_{>0}\underbrace{[-\kappa^2-3\kappa+1]}_{bimodal}\kappa + 8$$

Yet, $x(\kappa)$ takes at most the value of 2, and $2[2\kappa^2-2\kappa+1]+4[-\kappa^2-3\kappa+1]\kappa+8 = -16\kappa^2-3\kappa+8 > 0$ on $\kappa \in (0,1/2)$

Therefore, $\frac{\partial p(-1)}{\kappa} > 0$.

Simulation shows the unimodal behavior of $\frac{\partial p(1)}{\kappa}$.



Analytically, it can be shown that the relation is positive as $\kappa \to 0$:

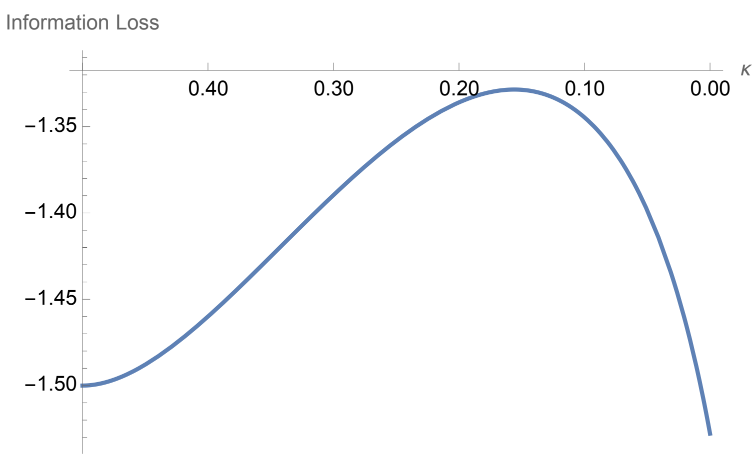
$$\frac{\partial p(1)}{\partial \kappa} = \frac{\big[-4+x'(\kappa)(2\kappa-1)+2x(\kappa)\big]\big[2+x(\kappa)(2\kappa^2-2\kappa+1)\big]-\big[2-4\kappa+x(\kappa)(2\kappa-1)\big]\big[x'(\kappa)(2\kappa-1)+x(\kappa)(4\kappa-2)\big]}{\big[2+x(\kappa)(2\kappa^2-2\kappa+1)\big]^2}$$

$$\propto 4\Big[(2\kappa^3-3\kappa^2+3\kappa-1)\,x'(\kappa)+(2\kappa^2-2\kappa+1)\,x(\kappa)+\kappa(1-\kappa)\,x(\kappa)^2-2\Big].$$

and $\lim_{\kappa\to 0} \frac{\partial p(1)}{\kappa} = 3 > 0$

Loss of information is reflected by the expected payoff of the investor:

$$\begin{aligned}
Information\ Loss &= -\mathbb{E}[p-i]^2 \\
&= -[p(1)-1]^2 * (1-\kappa+\kappa * c_1/2) - [p(-1)-1])^2 * (\kappa * c1/2) \\
&\quad - [p(-1)+1]^2 * ((1-c_{-1}/2)*(1-k)) - [p(1)+1]^2 * ((1-c_{-1}/2)*k + c_{-1}/2)
\end{aligned}$$



Finally, conditional on seeing a report $r = 1$, the information loss is decreasing and the conditional on seeing $r = -1$ the information loss is decreasing.