Anonymization and Information Loss

Ke Wu*

Baozhong Yang†

Zhenkun Ying‡

Dexin Zhou§

Abstract

We show that while anonymization effectively obscures firm identity, it significantly reduces the power of textual understanding, thereby diminishing models' ability to extract meaningful economic signals from financial texts. This information loss is particularly severe when numerical and object entities are removed from texts and is amplified in texts characterized by high linguistic uncertainty and firm specificity. Importantly, in the setting of sentiment extraction from earnings call transcripts, we find that information loss induced by anonymization is more pervasive and severe than the effects of look-ahead bias, suggesting that the costs of anonymization may outweigh its benefits in certain financial applications.

Keywords: Large Language Models, Look-ahead Bias, Memorization, Anonymization, Information Loss, Textual Analysis

---

*ke.wu@ruc.edu.cn, School of Finance, Renmin University of China.

†bzyang@gsu.edu, J. Mack Robinson College of Business, Georgia State University.

‡yingzhenkun0205@ruc.edu.cn, School of Finance, Renmin University of China.

§dexin.zhou@baruch.cuny.edu (Corresponding Author), Zicklin School of Business, Baruch College.

1 Introduction

Large language models (LLMs) have transformed how finance and accounting researchers extract information from textual data. Numerous studies underscore the power of LLMs compared to traditional methodologies like bag-of-words (e.g., Lopez-Lira and Tang, 2023). Researchers use LLMs to extract economic signals from earnings calls, interpret firm policies, and measure geopolitical exposure (Jha et al., 2024a,b; Siano, 2025; Kim et al., 2023; Clayton et al., 2025). Other applications include mapping global production networks using financial texts (Breitung and Müller, 2025), detecting investor sentiment on social media (Chen et al., 2025), and generating structured economic representations from news (Sarkar, 2025).

Despite their power in analyzing financial texts, several studies caution that LLMs may introduce look-ahead bias in the data extraction process. This bias is particularly relevant for finance and economics, where both extracted signals are leveraged to predict future economic performances or stock returns. This bias stems from LLMs' training stage, when they learn from vast corpora spanning long periods of time, potentially causing their outputs to incorporate information that would not have been accessible to an agent prior to the model's knowledge cutoff. (e.g., Levy, 2024; Sarkar and Vafa, 2024; Ludwig et al., 2025; Lopez-Lira et al., 2025).

Several strategies have been proposed to mitigate this bias. Among them, anonymization has gained significant popularity since it is relatively easy to implement while still allowing researchers to use state-of-the-art models. This methodology replaces identifiers such as firm names or tickers with placeholders (e.g., "FIRM_1") (Glasserman and Lin, 2023; Engelberg et al., 2025; Sarkar and Vafa, 2024). The premise is that obscuring these unique identifiers forces LLMs to analyze the document exclusively on its inherent linguistic and semantic content rather than firm-specific knowledge.[1]

---

[1]Other approaches include training models with a knowledge cutoff that predates the training chronologically consistent models that only use data before the date of the textual content of interest (e.g., Sarkar, 2024; He et al., 2025) or conducting robustness tests using texts beyond existing models' knowledge cutoff (e.g., Jha et al., 2024a; Siano, 2025).

In this study, we investigate an important limitation of the anonymization methodology: its potential to lead to significant information loss and degrade the extracted signal. The identity of an entity is not merely a label but a critical piece of contextual information essential for accurate interpretation. Consider the statement: "The firm's annual income is $10 million." For a multinational corporation like Apple Inc., this figure suggests poor performance, indicating a negative signal. In contrast, for a firm that had not been profitable previously, the same figure could signify remarkable growth, indicating a strongly positive signal. After anonymization, both scenarios collapse into an identical, ambiguous input: "FIRM_1's annual income is NUMBER_1." This procedure strips the model of the necessary context to make an accurate inference, potentially leading to noisy signals.

Empirically, we task LLMs to extract information from a sample of conference call transcripts that are dated after the models' knowledge cutoff. We compare the informativeness of signals extracted before and after popular anonymization procedures. As these conference call transcripts are not part of the models' pretraining datasets, any observed differences in the extracted signals' informativeness can be attributed directly to the changes in information content within the transcripts introduced by anonymization.

Our first analysis focuses on the extraction of firm-value-relevant signals using GPT-4o-mini. We prompt the model to classify whether a given conference call contains positive or negative sentiment regarding future firm value. Subsequently, we analyze the relationship between the extracted signal and short-term abnormal returns around the conference call date. This setting closely aligns with past research designs (e.g., Loughran and Mcdonald, 2011; Garcia et al., 2023; Siano, 2025).[2] An effective signal is expected to exhibit a stronger relationship with short-term abnormal returns.

We find that sentiment extracted from raw (unanonymized) earnings call transcripts exhibits a strong association with conference call-day stock returns, indicating the model's

---

[2]Our analyses are different from Lopez-Lira and Tang (2023); Chen et al. (2022); Engelberg et al. (2025), who focus on predicting future returns using LLM-extracted signals. Predicting future returns may require investors to underreact to value-relevant information. This requirement could confound the goal of evaluating LLMs' ability to extract textual sentiment.

capacity to capture value-relevant information. Subsequently, we anonymize transcripts using three strategies. The first two algorithms are from the SpaCy library, a powerful toolkit for identity recognition, and are designated as Small (SM) and Transformer (TRF).[3] We also consider the anonymization approach analyzed in Engelberg et al. (2025), which uses LLMs to identify and anonymize entities from financial texts. We find that the signals extracted from anonymized texts exhibit significantly attenuated informativeness. For instance, when using a standard Transformer-based anonymization model, the R^2 from a regression of returns on sentiment with controls drops by over 6% (from 0.132 to 0.124) and even drops by more than 10% (from 0.078 to 0.070) in the regression without controls. The information loss becomes even clearer in a direct "horse race" regression. When both raw and anonymized sentiment signals are included as explanatory variables, the coefficient on the raw signal remains large and highly significant, while the coefficient on the anonymized signal collapses from 2.331 to 0.775. This suggests that the sentiment information conveyed by the anonymized signal is largely subsumed by the information provided by the named entities.

We trace the source of this information loss to the removal of specific entity types. We classify entities into four main types: NUMBERS, OBJECTS, PLACES, and OTHERS. The degradation in signal quality is most severe when removing numbers (such as financial figures and dates) and objects (such as company and executive names), which act as critical anchors for contextual interpretation. In contrast, removing entities like places has a minimal impact.

Furthermore, we find that the divergence between raw and anonymized signals is greater in texts characterized by higher linguistic uncertainty, or those concerning smaller, more specific firms with less public exposure. This suggests that in ambiguous contexts, LLMs rely more heavily on concrete named entities to ground their analysis, making their removal particularly impactful. Similarly, because smaller and more specific firms are discussed less in public texts, LLMs have less training data for them compared to larger or widely known firms. This reduced data also forces the LLM to rely more heavily on named entities, causing

---

[3]TRF is considered a more stringent approach that identifies more entity-related information.

more severe information loss upon anonymization.

While our main analyses focus on GPT-4o-mini, we show that bigger models (e.g., GPT-4o) and reasoning-focused models (e.g., GPT-o3-mini) also exhibit information loss after conference calls are anonymized, indicating the generalizability of our results. In fact, we find that the information loss from anonymization is more pronounced when conducting information extraction using bigger models.

The loss of information associated with anonymization is also not confined to extracting signals related to firm value. We attempt to extract signals relating to uncertainty and find that anonymization leads to a significant reduction in the ability to predict future return volatility. Furthermore, we follow Jha et al. (2024a) and Jha et al. (2024b) to predict future corporate policy and firm performance. We find that anonymization similarly degrades models' ability in those prediction tasks.

Moreover, to confirm the generalizability of our conclusions, we replicate the sentiment analysis tests on a different corpus with distinct characteristics: short, entity-dense news headlines. The results are consistent, confirming that the loss of information is a fundamental consequence of the anonymization process itself, rather than a specific type of corporate disclosure.

Finally, we extend our analysis to include the pre-cutoff period (November 2022–October 2023), allowing us to compare information loss from anonymization across periods where look-ahead bias might differentially affect LLM performance. We find that the information loss resulting from anonymization is persistent and economically meaningful in both the pre- and post-cutoff periods, with the average magnitude of the loss remaining broadly similar on either side of the knowledge cutoff. Moreover, we do not find significant evidence that signals extracted from conference call transcripts before the knowledge cutoff dates are significantly more informative than those after the models' knowledge cutoff. These results suggest that look-ahead bias plays a limited role in this sentiment extraction setting, and the observed decline in predictive content predominantly reflects the direct impact of information loss due

to anonymization. Furthermore, our analysis indicates that metrics based on firm recognition rates are ineffective proxies for information leakage.

Our study contributes to an emerging literature that explores the use of LLMs to extract information from unstructured textual data. While LLMs are shown to have the capacity to extract information from unstructured data, many studies also caution about the risk of these models introducing look-ahead biases (e.g., Glasserman and Lin, 2023; Sarkar and Vafa, 2024; Ludwig et al., 2025; Levy, 2024; Lopez-Lira et al., 2025). Anonymization has been proposed and adopted as an easy-to-implement strategy to mitigate the look-ahead bias (e.g., Breitung and Müller, 2025). Recent studies, such as Engelberg et al. (2025), also attempt to improve anonymization strategies and show that LLMs themselves can be used to effectively eliminate entities that can help identify firms. Our study complements the earlier studies and empirically investigates the trade-off between mitigating look-ahead bias and preserving essential contextual information in financial text analysis. While anonymization may reduce look-ahead bias, our study empirically demonstrates that it does so at the cost of informational value. Thus, our results urge caution in interpreting the economic magnitudes of signals extracted from anonymized texts.

2 Methodology and Data

2.1 Anonymization Algorithms

We employ three approaches for text anonymization. The first two utilize specialized Named Entity Recognition (NER) models from the Python spaCy package, consistent with the non-LLM anonymization models employed by Breitung and Müller (2025). They include en-core-web-sm (SM), a Convolutional Neural Network (CNN) architecture utilizing GloVe word vectors, and en-core-web-trf (TRF), which is based on a Transformer architecture. The third approach follows Engelberg et al. (2025) and uses a commercial LLM, specifically GPT-4o-mini, in conjunction with custom prompts to perform anonymization. These three models

vary significantly in size, a difference that stems from their entirely different underlying architectures. The en-core-web-sm model is a lightweight, CNN-based system with a file size of only 12 MB. In contrast, en-core-web-trf is based on the RoBERTa-base Transformer architecture, resulting in a 436 MB file size,[4] and its RoBERTa-base core alone contains approximately 0.1 billion (0.1B) parameters.[5] Finally, the GPT-4o-mini is the largest by several orders of magnitude; while its exact parameters are undisclosed, it is widely estimated by the industry to be on the 8 billion (8B) parameters.[6]

Figure 1 Approach A illustrates a detailed processing workflow for spaCy NER models, using a sentence from an Apple Inc. earnings call transcript as an example. As illustrated in the figure, the model first processes the raw text and recognizes entities in sequence: "Apple", "Tim Cook", and "Luca Maestri." Based on their entity type and order of appearance, we generate a consistent global mapping between the anonymized and original entity.[7] Specifically, "Apple" is the first organization (ORG) identified, so it is mapped to ORG_1. "Tim Cook" is the first person (PERSON) identified and is mapped to PERSON_1. Subsequently, "Luca Maestri" is identified as the second person, and is thus mapped to PERSON_2. Once the entire text unit is processed and the map is complete, a global replacement is applied to produce the final anonymized output. This approach, in contrast to replacing all entities with a random placeholder string as employed by Glasserman and Lin (2023), is designed to maximally preserve textual information while maintaining the readability of the anonymized text.

For anonymization using the GPT model, we input each text unit along with a predefined prompt, which then automatically generates the anonymized output, as illustrated in Approach B of Figure 1. The model then processes the text, contextually identifying enti-

---

[4]https://spacy.io/models/en
[5]https://huggingface.co/roberta-base
[6]https://techcrunch.com/2024/07/18/openai-unveils-gpt-4o-mini-a-small-ai-model-powering-chatgpt
[7]As exhibited in Table A.4, we group these 18 types into four main categories to structure our anonymization: NUMBERS (which includes DATE, CARDINAL, MONEY, PERCENT, ORDINAL, TIME, and QUANTITY), PLACES (GPE, LOC, FAC), OBJECTS (PERSON, ORG, NORP), and OTHERS (PRODUCT, EVENT, LAW, WORK_OF_ART, and LANGUAGE).

ties and generating replacements sequentially. For instance, it identifies "Apple" as the first organization and replaces it with ORG_1, and "Tim Cook" as the first person, replacing him with PERSON_1. This results in an anonymized output format highly similar to the spaCy NER method's. This consistent output formatting not only ensures high readability but also establishes a common ground for a direct and fair comparison between this and earlier anonymization approaches. Our prompt design follows Engelberg et al. (2025) closely. The entity types we consider include those in Engelberg et al. (2025) and the 18 built-in entity labels from the spaCy NER model. The full prompt is available in Appendix B.1.

Table 1 presents a text excerpt from Apple Inc. 2024 Q1 earnings call transcript, along with the results from the three different anonymization methods. The SM model exhibits weaker performance, failing to identify "Apple" in the final paragraph as an organization, whereas the TRF model correctly identified it. Compared to the spaCy models, which are limited to identifying and replacing the 18 predefined entity types, the GPT model's replacement capability is more flexible. In this specific excerpt, the GPT model did not alter temporal expressions but instead anonymized regulatory filing terms like "10-K" and "8-K" into "FORM_1" and "FORM_2".

2.2 Data

2.2.1 Stock Data

The sample for this research consists of constituent stocks from the iShares Russell 3000 ETF for the period of November 2023 through December 2024. We obtain return data from CRSP, corporate accounting data from Compustat, and EPS forecast data from the I/B/E/S Detail History database. Detailed variable definitions can be found in Table A.2.

Our sample of earnings call transcripts comprises 7,382 transcripts from 1,831 unique firms over a period of 259 days, as detailed in Table A.3. The sample firms are tilted toward large-cap firms, with an average market capitalization of approximately $23,845 million, an

average book-to-market ratio of 0.55, and an average turnover rate of 2.77. 47.41% of these firms are listed on NASDAQ.

2.2.2 Earnings Call Transcripts

The primary textual source we examine in this study is corporate conference calls. They have been shown to carry rich value-relevant information about firms (e.g., Frankel et al., 1999; Brown et al., 2004). Many recent studies leverage LLMs to extract information from these texts (e.g., Jha et al., 2024a; Clayton et al., 2025). To isolate a clean effect unconfounded by look-ahead bias, we utilize data generated after the knowledge cutoff date of the Large Language Models (LLMs) employed. Models used in this study include GPT-4o-mini-2024-07-18, GPT-4o-2024-08-06, and GPT-o3-mini-2025-01-31, all of which have a knowledge cutoff date of October 2023[8]. Accordingly, we download earnings call transcripts from November 2023 onward from SeekingAlpha.

We remove duplicate entries and those with wrong time stamps. We also exclude transcripts exceeding 15,000 tokens. This threshold was established because the GPT-4o-mini model, used for the anonymization procedure, has a maximum output limit of 16,000 tokens; longer input transcripts would therefore result in truncated anonymized text. Furthermore, we matched the transcript data with corresponding returns, and EPS forecast errors using date and PERMNO. The detailed sample selection criteria are presented in Panel A of Table A.1.

2.2.3 News Headlines

Besides conference calls, news articles have also been shown to deliver rich information about firms (e.g., Tetlock et al., 2008; Tetlock, 2010). Moreover, many studies also attempt to extract signals from news headlines and articles (e.g., Bybee et al., 2023; Lopez-Lira and Tang, 2023; Engelberg et al., 2025). To ensure the generalizability of our results, we also

---

[8]https://github.com/HaoooWang/llm-knowledge-cutoff-dates

collect daily news headlines from the Finnhub API. Consistent with the out-of-knowledge-cutoff principle, our data collection begins on November 1, 2023. Following the methodology of Ke et al. (2019), we first exclude headlines associated with multiple firm tickers. To handle duplicate articles that arise from re-syndication across different news outlets, we retain news only from the most frequent source in our dataset, which is Yahoo Finance. Furthermore, we remove headlines containing the word "stock" or "stocks," as these titles may directly allude to stock price movements or trading volume, thus contaminating the signal.

We retain only headlines that feature a single organization. This is necessary because when we prompt the GPT model to assess the price implications for "the corresponding firm," headlines mentioning multiple or no specific firms would create ambiguity and confuse the model. Additionally, guided by Bybee et al. (2023), we exclude articles published on weekends. Finally, the filtered headlines are matched with other relevant stock and corporate financial data. The complete sample filtering procedure is detailed in Panel B of Table A.1.

2.2.4 LLM Generated Measurements

We design a suite of prompts to extract several measures from the aforementioned textual data, including sentiment, uncertainty, investment (the firm's outlook on future capital expenditures) and economy (the firm's outlook on the future macroeconomy). For each measure, we generate a quantitative score. The specific definitions for these scores and the complete prompts are available in the Appendix B.3 to B.6.

For the earnings call transcripts, we aggregate these scores by PERMNO and date. The matching rule is as follows: for calls held before the U.S. market close (16:00 ET), the announcement date (date 0) is set to the same calendar day. For calls held after the market close, the announcement date is set to the next trading day. This convention ensures that the stock return for the announcement date fully incorporates the information disclosed in the transcript.

Since a single stock may have multiple headlines on any given day, we average the scores

of all headlines for each PERMNO-date to create a daily score. We apply the same rule. Specifically, daily average scores are categorized as pre- or post-market-close based on publication times and are assigned an announcement date using the identical rule as for conference call transcripts.

3 Results

3.1 Anonymization Performance in Earnings Calls

We tokenize these call transcript data using the Xenova/GPT-4o tokenizer.[9] We then use three anonymization algorithms to identify entities in conference calls. In Panel A of Table 2, we report the conference call transcript characteristics. The average transcript length is 8,895 tokens, with 25th and 75th percentiles at 6,971 and 10,901 tokens, respectively. We also report the percentage of identified entities based on each anonymization algorithm. We find that the TRF approach identifies the highest percentage of entities, whereas LLM identifies the lowest percentage of entities.[10]

Next, we examine the effectiveness of various anonymization algorithms. Following past literature (e.g., Engelberg et al., 2025), we task large language models (LLMs) with identifying the masked firm and year identities of anonymized transcripts. Specifically, GPT-4o-mini attempts to deanonymize the data using the prompt detailed in the Appendix B.2, and we report these entity identification results in Panel B.

We find that transcripts anonymized by the SM algorithm offer the least protection; GPT-4o-mini successfully identifies 82.19% of firm identities and 44.51% of year identities. Transcripts anonymized by the TRF algorithm demonstrate improved anonymization performance, with GPT-4o-mini identifying only 57.42% of firm identities and 14.87% of year identities. Consistent with Engelberg et al. (2025), anonymization achieved through LLM-based

---

[9]https://huggingface.co/Xenova/gpt-4o

[10]It is worth noting that the percentage counts for the LLM approach is based on LLMs' output. The anonymized transcripts are sometimes shorter than the original transcripts due to truncation.

entity removal proves to be the most thorough method with respect to the comprehensive identification of firm and year, leaving only 19.32% of firm-level information and 20.69% of year information identifiable by GPT-4o-mini.

3.2 Sentiment Extraction

Our primary analysis investigates the impact of anonymization on the informational content of financial text. We extract the sentiment from transcripts using GPT-4o-mini for the original transcript, as well as the anonymized ones. We design a prompt that closely follows Engelberg et al. (2025). The prompt used to extract sentiment is detailed in the Appendix B.3. We tabulate the pairwise correlation between the extracted sentiment scores in Table 3. We find that the correlation between the sentiment score extracted from the raw and the anonymized transcripts ranges from 0.864 to 0.879, indicating that these signals remain highly correlated. Since the correlation between the anonymized and the raw score is not perfect, it potentially suggests the loss of information from the anonymization process.

To formally test severity of the information loss, we examine the association between sentiment scores, extracted from earnings call transcripts, and contemporaneous DGTW-adjusted stock returns ($DGTW_{i,t}$, as defined by Daniel et al. (1997)). Using the contemporaneous returns as the dependent variable is consistent with prior literature (e.g., Loughran and Mcdonald, 2011; Siano, 2025). A reduced relationship after anonymization would demonstrates the information loss resulting from anonymization.[11]

Empirically, we adopt the following regression model:

$$DGTW_{i,t} = \beta_1 \text{Sentiment}_{i,t} + \gamma X_{i,t} + \delta_t + \epsilon_{i,t} \tag{1}$$

The dependent variable is the announcement-day DGTW-adjusted stock return. The key independent variables are the sentiment scores extracted from the raw text, as well as from

---

[11]The use of contemporaneous stock returns is motivated by the fact that, compared to future returns, they capture the most complete and immediate market reaction to the sentiment expressed in the transcripts.

the texts anonymized using the SM, TRF, and LLM methods, respectively. Following (Tetlock et al., 2008; Engelberg et al., 2025), we include a standard set of control variables, $X_{i,t}$, in our model. These comprise analyst forecast error (FE), which prior studies have consistently shown to have significant explanatory power for abnormal returns (e.g., Bernard and Thomas, 1989; Hirshleifer et al., 2009), lagged DGTW-adjusted returns ($DGTW_{i,t-1}$, $DGTW_{i,t-2}$, $DGTW_{i,t-3}$, $DGTW_{i,t-22,t-4}$, and $DGTW_{i,t-253,t-23}$), and firm characteristics including the logarithm of firm size ($\ln(Size)$), the book-to-market ratio ($\ln(BM)$), and turnover ($\ln(Turnover)$). All regressions include date fixed effects. All independent variables are standardized, facilitating a direct comparison of coefficients across different text treatments and models.

In our regression analysis with controls (reported in Table 4), the sentiment score from the raw text, $Sentiment_{RAW}$, exhibits a strong and highly significant relationship with the contemporaneous stock return (coeff. = 2.487, t = 18.832), confirming that LLMs effectively capture sentiment signals. While sentiment scores from independently tested anonymized texts (SM, TRF, and LLM) also show a significant association with returns (see Columns 3–5), their coefficients are attenuated (SM coeff. = 2.373, TRF coeff. = 2.331, LLM coeff. = 2.435) and their respective adjusted R^2 values decline, indicating information loss across all three methods. This loss is particularly pronounced for the TRF-processed text, where the adjusted R^2 falls from 0.132 to 0.124 relative to the raw text.

To more formally compare the power of sentiment scores, we conduct pairwise regressions that include both raw and anonymized sentiment scores (see Columns 5–7). We find the coefficient of $Sentiment_{RAW}$ remains large and highly significant, whereas the coefficients on the anonymized sentiment scores contract sharply. For example, the $Sentiment_{TRF}$ coefficient plummets from 2.331 to 0.775 (see Column 6), representing a 67% informational loss. This result indicates that the sentiment-relevant information in anonymized text is largely subsumed by that of the original text, providing strong evidence that entity removal causes a significant loss of sentiment-relevant information. As reported in Table A.6, our

results are also robust without including control variables.

3.3 The Role of Different Types of Entities

While anonymization invariably results in information loss, the extent of this loss is likely heterogeneous across different types of information. As shown in Appendix Table A.4, entities such as NUMBERS and OBJECTS constitute a much larger proportion of our financial text corpus than PLACES and OTHERS. The varying roles and frequencies of different named entity types in the text suggest that their informational value may differ. Thus, we hypothesize that the degree of information loss from anonymization depends on the category of the entity being replaced. We anticipate that replacing core entities directly linked to firm fundamentals and quantitative performance, such as NUMBERS and OBJECTS, will result in a significantly greater loss of information than replacing entities that primarily provide contextual information, such as PLACES and OTHERS. To systematically test this hypothesis and identify the primary sources of information loss, we conduct the following analysis.

First, we mask each entity type, and task GPT-4o-mini to generate sentiment scores. In Table A.5, we tabulate the correlation of the sentiment scores extracted from the raw, fully anonymized, and partially anonymized transcripts. We find that the correlation between the raw and fully anonymized sentiment scores is lowest, while the correlation between the raw and place-masked sentiment scores is highest.

We further conduct horse race regression tests in Table 5 to identify the source of this information loss by selectively replacing different entity categories. Columns 1-5 present the regression results for the original text and for texts where NUMBERS, PLACES, OBJECTS, and OTHERS entities have been replaced. We observe a decline in both the regression coefficients and the adjusted R^2 values when the first three categories are replaced, indicating a tangible loss of information. Conversely, replacing the OTHERS category results in a

slight increase in the coefficient and R^2.[12] The results from the "horse race" regressions (Columns 6-9) reveal that the sources of information loss are not uniformly distributed. When only PLACE entities are replaced (Column 7), the sentiment score coefficients for the original versus the anonymized text are the closest (1.397 vs. 1.158), suggesting that the replacement of location information has the minimal impact on overall explanatory capability. In contrast, the information loss is most severe when replacing NUMBER entities (Column 6) and OBJECT entities (Column 8). In these cases, the coefficient for the original text's sentiment score is substantially larger than that of the anonymized text. This unveils a critical mechanism: in a financial context, numbers that directly quantify performance, scale, and expectations, as well as names directly associated with companies and key individuals, serve as the core information carriers for an LLM's assessment of market sentiment and corporate outlook.

3.4   The Role of Textual and Firm Characteristics

In our previous analyses, we establish that anonymization leads to a significant loss of explanatory and predictive value in text, a loss primarily attributable to the removal of numerical and object-class entities. We next investigate textual and firm characteristics that amplify or mitigate the extent of the information loss.

We measure the loss of information for each press release by the absolute difference between the two sentiment scores ($|\text{Sentiment}_{RAW} - \text{Sentiment}_{TRF}|$). This measure directly captures the magnitude of the "disagreement" in the LLM's judgment caused by the anonymization process. Table 6 presents the results of a series of univariate regressions using this divergence as the outcome.

Our regression variables include both textual and firm-level characteristics. The textual variables consist of the uncertainty scores extracted from both the raw text ($Uncertainty_{RAW}$)

---

[12]A potential explanation for this is that the LLM's information extraction from the original text may exhibit some degree of overfitting. By removing a small amount of information, this overfitting is mitigated, leading to a marginal improvement in explanatory power.

and the TRF-anonymized text ($Uncertainty_{TRF}$). The prompt used for extracting these scores was based on the definition provided by Loughran and Mcdonald (2011), which emphasizes a broad notion of linguistic imprecision rather than an exclusive focus on risk. The firm-level variables include the logarithm of firm size ($ln(Size)$), book-to-market ratio ($ln(BM)$), and turnover ($ln(Turnover)$). Additionally, we include a dummy variable indicating whether the company is listed on Nasdaq.

The findings reveal that textual uncertainty is a primary driver of divergence in sentiment judgments. As detailed in Columns 1 and 2 of Table 6, the uncertainty score derived from both RAW and TRF-anonymized texts is positively and significantly correlated with the sentiment score difference. This is supported by regression coefficients of 0.040 ($t = 12.080$) and 0.048 ($t = 13.452$), respectively. Furthermore, these models' adjusted R^2 values (0.042 and 0.059) significantly exceed those observed for all other variables.

These results suggest that when management statements are characterized by low linguistic ambiguity (e.g., "our revenue achieved 50% growth"), the sentiment signal is robust to anonymization. The removal of specific entities has a limited impact on the LLM's judgment. Conversely, when a text exhibits high levels of ambiguity, caution, or conditionality, access to knowledge about the entity help supply useful information to judge the sentiment of the texts. Entities, such as financial data or project names, function as "informational anchors." The anonymization process removes these anchors, compelling the LLM to make inferences from a semantically sparser context, which results in a greater deviation in its judgment. In essence, the informational value of entities is magnified in contexts of high uncertainty. Consequently, the impact of their removal is most pronounced under these conditions.

In addition to textual uncertainty, we explore whether firm fundamental characteristics and entity percentages moderate this information loss. The result in Column 3 indicates that firm size, measured as ln(Size), is negatively and significantly associated with the sentiment score difference (coeff. = -0.008). A plausible interpretation for this is the extensive representation of larger firms within the LLM's pre-training corpora. This exposure may provide

the model with a more developed prior understanding of these firms, thereby making its sentiment assessments more robust to the information loss induced by anonymization.

Furthermore, we explore how the specificity of a firm's product market influences this sentiment gap. We hypothesize that greater firm heterogeneity presents a dual challenge for LLMs. First, firms with highly unique products may be discussed less frequently in public texts, leading to a sparser representation within the LLM's training corpus. Second, due to their distinct market positions, the relationship between their operational narrative and financial outcomes can be more idiosyncratic and complex for an LLM to learn. To proxy for this heterogeneity, we employ two text-based measures from Hoberg and Phillips (2016), TNIC3TSIMM and TNIC3HHI, both constructed based on the Text-Based Network Industry Classification (TNIC) developed by Hoberg and Phillips (2016). TNIC3TSIMM measures a firm's total product similarity to its competitors (lower values indicate a more unique product portfolio), while TNIC3HHI is the Herfindahl-Hirschman Index of market concentration (higher values indicate a more specialized or concentrated market). According to Hoberg and Phillips (2010), firms that successfully differentiate their products in these ways tend to face less competition and achieve higher profitability. Given that these unique operational models are harder for an LLM to learn, its ability to accurately assess sentiment likely relies more heavily on named entities to anchor its analysis. Consequently, we expect the anonymization process to create a larger information gap for these firms. The results support this hypothesis. Column 4 shows that TNIC3TSIMM has a negative and significant coefficient (coeff. = -0.006), while Column 5 shows TNIC3HHI has a positive and significant coefficient (coeff. = 0.005). Both results indicate that firms with more unique product market characteristics experience a greater deviation in sentiment scores after anonymization.

Similar to the previous principle, analyst coverage serves as another indicator of a firm's exposure. A higher analyst coverage corresponds to increased attention and more public discourse surrounding the firm, which in turn reduces the sparsity of its representation in the LLM's training corpus, as discussed earlier. To quantify analyst coverage, we follow the

17

methodology outlined in He and Tian (2013). For each firm in our sample, we obtain the fiscal year-end date for 2021, and then calculate the average number of monthly earnings forecasts over the preceding 12 months. We use analyst forecast data for 2021 to avoid any look-ahead bias in our analysis. This average represents the firm's analyst coverage. Column 6 reports a negative and significant coefficient for analyst coverage (coeff. = -0.005), which corroborates our earlier hypothesis: greater exposure leads to a smaller gap and less information loss.

Finally, the proportion of a transcript composed of named entities is also an important factor. As shown in Column 7, a higher percentage of entities in a transcript is positively and significantly associated with the sentiment gap (coeff. = 0.004). This aligns with the intuition that a greater volume of entities corresponds to a larger amount of information being removed during the anonymization process, which in turn is likely to result in a larger sentiment deviation.

3.5 Alternative LLMs

To test whether our findings are contingent on a specific LLM, we further conduct the sentiment extraction task using two alternative LLMs: GPT-4o and GPT-o3-mini. Compared to the workhorse model in our prior analyses (i.e., GPT-4o-mini), GPT-4o is a larger model with more parameters. It also exhibits much stronger ability in textual tasks.[13] GPT-o3-mini is a new class of "thinking model" that is designed to address complicated reasoning problems. While it is unclear if financial textual analysis involves complex reasoning, we include this model to ensure that our prior results can be relevant for this new class of models.

We repeat the analyses conducted in Table 4 using the two alternative models. We report our results in Table 7. We also repeat our baseline results using GPT-4o-mini as a reference (Columns 1, 4, and 7 repeat the columns 1, 3, and 6 in Table 4). First, across all three mod-

---

[13]Based on the assessments provided by OpenAI, GPT-4o scores 88.7 on the MMLU benchmark, while GPT-4o-mini obtains 82.0. See https://openai.com/index/GPT-4o-mini-advancing-cost-efficient-intelligence/

els, the sentiment score derived from the original text consistently demonstrates significant explanatory power for stock returns (Columns 1-3). Second, when the analysis is performed on the TRF-anonymized text (Columns 4-6), we observe a uniform and marked decline in both the regression coefficients and the adjusted R^2. This indicates that information loss from anonymization is a widespread and robust phenomenon, detectable by all tested LLMs. Notably, the magnitude of this impact varies across models, indicating that different LLMs exhibit different sensitivities to information loss when processing the same anonymized text. The GPT-4o model is the most severely affected; in the respective regression, it loses approximately 20% of the sentiment's explanatory power, with its coefficient dropping from 2.528 to 2.005 and its adjusted R^2 decreasing from 0.131 to 0.108. Despite these variations in magnitude, the directional finding of information loss remains unequivocal. Importantly, the "horse race" regressions (Columns 7-9) provide the most decisive evidence. In this direct comparison, the sentiment coefficient for the raw text consistently and significantly outweighs that of the anonymized text, irrespective of the evaluation model. We therefore conclude that the observed information loss is a fundamental consequence of the anonymization process itself, rather than an artifact of a specific model's analytical limitations. In conclusion, our findings remain robust to the choice of the underlying LLM.

3.6 Can Anonymization Affect the Extraction of Other Information?

To comprehensively assess the breadth and economic significance of information loss induced by anonymization, our analysis must extend beyond the explanation of contemporaneous stock returns. If named entities are indeed primary carriers of information within the text, their removal should also impair the text's ability to predict a firm's future fundamentals and risk. This section broadens our investigation to three key domains to verify the pervasiveness of this information loss: future stock volatility, future capital investment, and future corporate operating performance. Table 8 shows all regression results.

First, prior research establishes a positive link between the textual uncertainty in 10-K reports and subsequent stock return volatility (Loughran and Mcdonald, 2011). Inspired by their findings, we investigate whether anonymization impairs the earnings call transcript text's ability to predict future stock volatility. Column 1 of Table 8 examines the predictive power of an Uncertainty score extracted from the text for future realized volatility over the subsequent 22 trading days (Vol_{post}). The uncertainty score variable is the same as that used in Section 3.4. It is extracted using a prompt designed based on the definition from Loughran and Mcdonald (2011). In our empirical specification, we follow Loughran and Mcdonald (2014) and choose pre-announcement volatility, alpha, absolute abnormal returns, the natural logarithm of the firm's market capitalization, and the book-to-market ratio as control variables. A detailed definition is provided in Table A.2.

As shown in Column 1, we include both $Uncertainty_{RAW}$ and $Uncertainty_{TRF}$ scores in a horse-race regression, the coefficient on $Uncertainty_{RAW}$ (0.026) remains significant, whereas the coefficient on $Uncertainty_{TRF}$ (0.012) becomes statistically insignificant ($t = 0.777$). The result is consistent with the findings of (Loughran and Mcdonald, 2011) that more uncertain language from management is associated with higher future stock price volatility. And more importantly, this result suggests that once specific entities such as project names, figures, and personal names are redacted, the LLM's ability to infer future uncertainty is severely weakened, rendering its signal value negligible in the presence of the signal from the original text.

Next, we turn our focus to corporate physical investment decisions, following the approach of Jha et al. (2024a). We extract corporate expectations regarding future investment (Investment) from earnings call transcripts. Column 2 compares the predictive power for future investments of Investment scores from RAW and TRF-anonymized management's forward-looking statements on capital expenditures, with the dependent variable being capital expenditures in quarter $t+2$ (two quarters ahead). The results show that the coefficient on $Investment_{RAW}$ (0.043) is larger than that on $Investment_{TRF}$ (0.033), with both re-

taining high statistical significance. The observed decline in the coefficient's magnitude is consistent with the findings of Jha et al. (2024a). This indicates that the anonymized text still contains valid information about future investment trajectories, but that information loss is also present.

Finally, we examine the predictive power of textual information for a firm's core future operating performance, drawing on Jha et al. (2024b). Columns 3 and 4 use future sales growth ($SaleChange_t$) and value-added growth ($ValueAddChange_t$), respectively, as dependent variables to test the predictive power of a macroeconomic sentiment (i.e., $Economy$) score extracted from the text. The findings in these tables are highly consistent and further corroborate the pattern observed in the volatility and capital expenditure analysis. Both the $Economy_{RAW}$ score from the original text and the $Economy_{TRF}$ score from the anonymized text demonstrate strong predictive power for future sales and value-added growth, and the predictive coefficient of $Economy_{RAW}$ is robustly higher than that of $Economy_{TRF}$. This once again confirms that anonymization leads to a degradation of informational value.

Taken together, these results provide compelling evidence that the information loss from anonymization is a pervasive phenomenon that extends beyond sentiment analysis. The value of other types of information extracted from earnings call transcripts is also compromised in this process, demonstrating the widespread nature of this degradation across various extraction tasks.

3.7 Alternative Textual Data

Our prior analyses focus on extracting information from conference call transcripts. In this subsection, we turn our attention to news headlines, another set of highly utilized textual data in finance setting (see e.g., Bybee et al., 2023; Lopez-Lira and Tang, 2023; Engelberg et al., 2025, for analyses using news or news headlines). Compared to the verbose nature of earnings call transcripts, news headlines are extremely short in length but feature a much higher density of information, particularly entity-related information.

We begin with a descriptive analysis of our news headline sample, with results presented in Table A.7. Our final sample comprises 52,716 news headlines from Yahoo Finance, corresponding to 30,950 firm-day observations (as a firm may have multiple news items on a given day), covering 2,052 unique firms over 292 trading days. The textual features of these headlines stand in stark contrast to those of earnings call transcripts. The average headline length is merely 15.55 tokens, with a highly concentrated distribution (a 25th percentile of 12 tokens and a 75th percentile of 18 tokens). Using the TRF model to anonymize all news headlines, we find that recognized entities account for an average of 38.30%, which is substantially higher than the 16.18% mean observed in earnings call transcripts under the same anonymization approach. This indicates that news headlines are highly condensed with entity-specific information, suggesting that anonymization could have an even more pronounced impact on their informational value. The detailed sample selection criteria for news headlines are provided in Panel B of Table A.1.

Following the methodology used for earnings call transcripts, we designed a specific prompt to extract sentiment from news headlines, which can be found in the Appendix B.7. For each firm-day, we average the sentiment scores from all associated news headlines to derive a daily news sentiment score.

Table 9 presents the regression results for explaining contemporaneous stock returns using sentiment from news headlines. In both univariate (Columns 1 and 2) and multivariate regressions with standard controls (Columns 4 and 5), sentiment scores derived from raw ($Sentiment_{RAW}^{News}$) and TRF-anonymized ($Sentiment_{TRF}^{News}$) texts are significant positive signals of contemporaneous stock returns. This indicates that LLMs can extract market sentiment signals from short-form texts such as news headlines. Consistent with the findings from our analysis of earnings call transcripts, the regression coefficients and adjusted R^2 values are lower for the TRF-anonymized texts compared to the raw texts.

A key finding emerges from the "horse race" regressions (Columns 3 and 6), which include sentiment scores from both raw and anonymized texts in the same model specification.

The results show clear evidence of information loss from anonymization. In the model without control variables (Column 3), the coefficient on $Sentiment_{RAW}^{News}$ (0.315) remains statistically significant, whereas the coefficient on $Sentiment_{TRF}^{News}$ (0.135) is attenuated to less than one-third of its magnitude in the standalone regression. This pattern persists after incorporating control variables (Column 6), where the coefficient on $Sentiment_{RAW}^{News}$ (0.314) remains substantially larger than that on $Sentiment_{TRF}^{News}$ (0.131).

Taken together, the results from the news headline analysis corroborate the findings from the earnings call transcripts. This suggests that the loss of information from anonymization is not an idiosyncratic phenomenon confined to a specific type of financial text. Whether the source is a lengthy earnings call transcript or a concise news headline, the named entities within the text are a significant source of its explanatory and predictive value. Consequently, removing these entities to mitigate potential look-ahead bias results in a corresponding loss of the text's intrinsic informational content.

3.8 Pre and Post Cut-off Date Sample

We next extend the earnings call transcript sample to include both the pre- and post-cutoff periods. Our earlier analyses focused on the post-cutoff period (i.e, November 2023–December 2024) to cleanly identify the effect of information loss resulting from anonymization.[14]

In this subsection, we extend our analysis to include the pre-cutoff period (November 2022–October 2023), allowing us to assess the impact of anonymization in a setting that is closer to prior research that uses anonymization to mitigate look-ahead bias. We hypothesize that while removing firm entities from this data may reduce look-ahead bias, it will also cause information loss during the extraction process. Consequently, we hypothesize that if look-ahead bias significantly inflates raw sentiment signals in the pre-cutoff period, the reduction in informativeness observed between raw and anonymized signals will be more pronounced in the pre-cutoff period compared to the post-cutoff period. This larger reduction would

---

[14]All LLMs used in this study share a knowledge cutoff of October 2023.

reflect both the inherent information loss from anonymization and the additional mitigation of look-ahead bias.

To ensure comparability between the pre- and post-cutoff periods, we retain only firms that appear in both samples. Throughout this section, we use GPT-4o-mini for information extraction and recognition testing to generate the main results. This is the same model used in previous sections, which facilitates a direct comparison. We also document the results of a robustness test in the Appendix Table A.8 and using GPT-4o for information extraction and recognition testing, which demonstrates a similar pattern.

We first visually present the variation in the relative performance of the raw and anonymized signals over time. We conduct quarter-by-quarter regressions following the specification of Table 4. Figure 2 plots the quarterly time series of coefficients on the sentiment score. The sentiment score was extracted from RAW and TRF-anonymized transcripts using GPT-4o-mini, and the coefficients were estimated separately in regressions with controls. All independent variables are standardized within each quarter, and both specifications are estimated quarter by quarter, so that the coefficients are comparable over time.

The knowledge cutoff (2023 Q4) is marked with a dashed vertical line. The coefficient of $Sentiment_{RAW}$ (in dark blue) consistently exceeds that of $Sentiment_{TRF}$ (in red), indicating persistent and economically meaningful information loss in both the pre- and post-cutoff periods. The figure also reports the coefficient difference. The average magnitude of the difference is broadly similar on either side of the knowledge cutoff.

We then formally assess the impact of the knowledge cutoff and firm recognition on the informativeness of sentiment scores by estimating the following regression:

$$DGTW_{i,t} = \beta_1 Sentiment_{i,t} + \beta_2(Sentiment_{i,t} \times Z_{i,t}) + \beta_3 Z_{i,t} + \gamma X_{i,t} + \delta_t + \epsilon_{i,t}, \quad (2)$$

This regression revises Regression (1) by introducing an interaction term between the sentiment signal and a proxy $Z_{i,t}$ for the effect of anonymization or look-ahead bias. The variable

$Sentiment_{i,t}$ denotes either $Sentiment_{RAW}$ or $Sentiment_{TRF}$, and $X_{i,t}$ represents the same set of control variables as in Table 4.

To evaluate the impact of the knowledge cutoff and whether models can correctly identify firm information, we consider two interaction variables: pre-cutoff indicator Pre, which equals 1 if the transcript falls in the pre-knowledge-cutoff period (November 2022 to October 2023) and 0 otherwise and the firm-recognition indicator $Recognition_{Firm}$, which equals 1 if the model correctly identifies the firm name in the TRF-anonymized transcript and 0 otherwise.

Table 10 implements specification (2) using GPT-4o-mini for sentiment extraction and firm recognition. In columns (1) and (2), we set $Z_{i,t} = Pre$ and use the full sample spanning both the pre- and post-cutoff periods in our regressions. In column (1), we regress returns on $Sentiment_{RAW}$ and its interaction with Pre. In column (1), the coefficient on $Sentiment_{RAW}$ is large and highly significant (2.510, $t = 19.559$), confirming that raw-text sentiment is strongly informative on average. The interaction $Sentiment_{RAW} \times Pre$ is significantly negative (-0.584, $t = -3.327$), implying that the predictive content of raw sentiment is somewhat weaker in the pre-cutoff period than in the post-cutoff period. Column (2) shows a similar pattern for $Sentiment_{TRF}$: the relationship between the extracted sentiment and stock returns is strongly positive (2.364, $t = 17.718$), and the interaction $Sentiment_{TRF} \times Pre$ is again negative and statistically significant (-0.489, $t = -2.745$).

These results suggest that look-ahead bias does not play an important role in our sentiment extraction setting. If look-ahead bias strongly affects the extracted signals, one would expect the coefficients on $Sentiment_{RAW} \times Pre$ (and possibly on $Sentiment_{TRF} \times Pre$) to be positive and significant, reflecting artificially inflated predictive power in the pre-cutoff period when the LLM's training data include future outcomes. Instead, the significantly negative interactions indicate that sentiment, both raw and anonymized, is slightly less informative in the pre-cutoff window. This pattern is difficult to reconcile with a dominant role for look-ahead bias.

Next, we explicitly examine whether models' ability to identify firms from anonymized transcripts leads to significantly improved power in sentiment extraction. Prior studies largely use the ability to identify firms from anonymized texts to assess the severity of look-ahead bias when extracting information from financial texts (Engelberg et al., 2025). We hypothesize that if the LLM successfully identifies the firm name, it would recover some lost context, thereby enhancing the informativeness of the extracted sentiment. In Columns (3) and (4) of Table 10, we interact extracted signals with $Recognition_{Firm}$ for pre- and post-knowledge cutoff samples, respectively. In the pre-cutoff period (reported in column (3)), $Sentiment_{TRF}$ remains highly significant (1.995, $t = 9.312$), whereas $Sentiment_{TRF} \times Recognition_{Firm}$ is statistically insignificant. In the post-cutoff period (reported in column (4)), $Sentiment_{TRF}$ continues to exhibit strong predictive power (2.324, $t = 12.677$). However, $Sentiment_{TRF} \times Recognition_{Firm}$ is statistically insignificant. The lack of a systematic relationship between firm-name recognition and the informativeness of anonymized sentiment suggests either that whether GPT-4o-mini can correctly identify the firm name is a poor proxy for the presence or severity of look-ahead bias or information loss from anonymization.[15]

Appendix Table A.8 applies the same specification (2) to sentiment scores and recognition outcomes generated by GPT-4o. The qualitative patterns closely mirror those in Table 10. Raw-text sentiment remains more informative than anonymized sentiment, the interaction $Sentiment_{RAW} \times Pre$ is again significantly negative, and interactions involving $Recognition_{Firm}$ are generally insignificant. The only marginal exception is a weakly positive main effect of $Recognition_{Firm}$ in the post-cutoff period, which is not robust once we consider interactions with the sentiment score.

---

[15]Appendix Table A.9 explores the impact of anonymization using the Gap measure (the absolute difference between sentiment scores from RAW and TRF-anonymized transcripts) as $Z_{i,t}$ in specification (2). We consistently find that the interaction $Sentiment_{TRF} \times Gap$ is negative and highly significant across various models (GPT-4o-mini and GPT-4o) and sample periods (pre- and post-cutoff). These results indicate that larger discrepancies between raw and anonymized sentiment scores systematically correlate with weaker explanatory power of anonymized sentiment, even in samples where forward-looking bias should be minimal. Unlike the limited role of firm recognition, the Gap measure proves to be a reliable and economically meaningful proxy for anonymization-induced information loss.

4 Conclusion

This study provides comprehensive empirical evidence that anonymization, a widely adopted technique to mitigate look-ahead bias in financial text analysis with LLMs, introduces a significant side effect: information loss. By systematically removing named entities, the process strips away essential context, thereby degrading the explanatory and predictive power of the extracted textual signals.

Our analysis, centered on earnings call transcripts, demonstrates that sentiment scores from raw text are strong signals of announcement-day stock returns. However, this explanatory ability is substantially attenuated after applying various anonymization methods. "Horse race" regressions confirm that the explanatory content of anonymized signals is largely subsumed by that of the raw text. We trace this degradation primarily to the removal of NUMBERS and OBJECTS (e.g., firm and executive names), which serve as critical anchors for contextual interpretation. Furthermore, we establish that the divergence between raw and anonymized signals is a reliable proxy for this loss, which we find is more severe in texts with high linguistic uncertainty or those from smaller, more specific firms.

These findings are robust across multiple dimensions. The information loss persists across different anonymization models, various LLMs used for signal extraction (GPT-4o-mini, GPT-4o, and GPT-o3-mini), and extends beyond return explanation to forecasting future stock volatility, capital expenditures, and sales growth. Crucially, we replicate our core results using an alternative text corpus of short, entity-dense news headlines, confirming that the observed information loss is a fundamental consequence of the anonymization process itself, rather than an artifact of a specific textual domain. Finally, the degree of information loss remains nearly constant in both pre- and post-cutoff periods. This not only provides strong evidence of anonymization's direct impact but also suggests that look-ahead bias may be less severe than previously expected.

Our research challenges the default reliance on full anonymization in financial applica-

27

tions of LLMs. It highlights an important trade-off between mitigating look-ahead bias and preserving the intrinsic informational content of text. Future research should focus on developing methodologies that can effectively balance these competing objectives, ensuring that efforts to achieve methodological rigor do not inadvertently discard valuable information.

References

Bernard, V. L. and Thomas, J. K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*.

Breitung, C. and Müller, S. (2025). Global business networks. *Journal of Financial Economics*.

Brown, S., Hillegeist, S. A., and Lo, K. (2004). Conference calls and information asymmetry. *Journal of Accounting and Economics*, 37(3):343–366.

Bybee, L., Kelly, B., and Su, Y. (2023). Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies*, 36(12):4759–4787.

Chen, S., Peng, L., and Zhou, D. (2025). Wisdom or whims? decoding the language of retail trading with social media and ai. *Asian Bureau of Finance and Economic Research*.

Chen, Y., Kelly, B. T., and Xiu, D. (2022). Expected returns and large language models. *Available at SSRN 4416687*.

Clayton, C., Coppola, A., Maggiori, M., and Schreger, J. (2025). Geoeconomic pressure. *Columbia Business School Research Paper Forthcoming, Stanford University Graduate School of Business Research Paper Forthcoming*.

Daniel, K., Grinblatt, M., Titman, S., and Wermers, R. (1997). Measuring mutual fund performance with characteristic-based benchmarks. *The Journal of Finance*.

Engelberg, J., Manela, A., Mullins, W., and Vulicevic, L. (2025). Entity neutering. *Available at SSRN*.

Frankel, R., Johnson, M., and Skinner, D. J. (1999). An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research*, 37(1):133–150.

Garcia, D., Hu, X., and Rohrer, M. (2023). The colour of finance words. *Journal of Financial Economics*.

Glasserman, P. and Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis. *arXiv preprint arXiv:2309.17322*.

He, J. J. and Tian, X. (2013). The dark side of analyst coverage: The case of innovation. *Journal of financial economics*, 109(3):856–878.

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

Hirshleifer, D., Lim, S. S., and Teoh, S. H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of political economy*, 124(5):1423–1465.

Jha, M., Qian, J., Weber, M., and Yang, B. (2024a). Chatgpt and corporate policies. *National Bureau of Economic Research*.

Jha, M., Qian, J., Weber, M., and Yang, B. (2024b). Harnessing generative ai for economic insights. *arXiv preprint arXiv:2410.03897*.

Ke, Z. T., Kelly, B. T., and Xiu, D. (2019). Predicting returns with text data. *National Bureau of Economic Research*.

Kim, A., Muhn, M., and Nikolaev, V. (2023). From transcripts to insights: Uncovering corporate risks using generative ai. *arXiv preprint arXiv:2310.17721*.

Levy, B. (2024). Caution ahead: Numerical reasoning and look-ahead bias in ai models. *Available at SSRN 5082861*.

Lopez-Lira, A. and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). The memorization problem: Can we trust llms' economic forecasts? *arXiv preprint arXiv:2504.14765*.

Loughran, T. and Mcdonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*.

Loughran, T. and Mcdonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. *National Bureau of Economic Research*.

Sarkar, S. K. (2024). Storieslm: A family of language models with time-indexed training data. *Available at SSRN 4881024*.

Sarkar, S. K. (2025). Economic representations.

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

Siano, F. (2025). The news in earnings announcement disclosures: Capturing word context using llm methods. *Management Science*.

Tetlock, P. C. (2010). Does public financial news resolve asymmetric information? *The Review of Financial Studies*, 23(9):3520–3557.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*.

Figure 1: **Anonymization Pipline**

This figure shows two text anonymization approaches applied to a sample from an Apple Inc. earnings call transcript. Approach A shows a Named Entity Recognition (NER) approach using spaCy models (SM and TRF). This method employs a global mapping strategy, where each unique entity is consistently replaced by a numbered placeholder (e.g., PERSON_1, ORG_2) according to its order of appearance to maintain readability. Approach B demonstrates an approach using a Large Language Model (LLM), specifically GPT-4o-mini, to perform the anonymization through designed prompts.

Figure 2: **Quarterly Time Series of Sentiment Coefficients**

This figure plots the quarterly time series of coefficients on sentiment scores. The scores are extracted from RAW and TRF-anonymized earnings call transcripts using GPT-4o-mini. The coefficients for the raw and anonymized samples are estimated separately from regressions with controls, following the specification in Table 4.

Table 1: **Comparison of Text Anonymization Models**

This table shows a comparative analysis of anonymization outputs, displaying the original raw text alongside versions anonymized by three distinct models: SM, TRF, and LLM (GPT-4o-mini). Replaced entities in the anonymized text are highlighted in bold for clarity.

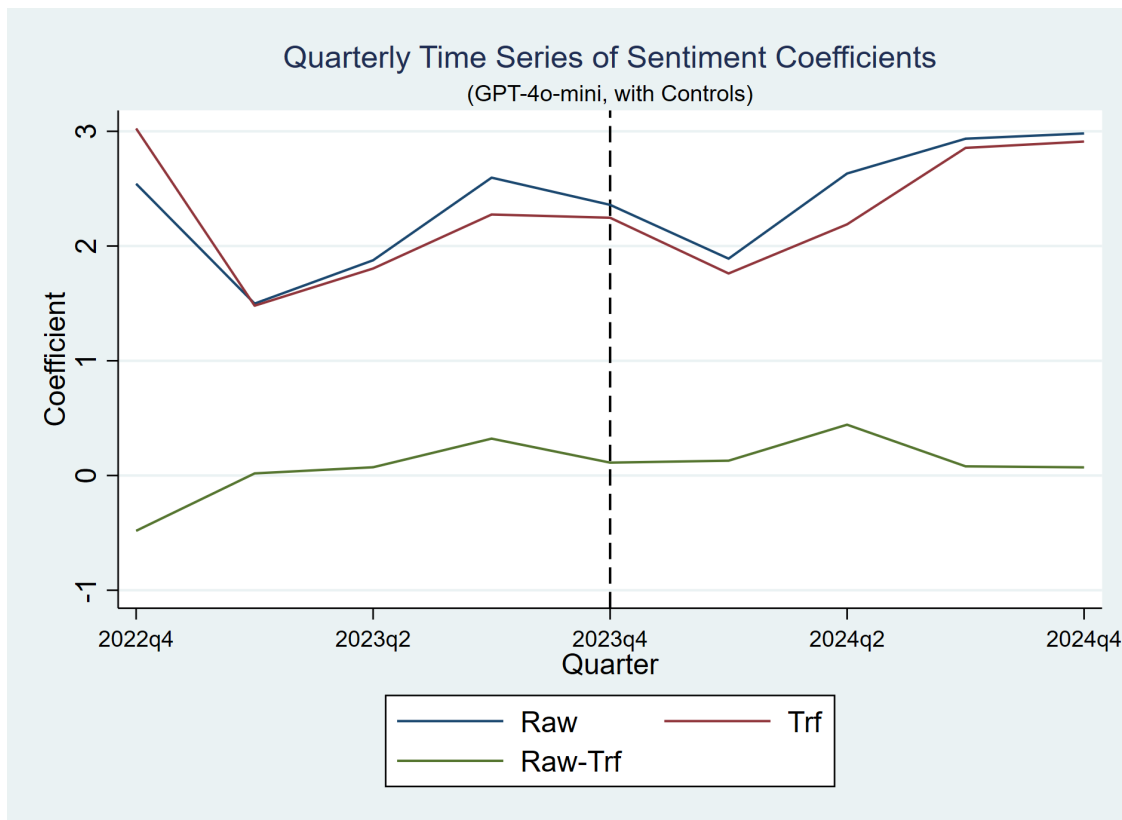| Raw Text	SM Anonymized	TRF Anonymized	LLM Anonymized
Please note that some of the information you'll hear during our discussion today will consist of forward-looking statements, including, without limitation, those regarding revenue, gross margin, operating expenses, other income and expense, taxes, capital allocation and future business outlook, including the potential impact of macroeconomic conditions on the firm's business and results of operations. These statements involve risks and uncertainties that may cause actual results or trends to differ materially from our forecast. For more information, please refer to the risk factors discussed in Apple's most recently filed annual report on Form 10-K and the form 8-K filed with the SEC today along with the associated press release. Apple assumes no obligation to update any forward-looking statements, which speak only as of the date they are made.	Please note that some of the information you'll hear during our discussion **DATE_1** will consist of forward-looking statements, including, without limitation, those regarding revenue, gross margin, operating expenses, other income and expense, taxes, capital allocation and future business outlook, including the potential impact of macroeconomic conditions on the firm's business and results of operations. These statements involve risks and uncertainties that may cause actual results or trends to differ materially from our forecast. For more information, please refer to the risk factors discussed in **ORG_1**'s most recently filed **DATE_2** report on Form 10-K and the form 8-K filed with the **ORG_2** **DATE_1** along with the associated press release. Apple assumes no obligation to update any forward-looking statements, which speak only as of the date they are made.	Please note that some of the information you'll hear during our discussion **DATE_1** will consist of forward-looking statements, including, without limitation, those regarding revenue, gross margin, operating expenses, other income and expense, taxes, capital allocation and future business outlook, including the potential impact of macroeconomic conditions on the firm's business and results of operations. These statements involve risks and uncertainties that may cause actual results or trends to differ materially from our forecast. For more information, please refer to the risk factors discussed in **ORG_1**'s most recently filed **DATE_2** report on Form 10-K and the form 8-K filed with the **ORG_2** **DATE_1** along with the associated press release. **ORG_1** assumes no obligation to update any forward-looking statements, which speak only as of the date they are made.	Please note that some of the information you'll hear during our discussion today will consist of forward-looking statements, including, without limitation, those regarding revenue, gross margin, operating expenses, other income and expense, taxes, capital allocation and future business outlook, including the potential impact of macroeconomic conditions on the firm's business and results of operations. These statements involve risks and uncertainties that may cause actual results or trends to differ materially from our forecast. For more information, please refer to the risk factors discussed in **ORG_1**'s most recently filed annual report on **FORM_1** and the **FORM_2** filed with the SEC today along with the associated press release. **ORG_1** assumes no obligation to update any forward-looking statements, which speak only as of the date they are made.

Table 2: **Conference Call Entity Count and Anonymization Performance**

This table presents the summary statistics for conference call transcripts. In Panel A, we report the mean, standard deviation, and percentile distributions of earnings call transcript lengths and entity percentages identified by three methods: SM, TRF, and LLM (GPT-4o-mini). Lengths are measured in number of tokens. Entity percentages for SM and TRF methods are calculated by dividing the token count of identified entities in the raw text by the total token count of the entire raw transcript. Entity% (LLM) is calculated using the anonymized transcript, where the percentage represents the ratio of anonymized entity tokens to the total tokens in the anonymized version. In Panel B, we report the percentage of correctly recognized items, including Firm, Year, Firm and Year, and Firm or Year—from the transcripts under different anonymization approaches, entity replacement categories, and recognition methods.

Panel A: Transcript Length and Entity Percentages					
	Mean	**Std. Dev.**	**25th Pct.**	**Median**	**75th Pct.**
Length	8,895	2,532	6,971	9,016	10,901
Entity% (SM)	15.64	2.73	13.76	15.41	17.15
Entity% (TRF)	16.18	2.80	14.26	15.94	17.79
Entity% (LLM)	13.46	2.91	11.51	13.24	15.11

Panel B: Recognition Ratio			
Firm	82.19	57.42	19.32
Year	44.51	14.87	20.69
Firm and Year	36.82	8.64	4.21
Firm or Year	89.88	63.65	35.79
Anonymization Approach	SM	TRF	GPT-4o-mini
Entity Remove	ALL	ALL	ALL
Recognition Approach	GPT-4o-mini	GPT-4o-mini	GPT-4o-mini

Table 3: **Correlations of Sentiment Signals**

This table shows correlations among sentiment scores extracted from raw transcripts (RAW) and anonymized transcripts from different anonymization approaches (SM, TRF, and LLM). All sentiment measures are extracted using the GPT-4o-mini model. All values are Pearson correlation coefficients.

	$Sentiment_{RAW}$	$Sentiment_{SM}$	$Sentiment_{TRF}$	$Sentiment_{LLM}$
$Sentiment_{RAW}$	1.000			
$Sentiment_{SM}$	0.879	1.000		
$Sentiment_{TRF}$	0.864	0.910	1.000	
$Sentiment_{LLM}$	0.872	0.880	0.869	1.000

Table 4: **Explanatory Power Comparison of Sentiment Scores with controls**

This table presents the results of regressions with DGTW-adjusted returns on the earnings conference call day as dependent variables. The key independent variables are sentiment measures extracted using GPT-4o-mini from earnings conference call transcripts. Our control variables including standard drivers of stock returns, including analyst forecast error (FE), past returns at various horizons, and firm characteristics such as size, book-to-market ratio, and turnover. Columns (1) through (4) augment this baseline model by individually adding the sentiment scores derived from the four text sources. Columns (5) through (7) include both the sentiment extracted from raw transcripts and from an entity-removed transcripts. All variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized. All regressions include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

| VARIABLES | Dependent Variable: $DGTW_{i,t}$ | | | | | | |
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$Sentiment_{RAW}$	2.487***				1.743***	1.822***	1.484***
	(18.832)				(6.859)	(8.475)	(5.467)
$Sentiment_{SM}$		2.373***			0.854***		
		(17.635)			(3.418)		
$Sentiment_{TRF}$			2.331***			0.775***	
			(17.287)			(3.662)	
$Sentiment_{LLM}$				2.435***			1.156***
				(18.522)			(4.374)
FE	2.054***	2.064***	2.073***	2.050***	2.046***	2.048***	2.038***
	(12.622)	(12.550)	(12.595)	(12.426)	(12.589)	(12.608)	(12.532)
$DGTW_{i,t-1}$	-0.335***	-0.354***	-0.345***	-0.345***	-0.347***	-0.344***	-0.346***
	(-2.838)	(-2.975)	(-2.877)	(-2.852)	(-2.929)	(-2.901)	(-2.900)
$DGTW_{i,t-2}$	-0.050	-0.005	-0.009	-0.028	-0.040	-0.043	-0.047
	(-0.350)	(-0.033)	(-0.060)	(-0.190)	(-0.279)	(-0.299)	(-0.324)
$DGTW_{i,t-3}$	0.132	0.135	0.165	0.145	0.125	0.135	0.129
	(1.039)	(1.056)	(1.301)	(1.140)	(0.991)	(1.069)	(1.023)
$DGTW_{i,t-22,t-4}$	-0.419***	-0.402***	-0.393***	-0.386***	-0.426***	-0.425***	-0.419***
	(-3.107)	(-2.944)	(-2.875)	(-2.849)	(-3.150)	(-3.140)	(-3.098)
$DGTW_{i,t-253,t-23}$	-0.331**	-0.300**	-0.290**	-0.277**	-0.347**	-0.347**	-0.338**
	(-2.348)	(-2.195)	(-2.120)	(-1.985)	(-2.499)	(-2.494)	(-2.411)
$ln(Size)$	-0.434***	-0.388***	-0.364***	-0.392***	-0.428***	-0.422***	-0.427***
	(-3.673)	(-3.327)	(-3.156)	(-3.430)	(-3.634)	(-3.574)	(-3.665)
$ln(BM)$	-0.194*	-0.201*	-0.194*	-0.198*	-0.194	-0.191	-0.192*
	(-1.657)	(-1.695)	(-1.656)	(-1.703)	(-1.650)	(-1.633)	(-1.654)
$ln(Turnover)$	0.201*	0.204*	0.196*	0.185	0.204*	0.201*	0.195*
	(1.810)	(1.838)	(1.746)	(1.603)	(1.838)	(1.807)	(1.728)
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7352	7352	7352	7352	7352	7352	7352
Adj.R^2	0.132	0.126	0.124	0.129	0.133	0.133	0.135

Table 5: **Comparing Explanatory Power Under Anonymizing Different Categories of Entities**

This table investigates the source of information loss by examining the impact of anonymizing specific entity categories. The selective replacement of these entities is performed using the TRF model, based on the definitions for each category provided in Appendix Table A.4. The dependent variable is the DGTW-adjusted stock return ($DGTW_{i,t}$). Columns (1) through (5) present results for sentiment scores derived from texts where only one category of entities—NUMBERS, PLACES, OBJECTS, or OTHERS—has been replaced, benchmarked against the original RAW text score. Columns (6) through (9) conduct "horse-race" regressions, directly comparing the explanatory power of the sentiment from the original RAW text against the sentiment from each of the selectively anonymized texts. Control variables are identical to those in Table 4. All variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized. All regressions include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

VARIABLES	Dependent Variable: $DGTW_{i,t}$								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$Sentiment_{RAW}$	2.487***					1.845***	1.397***	1.767***	1.009***
	(18.832)					(7.998)	(4.499)	(6.317)	(2.666)
$Sentiment_{NUM}$		2.336***				0.737***			
		(16.934)				(3.155)			
$Sentiment_{PLC}$			2.476***				1.158***		
			(18.194)				(3.687)		
$Sentiment_{OBJ}$				2.397***				0.795***	
				(18.375)				(2.993)	
$Sentiment_{OTH}$					2.527***				1.583***
					(20.124)				(4.442)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7352	7352	7352	7352	7352	7352	7352	7352	7352
Adj.R^2	0.132	0.124	0.131	0.127	0.134	0.133	0.133	0.133	0.135

Table 6: **Explanation of Sentiment Score Gap**

This table presents an analysis of the determinants of absolute difference between sentiment scores from RAW and TRF-anonymized transcripts. The dependent variable is $Gap_{i,t}$, defined as the absolute difference between the two scores ($|Sentiment_{RAW} - Sentiment_{TRF}|$). Each column reports a separate univariate regression of this sentiment difference on a potential explanatory variable. Columns (1) and (2) test the role of textual uncertainty (derived from both RAW and TRF-anonymized texts, respectively). Columns (3) through (7) examine the association with firms' information environment, including firm size, total product similarity (TNIC3TSIMM), product market Herfindahl-Hirschman Index (TNIC3HHI), analyst coverage, and the percentage of entities identified by the TRF model. All variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized. All regressions include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses.. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Gap_{i,t}$						
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$Uncertainty_{RAW}$	0.040***						
	(12.080)						
$Uncertainty_{TRF}$		0.048***					
		(13.452)					
$ln(Size)$			-0.008***				
			(-3.337)				
$TNIC3TSIMM$				-0.006***			
				(-3.416)			
$TNIC3HHI$					0.005**		
					(2.139)		
$Coverage$						-0.005**	
						(-2.003)	
$Entity\% (TRF)$							0.004*
							(1.682)
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7352	7352	7352	6860	6792	7352	7352
Adj.R^2	0.042	0.059	0.005	0.005	0.006	0.004	0.004

Table 7: **Information Loss of Alternative LLMs**

This table compares sentiment scores generated by three distinct models: GPT-4o-mini, GPT-4o, and GPT-o3-mini. All regressions include the full set of control variables and use the DGTW-adjusted stock return ($DGTW_{i,t}$) as the dependent variable. Columns (1) through (3) report the results for sentiment extracted from the raw transcripts by each LLM. Columns (4) through (6) report the results for sentiment extracted from the TRF-anonymized text. Finally, Columns (7) through (9) conduct "horse-race" regressions for each LLM-extracted signals. Control variables are identical to Table 4. All variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized. All regressions include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $DGTW_{i,t}$								
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$Sentiment_{RAW}$	2.487***	2.528***	2.400***				1.822***	2.105***	1.660***
	(18.832)	(19.972)	(18.968)				(8.475)	(13.157)	(9.504)
$Sentiment_{TRF}$				2.331***	2.005***	2.229***	0.775***	0.611***	0.942***
				(17.287)	(16.879)	(16.412)	(3.662)	(4.337)	(5.069)
LLM	4o-mini	4o	o3-mini	4o-mini	4o	o3-mini	4o-mini	4o	o3-mini
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7352	7352	7352	7352	7352	7352	7352	7352	7352
Adj.R^2	0.132	0.131	0.125	0.124	0.108	0.118	0.133	0.133	0.129

Table 8: **Information Loss in Alternative Tasks**

This table compares the predictive power of textual scores extracted from RAW and TRF-anonymized earnings-call transcripts in forecasting four forward-looking outcomes: future stock volatility, future capital investment, sales growth, and value-added growth. Each column reports the results from a "horse-race" regression where both the RAW and TRF-anonymized scores are included simultaneously. Column (1) examines realized post-announcement stock return volatility over the 22 trading days following the call (Vol_{post}) using the textual measure of managerial uncertainty ($Uncertainty_{RAW}$ and $Uncertainty_{TRF}$) as predictors, controlling for pre-announcement volatility (Vol_{pre}), pre-announcement alpha ($Alpha_{pre}$), absolute abnormal return ($|Abnormal\ Ret|$), firm size ($\ln(Size)$), and book-to-market ratio ($\ln(BM)$). Column (2) predicts firms' capital expenditures two quarters ahead ($Capx_{t+2}$) using investment expectations extracted from raw and anonymized texts ($Investment_{RAW}$ and $Investment_{TRF}$), controlling for current capital expenditure ($Capx_t$), financial leverage ($Leverage$), and firm size measured by the logarithm of total assets ($\ln(Total\ Asset)$). Column (3) investigates two-quarter sales growth ($SaleChange_t$) with firms' expressed outlook on the economy ($Economy_{RAW}$ and $Economy_{TRF}$) as key explanatory variables, controlling for lagged sales growth ($SaleChange_{t-2}$), book-to-market ratio (BM), firm size ($\ln(Total\ Asset)$), and asset tangibility ($Tangibility$). Column (4) presents analogous predictive regression of two-quarter value-added growth ($ValueAddChange_t$), using the same set of control variables as in Column (3), except that lagged sales growth ($SaleChange_{t-2}$) is replaced with lagged value-added growth ($ValueAddChange_{t-2}$). All variable detailed definitions are listed in Table A.2. All regressions include date fixed effects. All variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

VARIABLES	Vol_{post} (1)	$Capx_{t+2}$ (2)	$SaleChange_t$ (3)	$ValueAddChange_t$ (4)
$Uncertainty_{RAW}$	0.026**			
	(2.028)			
$Uncertainty_{TRF}$	0.012			
	(0.777)			
$Investment_{RAW}$		0.043***		
		(3.482)		
$Investment_{TRF}$		0.033**		
		(2.502)		
$Economy_{RAW}$			1.633***	2.889***
			(4.207)	(5.922)
$Economy_{TRF}$			1.380***	1.409***
			(3.451)	(3.150)
Controls	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	7255	5692	6630	6365
Adj.R^2	0.776	0.730	0.212	0.249

Table 9: **Anonymization and Information Loss from News Headlines**

This table examines if information loss also affects news headlines. The dependent variable is DGTW-adjusted stock returns ($DGTW_{i,t}$) on news days. The sentiment scores are derived from RAW and TRF-anonymized headlines. Columns (1) and (2) show univariate regressions for the RAW and TRF-anonymized scores, respectively. Column (3) conducts a horse-race regression between the RAW and TRF-anonymized sentiment scores. Columns (4) through (6) replicate this analysis with control variables, which include past returns at various horizons, and firm characteristics such as size, book-to-market ratio, and turnover. All variables are winsorized at the 1st and 99th percentiles and standardized. All regressions include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

VARIABLES	Dependent Variable: $DGTW_{i,t}$					
	(1)	(2)	(3)	(4)	(5)	(6)
$Sentiment_{RAW}^{News}$	0.444***		0.315***	0.439***		0.314***
	(17.477)		(4.664)	(17.353)		(4.633)
$Sentiment_{TRF}^{News}$		0.437***	0.135**		0.432***	0.131**
		(17.325)	(2.036)		(17.212)	(1.974)
$DGTW_{i,t-1}$				-0.036	-0.033	-0.036
				(-1.061)	(-0.981)	(-1.060)
$DGTW_{i,t-2}$				-0.041	-0.041	-0.041
				(-1.237)	(-1.254)	(-1.251)
$DGTW_{i,t-3}$				0.006	0.007	0.006
				(0.191)	(0.230)	(0.199)
$DGTW_{i,t-22,t-4}$				0.036	0.038	0.036
				(0.998)	(1.046)	(1.008)
$DGTW_{i,t-253,t-23}$				0.030	0.032	0.030
				(0.977)	(1.023)	(0.980)
$ln(Size)$				-0.075***	-0.074***	-0.073***
				(-2.984)	(-2.923)	(-2.910)
$ln(BM)$				-0.038*	-0.040*	-0.038*
				(-1.737)	(-1.805)	(-1.740)
$ln(Turnover)$				0.043	0.044	0.044
				(1.531)	(1.554)	(1.566)
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	30950	30950	30950	30950	30950	30950
Adj.R^2	0.016	0.015	0.016	0.016	0.016	0.016

Table 10: **Knowledge Cutoff, Firm Recognition, and the Explanatory Power of Sentiment Scores**

This table examines how the GPT-4o-mini knowledge cutoff and firm-name recognition affect the informativeness of sentiment scores extracted from earnings call transcripts. The dependent variable is the contemporaneous DGTW-adjusted stock return ($DGTW_{i,t}$) on the earnings-announcement day. $Sentiment_{RAW}$ and $Sentiment_{TRF}$ denote sentiment scores computed from the original and TRF-anonymized transcripts, respectively. The indicator Pre equals 1 if the transcript falls in the pre-knowledge-cutoff period of GPT-4o-mini (November 2022 to October 2023), and 0 otherwise. The indicator $Recognition_{Firm}$ equals 1 if GPT-4o-mini correctly recognizes the firm name in a given TRF-anonymized transcript, and 0 otherwise. Columns (1)–(2) use the full sample and regress returns on $Sentiment_{RAW}$ or $Sentiment_{TRF}$, respectively, together with their interactions with Pre to capture differences between the pre- and post-cutoff periods. Columns (3)–(4) restrict the sample to the pre- and post-cutoff periods, respectively, and use $Sentiment_{TRF}$, $Recognition_{Firm}$, and their interaction to test whether the predictive power of anonymized sentiment depends on firm-name recognition. Control variables are identical to those in Table 4. All variables are winsorized at the 1st and 99th percentiles and standardized. All specifications include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

VARIABLES	Dependent Variable: $DGTW_{i,t}$			
	(1)	(2)	(3)	(4)
$Sentiment_{RAW}$	2.510***			
	(19.559)			
$Sentiment_{RAW} \times Pre$	-0.584***			
	(-3.327)			
$Sentiment_{TRF}$		2.364***	1.995***	2.324***
		(17.718)	(9.312)	(12.677)
$Sentiment_{TRF} \times Pre$		-0.489***		
		(-2.745)		
$Recognition_{Firm}$			0.138	-0.014
			(0.614)	(-0.064)
$Sentiment_{TRF} \times Recognition_{Firm}$			-0.177	0.035
			(-0.715)	(0.177)
Period	Full	Full	Pre	Post
Controls	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	12852	12852	5746	7106
Adj.R^2	0.131	0.125	0.121	0.129

A Appendix

Table A.1: **Sample Selection Criteria**

This table outlines the filtering procedure used to construct the final samples. Panel A details the selection process for earnings call transcripts, and Panel B for news headlines. Each row sequentially applies a filter, reporting the number of observations removed and the resulting sample size.

Panel A: Earnings Call Transcripts		
Filter	**Sample Size**	**Obs. Removed**
Earnings call transcripts from Seeking Alpha for iShares Russell 3000 ETF constituents (Nov 2023 - Dec 2024)	10,812	
Remove duplicates, match dates and PERMNOs, remove errors	10,628	184
Remove transcripts with > 15,000 tokens	10,529	99
Match with DGTW returns (Nov 2023 - Dec 2024)	8,479	2,050
Match with analyst EPS forecast errors	7,596	883
Removing observations with missing controls	7,382	214
Panel B: News Headlines		
Filter	**Sample Size**	**Obs. Removed**
Firm news headlines from Finnhub for iShares Russell 3000 ETF constituents (Nov 2023 - Dec 2024)	1,612,948	
Remove duplicates, require valid dates and non-missing headlines	969,077	643,871
Retain headlines from Yahoo, exclude "stock" or "stocks"	271,194	697,883
Retain headlines with a single organization mention	118,797	152,397
Remove weekend news and match with PERMNO	112,835	5,962
Aggregate by PERMNO and date	80,337	32,498
Match with DGTW returns (Nov 2023 - Dec 2024)	57,603	22,734
Removing observations with missing controls	30,950	26,653

Table A.2 **Variable Definitions**

This table defines the variables used in our analysis. Panel A lists measures derived from raw and anonymized earnings call transcripts and news headlines; all measures are extracted by LLMs through designed prompts listed in Appendix B.3 to B.7. Panel B lists the dependent variables. Panel C lists the control variables. Unless otherwise specified, stock return data are from CRSP, and firm-level accounting data are from Compustat.

Variable	Definition
Panel A: LLM Measures	
$Sentiment_{RAW}$	Sentiment of the original earnings call transcript text. It is defined as a measure of its positive or negative implication for stock returns. The sentiment score, which ranges from -1 to 1, is computed with the formula Sentiment = (2 * Direction - 1) * Magnitude, where Direction captures the binary tone and Magnitude captures its intensity.
$Sentiment_{SM}$	Sentiment of the earnings call transcript text after anonymization using the spaCy en-core-web-sm (SM) model.
$Sentiment_{TRF}$	Sentiment of the earnings call transcript text after anonymization using the spaCy en-core-web-trf (TRF) model.
$Sentiment_{LLM}$	Sentiment of the earnings call transcript text after anonymization using GPT-4o-mini.
$Sentiment_{NUM}$	Sentiment of the transcript where only NUMBERS entity types are anonymized by the TRF model.
$Sentiment_{PLC}$	Sentiment of the transcript where only PLACES entity types are anonymized by the TRF model.
$Sentiment_{OBJ}$	Sentiment of the transcript where only OBJECTS entity types are anonymized by the TRF model.
$Sentiment_{OTH}$	Sentiment of the transcript where only OTHERS entity types are anonymized by the TRF model.
$Uncertainty_{RAW}$	Uncertainty score of the original earnings call transcript. It is defined as a measure of the level of imprecision and cautious tone conveyed in the text. The score, which ranges from 0 (absolute certainty) to 1 (maximum uncertainty), is extracted based on the prompt derived from Loughran and Mcdonald (2011).
$Uncertainty_{TRF}$	Uncertainty score of the TRF-anonymized earnings call transcript.
$Investment_{RAW}$	Firm's investment score of the original earnings call transcript. It is defined as a classification of the firm's planned capital spending changes over the next year. The score is categorized into one of five levels (Increase substantially, increase, no change, decrease, decrease substantially) based on the prompt derived from Jha et al. (2024a).
$Investment_{TRF}$	Firm's investment score of the TRF-anonymized earnings call transcript.

Variable	Definition
$Economy_{RAW}$	Firm's view on macroeconomic trends of the original earnings call transcript. It is defined as a classification of the firm's anticipated change in optimism about the US economy over the next quarter. The score is categorized into one of five levels (Increase substantially, increase, no change, decrease, decrease substantially) based on the prompt derived from Jha et al. (2024b).
$Economy_{TRF}$	Firm's view on macroeconomic trends of the TRF-anonymized earnings call transcript.
$Sentiment_{RAW}^{News}$	Sentiment of original Yahoo news headlines, which is defined as a measure of its positive or negative implication for stock returns. The sentiment score, which ranges from -1 to 1, is computed with the formula Sentiment = (2 * Direction - 1) * Magnitude, where Direction captures the binary tone and Magnitude captures its intensity.
$Sentiment_{TRF}^{News}$	Sentiment of the TRF-anonymized Yahoo news headlines.

Panel B: Dependent Variables	
$DGTW_{i,t}$	DGTW-adjusted stock return on the announcement date (day 0), defined as Daniel et al. (1997).
Vol_{post}	Standard deviation of stock returns over the post-announcement window [0, 22], requiring at least 10 daily observations.
$Capx_{t+2}$	Capital expenditure (CAPX) at the end of quarter $t+2$, scaled by total assets.
$SaleChange_t$	Defined as $\log(Sale_{t+1}/Sale_{t-1})$; represents two-quarter sales growth.
$ValueAddChange_t$	Defined as $\log(ValueAdd_{t+1}/ValueAdd_{t-1})$; represents two-quarter value-added growth.
$Gap_{i,t}$	Absolute difference between the sentiment of original and TRF-anonymized transcripts.

Panel C: Control Variables	
FE	Analyst forecast error, defined as Hirshleifer et al. (2009).
$DGTW_{i,t-1}$	DGTW-adjusted stock return on event day -1.
$DGTW_{i,t-2}$	DGTW-adjusted stock return on event day -2.
$DGTW_{i,t-3}$	DGTW-adjusted stock return on event day -3.
$DGTW_{i,t-22,t-4}$	DGTW-adjusted stock return over the event window [-22, -4].
$DGTW_{i,t-253,t-23}$	DGTW-adjusted stock return over the event window [-253, -23].
$Size$	Market capitalization at the end of June, calculated as CRSP price times shares outstanding.
BM	Book-to-market ratio at the end of June, calculated as book equity (fiscal year-end) divided by market equity (fiscal year-end).
$Turnover$	Share turnover at the end of June, calculated as prior year's trading volume divided by shares outstanding.

Variable	Definition
Vol_{pre}	Standard deviation of stock returns over the pre-announcement window [-252, -1], requiring at least 60 daily observations.
$Alpha_{pre}$	Alpha from a market model estimated over the window [-252, -6], requiring at least 60 daily observations.
|Abnormal Ret|	Absolute abnormal return, measured as the two-day (day 0 to +1) buy-and-hold return minus the buy-and-hold return of the CRSP value-weighted index.
$Capx_t$	Capital expenditure (CAPX) at the end of quarter t, scaled by total assets.
Leverage	Financial leverage, calculated as total debt (dlttq + dlcq) divided by total debt plus market value of equity.
Total Asset	Logarithm of quarterly total assets from Compustat (atq).
Tangibility	Plant, Property, and Equipment (PPENTQ) scaled by total assets (atq) at the end of the quarter.
$TNIC3TSIMM$	A firm-specific measure of total product similarity against rivals in 2023 based on the Text-Based Network Industry Classification (TNIC), defined as Hoberg and Phillips (2016).
$TNIC3HHI$	A firm-specific Herfindahl-Hirschman Index measuring market concentration among rivals in 2023 based on the Text-Based Network Industry Classification (TNIC), defined as Hoberg and Phillips (2016).
Coverage	The average of the 12 monthly numbers of earnings forecasts from I/B/E/S, defined as He and Tian (2013).

Table A.3: **Summary Statistics of Earnings Call Transcripts Sample**

This table presents summary statistics for earnings call transcripts from November 2023 through December 2024 in our sample. Panel A provides an overview of the sample, including the number of earnings call transcripts, firms, and dates, as well as the percentage of earnings calls held before market close and the percentage of firms listed on NASDAQ. Panel B presents detailed distributional statistics for firm-level fundamental variables, such as firm size (market capitalization), book-to-market ratio, and stock turnover. Firm size is reported in millions of dollars.

Panel A: Sample Overview				
Transcripts	Firms	Dates	Before Mkt Close (%)	NASDAQ (%)
7,382	1,831	259	64.21	47.41

Panel B: Fundamental Statistics			
	Size (Million $)	Book-to-Market	Turnover
Mean	23,845	0.55	2.77
Std. Dev.	137,878	0.52	8.04
25th Percentile	1,119	0.23	1.37
Median	3,293	0.44	1.92
75th Percentile	10,632	0.75	2.91

Table A.4: **Entity Category**

This table presents the 18 mutually exclusive entity types identified by the en-core-web-trf NER model from Python's spaCy library. These entity types are categorized into four primary groups: NUMBERS, PLACES, OBJECTS, and OTHERS. For each entity type, the table includes its description, an example from an earnings call transcript, the total frequency of occurrences in the transcript corpus, and the average frequency of each of the four main categories per transcript. The last two columns show the total frequency of each entity type in news headlines, along with the average frequency of each of the four main categories per headline.

Category	Entity Type	Description	Example	Total Frequency (Transcript)	Frequency per Transcript	Total Frequency (Headlines)	Frequency per Headline
NUMBERS	DATE	Absolute or relative dates or periods	Q1 2024	1247244		21862	
	CARDINAL	Numerals that do not fall under another type	one	274350		4670	
	MONEY	Monetary values, including unit	$69.7 billion	261834		5419	
	PERCENT	Percentage, including "%"	6%	250714	300.72	2322	0.678
	ORDINAL	"first", "second", etc.	first	92093		1214	
	TIME	Times smaller than a day	evening	77944		120	
	QUANTITY	Measurements, as of weight or distance	100-foot	15736		142	
PLACES	GPE	Countries, cities, states	Korea	115406		6528	
	LOC	Non-GPE locations, mountain ranges, bodies of water	the Middle East	43917	22.86	773	0.146
	FAC	Buildings, airports, highways, bridges, etc.	San Ciprián	9393		384	
OBJECTS	PERSON	People, including fictional	Tim Cook	771286		5741	
	ORG	firms, agencies, institutions, etc.	Apple Inc.	555867	181.37	52714	1.125
	NORP	Nationalities or religious or political groups	Chinese	11653		863	
OTHERS	PRODUCT	Objects, vehicles, foods, etc. (not services)	Watch Ultra	97619		3100	
	EVENT	Named hurricanes, battles, wars, sports events, etc.	Super Bowl	10825		834	
	LAW	Named documents made into laws	the IRA Inflation Reduction Act	7372	16.07	76	0.084
	WORK_OF_ART	Titles of books, songs, etc.	America's Greatest Work-places 2024	2492		395	
	LANGUAGE	Any named language	English	315		7	

Table A.5: **Correlations of Sentiment Signals After Removing Different Categories of Entity Information**

This table shows correlations among sentiment scores extracted from raw transcripts (RAW), TRF-anonymized transcripts (TRF) and transcripts where one category of entities has been replaced (NUMBERS, PLACES, OBJECTS or OTHERS). All sentiment measures are extracted using the GPT-4o-mini model. All values are Pearson correlation coefficients.

	Entity Replacement Sentiment Correlation					
	$Sentiment_{RAW}$	$Sentiment_{TRF}$	$Sentiment_{NUM}$	$Sentiment_{PLC}$	$Sentiment_{OBJ}$	$Sentiment_{OTH}$
$Sentiment_{RAW}$	1.000					
$Sentiment_{TRF}$	0.864	1.000				
$Sentiment_{NUM}$	0.876	0.899	1.000			
$Sentiment_{PLC}$	0.946	0.854	0.868	1.000		
$Sentiment_{OBJ}$	0.912	0.863	0.872	0.917	1.000	
$Sentiment_{OTH}$	0.939	0.860	0.866	0.942	0.910	1.000

Table A.6: **Explanatory Power Comparison of Sentiment Scores Without Controls**

This table presents the results of univariate regressions with DGTW-adjusted returns on the earnings conference call day as dependent variables. The independent variables are sentiment measures extracted using GPT-4o-mini from earnings conference call transcripts. Columns (1) through (4) augment this baseline model by individually adding the sentiment scores derived from the four text sources. Columns (5) through (7) include both the sentiment extracted from raw transcripts and from an entity-removed transcripts. All variables are winsorized at the 1st and 99th percentiles. All independent variables are standardized. All regressions include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

VARIABLES	Dependent Variable: $DGTW_{i,t}$						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$Sentiment_{RAW}$	2.549***				1.759***	1.842***	1.441***
	(19.602)				(6.688)	(8.741)	(5.091)
$Sentiment_{SM}$		2.450***			0.901***		
		(18.282)			(3.474)		
$Sentiment_{TRF}$			2.413***			0.820***	
			(17.329)			(3.766)	
$Sentiment_{LLM}$				2.532***			1.273***
				(19.351)			(4.637)
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7352	7352	7352	7352	7352	7352	7352
Adj.R^2	0.078	0.072	0.070	0.077	0.080	0.080	0.083

Table A.7: **Summary Statistics of Yahoo News Headlines**

This table presents the summary statistics for Yahoo news headlines from November 2023 to December 2024. Panel A provides an overview of the sample, including the total number of news headlines, firm-date level observations, as well as the number of firms and dates, along with the percentage of headlines posted before market close. Panel B presents detailed distributional statistics of both textual and firm-level characteristics. Specifically, it includes headline length (in tokens), the proportion of entities identified by the TRF model, and firm-level variables such as market capitalization (measured in millions of dollars), book-to-market ratio, and stock turnover.

Panel A: Sample Overview				
Headlines	Firm-Date-Obs	Firms	Dates	Before Mkt Close (%)
52,716	30,950	2,052	292	64.66

Panel B: Textual and Fundamental Statistics					
Statistics	Length	Entity%(TRF)	Size (Million $)	Book-to-Market	Turnover
Mean	15.55	38.30	121,493	0.50	2.99
Std. Dev.	6.13	20.45	419,148	0.50	9.63
25th Percentile	12.00	22.22	2,154	0.18	1.27
Median	14.00	35.71	9,275	0.38	1.85
75th Percentile	18.00	52.63	53,449	0.70	2.91

Table A.8: **Knowledge Cutoff, Firm Recognition, and the Explanatory Power of Sentiment Scores Using GPT-4o**

This table examines how the GPT-4o knowledge cutoff and firm-name recognition affect the informativeness of sentiment scores extracted from earnings call transcripts. The dependent variable is the contemporaneous DGTW-adjusted stock return ($DGTW_{i,t}$) on the earnings-announcement day. $Sentiment_{RAW}$ and $Sentiment_{TRF}$ denote sentiment scores computed from the original and TRF-anonymized transcripts, respectively. The indicator Pre equals 1 if the transcript falls in the pre-knowledge-cutoff period of GPT-4o-mini (November 2022 to October 2023), and 0 otherwise. The indicator $Recognition_{Firm}$ equals 1 if GPT-4o-mini correctly recognizes the firm name in a given TRF-anonymized transcript, and 0 otherwise. Columns (1)–(2) use the full sample and regress returns on $Sentiment_{RAW}$ or $Sentiment_{TRF}$, respectively, together with their interactions with Pre to capture differences between the pre- and post-cutoff periods. Columns (3)–(4) restrict the sample to the pre- and post-cutoff periods, respectively, and use $Sentiment_{TRF}$, $Recognition_{Firm}$, and their interaction to test whether the predictive power of anonymized sentiment depends on firm-name recognition. Control variables are identical to those in Table 4. All variables are winsorized at the 1st and 99th percentiles and standardized. All specifications include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $DGTW_{i,t}$			
VARIABLES	(1)	(2)	(3)	(4)
$Sentiment_{RAW}$	2.635***			
	(20.709)			
$Sentiment_{RAW} \times Pre$	-0.471***			
	(-2.663)			
$Sentiment_{TRF}$		2.097***	1.818***	2.220***
		(17.926)	(6.818)	(7.106)
$Sentiment_{TRF} \times Pre$		-0.247		
		(-1.448)		
$Recognition_{Firm}$			-0.014	0.525*
			(-0.040)	(1.837)
$Sentiment_{TRF} \times Recognition_{Firm}$			0.082	-0.179
			(0.321)	(-0.561)
Period	Full	Full	Pre	Post
Controls	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	12852	12852	5746	7106
Adj.R^2	0.136	0.116	0.122	0.114

Table A.9: **Anonymization Gap, Information Loss, and the Explanatory Power of TRF Sentiment**

This table examines whether the gap between sentiment scores from RAW and TRF-anonymized transcripts is informative about anonymization-induced information loss in explaining contemporaneous stock returns. The dependent variable in all columns is the DGTW-adjusted stock return on the earnings-announcement date ($DGTW_{i,t}$). $Sentiment_{TRF}$ denotes the sentiment score computed from TRF-anonymized transcripts. The variable Gap is defined as the absolute difference between the raw-text and anonymized sentiment scores, $Gap = |Sentiment_{RAW} - Sentiment_{TRF}|$. Each specification includes $Sentiment_{TRF}$, Gap, and their interaction $Sentiment_{TRF} \times Gap$. Columns (1)–(3) use sentiment scores generated by GPT-4o-mini, whereas columns (4)–(6) use scores generated by GPT-4o. The row "Period" indicates whether each regression is estimated on the pre- or post–knowledge-cutoff subsample. The row "Filter" reports whether the regression is run on the full sample ("–") or on a restricted anonymized subsample ("Anonymized"). In the anonymized subsample used in columns (3) and (6), we keep only transcripts for which, after TRF anonymization, the corresponding model (GPT-4o-mini in column (3) and GPT-4o in column (6)) fails to recognize the firm name, so as to better isolate anonymization-related information loss and mitigate forward-looking bias. Control variables are identical to those in Table 4. All variables are winsorized at the 1st and 99th percentiles, and all independent variables are standardized. All specifications include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

VARIABLES	Dependent Variable: $DGTW_{i,t}$					
	GPT-4o-mini			GPT-4o		
	(1)	(2)	(3)	(4)	(5)	(6)
$Sentiment_{TRF}$	1.997***	2.519***	2.164***	1.809***	2.038***	1.731***
	(14.812)	(17.958)	(8.486)	(13.887)	(16.775)	(6.240)
Gap	-0.647***	-0.865***	-0.629**	-0.936***	-0.777***	-0.706**
	(-4.148)	(-5.210)	(-2.334)	(-7.067)	(-5.749)	(-2.106)
$Sentiment_{TRF} \times Gap$	-0.357***	-0.582***	-0.420***	-0.825***	-1.364***	-0.976***
	(-5.286)	(-7.585)	(-4.098)	(-7.179)	(-9.847)	(-4.227)
Period	Pre	Post	Pre	Pre	Post	Pre
Filter	–	–	Anonymized	–	–	Anonymized
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5746	7106	2354	5746	7106	810
Adj.R^2	0.125	0.137	0.158	0.139	0.130	0.168

B Prompts

B.1 Anonymization Prompt

We developed an anonymization prompt for earnings call transcripts based on Engelberg et al. (2025) and the 18 entity types and definitions from the spaCy package, utilizing the following prompt to anonymize the transcripts:

Your role is to ANONYMIZE all text that is provided by the user. After you have anonymized a text, NOBODY, not even an expert financial analyst should be able to read the text and know the identity of the company nor the industry the company operates in.

For example, if the text is: "Today, Apple is reporting revenue of $119.6 billion for the December quarter, up 2% from a year ago despite having one less week in the quarter. ", then you should ANONYMIZE it to: "DATE_1, ORG_1 is reporting revenue of MONEY_1 for DATE_2, up PERCENT_1 from DATE_3 despite having DATE_4 in DATE_5." You should also ANONYMIZE any other information which one could use to identify the company or make an educated guess at its identity. These pieces of information include the following definitions and replacement rules:

Convert "People, including fictional" to "PERSON_x".
Convert "Nationalities or religious or political groups" to "NORP_x".
Convert "Buildings, airports, highways, bridges, etc." to "FAC_x".
Convert "Companies, agencies, institutions, etc." to "ORG_x".
Convert "Countries, cities, states" to "GPE_x".
Convert "Non-GPE locations, mountain ranges, bodies of water" to "LOC_x".
Convert "Industries" to "INDUSTRY_x".
Convert "Sectors" to "SECTOR_x".
Convert "Objects, vehicles, foods, etc. (NOT services)" to "PRODUCT_x".
Convert "Named hurricanes, battles, wars, sports events, etc." to "EVENT_x".
Convert "Titles of books, songs, etc." to "WORK_OF_ART_x".
Convert "Named documents made into laws" to "LAW_x".
Convert "Any named language" to "LANGUAGE_x".
Convert "Absolute or relative dates or periods" to "DATE_x".
Convert "Times smaller than a day" to "TIME_x".
Convert "Percentage, including '%'" to "PERCENT_x".
Convert "Monetary values, including unit" to "MONEY_x".
Convert "Measurements, as of weight or distance" to "QUANTITY_x".
Convert "'first', 'second', etc. " to "ORDINAL_x".
Convert "Numerals that do not fall under another type" to "CARDINAL_x".
Convert "Website or internet links" to "LINK_x".

You should never just delete an identifier; instead, always replace it with an anonymous analog. After you read and ANONYMIZE the text, you should output the anonymized text and nothing else.

B.2 Anonymization test Prompt

We refer to Engelberg et al. (2025) and use the following prompt to test whether an anonymized earnings call transcript can be identified as belonging to a specific firm and the time of the call:

You will receive a body of text which has been anonymized. You are omniscient. Use all your knowledge and the context to identify which company the text is about, as well as the year it was written. Make your best guess based on information and context if you are unsure. Please only provide the ticker of the company you have identified. Provide your estimate exactly in the following format, with no other text at all (TIK is your estimate of the ticker, Y is your estimate of the year): **Company Estimate: TIK**,**Year Estimate: Y**.

B.3 Sentiment Prompt(For Transcript)

In designing the prompt for our analysis, we reference the methodology for controlling LLM output tokenization from Engelberg et al. (2025). They demonstrate that by forcing the model to generate a number in a structured format—such as 0.X for a sentiment score—one can isolate the magnitude of the estimate into the log probability of a single, unambiguous token (the digit X). We adopt this principle to ensure that the numerical values generated by our prompt are free from tokenization artifacts, such as leading or trailing spaces. This structured approach allows us to reliably use the model's log probabilities as a direct measure of its assessment. We use the following prompt:

You are an expert Financial Analyst. Your task is to read quarterly company earnings call transcripts and infer if the information in the transcripts is a good or bad signal for the future returns of the security of the company whose earnings call is being transcribed. Based on your answer, I will choose to buy or sell that security. After reading the transcripts, provide a bull or bear signal as two numbers: the direction and the magnitude. For direction, output only 0,1, or NA, where 0 is bearish, 1 is bullish, and NA means there is no information relevant to security prices. For magnitude, output a number ranging from 0 to 1 ONLY, where numbers near 0 indicate slightly bearish (if the direction is 0) or slightly bullish (if the direction is 1), and numbers near 1 indicate highly bearish (if the direction is 0) or highly bullish (if the direction is 1). Provide your output exactly in the following format, with no other text at all: **Direction Estimate: DIRECTION**,**Magnitude Estimate: MAGNITUDE**.

B.4 Uncertainty Score Prompt

We refer to the definition in Loughran and Mcdonald (2011) to extract the uncertainty score from the transcript, with the following prompt:

Analyze the following conference call transcript. As a financial analyst specializing in textual analysis, your task is to evaluate the level of Uncertainty. Your evaluation must be on a

continuous scale from 0 (representing absolute certainty) to 1 (representing maximum uncertainty) ONLY. This assessment should not be based on a simple count of keywords. Instead, you must analyze the contextual meaning, usage, and significance of words and phrases that convey a cautious or uncertain tone, emphasizing the general notion of imprecision rather than exclusively focusing on risk. Your goal is to determine the genuine, unmitigated uncertainty conveyed by the transcript. Provide your output only in the following format, with no other text: **Uncertainty Score: UNCERTAINTY**.

B.5 Investment Score

We refer to the prompt in Jha et al. (2024a) to extract the Investment score from the transcript, with the following prompt:

The following text is a company's earnings call transcript. You are a finance expert. Based on this text only, please answer the following question. How does the firm plan to change its capital spending over the next year? There are five choices: Increase substantially, increase, no change, decrease, and decrease substantially. Please select one of the above five choices for each question and provide a one-sentence explanation of your choice for each question. The format for the answer to each question should be **choice - explanation**. If no relevant information is provided related to the question, answer **no information is provided**.

B.6 Economy Score

We refer to the prompt in Jha et al. (2024b) to extract the Economy score from the transcript, with the following prompt:

The following text is a company's earnings call transcript. You are a finance expert. Based on this text only, please answer the following question. Over the next quarter, how does the firm anticipate a change in optimism about the US economy? There are five choices: Increase substantially, increase, no change, decrease, and decrease substantially. Please select one of the above five choices for each question and provide a one-sentence explanation of your choice for each question. The format for the answer to each question should be **choice - explanation**. If no relevant information is provided related to the question, answer **no information is provided**.

B.7 Sentiment Prompt(For News)

We used a similar LLM prompt to extract sentiment from news:

You are an expert Financial Analyst. Your task is to read news headlines and infer if the reported headline is a good or bad signal for the future returns of the security being written about. Based on your answer, I will choose to buy or sell that security. After reading the headline, provide a bull or bear signal as two numbers: the direction and the magnitude. For

direction, output only 0,1, or NA, where 0 is bearish, 1 is bullish, and NA means there is no information relevant to security prices. For magnitude, output a number ranging from 0 to 1 ONLY, where numbers near 0 indicate slightly bearish (if the direction is 0) or slightly bullish (if the direction is 1), and numbers near 1 indicate highly bearish (if the direction is 0) or highly bullish (if the direction is 1). Provide your output exactly in the following format, with no other text at all: **Direction Estimate: DIRECTION**,**Magnitude Estimate: MAGNITUDE**.