Anonymization and Information Loss*

Ke Wu

Baozhong Yang

Zhenkun Ying

Dexin Zhou

Abstract

Anonymizing financial texts is a widely used technique to reduce look-ahead bias when extracting textual information using large language models. We demonstrate that although anonymization reduces models' ability to recover selected identifiers, it compromises textual understanding, diminishing their ability to extract meaningful economic signals from financial texts. Among the four entity categories, preserving numerical or agent entities recovers more of the raw text's predictive content than preserving place or other entities. Notably, we find little evidence of significant look-ahead bias in our setting, whereas the information loss induced by anonymization remains stable across the model's knowledge cutoff. We further show that the attenuation-adjusted absolute difference between signals extracted from raw and anonymized texts serves as a proxy for the document-level distortion introduced by anonymization. A ten-factor credit-risk audit shows that information loss arises through both evidence displacement and interpretive distortion. We further show that this information loss is pervasive across other LLMs, alternative prediction tasks, and different types of financial text.

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

*We are grateful for comments and feedback from Yi Cao, Wenxi Jiang, Yu-Jane Liu, Markus Pelger, Suproteem Sarkar, Olivier Scaillet, Xudong Wen, Zhengyang Xu, Bernard Yeung, Bohui Zhang, and Shunming Zhang, as well as from seminar and conference participants at the Wolfe Research QES 8th Annual AI in Finance Conference, the 2026 China Fintech Research Conference (CFTRC), the 2026 Jingshi Scholars Forum (Tsinghua–Peking–Renmin Joint Graduate Forum in Economics and Management), the 5th Annual Hong Kong Conference on FinTech and AI in Finance, the 2026 China International Conference in Finance (CICF), the 2nd Future Scholars in Finance Forum, and the 2026 CES China Annual Conference. Contact information: Ke Wu: ke.wu@ruc.edu.cn; Baozhong Yang: bzyang@gsu.edu; Zhenkun Ying: yingzhenkun0205@ruc.edu.cn; Dexin Zhou: dexin.zhou@baruch.cuny.edu.

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, 2026). 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., 2026), and generating structured economic representations from news (Sarkar, 2026).

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 extracted signals are used to predict future economic performance or stock returns. It arises during training, when LLMs 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, 2026; Sarkar and Vafa, 2024; Ludwig et al., 2026; 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, 2024; Engelberg et al., 2025). 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]

In this study, we investigate an important limitation of the anonymization methodology:

[1]Other approaches include 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).

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 announced it will lay off 10,000 employees." For a company undergoing financial distress, such as a struggling retailer facing bankruptcy, this announcement likely reinforces a negative outlook, signaling deepening trouble. Conversely, for a highly profitable technology giant undertaking strategic restructuring, the same announcement could be interpreted positively by the market as a sign of disciplined cost management. After anonymization, both scenarios collapse into an identical, ambiguous input: "FIRM_1 announced it will lay off NUMBER_1 employees." This procedure strips the model of the context necessary to make an accurate inference, potentially leading to noisy signals.

To organize this trade-off, we develop a parsimonious framework in which the anonymized score combines an attenuated version of the raw assessment with document-specific distortion. Under the benchmark in which the raw score is a sufficient summary of the text, anonymization weakens predictive power in a standalone regression and contributes no incremental information once the raw score is included. The framework separately allows pre-cutoff scores to absorb future innovations from the model's training data, thereby distinguishing the removal of look-ahead contamination from the destruction of genuine context. It also motivates an attenuation-adjusted gap between the raw and anonymized scores as a document-level measure of interpretive distortion.

Empirically, we ask LLMs to extract information from a sample of conference call transcripts 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.

We consider two widely used anonymization strategies. The first relies on a named-entity-

recognition model from the spaCy library—its most capable Transformer-based model, which we denote TRF—to detect entities and replace them with placeholders such as "FIRM_1". The second is the iterative procedure of Engelberg et al. (2025), which uses an LLM to mask entities and paraphrase the surrounding text.

Our main analysis examines a salient and economically important prediction task: forecasting future S&P credit-rating downgrades. Following Donovan et al. (2021), we prompt GPT-4o-mini to extract a downgrade-probability score from each transcript and from its anonymized versions, and we relate these scores to actual rating downgrades over the subsequent twelve months. The score extracted from the raw text is a strong predictor of downgrades; the anonymized scores predict downgrades on their own, but in a direct "horse race" that includes both, the raw signal remains large and highly significant while the anonymized signal collapses to a coefficient that is statistically indistinguishable from zero. The credit-risk content of the anonymized text is thus almost entirely subsumed by that of the raw text—our first and clearest evidence that entity removal destroys genuine information.

We trace this degradation to specific entity types, which we group into four categories: NUMBERS, AGENTS, PLACES, and OTHERS. To assess the information content of each category, we extract the signal from transcripts in which only one category is preserved while the other three are anonymized, and ask how much of the raw text's predictive content each single category recovers. No single preserved category is sufficient to recover the raw text's predictive content. Moreover, preserving NUMBERS (such as financial figures and dates) or AGENTS (such as company and executive names) recovers substantially more of the raw signal than preserving PLACES, which adds little. Numerical and agent entities thus serve as the critical anchors that LLMs rely on to interpret financial text.

A natural concern is that the raw text's advantage reflects look-ahead bias rather than genuine information: an LLM may already "know" a firm's subsequent downgrade from its training data. To address this, we extend the sample to the pre-knowledge-cutoff period (November 2022–October 2023) and compare the raw signal's predictive edge on either side

of the cutoff. If look-ahead bias were inflating the raw signal, its advantage would be concentrated before the cutoff; instead, the raw signal's advantage over the anonymized signals is comparable across the two periods, and its interaction with a pre-cutoff indicator is statistically insignificant. Look-ahead bias therefore plays only a limited role, and the raw text's advantage reflects genuine, contemporaneously available information.

We further develop a document-level diagnostic for the severity of this loss. Defining the anonymization gap as the attenuation-adjusted absolute difference between the raw and anonymized signals, we find that the anonymized signal predicts downgrades more weakly precisely when this gap is larger. The interaction is negative in both windows and is statistically significant for both anonymization methods in the post-cutoff sample, where look-ahead bias is precluded. The gap thus provides a simple, economically meaningful proxy for the distortion introduced by anonymization.

The information loss introduced by anonymization is not specific to a single model. Repeating the extraction with a larger model (GPT-4o) and a reasoning-focused model (o3-mini) yields the same qualitative pattern: the raw signal remains significant in every horse race, whereas the anonymized signal is either insignificant or, for GPT-4o, only marginally significant. Information loss is therefore not an artifact of one model's limitations.

To look inside this aggregate loss, we conduct a quotation-backed audit organized around ten factors of S&P Global Ratings' Corporate Methodology. The audit reveals two distinct channels. *Evidence displacement* occurs when anonymization removes a credit-risk reason or redirects the model toward a different supporting passage; *interpretive distortion* occurs when the model continues to cite the same passage but assigns it a different risk assessment. A composite formed from the ten-factor assessments reproduces the predictive loss in the one-shot score, and channel-specific horse races show that both the displaced evidence and the altered reading of shared evidence matter. A Shapley decomposition assigns 40.6% of the loss under TRF anonymization and 41.3% under LLM anonymization to evidence displacement; interpretive distortion accounts for the remaining 59.4% and 58.7%, respectively.

The information loss extends well beyond credit risk. Because predicting stock returns is central to asset-pricing research and concerns about look-ahead bias are especially salient in this setting, we first examine whether anonymization weakens the relation between sentiment extracted by LLMs and stock returns. Sentiment extracted from raw earnings call transcripts strongly predicts announcement-window DGTW-adjusted returns. In standalone regressions, sentiment extracted from TRF- and LLM-anonymized transcripts produces smaller coefficients and lower adjusted R^2 than raw sentiment. When the raw and anonymized sentiment measures enter jointly, the coefficient on TRF-anonymized sentiment declines from 2.331 to 0.775, while the coefficient on LLM-anonymized sentiment declines from 2.329 to 0.798. In both specifications, raw sentiment remains large and highly significant. Thus, although the anonymized sentiment measures retain some incremental predictive content, much of their information is subsumed by raw sentiment.

Beyond stock returns, we examine four additional prediction tasks using earnings call transcripts. Following the designs of Loughran and McDonald (2011), Jha et al. (2024a), and Jha et al. (2024b), we use textual measures of managerial uncertainty, investment expectations, and macroeconomic outlook to predict future return volatility, capital expenditures, sales growth, and value-added growth. Across all four outcomes, the coefficient on the raw-text measure exceeds that on its TRF-anonymized counterpart. The anonymized uncertainty measure becomes statistically insignificant once the raw measure is included, while the anonymized measures retain significant but weaker predictive content in the other three tasks. Together, these results show that the information loss induced by anonymization extends across a broad range of economically important outcomes rather than being confined to credit risk.

Finally, to confirm that our conclusions are not specific to earnings calls, we replicate the return analysis on short, entity-dense news headlines. The results are consistent—raw-text sentiment dominates anonymized sentiment in horse races—confirming that information loss is a fundamental consequence of the anonymization process itself rather than a feature of a

particular type of disclosure.

Our study contributes to the emerging literature on LLM-based information extraction by quantifying a critical methodological trade-off. While prior work cautions against look-ahead bias (Glasserman and Lin, 2024; Sarkar and Vafa, 2024) and proposes anonymization as a remedy (Breitung and Müller, 2025; Engelberg et al., 2025), we show that anonymization carries a significant hidden cost: it removes genuine, value-relevant information along with firm identity. Because we document this cost in settings where look-ahead bias turns out to play only a limited role, our results urge caution—absent a post-cutoff sample for comparison, researchers may mistakenly attribute a reduction in signal informativeness to the successful mitigation of bias when it is in fact a byproduct of information loss. More constructively, the anonymization gap we introduce offers a simple diagnostic for the severity of this loss.

2 Econometric Framework

Anonymization does more than remove information tied to identifiable entities. By stripping language of the contextual anchors that give it economic meaning, it may also render the remaining text ambiguous and thereby distort its interpretation. We formalize these two channels in a framework in which anonymization both attenuates the underlying signal and introduces document-specific distortion. The framework then distinguishes this loss of genuine context from the removal of look-ahead contamination and develops a document-level diagnostic for the distortion that remains after average attenuation is removed. Supporting derivations are reported in Appendix A.

2.1 Modeling Information Loss

Let Y denote a future outcome. Applying the same model and prompt to the raw and anonymized transcripts yields standardized scores R and A. Their correlation,

$$\rho \equiv \mathrm{Corr}(R, A) \in (0, 1), \tag{1}$$

measures their average agreement. We write the anonymized score as

$$A = \rho R + u. \tag{2}$$

Because both scores are standardized, this decomposition implies $\mathbb{E}[u] = 0$, $\mathrm{Cov}(R, u) = 0$, and $\mathrm{Var}(u) = 1 - \rho^2$. The loading ρ captures the common attenuation of the raw assessment, whereas u captures the document-specific distortion introduced by anonymization. Section 5 gives these statistical objects observable textual counterparts.

To connect the decomposition to predictive content, suppose that the raw score is sufficient for the information in the text:

$$Y = \mu + \beta R + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid R, A] = 0. \tag{3}$$

The separate regression slopes are then $b_R = \beta$ and $b_A = \rho\beta$, with $R_A^2 = \rho^2 R_R^2$. Moreover, once R is included in a horse race, the population coefficient on A is zero. Anonymization therefore appears in two related ways: attenuation weakens the anonymized score in a standalone regression, and document-specific distortion leaves it with no incremental predictive content relative to the raw score.

2.2 Look-Ahead Contamination

The comparison is more delicate for documents dated before the model's knowledge cutoff. In that period, a score may partly reflect outcomes that entered the model's training data rather than information contained in the document. Let P equal one for pre-cutoff documents and zero otherwise. For either leakage-free score $S \in \{R, A\}$, let $\widetilde{S}$ denote the score we observe. Following Gao et al. (2025), we represent contamination as

$$\widetilde{S} = S + \gamma_S P \varepsilon, \qquad 0 \leq \gamma_A \leq \gamma_R. \tag{4}$$

Here γ_S measures the amount of the future innovation absorbed by the observed score. The ordering allows anonymization to weaken the link between the document and information the model may have learned about the firm.

The cutoff comparison follows from

$$Y = d_0 + d_1 \widetilde{S} + d_2 P + d_3 (\widetilde{S} \times P) + e. \tag{5}$$

Holding the leakage-free relation between the score and the outcome fixed across the cutoff, d_1 is the post-cutoff predictive slope and d_3 measures the additional pre-cutoff slope. A contaminated score loads mechanically on the future innovation and is therefore unusually predictive before the cutoff; for the empirically relevant case of moderate contamination, this implies $d_3 > 0$. If that interaction appears for the raw score but not for its anonymized counterpart, the pattern is consistent with anonymization removing leakage. If neither score exhibits a positive interaction, there is little leakage to remove, and weaker performance after anonymization instead points to the loss of genuine context. Appendix A gives the exact interaction coefficient and its sign condition.

2.3 The Anonymization Gap

Average attenuation can conceal substantial heterogeneity. From equation (2), define the rescaled distortion v and its absolute magnitude as

$$v \equiv \frac{A}{\rho} - R = \frac{u}{\rho}, \qquad Gap \equiv |v| = \left| \frac{A}{\rho} - R \right|. \tag{6}$$

Dividing by ρ removes the common attenuation component, leaving the signed document-specific departure v; Gap records its magnitude. The gap has a direct predictive implication. In a leakage-free sample, if the rescaled distortion is symmetric and independent of the raw score, the predictive slope of the anonymized score conditional on $Gap = g$ is

$$b(g) \equiv \frac{\mathrm{Cov}(Y, A \mid Gap = g)}{\mathrm{Var}(A \mid Gap = g)} = \frac{\beta}{\rho(1 + g^2)}. \tag{7}$$

For the positively oriented scores used in our tests, this slope decreases as the gap rises: anonymization adds variation to the assessment without adding covariance with the outcome, which indicates that Gap can work as a proxy of document-level heterogeneous interpretive distortion of the anonymized assessment. This implication motivates the interaction regression

$$Y = c_0 + c_1 A + c_2 Gap + c_3 (A \times Gap) + e. \tag{8}$$

The framework predicts $c_3 < 0$. A negative interaction in the post-cutoff sample is especially informative: because those documents cannot contain memorized future outcomes, the interaction isolates document-level distortion rather than the removal of look-ahead contamination.

3 Methodology and Data

3.1 Anonymization Algorithms

We employ two approaches for text anonymization. The first utilizes a specialized named-entity-recognition (NER) model from the Python spaCy package, `en_core_web_trf` (TRF), which is based on a Transformer architecture. The second approach follows Engelberg et al. (2025) and uses a commercial LLM, specifically GPT-4o-mini, in conjunction with custom prompts to perform anonymization.

These two models differ substantially in size because of their distinct underlying architectures. The `en_core_web_trf` model is based on the RoBERTa-base Transformer architecture and contains approximately 0.1 billion parameters.[2] By contrast, GPT-4o-mini is larger by roughly two orders of magnitude; although its exact parameter count is undisclosed, industry estimates place it on the order of 8 billion parameters.[3]

Approach A in Figure 1 illustrates the workflow for the spaCy NER model (`en_core_web_trf`). The model first scans the raw text to recognize specific entities, such as "Apple," "Tim Cook," and "Luca Maestri" in the example shown. It then generates a consistent global map based on the entity type and order of appearance—for instance, mapping "Apple" to ORG_1 as the first organization identified, and "Tim Cook" to PERSON_1 as the first person.[4] Finally, the system applies these mappings globally to the entire text unit, producing an anonymized output that preserves textual structure while obscuring identities.

Approach B in Figure 1 describes the strategy using a large language model, specifically GPT-4o-mini. Unlike the linear NER approach, this method employs an iterative verification loop. The model first attempts to mask entities by replacing them with placeholders. The

[2] https://huggingface.co/roberta-base

[3] https://techcrunch.com/2024/07/18/openai-unveils-gpt-4o-mini-a-small-ai-model-powering-chatgpt

[4] As reported in Appendix Table B.3, 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), AGENTS (PERSON, ORG, NORP), and OTHERS (PRODUCT, EVENT, LAW, WORK_OF_ART, and LANGUAGE).

system then evaluates whether the entities in this masked output can still be identified by an LLM.[5] If the guess doesn't match the actual name (based on a string similarity threshold) and the actual date, the system loops move to the final output. Otherwise the text can't pass this initial anonymization check, it undergoes a paraphrasing process in which sentence structures are rewritten to remove stylistic identifiers (e.g., changing "Speaking first today..." to "We will begin today with remarks from..."). A final identification check is performed on this paraphrased output. If the firm or date can still be identified, the process restarts from the beginning. This cycle of masking and paraphrasing repeats for a maximum of three rounds, as the marginal benefit of anonymization diminishes significantly beyond this point. Only when the text passes the final verification or reaches the round limit is the final LLM-anonymized text generated. Our prompt design follows Engelberg et al. (2025), and the full prompt is available in Appendix D.1.

Table 1 presents a text excerpt from Apple Inc.'s 2024 Q1 earnings call transcript, along with the anonymization results from the TRF and LLM approaches. The TRF model exhibits a rigid, category-based replacement strategy; for instance, it anonymizes temporal expressions like "today" and "annual" into "DATE_1" and "DATE_2," and correctly identifies "SEC" as an organization ("ORG_2"). In contrast, the LLM approach offers a distinct progression. After several rounds of masking and paraphrasing, it demonstrates superior flexibility by altering sentence structure, for example, by changing "Please note" to "Please be aware," and by replacing specific identifiers with semantic descriptions, such as converting "SEC" into "the regulatory body," thereby mitigating identification risks associated with specific stylistic patterns.

[5]We consider the text to be "identified" if the LLM's guessed firm name has a `fuzz.token_sort_ratio` score of 75 or higher when matched against the actual firm name, or if the estimated conference call date falls within 7 days of the actual date.

3.2 Data

3.2.1 Firm 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, credit-downgrade dates from S&P, and EPS forecast data from the I/B/E/S Detail History database. Detailed variable definitions can be found in Appendix Table B.2.

3.2.2 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. The models used in this study—GPT-4o-mini-2024-07-18, GPT-4o-2024-08-06, and o3-mini-2025-01-31—all share a knowledge-cutoff date of October 2023.[6] Accordingly, we download earnings call transcripts dated from November 2023 onward from Seeking Alpha.

We employ two distinct sample-filtering schemes tailored to the data requirements of our prediction tasks. The first applies to the credit-downgrade prediction; the full procedure is detailed in Panel A of Appendix Table B.1. The second applies to the earnings-call return analysis and the alternative transcript-based prediction tasks in Section 6. Because these tasks rely on different matched data and matching procedures, we adopt a separate filtering scheme, detailed in Panel B of Appendix Table B.1.

The two schemes share the same initial steps. Both are strictly restricted to the November

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

2023–December 2024 window. We remove duplicate entries and those with incorrect timestamps, and we exclude transcripts exceeding 15,000 tokens. This threshold is 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.

The schemes diverge thereafter. For the downgrade prediction, following Donovan et al. (2021), we exclude utilities and financial firms (SIC 4000–4999 and 6000–6299), require the full set of control variables, and require an S&P credit rating observed before the earnings call. These sequential filters leave 2,488 post-cutoff transcripts.[7] For the return-prediction and other transcript-based tasks, we match the transcript data with the relevant returns, analyst EPS forecast errors, and accounting outcomes using date and PERMNO, and drop observations with missing controls for the corresponding specification.

To facilitate our analysis of the pre-cutoff period, we extend our dataset to include earnings call transcripts from November 2022 to October 2023. We apply the same filtering criteria to this earlier sample for both the downgrade-prediction and return-prediction tasks. To ensure a consistent comparison across the pre- and post-cutoff periods and to mitigate potential selection bias, our combined analysis is restricted to a balanced panel of firms that have valid observations in both the pre-cutoff (November 2022–October 2023) and post-cutoff (November 2023–December 2024) periods.

3.2.3 News Headlines

In addition to 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 attempt to extract signals from news headlines and articles (e.g., Bybee et al., 2023; Lopez-Lira and Tang, 2026; Engelberg et al., 2025). To ensure the generalizability of our results, we also

[7]The sequential sample selection in Appendix Table B.1 leaves 2,488 transcripts. The main regressions with SIC3 × Quarter fixed effects report 2,243 observations because the fixed-effect estimator drops 245 singleton observations. The same singleton-removal procedure applies to the other fixed-effects regressions in the paper, although the number of observations excluded may vary across specifications.

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. (2026), we first exclude headlines associated with multiple firm tickers. To handle duplicate articles that arise from re-syndication across different news outlets, we retain headlines 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 C of Appendix Table B.1.

3.2.4 LLM-Generated Measurements

We design a suite of prompts to extract several measures from the aforementioned textual data, including the probability of a credit-rating downgrade, uncertainty, investment (the firm's outlook for future capital expenditures), economy (the firm's outlook for the future macroeconomy), and sentiment. For each measure, we generate a quantitative score. The specific definitions of these scores and the complete prompts are reported in Appendix D, Sections D.3–D.8.

For the credit-risk mechanism analysis, we additionally conduct a structured score-and-quote extraction based on ten factors of S&P Global Ratings' Corporate Methodology using GPT-4o-mini-2024-07-18, and Appendix D.9 gives the complete prompt.

4 Main Results

4.1 Anonymization Performance in Earnings Calls

We examine the effectiveness of various anonymization algorithms by analyzing whether LLMs can recognize firm and date entities from the anonymized texts. First, we analyze the number of entities that were identified by anonymization algorithms. To achieve this goal, we tokenize these call transcript data using the Xenova/GPT-4o tokenizer.[8] We then use two 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 9,539 tokens, with 25th and 75th percentiles at 7,977 and 11,237 tokens, respectively. We also report the percentage of identified entities based on each anonymization algorithm. Because the LLM generates plain text rather than structurally tagged tokens, we calculate the entity percentage by measuring the proportion of anonymized placeholders in the final output for both methods. This ensures the metrics are comparable.

Next, we examine the effectiveness of the two anonymization algorithms. Following prior literature (e.g., Engelberg et al., 2025), we ask GPT-4o-mini to identify the masked date, firm name, ticker, and industry associated with each anonymized transcript. The deanonymization prompt is detailed in Appendix D.2, and Panel B reports the resulting identification rates.

We find that the TRF algorithm exhibits weaker anonymization performance. GPT-4o-mini successfully identifies 66.88% of firm names and 53.90% of tickers, although date identification remains low at 6.03%. In contrast, the LLM-based approach is substantially more thorough. Consistent with Engelberg et al. (2025), it drastically reduces re-identification rates, leaving only 9.04% of firm names and 2.81% of dates identifiable. This result underscores the superior capability of the iterative LLM approach to obscure firm-specific

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

information compared with standard NER-based substitution.

4.2 Predicting Credit Rating Downgrades

Our headline application examines whether anonymization degrades the predictive content of earnings call transcripts for a salient and economically important outcome: future credit-rating downgrades. Following Donovan et al. (2021), who show that machine-learning methods can extract qualitative credit-risk signals from conference calls, we use GPT-4o-mini to construct a downgrade-probability score from the raw transcript and from its TRF- and LLM-anonymized versions; the extraction prompt is detailed in Appendix D.3. Panel A of Table 3 reports the pairwise correlations among these three scores. The raw-text score correlates 0.9065 with both the TRF- and the LLM-anonymized scores, and the two anonymized scores correlate 0.9311 with each other. Panel B reports correlations between the raw-text score and scores extracted from transcripts in which only one entity category is preserved. The raw score correlates 0.9459, 0.9159, 0.9324, and 0.9138 with scores preserving only NUMBERS, PLACES, AGENTS, and OTHERS, respectively. These high but imperfect correlations already hint at a loss of information through anonymization.

Before turning to formal regressions, we present visual evidence of this information loss in Figure 2, which plots the receiver operating characteristic (ROC) curves for predicting the realized 12-month downgrade using the RAW, TRF-, and LLM-anonymized downgrade scores. All three curves bow well above the 45-degree reference line, confirming that each score carries genuine predictive content for downgrades. The ordering among them, however, is unambiguous: the RAW curve lies above both anonymized curves across essentially the entire range of the false-positive axis and therefore has a larger estimated area under the curve (AUC). The TRF and LLM curves, by contrast, are nearly indistinguishable from each other and sit consistently below RAW, indicating that the two anonymization procedures discard a similar amount of predictive content despite their very different underlying architectures. Crucially, the transcripts underlying this figure are drawn entirely from the

post-cutoff window (November 2023–December 2024), which postdates the October 2023 knowledge cutoff of the extraction model. The RAW curve's dominance therefore cannot reflect look-ahead bias, and instead constitutes direct visual evidence that the gap between the raw and anonymized curves is attributable to the information lost when named entities are removed. We now quantify this loss in a regression framework.

To formally test for information loss, we regress the realized downgrade indicator $Downgrade_{i,t+1}$—equal to one if S&P downgrades the firm's credit rating within the twelve months following the earnings call, and zero otherwise—on the extracted downgrade-probability scores:

$$Downgrade_{i,t+1} = \beta \widehat{Downgrade}_{i,t} + \gamma X_{i,t} + \delta_t + \epsilon_{i,t}, \tag{9}$$

where $\widehat{Downgrade}_{i,t}$ is the score extracted from the raw, TRF-, or LLM-anonymized text, and δ_t denotes SIC3 × Quarter fixed effects. The control vector $X_{i,t}$ contains $\ln(FirmValue)$, $Leverage_{Asset}$, MTB, ROA, $RevGrowth$, $Std(Income)$, $Std(Revenue)$, $Zscore$, and $Oscore$. All independent variables are standardized to facilitate comparison across text treatments. We report the results in Table 4, without firm-level controls in Panel A and with the full set of controls in Panel B.

In the standalone specifications (Columns 1–3 of Panel B), the raw-text score is a strong and highly significant predictor of downgrades (coefficient 0.061, $t = 5.52$), and the TRF- and LLM-anonymized scores are also significant (0.047, $t = 4.80$; and 0.049, $t = 4.92$) but carry smaller coefficients and lower adjusted R^2—a first indication of information loss. The horse-race specifications (Columns 4 and 5) are decisive. When the raw-text score and an anonymized score enter the same regression, the coefficient on $\widehat{Downgrade}_{RAW}$ remains large and significant (0.073, $t = 3.58$ in Column 4), whereas the coefficient on $\widehat{Downgrade}_{TRF}$ becomes economically negligible and statistically indistinguishable from zero (-0.012, $t = -0.72$). The same pattern holds for the LLM-anonymized score (-0.007, $t = -0.52$ in Column 5). The predictive content of the anonymized scores is thus almost entirely subsumed

by the raw-text score, providing strong evidence that entity removal strips the transcript of credit-risk-relevant information. As shown in Panel A, the results are robust to excluding control variables.

4.3 The Role of Different Types of Entities

To localize the source of this information loss, we re-extract the downgrade score from transcripts in which the TRF model preserves only one entity category—NUMBERS, PLACES, AGENTS, or OTHERS—while anonymizing the other three. As shown in Appendix Table B.3, NUMBERS and AGENTS are far more frequent in our corpus than PLACES and OTHERS, so their removal might be expected to matter most. Table 5 reports horse-race regressions of $Downgrade_{i,t+1}$ on the raw-text score together with each single-category-preserved variant. In the standalone columns (1–5) every variant is statistically significant, but in the horse races (Columns 6–9) all four are driven to insignificance once the raw-text score is included: the coefficient is 0.019 ($t = 1.06$) on $\widehat{Downgrade}_{NUM}$, -0.011 ($t = -0.65$) on $\widehat{Downgrade}_{PLC}$, 0.017 ($t = 0.90$) on $\widehat{Downgrade}_{AGT}$, and 0.003 ($t = 0.16$) on $\widehat{Downgrade}_{OTH}$, while $\widehat{Downgrade}_{RAW}$ remains significant throughout. Preserving any single entity category is therefore insufficient to recover the credit-risk signal embedded in the full, un-anonymized transcript.

4.4 Sample Period and the Limited Role of Look-Ahead Bias

A natural concern is that the raw transcript's apparent advantage reflects look-ahead bias. An LLM trained on data through its knowledge cutoff may already "know" a firm's subsequent downgrade, in which case anonymization would be valuable because it removes this leakage rather than costly because it destroys genuine information.[9] To assess this possibility, we extend the downgrade sample to include a pre-knowledge-cutoff window from November 2022 through October 2023. We retain only firms appearing in both the pre-

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

and post-cutoff windows and ask whether the predictive content of the extracted downgrade scores is systematically greater before the cutoff.

We first present the evidence visually. Figure 3 plots quarter-by-quarter coefficients on the downgrade score. The coefficient on the raw-text score consistently exceeds those on the anonymized scores on both sides of the cutoff, marked by the dashed vertical line, with no visible increase as the sample moves into the pre-cutoff period. We then estimate

$$Downgrade_{i,t+1} = \beta_1 \widehat{Downgrade}_{i,t} + \beta_2 (\widehat{Downgrade}_{i,t} \times Z_{i,t}) + \beta_3 Z_{i,t} + \gamma X_{i,t} + \delta_t + \epsilon_{i,t}, \quad (10)$$

where $Z_{i,t}$ is a moderating variable. Setting $Z_{i,t} = Pre$, an indicator equal to one for transcripts in the pre-cutoff window, Table 6 reports the estimates. The downgrade score is positive and significant in all three specifications: the coefficients are 0.061 ($t = 5.35$), 0.047 ($t = 4.55$), and 0.050 ($t = 4.82$) for the RAW, TRF, and LLM scores, respectively. By contrast, the corresponding interactions with Pre are small and statistically insignificant: -0.005 ($t = -0.31$), -0.009 ($t = -0.59$), and -0.010 ($t = -0.63$).

One feature of this test makes it conservative. The baseline outcome records downgrades realized within twelve months after each earnings call. Because the pre-cutoff sample extends through October 2023, every twelve-month prediction window overlaps, to some degree, with dates after the model's knowledge cutoff. Downgrades realized during this post-cutoff portion could not have entered the model's training data. Consequently, even if look-ahead bias were present for downgrades realized sufficiently early, the inclusion of subsequently realized outcomes could attenuate the estimated interaction with Pre.

We address this timing concern by shortening the prediction horizon from twelve months to three months. We redefine the dependent variable to indicate whether S&P downgrades the firm within three months after the earnings call and re-extract the downgrade-probability scores using an otherwise identical prompt that replaces "twelve months" with "three months." The shorter horizon sharply limits the overlap between the pre-cutoff pre-

diction windows and post-cutoff dates, making realized pre-cutoff downgrades substantially more likely to have been available in the model's training data.

Appendix Table B.4 reports the results. The three-month downgrade scores remain positive and significant in the post-cutoff period, with coefficients of 0.026 ($t = 4.52$), 0.023 ($t = 3.86$), and 0.020 ($t = 3.71$) for the RAW, TRF, and LLM scores, respectively. More importantly, none of the interactions with Pre is positive or statistically significant. The interaction coefficients are -0.006 ($t = -0.68$) for RAW, -0.010 ($t = -1.30$) for TRF, and -0.002 ($t = -0.30$) for LLM. Thus, tightening the outcome window in a way that makes the test more sensitive to memorized downgrade outcomes produces essentially the same conclusion as the twelve-month specification.

If look-ahead bias materially inflated the extracted scores, their predictive power should be stronger before the knowledge cutoff, producing a positive interaction with Pre. We find no such pattern in either the twelve-month or the three-month specification. Under the maintained assumption that the mapping between contemporaneous credit-risk information and future downgrades is otherwise stable across the adjacent sample periods, the evidence suggests that look-ahead bias plays only a limited role in our setting. The raw transcript's predictive advantage instead appears to arise primarily from genuine credit-risk information that anonymization removes or distorts.

4.5 The Anonymization Gap and Information Loss

We next implement the gap in equation (6). Let R and A denote the winsorized and standardized raw and anonymized downgrade scores. For each anonymization method, we estimate ρ over the pooled balanced sample and construct $Gap = |A/\rho - R|$; we then winsorize and standardize Gap and interact it with the standardized anonymized score. The estimated values of ρ are 0.8993 for TRF and 0.8997 for LLM. Setting $Z_{i,t} = Gap$ in Equation (10), Table 7 reports the estimates for the full sample and the pre- and post-cutoff subsamples.

The interaction is negative throughout. In the full sample, the coefficients are -0.019

($t = -3.62$) for TRF and -0.012 ($t = -2.37$) for LLM. Because both the score and *Gap* are standardized, a one-standard-deviation increase in *Gap* reduces the marginal predictive effect from its value at the mean gap by approximately 37% for TRF (0.019/0.051) and 24% for LLM (0.012/0.049). The pre-cutoff interactions are -0.015 ($t = -1.88$) and -0.013 ($t = -1.46$), while the post-cutoff interactions are -0.023 ($t = -3.02$) and -0.013 ($t = -2.19$), respectively. Thus, the anonymized score predicts downgrades least effectively where its attenuation-adjusted deviation from the raw score is greatest. In particular, both post-cutoff estimates are statistically significant and cannot be attributed to the removal of look-ahead bias; the same negative signs in the pre-cutoff sample are directionally consistent with this information-loss channel, although they are estimated less precisely.

4.6 Alternative LLMs

To verify that our findings are not an artifact of a single extraction model, we repeat the downgrade extraction using GPT-4o and o3-mini in addition to our workhorse GPT-4o-mini. Table 8 reports the results. Across all three models, the raw-text score predicts downgrades (Columns 1–3), while the TRF-anonymized score has a smaller coefficient and lower adjusted R^2 in every standalone specification (Columns 4–6). In the horse races (Columns 7–9), the anonymized coefficient is -0.012 ($t = -0.72$) for GPT-4o-mini, 0.025 ($t = 1.73$) for GPT-4o, and approximately zero ($t = -0.04$) for o3-mini; the raw-text score remains significant in every case. Thus, the anonymized signal is either fully subsumed or, for GPT-4o, only marginally significant once the raw signal is included. The information loss from anonymization is robust to the choice of extraction LLM.

5 Dissecting Information Loss: A Ten-Factor Credit-Risk Audit

Section 2 describes anonymization as the combination of two operations. It can remove part of the raw signal, lowering the retention parameter ρ, and it can distort the remaining assessment through u. The one-shot downgrade score establishes that the two operations jointly destroy predictive content, but it cannot show where the loss occurs. To make that distinction observable, we give the two terms textual counterparts.

We call the first channel *evidence displacement*. It occurs when a credit-risk reason found in the raw transcript disappears after anonymization or when the model replaces its supporting passage with a different one. The term is deliberately broader than "information removal": anonymization may leave the number of extracted reasons unchanged while redirecting the model toward less informative evidence. We call the second channel *interpretive distortion*. It occurs when the model continues to cite the same passage but changes the risk assessment attached to it after identifying information and quantitative context have been removed. Evidence displacement is the textual counterpart of lower signal retention; interpretive distortion is the textual counterpart of u.

This distinction organizes the analysis that follows. We first build a structured, quotation-backed credit-risk audit and use it to document both channels in the text. We then ask whether the ten factors jointly reproduce the information loss in the one-shot score, whether each channel separately accounts for lost predictive content, and finally how much of the total loss should be attributed to each channel.

5.1 A Two-Channel Credit-Risk Audit

The audit requires a common analytical structure and a textual anchor for each assessment. We obtain the structure from ten factors of S&P Global Ratings' Corporate Method-

ology. Industry Risk, Country Risk, and Competitive Position describe the business risk profile; Cash Flow/Leverage describes the financial risk profile; Diversification/Portfolio Effect, Capital Structure, Financial Policy, Liquidity, and Management and Governance operate as post-anchor modifiers; and Comparable Ratings Analysis provides a final relative-positioning assessment. Appendix C.1 reports the S&P credit rating framework and definitions of ten factors.

This taxonomy disciplines the extraction without attempting to reproduce an official S&P rating or stand-alone credit profile, which also depends on information outside the call, analytical adjustments, forecasts, peer comparisons, and committee judgment. Our narrower question is whether the transcript contains issuer-specific evidence relevant to downgrade risk over the next twelve months and, if so, how adversely the model reads that evidence.

We apply the same GPT-4o-mini-2024-07-18 snapshot, temperature zero, prompt, and output schema separately to the raw, TRF-anonymized, and LLM-anonymized versions of every transcript. For each factor, the model returns a score in $[0, 1]$ and one exact supporting quote; a higher score represents a higher probability of being downgraded in the future based on the factor. Requiring a quote ties an otherwise opaque score to a passage that can be tracked through anonymization. If the transcript contains no qualifying issuer-specific evidence for a factor, both fields are null. The full prompt appears in Appendix D.9.

The descriptive audit covers 2,488 post-cutoff transcripts with valid extractions, drawn from the post-cutoff sample used in the main analysis. Because every call postdates the model's October 2023 knowledge cutoff, differences between the raw and anonymized readings cannot plausibly be attributed to the removal of memorized future outcomes.

5.2 Measuring Displacement and Distortion

Let j index transcripts, k index credit-risk factors, and $v \in \{RAW, TRF, LLM\}$ denote the transcript version. Let $\mathcal{K}_j^v$ be the set of factors for which version v returns both a non-null score and a non-null supporting quote. Let s_{jk}^v denote the score assigned to factor k in

version v. The notation $|\mathcal{K}_j^v|$ denotes the cardinality of $\mathcal{K}_j^v$—that is, the number of factors with valid score–quote pairs. The composite is the average score across these factors:

$$C_j^v = \frac{1}{|\mathcal{K}_j^v|} \sum_{k \in \mathcal{K}_j^v} s_{jk}^v. \tag{11}$$

A factor enters the average only when both its score and quote are non-null. The composite therefore measures average assessed probability of the downgrade across the factors for which the transcript supplies qualifying evidence.

To separate the two channels, we next compare the quoted passages. We remove entities, placeholders, and numerical tokens that would mechanically prevent a match, normalize the remaining text, and compute a token-sort similarity score. This comparison divides the extracted evidence into three intuitive groups: *shared reasons*, for which both versions cite the same normalized passage; *RAW-only reasons*, which disappear or switch passages after anonymization; and *anonymized-only reasons*, which replace them.

We then calculate four variables whose names describe their construction: *SharedRaw* and *SharedAnon* are the average raw and anonymized scores on shared reasons; *RawOnly* is the average score on RAW-only reasons; and *AnonOnly* is the average score on anonymized-only reasons. Comparing *SharedRaw* with *SharedAnon* holds the cited passage fixed and captures interpretive distortion. Comparing *RawOnly* with *AnonOnly* captures evidence displacement. The overall composite is a weighted average of the relevant shared and version-specific reasons, not the sum of their group means. Appendix C.2 gives the detailed procedure and the exact set-based formulas.

Figure 4 reports evidence coverage by factor. The total number of reasons falls from 18,985 in the raw extractions to 18,245 under TRF and 18,293 under LLM anonymization—declines of 3.9% and 3.6%. This aggregate decline provides initial evidence that anonymization reduces evidence coverage. Some dimensions lose evidence outright: Country Risk reasons fall by 46% under TRF and 40% under LLM anonymization, while Comparable Rat-

ings Analysis falls by 42% and 32%. More importantly, replacement of the cited evidence is far more pervasive than the aggregate counts suggest. Competitive Position, for example, is populated almost as often after anonymization as before, yet only 34% of raw reasons retain the same TRF quote anchor and only 30% retain the same LLM anchor. The same broad pattern appears across the factor bars: many dimensions preserve their reason counts while replacing a large share of the passages used to support those reasons. Anonymization therefore changes what the model treats as evidence even when it does not reduce how much evidence the model reports.

Anonymization also changes how the selected evidence is read. Figure 5 compares factor scores when both versions contain evidence. The changes are predominantly positive: nine of the ten TRF differences and seven of the ten LLM differences remain significant at the 5% family-wise level after the Holm correction. Under TRF anonymization, the average Cash Flow/Leverage, Capital Structure, Financial Policy, and Liquidity scores rise by roughly three to four percentage points; many LLM-anonymized scores rise by one to two percentage points. Removing identity thus shifts the model's factor-level assessments systematically upward. These score and reason comparisons are descriptive outputs of the extraction model. They show that anonymization changes the structure of the audit, but do not by themselves establish that the changes destroy economically relevant information. We test that implication next using realized downgrades.

5.3 Does the Composite Preserve Predictive Content?

The first regression asks whether the ten-factor audit reproduces the information loss in the one-shot downgrade score. Let j continue to index a transcript, and let D_j indicate whether its firm is downgraded over the next twelve months. C_j^{RAW} and C_j^{Anon} represent the average score across ten factors from raw and anonymized text separately, and *Anon* denotes TRF or LLM. For each anonymization method, we estimate standalone specifications and

the horse race

$$D_j = \alpha + \beta_R C_j^{RAW} + \beta_A C_j^{\text{Anon}} + \gamma X_j + \delta_t + \varepsilon_j. \tag{12}$$

The control vector X_j, SIC3 $\times$ Quarter fixed effects δ_t, standardization, and SIC3-clustered standard errors are identical to those in equation (9). Standalone specifications include one composite at a time.

Before turning to the regressions in Table 9, we validate the composite against the one-shot downgrade score. The ten-factor composites correlate 0.819, 0.798, and 0.809 with the corresponding one-shot downgrade scores for raw, TRF-anonymized, and LLM-anonymized transcripts, respectively. These high correlations indicate that the factor-based decomposition largely recovers the information captured by the one-shot scores, suggesting that the ten factors provide a meaningful representation of the underlying credit-risk dimensions. In the standalone regressions, a one-standard-deviation increase in the raw composite predicts a 4.5-percentage-point increase in downgrade probability ($t = 4.41$), compared with 3.5 percentage points for both TRF ($t = 4.22$) and LLM anonymization ($t = 4.10$).

The horse races reveal the lost information. Against TRF, the raw coefficient is 0.045 and remains significant, while the anonymized coefficient is approximately zero and statistically insignificant. Against LLM, the raw coefficient is 0.046, while the anonymized coefficient falls to -0.001 and is again insignificant. The ten-factor composite therefore reproduces the information loss found with the one-shot score. This result gives economic content to the descriptive audit: the disappearance and replacement of credit-risk reasons, together with the systematic upward drift in factor assessments, are not merely changes in model output. Collectively, they accompany a loss of predictive information about future downgrades.

5.4 Which Channel Destroys Predictive Content?

The second regression replaces the two composites with the four conditional means:

$$\begin{aligned} D_j = \alpha &+ \beta_1 \text{SharedRaw}_j + \beta_2 \text{SharedAnon}_j \\ &+ \beta_3 \text{RawOnly}_j + \beta_4 \text{AnonOnly}_j + \gamma X_j + \delta_t + \varepsilon_j. \end{aligned} \tag{13}$$

The shared pair compares how the same evidence is interpreted; the version-specific pair compares the average predictive content of the evidence displaced by and introduced through anonymization. The controls, fixed effects, and error clustering are the same as in equation (12); the four conditional means are winsorized but retain their original $[0, 1]$ units.

Table 10 shows that interpretive distortion matters even when the evidence anchor is held fixed. For reasons shared between the raw and TRF readings, the two conditional means have positive standalone coefficients of 0.293 and 0.234. When both enter together, the raw-reading coefficient is 0.287 ($t = 3.74$), while the anonymized-reading coefficient falls to 0.007 ($t = 0.12$). The LLM comparison is similar: standalone coefficients of 0.364 and 0.272 become 0.385 ($t = 4.29$) and -0.026 ($t = -0.35$) in the horse race. The model continues to point to the same broad passage, but the assessment formed without the original context contains no positive incremental information.

Evidence displacement produces the same loss through a different route. In the raw–TRF comparison, the RAW-only and TRF-only means have standalone coefficients of 0.331 and 0.270. In their horse race, the RAW-only coefficient remains 0.287 ($t = 2.80$), whereas the TRF-only coefficient falls to 0.056 ($t = 0.94$). Under LLM anonymization, standalone coefficients of 0.314 and 0.258 become 0.247 ($t = 2.08$) and 0.074 ($t = 0.90$). The reasons that anonymization displaces retain incremental predictive content; the reasons that replace them do not.

The full specifications bring both channels together in the last column. For TRF, *SharedRaw* remains positive and significant at 0.216, while *RawOnly* remains positive at 0.151 but

is estimated less precisely; both anonymized measures are insignificant. For LLM, the shared RAW mean remains positive and significant at 0.352; the RAW-only mean is 0.035 and imprecisely estimated, and both anonymized means are again insignificant. The significance of *SharedRaw* but not *RawOnly* suggests that the predictive signal is concentrated in evidence that remains recognizable after anonymization, whereas RAW-only evidence provides limited incremental information once the shared component is accounted for. The negative coefficients on some anonymized means should not be interpreted as causal protective effects; they capture residual variation among highly correlated signals. The central conclusion is simpler: both changing the evidence set and changing the reading of preserved evidence contribute to information loss.

5.5 How Much Does Each Channel Matter?

The component regressions establish that both channels matter, but their coefficients do not provide an additive attribution of the total loss. The conditional means are defined on different reason sets and are highly correlated, and the measured contribution of one channel may depend on whether it is introduced before or after the other. We address this path dependence with a Shapley decomposition of the decline in adjusted R^2.

For each anonymization method, we construct four states with descriptive names. *Raw* is the original composite score based on the unaltered transcript. *Replace* substitutes anonymized-only reasons for RAW-only reasons while retaining the raw interpretation of shared reasons; it therefore introduces only evidence displacement. *Reread* retains the raw evidence set but applies the anonymized interpretation to shared reasons; it therefore introduces only interpretive distortion. *Anon* is the fully anonymized composite score, incorporating both evidence displacement and interpretive distortion. For each state, we aggregate the underlying factor-level scores into a transcript-level composite score. We then estimate the same predictive regression separately using each composite score as the primary explanatory variable and collect the corresponding R^2. Within each anonymization method, all four

regressions use the common sample for which the four composites, the outcome, and the controls are observed; the TRF and LLM common samples are constructed separately. These four predictive performances provide the basis for the subsequent Shapley decomposition of information loss. The detailed procedure is described in Appendix Section C.3.

Let R^2_{Raw}, R^2_{Replace}, R^2_{Reread}, and R^2_{Anon} denote the adjusted R^2 from entering each state separately into the same downgrade regression. Total information loss is $L = R^2_{\text{Raw}} - R^2_{\text{Anon}}$. The Shapley contributions of evidence displacement (D) and interpretive distortion (I) are

$$\phi_D = \frac{1}{2}\left[\left(R^2_{\text{Raw}} - R^2_{\text{Replace}}\right) + \left(R^2_{\text{Reread}} - R^2_{\text{Anon}}\right)\right],$$
$$\phi_I = \frac{1}{2}\left[\left(R^2_{\text{Raw}} - R^2_{\text{Reread}}\right) + \left(R^2_{\text{Replace}} - R^2_{\text{Anon}}\right)\right].$$

Each contribution averages the channel's marginal loss across the two possible orderings, and $\phi_D + \phi_I = L$ by construction.

Figure 6 shows that the two channels are economically comparable, although interpretive distortion accounts for the larger share. Evidence displacement accounts for 40.6% of the adjusted R^2 loss under TRF anonymization and 41.3% under LLM anonymization. Interpretive distortion accounts for the remaining 59.4% and 58.7%, respectively. Anonymization therefore does not merely alter which passages the model treats as evidence. Approximately three-fifths of the loss arises even when the evidence anchor is preserved, because stripping away entity and numerical context changes what the model believes the surviving passage means.

6 Information Loss Beyond Credit Risk

We next examine whether the information loss documented for credit risk extends to other economically important outcomes and textual sources. We begin with earnings-call sentiment and announcement returns, then study transcript-based measures of uncertainty, investment plans, and macroeconomic outlook, and finally replicate the return analysis using

news headlines. The return setting is especially useful because mitigating look-ahead bias is a central concern in asset-pricing applications of LLMs (e.g., Tetlock et al., 2008; Engelberg et al., 2025). Across these settings, the comparison remains the same: we evaluate whether the signal extracted from anonymized text retains the predictive content of the corresponding raw-text signal.

6.1 Earnings Call Sentiment and Stock Returns

We extract sentiment from the original and anonymized transcripts using GPT-4o-mini, with a prompt that closely follows Engelberg et al. (2025) (detailed in Appendix D.7). We tabulate the pairwise correlation between the extracted sentiment scores in Appendix Table B.5. The correlation between the sentiment score extracted from the raw and the TRF-anonymized transcripts is 0.864, while the correlation with the LLM-anonymized transcripts is 0.858, indicating that these signals remain highly correlated. As the correlation between the anonymized and the raw score is not perfect, it already suggests some loss of information from the anonymization process. To formally test the severity of this information loss, we examine the association between the extracted sentiment scores and the DGTW-adjusted announcement return ($DGTW_{i,t+1}$, as defined by Daniel et al. (1997)).

For the return-prediction analysis, we first assign each conference call to an announcement trading date based on when its information becomes available to the market. A call held before the U.S. market closes at 16:00 ET is assigned to the same trading day, whereas a call held after the market closes is assigned to the next trading day. Thus, for example, a call held at 15:00 ET on Monday is matched to Monday's return, while a call held at 17:00 ET on Monday is matched to Tuesday's return. This convention ensures that the closing price used to calculate the announcement return is observed after the transcript information has been disclosed, and the DGTW-adjusted announcement return $DGTW_{i,t+1}$ fully incorporates the information disclosed in the transcript.

Empirically, we adopt the following regression model:

$$DGTW_{i,t+1} = \beta\,\text{Sentiment}_{i,t} + \gamma X_{i,t} + \delta_t + \epsilon_{i,t} \tag{14}$$

The key independent variables are the sentiment scores extracted from the raw text and from the texts anonymized using the TRF and LLM methods, respectively. Following Tetlock et al. (2008) and Engelberg et al. (2025), we include a standard set of control variables, $X_{i,t}$: analyst forecast error (FE), shown by prior work to explain abnormal returns (Hirshleifer et al., 2009); DGTW-adjusted returns over the run-up window ($DGTW_{i,t}$, $DGTW_{i,t-1}$, $DGTW_{i,t-2}$, $DGTW_{i,t-21,t-3}$, $DGTW_{i,t-252,t-22}$); 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, and all independent variables are standardized to allow a direct comparison of coefficients across text treatments. In the specification with controls (reported in Panel B of Table 11), the sentiment score from the raw text exhibits a strong and highly significant relationship with the return: as reported in Column 1, the coefficient on the raw signal is 2.487 ($t = 18.832$), confirming that LLMs effectively capture sentiment signals. As shown in Columns 2 and 3, sentiment scores from anonymized texts (TRF and LLM) remain significantly associated with returns, but their coefficients are attenuated to 2.331 and 2.329, respectively, accompanied by a decline in adjusted R^2 from 0.132 to 0.124, indicating information loss across both anonymization methods.

To compare the signals more formally, we conduct pairwise regressions that include both the raw and an anonymized sentiment score (Columns 4 and 5). The coefficient on $Sentiment_{RAW}$ remains large and highly significant, whereas the coefficients on the anonymized scores contract sharply: the $Sentiment_{TRF}$ coefficient falls from 2.331 to 0.775 (Column 4), a 67% reduction, and the $Sentiment_{LLM}$ coefficient drops to 0.798 (Column 5). The sentiment-relevant information in the anonymized text is thus largely subsumed by that of the original text. As shown in Panel A of Table 11, the results are robust to excluding

control variables.

6.2 Alternative Transcript-Based Prediction Tasks

We next ask whether anonymization degrades signals constructed for outcomes other than credit ratings and stock returns. If identifying context helps an LLM interpret forward-looking statements, removing that context should weaken the text's ability to explain subsequent risk, investment, and operating performance. Table 12 reports horse-race regressions that include both the raw and TRF-anonymized score in each task.

First, prior research establishes a positive relation between textual uncertainty in corporate disclosures and subsequent return volatility (Loughran and McDonald, 2011). In Column 1, the raw uncertainty score is positive and significant (0.026, $t = 2.03$), whereas the TRF-anonymized score is smaller and statistically insignificant (0.012, $t = 0.78$). Second, following Jha et al. (2024a), Column 2 uses management's investment outlook to predict capital expenditures two quarters ahead. Both scores are significant, but the raw-text coefficient is larger (0.043 versus 0.033).

Third, drawing on Jha et al. (2024b), Columns 3 and 4 use the firm's macroeconomic outlook to predict sales growth and value-added growth two fiscal quarters ahead. Both raw and anonymized scores remain significant, but the raw coefficient is larger in each task: 1.803 versus 1.550 for sales growth and 2.460 versus 1.787 for value-added growth. Thus, anonymization weakens predictive content across alternative outcomes, with particularly pronounced attenuation for uncertainty.

6.3 News Headlines

We next turn to news headlines, another set of widely used textual data in finance (see, e.g., Bybee et al., 2023; Lopez-Lira and Tang, 2026; Engelberg et al., 2025). Compared with the verbose nature of earnings call transcripts, news headlines are extremely short but feature a much higher density of information, particularly entity-related information.

Because a single stock may have multiple headlines on a given day, we average all headline scores for each PERMNO-date to create a daily score. Our final sample comprises 52,716 news headlines from Yahoo Finance, corresponding to 30,950 firm-day observations , with the detailed sample-selection criteria reported in Panel C of Appendix Table B.1. The average headline length is merely 15.55 tokens, and entities recognized by the TRF model account for 38.30% of headline tokens on average, substantially higher than the 18.92% observed in earnings call transcripts under the same approach. Following the methodology used for transcripts, we design a prompt to extract sentiment from headlines (Appendix D.8) and average the scores across all headlines for a given firm-day.

Table 13 presents regressions explaining the DGTW-adjusted announcement return using news sentiment, where the return follows the same timing convention as for earnings calls. In both the single-score specifications (Columns 1 and 2) and the specifications with standard controls (Columns 4 and 5), sentiment derived from the raw ($Sentiment^{News}_{RAW}$) and TRF-anonymized ($Sentiment^{News}_{TRF}$) headlines is a significant positive predictor of returns, and the coefficients and adjusted R^2 are lower for the anonymized text, consistent with the transcript evidence. The horse-race regressions (Columns 3 and 6) are again decisive: without controls (Column 3) the coefficient on $Sentiment^{News}_{RAW}$ (0.315) remains significant while that on $Sentiment^{News}_{TRF}$ (0.135) is attenuated to less than a third of its standalone magnitude, and the pattern persists with controls (Column 6), where $Sentiment^{News}_{RAW}$ (0.314) remains substantially larger than $Sentiment^{News}_{TRF}$ (0.131). The news evidence thus corroborates the transcript findings: whether the source is a lengthy transcript or a concise headline, named entities are a significant source of the text's explanatory value, and removing them to mitigate look-ahead bias entails a corresponding loss of informational content.

6.4 Mechanisms and Robustness

The return setting reproduces several features of the credit-downgrade analysis, so we summarize the supporting evidence briefly.

First, when only one entity category is preserved at a time, none of the resulting sentiment scores recovers the full predictive content of raw sentiment (Appendix Table B.6). The NUMBERS- and AGENTS-preserved variants retain somewhat more incremental predictive content than the PLACES- and OTHERS-preserved variants, consistent with numerical and agent entities serving as particularly important contextual anchors.

Second, look-ahead bias again plays a limited role. Quarter-by-quarter coefficients (Appendix Figure B.1) show raw-text sentiment exceeding its anonymized counterparts on both sides of the knowledge cutoff. Estimating the interaction with the pre-cutoff indicator (Appendix Table B.7), we find that the $Sentiment \times Pre$ term is, if anything, negative—the opposite of the positive sign that look-ahead bias would predict—so the raw signal's advantage does not stem from leakage of future returns.

Third, the gap is once more a reliable proxy for information loss. Using $Gap = |A/\rho - R|$, with $\rho = 0.8529$ for TRF and 0.8510 for LLM, the $Sentiment \times Gap$ coefficients are -0.543 and -0.541 in the full sample, -0.431 and -0.401 in the pre-cutoff sample, and -0.635 and -0.666 in the post-cutoff sample; all are significant at the 1% level (Appendix Table B.8). Anonymized sentiment is therefore least informative exactly where its attenuation-adjusted deviation from raw sentiment is largest.

Finally, the information loss is robust to the extraction model—GPT-4o-mini, GPT-4o, and o3-mini all exhibit attenuated anonymized coefficients and a raw-text signal that dominates in horse races (Appendix Table B.9). These results echo the credit-downgrade evidence and require no separate interpretation.

7 Conclusion

This study provides comprehensive empirical evidence that anonymization, a widely used technique for mitigating look-ahead bias in LLM-based financial text analysis, carries a significant and underappreciated cost: information loss. By removing named entities and

related contextual information, anonymization can weaken the ability of LLMs to extract economically meaningful signals from financial texts.

Our econometric framework clarifies the source and interpretation of this decline. It represents the anonymized score as an attenuated version of the raw assessment plus document-specific distortion, while separately allowing scores extracted before the model's knowledge cutoff to contain look-ahead contamination. This framework connects the standalone and horse-race regressions, the knowledge-cutoff tests, and the anonymization gap within a common analytical structure.

Our main analysis examines the prediction of future S&P credit-rating downgrades. The downgrade probability extracted from raw earnings call transcripts strongly predicts subsequent downgrades. Scores extracted from anonymized transcripts remain informative in standalone regressions but contain little incremental predictive information once the raw score is included. The credit-risk content of the anonymized signals is therefore largely subsumed by that of the raw signal.

We next examine which types of entities carry this predictive information. Preserving any single category of entities is insufficient to recover the full predictive content of the raw transcript. Nevertheless, preserving numerical entities, such as financial figures and dates, or agent entities, such as firm and executive names, recovers substantially more of the raw signal than preserving place or other entities. We then extend the analysis across the models' knowledge cutoff. The raw signal's advantage remains comparable across the pre- and post-cutoff periods, and we find no significant positive interaction between the raw score and the pre-cutoff indicator. These results provide little evidence that the raw signal's advantage is concentrated in the period potentially exposed to look-ahead bias. We also introduce an attenuation-adjusted gap between the raw and anonymized signals as a document-level diagnostic of information loss. The anonymized signal predicts downgrades less effectively when this gap is larger, including in the post-cutoff sample. Finally, repeating the downgrade analysis using GPT-4o and o3-mini produces the same qualitative pattern, showing that the

main result is not specific to GPT-4o-mini.

Having established the aggregate information loss and its robustness, we use a ten-factor credit-risk audit to examine its underlying mechanisms. The audit shows that anonymization changes both the set of credit-risk reasons identified by the model and the assessments assigned to those reasons. We refer to these two channels as evidence displacement and interpretive distortion. A composite constructed from the ten-factor assessments reproduces the predictive loss found using the one-shot downgrade score. When the raw and anonymized assessments are compared directly, both the replacement of credit-risk evidence and the altered interpretation of preserved evidence contribute to the loss of predictive content. A Shapley decomposition attributes 40.6% of the adjusted R^2 loss under TRF anonymization and 41.3% of the loss under LLM anonymization to evidence displacement; interpretive distortion accounts for the remaining 59.4% and 58.7%, respectively. Anonymization therefore changes both what the model treats as relevant evidence and what it infers from the evidence that remains.

The information loss extends beyond credit risk. Sentiment extracted from raw earnings call transcripts strongly predicts announcement-window stock returns, while sentiment extracted from anonymized transcripts exhibits weaker predictive content and is substantially attenuated once raw sentiment is included. In additional transcript-based tasks, the raw-text measures have greater predictive power for future return volatility, capital expenditures, sales growth, and value-added growth than their anonymized counterparts. We obtain similar results using news headlines. Although sentiment extracted from anonymized headlines remains informative, its explanatory power is weaker than that of sentiment extracted from raw headlines. Additional analyses based on entity categories, the knowledge cutoff, the anonymization gap, and alternative extraction models confirm that these patterns are not confined to the credit-rating setting.

Our findings highlight a fundamental methodological trade-off. Anonymization may reduce an LLM's exposure to memorized firm-specific information, but it can also remove

genuine economic context that is necessary for accurate textual interpretation. Researchers should therefore not interpret a decline in predictive power after anonymization as evidence that look-ahead bias has been successfully eliminated. Without a post-cutoff benchmark, the same decline may instead reflect the loss of economically relevant information. Future work should develop methods that mitigate look-ahead bias while preserving the contextual information required for meaningful financial text analysis.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*.

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*.

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

Chen, S., Peng, L., and Zhou, D. (2026). Wisdom or whims? Investor opinions, trading strategies, and market outcomes. *Available at SSRN 4867401*.

Clayton, C., Coppola, A., Maggiori, M., and Schreger, J. (2025). Geoeconomic pressure. *National Bureau of Economic Research*.

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

Donovan, J., Jennings, J., Koharki, K., and Lee, J. (2021). Measuring credit risk using qualitative disclosure. *Review of Accounting Studies*.

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

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

Gao, Z., Jiang, W., and Yan, Y. (2025). Detecting lookahead bias in LLM forecasts. *arXiv preprint arXiv:2512.23847*.

Glasserman, P. and Lin, C. (2024). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *The Journal of Financial Data Science*.

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*.

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). Generative AI, managerial expectations, and economic activity. *arXiv preprint arXiv:2410.03897*.

Ke, Z. T., Kelly, B. T., and Xiu, D. (2026). Enhancing investment decisions with sentiment analysis: A probabilistic ranking framework. *Journal of the American Statistical Association*.

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. (2026). Caution ahead: Numerical reasoning and look-ahead bias in AI models. *Journal of Accounting Research*.

Lopez-Lira, A. and Tang, Y. (2026). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Journal of Financial Economics*.

Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). The memorization problem: Can we trust LLMs' economic forecasts? *Available at SSRN 5217505*.

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

Ludwig, J., Mullainathan, S., and Rambachan, A. (2026). Large language models: An applied econometric framework. *Annual Review of Economics*.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*.

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

Sarkar, S. K. (2026). Economic representations. *Available at SSRN 6999938*.

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*.

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

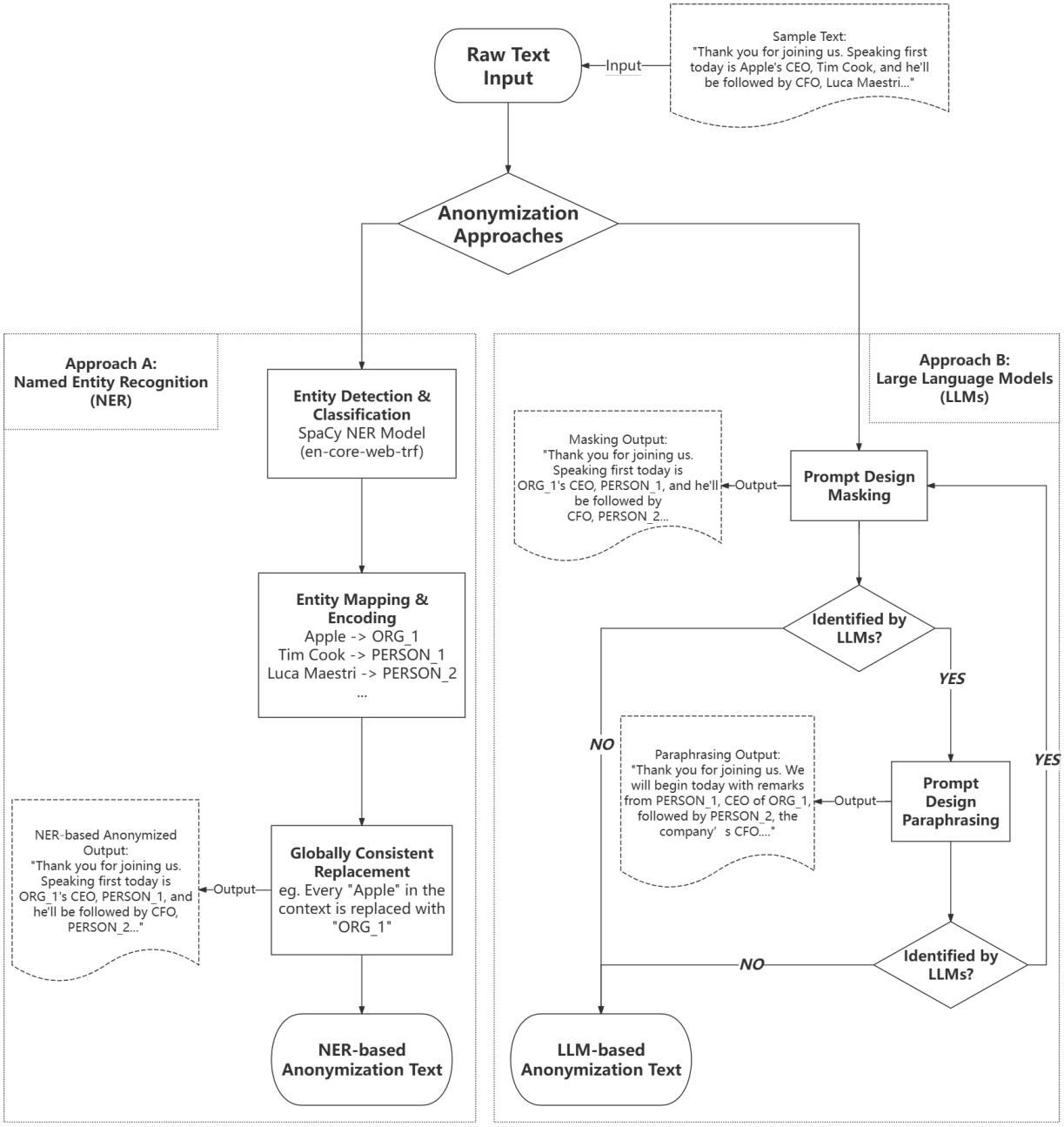


Figure 1: **Anonymization Pipeline**

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 model (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 utilizes a Large Language Model (LLM), specifically GPT-4o-mini, to perform anonymization through carefully designed prompts and an iterative process.

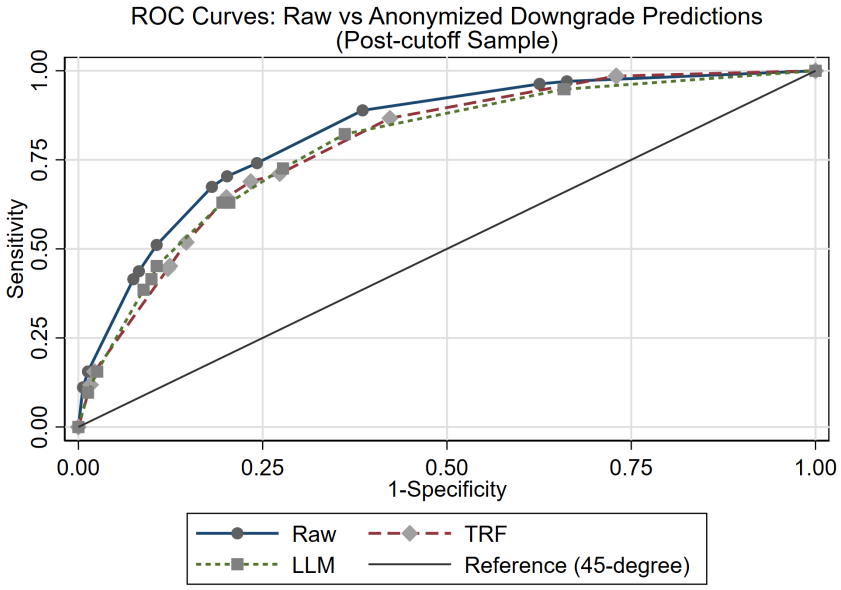
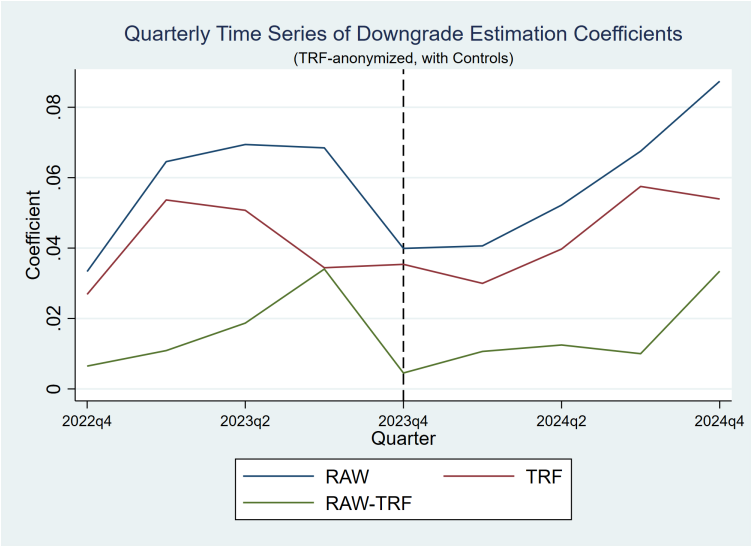
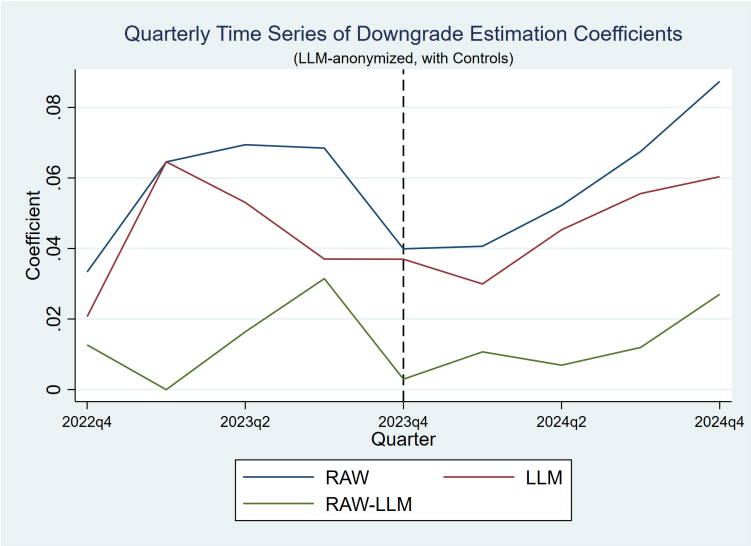


Figure 2: **ROC Curves of Downgrade Predictions Across Anonymization Methods**
This figure plots the receiver operating characteristic (ROC) curves for predicting the realized 12-month S&P credit rating downgrade using the GPT-4o-mini-extracted downgrade probability from RAW, TRF-anonymized, and LLM-anonymized transcripts. The sample contains 2,488 post-cutoff transcripts from November 2023 through December 2024.



(a) Quarterly Downgrade Predictive Coefficients (RAW and TRF-anonymized)



(b) Quarterly Downgrade Predictive Coefficients (RAW and LLM-anonymized)

Figure 3: **Quarterly Time Series of Downgrade Predictive Coefficients**

Each subfigure plots three quarterly time series on the balanced panel of firms appearing in both the pre- and post-cutoff samples: the quarterly regression coefficient of $Downgrade_{i,t+1}$ on $\widehat{Downgrade}_{RAW,i,t}$ (dark blue), on the anonymized counterpart (red), and the difference between the two coefficients. Each quarterly regression includes the same firm-level controls as Panel B of Table 4 and 3-digit SIC industry fixed effects, with standard errors clustered at the 3-digit SIC level. The vertical dashed line marks the LLM knowledge cutoff (2023 Q4). Subfigure (a) compares the coefficients derived from the RAW transcripts against those from the TRF-anonymized samples. Subfigure (b) compares the coefficients from the RAW transcripts against those from the LLM-anonymized samples.

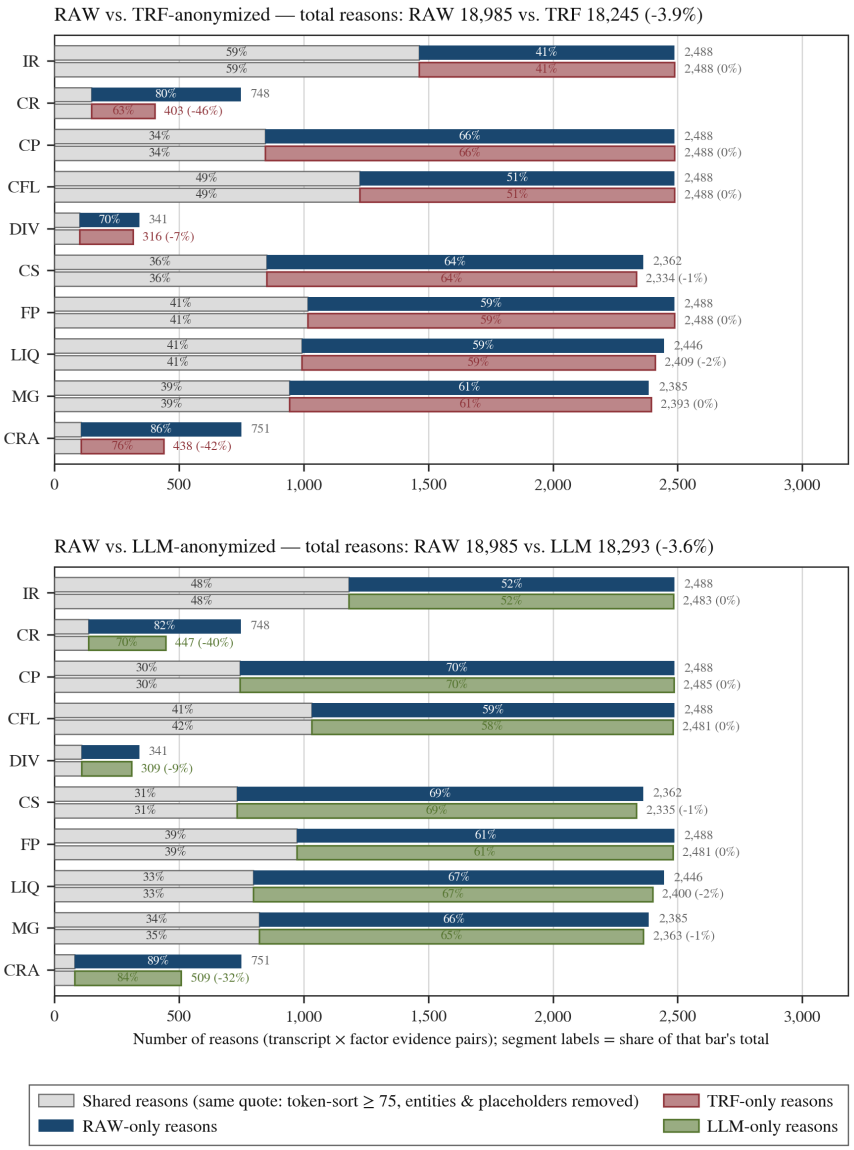


Figure 4: **S&P Ten-Factor Reason Coverage and Turnover After Anonymization**
This figure compares transcript-grounded credit-risk reasons extracted from the raw and anonymized versions of 2,488 post-cutoff earnings call transcripts. A reason is an evidence-backed transcript–factor assessment for which the model returns both a non-null score and a supporting quote under the ten-factor taxonomy described in Appendix C. For each anonymization method, the upper bar for a factor reports the raw-text reasons and the lower bar reports the anonymized-text reasons. Evidence is classified as matched when the raw and anonymized quotes have a token-sort similarity of at least 75 after spaCy-recognized entities are removed from the raw quote and placeholders and tokens containing numbers are removed from the anonymized quote. The remaining segments constitute displaced evidence: the raw segment combines evidence that disappears after anonymization and the raw side of a pair that cites a nonmatched anchor, while the anonymized segment is defined symmetrically. Segment labels report shares within each bar; numbers at the right report factor totals and the anonymized percentage change. IR, CR, CP, CFL, DIV, CS, FP, LIQ, MG, and CRA denote Industry Risk, Country Risk, Competitive Position, Cash Flow/Leverage, Diversification/Portfolio Effect, Capital Structure, Financial Policy, Liquidity, Management and Governance, and Comparable Ratings Analysis, respectively.

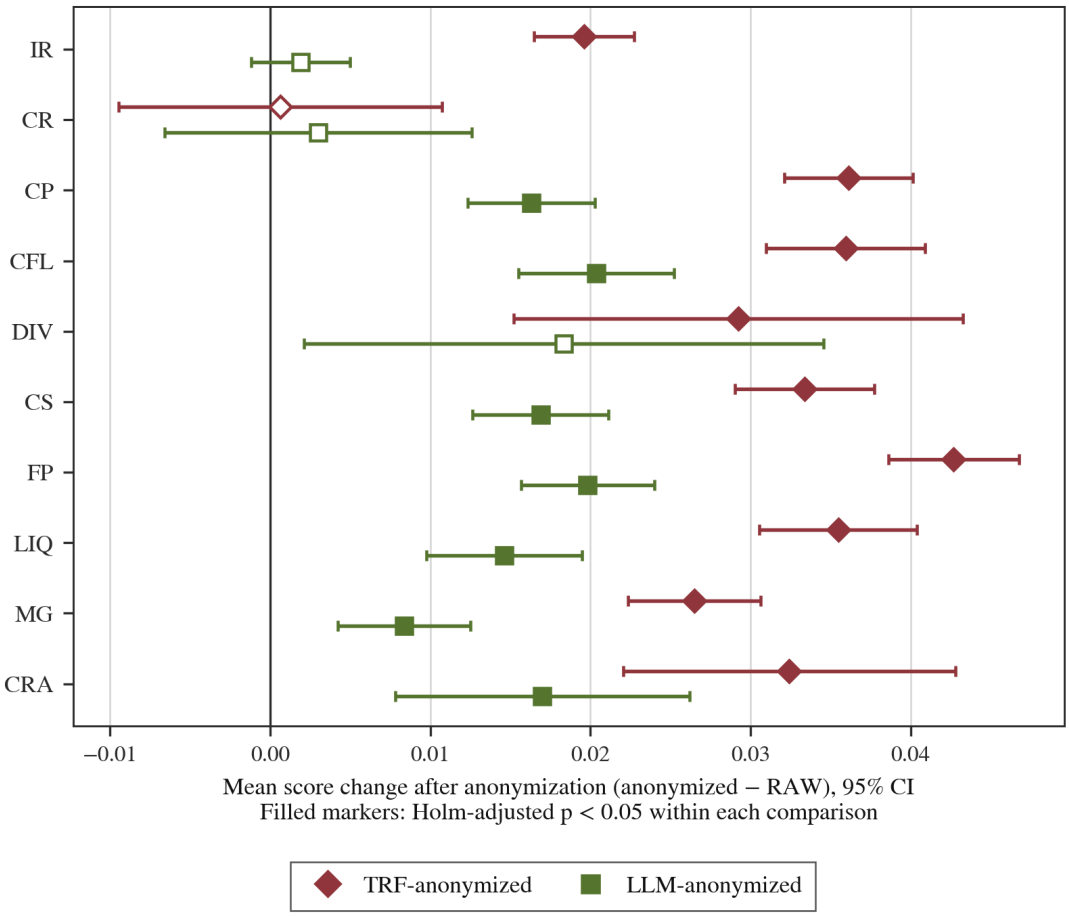


Figure 5: **Conditional Drift in S&P-Aligned Factor Scores After Anonymization**
This figure plots paired mean changes in the researcher-constructed factor-level downgrade-risk score after anonymization, computed as the anonymized score minus the raw score. The sample for each estimate is restricted to transcript–factor pairs for which both text versions return a non-null score and supporting quote. Scores range from zero to one, with larger values indicating a more adverse transcript-grounded credit-risk assessment within the factor's scope. Horizontal bars show 95% confidence intervals. Filled markers indicate that the paired mean difference remains significant at the 5% level after Holm adjustment across the ten factor-level tests within each raw–anonymized comparison. The factors and sample are defined as in Figure 4.

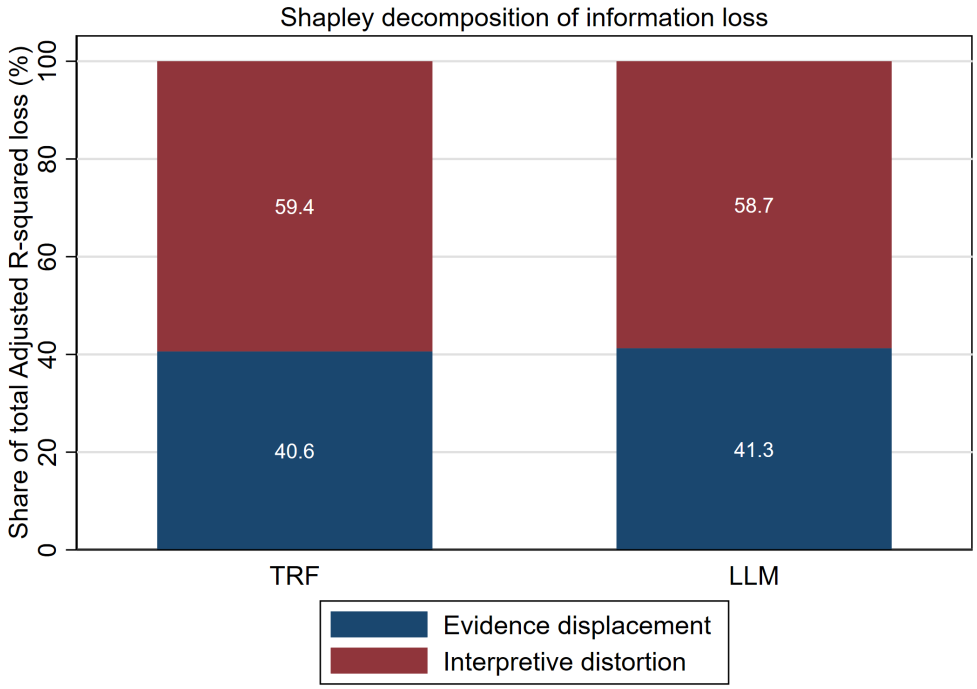


Figure 6: **Shapley Decomposition of Credit-Risk Information Loss**
This figure decomposes the decline in adjusted R^2 from the raw ten-factor composite to the fully anonymized composite into two channels: evidence displacement and interpretive distortion. Each stacked bar reports the Shapley share of the total adjusted R^2 decline for the TRF- and LLM-anonymized transcripts, respectively; the two components sum to 100 percent within each anonymization method. The values shown inside the bars are percentage shares. The decomposition procedure is described in detail in Appendix Section C.3.

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

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

Raw Text	**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. **ORG_1** assumes no obligation to update any forward-looking statements, which speak only as of the date they are made.	Please be aware that some of the information shared during our conversation today will include forward-looking statements, which may encompass, but are not limited to, projections regarding revenue, gross margin, operating expenses, other income and expenses, taxes, capital allocation, and future business outlook, including potential effects of macroeconomic conditions on the entity's operations and results. These statements carry risks and uncertainties that could lead to actual outcomes differing significantly from our predictions. For further details, please refer to the risk factors outlined in **Entity_A**'s most recent annual report on Form 10-K and the Form 8-K submitted to the regulatory body today, along with the related press release. **Entity_A** has no obligation to update any forward-looking statements, which are valid only as of the date they are made.

Table 2: **Conference Call Entity Count and Anonymization Performance (Downgrade Prediction Subsample)**

This table reports the entity count and anonymization performance for the downgrade-prediction post-cutoff subsample of 2,488 transcripts. Panel A reports descriptive statistics on transcript length and the share of tokens that are entities after anonymization, for the TRF and LLM (GPT-4o-mini) approaches. Length is measured in number of tokens. Entity% is calculated using the anonymized transcript, where the percentage represents the ratio of anonymized entity tokens to the total tokens in the anonymized version. Panel B reports the re-identification rates for four key attributes (Date, Firm Name, Ticker, and Industry) under the two anonymization approaches. We define a correct match using the following criteria: the estimated Date must fall within 7 days of the actual conference call date; Firm Name and Industry are evaluated using fuzzy string matching (Levenshtein token sort ratio), where a score of 75 or higher after text normalization is considered a match; and Ticker symbols require an exact case-insensitive match with the ground truth.

Panel A: Transcript Length and Entity Percentages

	Mean	**Std. Dev.**	**25th Pct.**	**Median**	**75th Pct.**
Length	9,539	2,323	7,977	9,888	11,237
Entity% (TRF)	18.92	2.53	17.20	18.71	20.55
Entity% (LLM)	11.51	3.42	9.47	11.26	13.19

Panel B: Identification Ratio %

Date	6.03	2.81
Firm Name	66.88	9.04
Ticker	53.90	9.12
Industry	50.56	44.01
Anonymization Approach	TRF	GPT-4o-mini
Entity Remove	ALL	ALL
Recognition Approach	GPT-4o-mini	GPT-4o-mini

Table 3: **Correlations of Downgrade Prediction Signals**

This table reports Pearson correlations among downgrade probabilities extracted using GPT-4o-mini. Panel A compares scores extracted from raw transcripts (RAW) with scores extracted from transcripts fully anonymized using TRF or LLM. Panel B compares the RAW score with scores extracted from transcripts in which only one entity category is preserved and the remaining categories are anonymized using TRF. NUM, PLC, AGT, and OTH denote specifications preserving only NUMBERS, PLACES, AGENTS, and OTHERS, respectively. All correlations are computed on the 2,488-transcript post-cutoff sample.

Panel A: Raw and Fully Anonymized Downgrade Scores

	$\widehat{Downgrade}_{RAW}$	$\widehat{Downgrade}_{TRF}$	$\widehat{Downgrade}_{LLM}$
$\widehat{Downgrade}_{RAW}$	1.000		
$\widehat{Downgrade}_{TRF}$	0.907	1.000	
$\widehat{Downgrade}_{LLM}$	0.907	0.931	1.000

Panel B: Raw and Single-Category-Preserved Downgrade Scores

	$\widehat{Downgrade}_{RAW}$	$\widehat{Downgrade}_{NUM}$	$\widehat{Downgrade}_{PLC}$	$\widehat{Downgrade}_{AGT}$	$\widehat{Downgrade}_{OTH}$
$\widehat{Downgrade}_{RAW}$	1.000				
$\widehat{Downgrade}_{NUM}$	0.946	1.000			
$\widehat{Downgrade}_{PLC}$	0.916	0.939	1.000		
$\widehat{Downgrade}_{AGT}$	0.932	0.899	0.912	1.000	
$\widehat{Downgrade}_{OTH}$	0.914	0.935	0.976	0.916	1.000

Table 4: **Horse Race for Downgrade Predictions**

This table reports regressions of the realized S&P credit-rating downgrade indicator $Downgrade_{i,t+1}$ on the model-extracted downgrade probability $\widehat{Downgrade}_{RAW}$, $\widehat{Downgrade}_{TRF}$, and $\widehat{Downgrade}_{LLM}$. Panel A reports regressions without firm-level controls; Panel B includes the full set of firm-level controls: $\ln(FirmValue)$, $Leverage_{Asset}$, MTB, ROA, $RevGrowth$, $Std(Income)$, $Std(Revenue)$, $Zscore$, and $Oscore$, all defined as in Appendix Table B.2. Within each panel, Columns (1)–(3) report standalone single-score regressions; Columns (4) and (5) report horse-race specifications that include the raw-text score together with one anonymized score. The coefficient on the anonymized score becomes statistically indistinguishable from zero once $\widehat{Downgrade}_{RAW}$ is included, indicating that the anonymized score's predictive content is fully subsumed by the raw-text score. The sample is the downgrade-task post-cutoff subsample (Nov 2023–Dec 2024). All variables are winsorized at the 1st and 99th percentiles, and all independent variables are standardized within the regression sample. All regressions include SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Downgrade_{i,t+1}$				
VARIABLES	(1)	(2)	(3)	(4)	(5)
Panel A: Without Controls					
$\widehat{Downgrade}_{RAW}$	0.077***			0.099***	0.095***
	(6.471)			(4.520)	(5.483)
$\widehat{Downgrade}_{TRF}$		0.062***		-0.024	
		(5.820)		(-1.350)	
$\widehat{Downgrade}_{LLM}$			0.064***		-0.019
			(5.864)		(-1.520)
Controls	No	No	No	No	No
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes
Observations	2,243	2,243	2,243	2,243	2,243
Adj. R^2	0.223	0.195	0.198	0.225	0.224
Panel B: With Controls					
$\widehat{Downgrade}_{RAW}$	0.061***			0.073***	0.068***
	(5.515)			(3.576)	(4.215)
$\widehat{Downgrade}_{TRF}$		0.047***		-0.012	
		(4.799)		(-0.719)	
$\widehat{Downgrade}_{LLM}$			0.049***		-0.007
			(4.919)		(-0.522)
Controls	Yes	Yes	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes
Observations	2,243	2,243	2,243	2,243	2,243
Adj. R^2	0.246	0.232	0.234	0.246	0.245

Table 5: **Comparing Downgrade Predictions Under Preserving Different Categories of Entities**

This table reports regressions of $Downgrade_{i,t+1}$ on the downgrade probability extracted from transcripts where only one of the four entity categories is preserved at a time, while the other three categories are anonymized using the TRF model. The category definitions are provided in Appendix Table B.3. Specifically, $\widehat{Downgrade}_{NUM}$, $\widehat{Downgrade}_{PLC}$, $\widehat{Downgrade}_{AGT}$, and $\widehat{Downgrade}_{OTH}$ are extracted from transcripts where the TRF model preserves only the NUMBERS, PLACES, AGENTS, or OTHERS category, respectively, while anonymizing all entities belonging to the other three categories. Columns (1)–(5) report standalone single-score regressions for the RAW score and the four single-category-preserved variants; Columns (6)–(9) include the RAW score together with each single-category-preserved variant. All four single-category-preserved variants lose statistical significance once RAW is also included, suggesting that the raw transcript subsumes the predictive content of any variant in which only one entity category is preserved. Control variables are identical to those in Panel B of Table 4. All variables are winsorized at the 1st and 99th percentiles, and all independent variables are standardized. All regressions include SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Downgrade_{i,t+1}$								
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\widehat{Downgrade}_{RAW}$	0.061***					0.044**	0.072***	0.045**	0.059***
	(5.515)					(2.380)	(3.535)	(2.272)	(2.976)
$\widehat{Downgrade}_{NUM}$		0.058***				0.019			
		(5.323)				(1.060)			
$\widehat{Downgrade}_{PLC}$			0.049***				-0.011		
			(4.890)				(-0.653)		
$\widehat{Downgrade}_{AGT}$				0.058***				0.017	
				(5.227)				(0.895)	
$\widehat{Downgrade}_{OTH}$					0.051***				0.003
					(5.075)				(0.157)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,243	2,243	2,243	2,243	2,243	2,243	2,243	2,243	2,243
Adj. R^2	0.246	0.243	0.233	0.242	0.237	0.246	0.246	0.246	0.245

Table 6: **Sample Period and Predictive Power of Downgrade Estimations**

This table examines whether the predictive power of downgrade-probability scores extracted from raw and anonymized earnings call transcripts differs between pre- and post-knowledge-cutoff subsamples. The dependent variable is $Downgrade_{i,t+1}$, an indicator equal to 1 if S&P downgrades the firm's credit rating within the following 12 months from the earnings call and 0 otherwise. $\widehat{Downgrade}_{RAW}$, $\widehat{Downgrade}_{TRF}$, and $\widehat{Downgrade}_{LLM}$ denote the predicted downgrade probabilities from raw, TRF-anonymized, and LLM-anonymized transcripts. Pre is an indicator variable equal to 1 if the transcript falls in the pre-cutoff window (Nov 2022–Oct 2023) and 0 in the post-cutoff window (Nov 2023–Dec 2024). Columns (1)–(3) use the full balanced panel of firms appearing in both periods. All variables are winsorized at the 1st and 99th percentiles, and all independent variables are standardized (except dummy variable Pre). Control variables are identical to those in Panel B of Table 4. All regressions include SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Downgrade_{i,t+1}$		
VARIABLES	(1)	(2)	(3)
$\widehat{Downgrade}_{RAW}$	0.061***		
	(5.351)		
$\widehat{Downgrade}_{RAW} \times Pre$	-0.005		
	(-0.306)		
$\widehat{Downgrade}_{TRF}$		0.047***	
		(4.551)	
$\widehat{Downgrade}_{TRF} \times Pre$		-0.009	
		(-0.587)	
$\widehat{Downgrade}_{LLM}$			0.050***
			(4.821)
$\widehat{Downgrade}_{LLM} \times Pre$			-0.010
			(-0.631)
Controls	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes
Observations	3,943	3,943	3,943
Adj. R^2	0.218	0.202	0.205

Table 7: **Anonymization Gap and Predictive Power of Downgrade Estimates**

This table examines whether the attenuation-adjusted gap between raw and anonymized downgrade scores identifies where anonymization weakens predictive content. The dependent variable is $Downgrade_{i,t+1}$, an indicator equal to one if S&P downgrades the firm's credit rating within twelve months after the earnings call. For each anonymization method, the raw score R and anonymized score A are winsorized at the 1st and 99th percentiles and standardized over the pooled balanced sample. We estimate $\rho = \text{Corr}(R, A)$ on that sample and define $Gap = |A/\rho - R|$. The gap is then winsorized at the 1st and 99th percentiles and standardized. Each specification includes the standardized anonymized score, the standardized gap, and their product; the interaction itself is not re-standardized. The estimated values are $\rho_{TRF} = 0.8993$ and $\rho_{LLM} = 0.8997$, and the same pooled-sample estimate of ρ is used in all six columns. Columns (1)–(2) use the full balanced sample, Columns (3)–(4) the pre-cutoff window (Nov. 2022–Oct. 2023), and Columns (5)–(6) the post-cutoff window (Nov. 2023–Dec. 2024). Controls are identical to Panel B of Table 4. All regressions include SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Downgrade_{i,t+1}$					
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
$\widehat{Downgrade}_{TRF}$	0.051***		0.049***		0.054***	
	(4.755)		(3.280)		(4.338)	
$\widehat{Downgrade}_{TRF} \times Gap_{TRF}$	-0.019***		-0.015*		-0.023***	
	(-3.615)		(-1.881)		(-3.015)	
Gap_{TRF}	0.006		-0.002		0.010	
	(0.830)		(-0.188)		(1.139)	
$\widehat{Downgrade}_{LLM}$		0.049***		0.043***		0.054***
		(4.584)		(2.821)		(4.307)
$\widehat{Downgrade}_{LLM} \times Gap_{LLM}$		-0.012**		-0.013		-0.013**
		(-2.371)		(-1.461)		(-2.194)
Gap_{LLM}		0.006		0.013		-0.000
		(0.877)		(1.065)		(-0.032)
Sample	Full	Full	Pre	Pre	Post	Post
Controls	Yes	Yes	Yes	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,943	3,943	1,760	1,760	2,133	2,133
Adj. R^2	0.211	0.208	0.181	0.179	0.250	0.245

Table 8: **Information Loss of Alternative LLMs for Downgrade Predictions**

This table compares downgrade probabilities generated by three distinct models: GPT-4o-mini-2024-07-18, GPT-4o-2024-08-06, and o3-mini-2025-01-31. All regressions include the full set of control variables and use the realized S&P credit-rating downgrade indicator $Downgrade_{i,t+1}$ as the dependent variable. Columns (1) through (3) report the results for $\widehat{Downgrade}_{RAW}$ extracted from the raw transcripts by each LLM. Columns (4) through (6) report the results for $\widehat{Downgrade}_{TRF}$ extracted from the TRF-anonymized text. Columns (7) through (9) conduct horse-race regressions for each LLM-extracted signal. The qualitative pattern from Panel B of Table 4 (Column 4) persists across the extraction LLMs: the raw score remains significant, while the anonymized coefficient is substantially attenuated and is at most marginally significant once the raw score is included. Control variables are identical to those in Panel B of Table 4. All variables are winsorized at the 1st and 99th percentiles, and all independent variables are standardized. All regressions include SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Downgrade_{i,t+1}$								
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\widehat{Downgrade}_{RAW}$	0.061***	0.058***	0.070***				0.073***	0.039**	0.071***
	(5.515)	(4.782)	(5.905)				(3.576)	(2.488)	(5.569)
$\widehat{Downgrade}_{TRF}$				0.047***	0.049***	0.046***	-0.012	0.025*	-0.000
				(4.799)	(4.321)	(4.225)	(-0.719)	(1.731)	(-0.040)
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
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,243	2,243	2,243	2,243	2,243	2,243	2,243	2,243	2,243
Adj. R^2	0.246	0.237	0.260	0.232	0.233	0.229	0.246	0.241	0.260

Table 9: **S&P Ten-Factor Composite Scores: Standalone and Horse-Race Regressions**

This table reports regressions of the realized twelve-month downgrade indicator on researcher-constructed composite scores from the S&P-aligned ten-factor audit. For each text version, the composite is the average score across factors for which the model returns both a score and a supporting quote. Evidence coverage is reported separately in Figure 4. The construction is an audit measure and does not reproduce S&P's rating aggregation. Composite scores are winsorized at the 1st and 99th percentiles and standardized. Columns (1)–(3) report standalone regressions for the raw, TRF-anonymized, and LLM-anonymized composites. Columns (4)–(5) include the raw composite with one anonymized composite. All specifications include the controls in Panel B of Table 4 and SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Downgrade_{i,t+1}$				
VARIABLES	(1) Raw	(2) TRF	(3) LLM	(4) Raw + TRF	(5) Raw + LLM
$\widetilde{C}^{RAW}$	0.045***			0.045***	0.046***
	(4.405)			(3.648)	(3.068)
$\widetilde{C}^{TRF}$		0.035***		-0.000	
		(4.217)		(-0.040)	
$\widetilde{C}^{LLM}$			0.035***		-0.001
			(4.099)		(-0.093)
Controls	Yes	Yes	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes
Observations	2,243	2,243	2,241	2,243	2,241
Adj. R^2	0.226	0.218	0.218	0.225	0.225

Table 10: **Predictive Content of Shared and Version-Specific Credit-Risk Reasons**

This table tests the two channels defined in Section 5 and Appendix C. Shared reasons are factor assessments for which the raw and anonymized versions cite matched normalized quote anchors. *SharedRaw* and *SharedAnon* are the mean raw and anonymized scores across those reasons; their comparison captures interpretive distortion. RAW-only reasons comprise raw evidence that disappears after anonymization and the raw side of a factor for which the cited passage changes. Anonymized-only reasons are defined symmetrically. *RawOnly* and *AnonOnly* are the corresponding means, and their comparison captures evidence displacement. A mean is missing when its reason set is empty, so observations vary across columns. Panel A compares raw and TRF-anonymized transcripts, and Panel B compares raw and LLM-anonymized transcripts. Columns (1)–(2) report separate regressions for *SharedRaw* and *SharedAnon*, and Column (3) includes both. Columns (4)–(5) report separate regressions for *RawOnly* and *AnonOnly*, and Column (6) includes both. Column (7) includes all four measures. The variables are winsorized at the 1st and 99th percentiles and retain their original $[0, 1]$ units. All specifications include the controls in Panel B of Table 4 and SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	(1) Shared RAW	(2) Shared Anon.	(3) Shared Joint	(4) RAW-only	(5) Anon.-only	(6) Only Joint	(7) All Four
Panel A: RAW versus TRF-anonymized transcripts							
SharedRaw	0.293***		0.287***				0.216***
	(4.240)		(3.744)				(2.677)
SharedAnon		0.234***	0.007				-0.005
		(3.734)	(0.118)				(-0.085)
RawOnly				0.331***		0.287***	0.151
				(3.984)		(2.802)	(1.372)
AnonOnly					0.270***	0.056	0.004
					(4.163)	(0.938)	(0.061)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,172	2,172	2,172	2,227	2,231	2,227	2,155
Adj. R^2	0.222	0.215	0.222	0.226	0.219	0.225	0.230
Panel B: RAW versus LLM-anonymized transcripts							
SharedRaw	0.364***		0.385***				0.352***
	(5.139)		(4.288)				(4.172)
SharedAnon		0.272***	-0.026				-0.062
		(4.263)	(-0.347)				(-0.859)
RawOnly				0.314***		0.247**	0.035
				(3.601)		(2.076)	(0.274)
AnonOnly					0.258***	0.074	0.079
					(3.805)	(0.895)	(0.988)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,015	2,015	2,015	2,233	2,228	2,227	2,001
Adj. R^2	0.225	0.212	0.224	0.207	0.201	0.205	0.209

Table 11: **Predictive Power of Earnings Call Sentiment for DGTW Returns**

This table presents regressions of DGTW-adjusted stock returns ($DGTW_{i,t+1}$) on sentiment measures extracted using GPT-4o-mini from earnings conference call transcripts. Following the matching convention described in the text, the dependent variable captures the market's reaction to the call: for calls held before the U.S. market close (16:00 ET) it is measured on the same trading day, and for calls held after the close on the next trading day, so that it fully incorporates the information disclosed in the transcript; we denote it $DGTW_{i,t+1}$. Panel A reports specifications without control variables, and Panel B adds the full set of controls. The control variables are collapsed into a single "Controls" row for brevity; they comprise analyst forecast error (FE), DGTW-adjusted returns over the run-up window ($DGTW_{i,t}$, $DGTW_{i,t-1}$, $DGTW_{i,t-2}$, $DGTW_{i,t-21,t-3}$, $DGTW_{i,t-252,t-22}$), and firm characteristics including the logarithm of firm size ($\ln(Size)$), book-to-market ratio ($\ln(BM)$), and turnover ($\ln(Turnover)$). In each panel, Columns (1) through (3) individually add the sentiment scores derived from the raw, TRF-anonymized, and LLM-anonymized transcripts. Columns (4) and (5) include both the sentiment extracted from the raw transcript and from an anonymized transcript. 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: $DGTW_{i,t+1}$				
VARIABLES	(1)	(2)	(3)	(4)	(5)
Panel A: Without Controls					
$Sentiment_{RAW}$	2.549***			1.842***	1.833***
	(19.602)			(8.741)	(7.237)
$Sentiment_{TRF}$		2.413***		0.820***	
		(17.329)		(3.766)	
$Sentiment_{LLM}$			2.410***		0.838***
			(17.521)		(3.212)
Controls	No	No	No	No	No
Date FE	Yes	Yes	Yes	Yes	Yes
Observations	7352	7352	7352	7352	7352
Adj. R^2	0.078	0.070	0.070	0.080	0.080
Panel B: With Controls					
$Sentiment_{RAW}$	2.487***			1.822***	1.808***
	(18.832)			(8.475)	(7.234)
$Sentiment_{TRF}$		2.331***		0.775***	
		(17.287)		(3.662)	
$Sentiment_{LLM}$			2.329***		0.798***
			(17.466)		(3.247)
Controls	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Observations	7352	7352	7352	7352	7352
Adj. R^2	0.132	0.124	0.124	0.133	0.133

Table 12: **Information Loss in Alternative Tasks**

This table compares the predictive content of textual scores extracted from RAW and TRF-anonymized earnings-call transcripts across four forward-looking outcomes. Each column reports a horse-race regression that includes both the RAW and TRF-anonymized scores. Column (1) examines realized post-announcement return volatility (Vol_{post}) using managerial uncertainty, 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 capital expenditures two quarters ahead ($Capx_{t+2}$) using investment expectations, controlling for current capital expenditures ($Capx_t$), market leverage ($Leverage_{Market}$), and firm size ($\ln(Total\ Asset)$). Columns (3) and (4) predict two-quarter-ahead sales growth ($SaleChange_{t+2}$) and value-added growth ($ValueAddChange_{t+2}$), respectively, using firms' expressed macroeconomic outlook and controlling for current sales growth ($SaleChange_t$) or value-added growth ($ValueAddChange_t$), as well as book-to-market ratio (BM), firm size ($\ln(Total\ Asset)$), and asset tangibility ($Tangibility$). The post-announcement window underlying Vol_{post} is matched on the calendar (trading-day) window following each call, whereas $Capx_{t+2}$, $SaleChange_{t+2}$, and $ValueAddChange_{t+2}$ are matched on the firm's fiscal quarters following the call, ensuring that each outcome is realized strictly after, and is therefore genuinely predicted by, the earnings call. Detailed variable definitions are reported in Appendix Table B.2. All variables are winsorized at the 1st and 99th percentiles, and all independent variables are standardized. All regressions include date fixed effects. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Vol_{post}	$Capx_{t+2}$	$SaleChange_{t+2}$	$ValueAddChange_{t+2}$
VARIABLES	(1)	(2)	(3)	(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)		
$Economy_{RAW}$			1.803***	2.460***
			(5.473)	(5.506)
$Economy_{TRF}$			1.550***	1.787***
			(4.376)	(3.491)
Controls	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	7,255	5,692	6,650	6,398
Adj. R^2	0.776	0.730	0.237	0.278

Table 13: **Predictive Power of News Headlines Sentiment for DGTW Returns**

This table examines whether information loss also affects news headlines. The dependent variable is the DGTW-adjusted stock return ($DGTW_{i,t+1}$). Following the same matching convention used for earnings calls, this return is measured on the same trading day for news released before the U.S. market close (16:00 ET) and on the next trading day for news released after the close, so that it fully incorporates the information in the headline. The sentiment scores are derived from RAW and TRF-anonymized headlines. Columns (1) and (2) show single-score 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. The control variables are collapsed into a single "Controls" row for brevity; they comprise past returns over the run-up window ($DGTW_{i,t}$, $DGTW_{i,t-1}$, $DGTW_{i,t-2}$, $DGTW_{i,t-21,t-3}$, $DGTW_{i,t-252,t-22}$) and firm characteristics ($\ln(Size)$, $\ln(BM)$, $\ln(Turnover)$). 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $DGTW_{i,t+1}$					
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
$Sentiment^{News}_{RAW}$	0.444***		0.315***	0.439***		0.314***
	(17.477)		(4.664)	(17.353)		(4.633)
$Sentiment^{News}_{TRF}$		0.437***	0.135**		0.432***	0.131**
		(17.325)	(2.036)		(17.212)	(1.974)
Controls	No	No	No	Yes	Yes	Yes
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

A Supporting Derivations for the Econometric Framework

This appendix provides the algebra derivations behind Section 2.

A.1 Attenuation and Horse-Race Regressions

Equation (2) gives $u = A - \rho R$. Because R and A have unit variance and correlation ρ,

$$\mathrm{Cov}(R, u) = \mathrm{Cov}(R, A) - \rho \,\mathrm{Var}(R) = \rho - \rho = 0, \qquad \mathrm{Var}(u) = 1 - \rho^2.$$

Thus u is the part of the anonymized score that does not move with the raw score.

Now use $Y = \mu + \beta R + \varepsilon$. In a regression with a single standardized score, the slope is its covariance with Y. Therefore,

$$b_R = \mathrm{Cov}(Y, R) = \beta, \qquad b_A = \mathrm{Cov}(Y, A) = \rho\beta.$$

The anonymized slope is the raw slope multiplied by ρ. Squaring this attenuation gives the corresponding fit:

$$R^2_A = \rho^2 R^2_R.$$

In a horse race containing both R and A, write $A = \rho R + u$. Once R is included, only u is left to identify the coefficient on A. Under equation (3), u is unrelated to the future outcome, so the population coefficient on A is zero and the coefficient on R is β.

A.2 Knowledge-Cutoff Interaction

Equation (5) is fully interacted with the binary indicator P. Its population coefficients therefore reproduce the two within-period regressions: d_1 is the slope of Y on $\widetilde{S}$ when $P = 0$,

and $d_1 + d_3$ is the corresponding slope when $P = 1$.

Let b_S be the predictive slope of the leakage-free score. After the cutoff, $P = 0$ and $\widetilde{S} = S$, so $d_1 = b_S$. Before the cutoff, $\widetilde{S} = S + \gamma_S \varepsilon$. We assume that b_S and σ^2 are stable across the two periods and that the leakage-free score is uncorrelated with ε. Basic covariance and variance formulas then give

$$\mathrm{Cov}(Y, \widetilde{S} \mid P = 1) = b_S + \gamma_S \sigma^2, \qquad \mathrm{Var}(\widetilde{S} \mid P = 1) = 1 + \gamma_S^2 \sigma^2.$$

Hence the pre-cutoff slope is

$$d_1 + d_3 = \frac{b_S + \gamma_S \sigma^2}{1 + \gamma_S^2 \sigma^2},$$

so

$$d_3 = \frac{\gamma_S \sigma^2 (1 - b_S \gamma_S)}{1 + \gamma_S^2 \sigma^2}.$$

If there is no leakage, $\gamma_S = 0$ and therefore $d_3 = 0$. For a positively oriented score, moderate leakage means $0 < \gamma_S < 1/b_S$, which gives $d_3 > 0$. Intuitively, the pre-cutoff score becomes unusually predictive because it contains part of the future innovation itself.

A.3 The Anonymization Gap

Equation (6) defines

$$v = \frac{A}{\rho} - R = \frac{u}{\rho}, \qquad Gap = |v|,$$

so the anonymized score can be written as

$$A = \rho(R + v).$$

Suppose that v is symmetric and independent of R. Conditional on $Gap = g$, we have $v^2 = g^2$, while symmetry implies $\mathbb{E}[v \mid Gap = g] = 0$. Independence from the standardized

raw score further implies

$$\mathbb{E}[R \mid Gap = g] = 0, \qquad \mathbb{E}[R^2 \mid Gap = g] = 1, \qquad \mathbb{E}[Rv \mid Gap = g] = 0.$$

These relations make the conditional variance of A particularly simple. Because its conditional mean is zero,

$$\begin{aligned}\mathrm{Var}(A \mid Gap = g) &= \rho^2 \mathbb{E}[(R+v)^2 \mid Gap = g] \\ &= \rho^2 \left(\mathbb{E}[R^2 \mid Gap = g] + 2\mathbb{E}[Rv \mid Gap = g] + \mathbb{E}[v^2 \mid Gap = g]\right) \\ &= \rho^2(1+g^2).\end{aligned}$$

Next use the outcome equation $Y = \mu + \beta R + \varepsilon$. The condition $\mathbb{E}[\varepsilon \mid R, A] = 0$ removes the covariance between the future innovation and the anonymized score. Hence

$$\begin{aligned}\mathrm{Cov}(Y, A \mid Gap = g) &= \beta \,\mathrm{Cov}(R, \rho(R+v) \mid Gap = g) \\ &= \rho\beta.\end{aligned}$$

The conditional regression slope of Y on A is therefore

$$b(g) = \frac{\mathrm{Cov}(Y, A \mid Gap = g)}{\mathrm{Var}(A \mid Gap = g)} = \frac{\beta}{\rho(1+g^2)},$$

which is equation (7). For a positively oriented signal, $\beta > 0$, and

$$b'(g) = -\frac{2\beta g}{\rho(1+g^2)^2} < 0 \qquad \text{for } g > 0.$$

The covariance with the outcome is unchanged across gap levels, but the variance of the anonymized score increases with g. The signal therefore becomes less predictive as document-specific distortion grows.

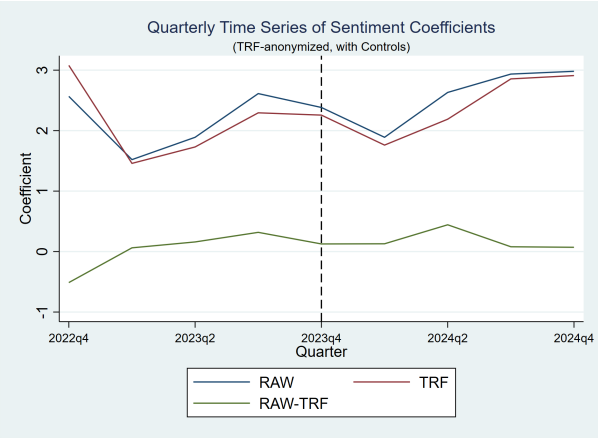
Finally, equation (8) approximates the conditional slope $b(g)$ by the linear function $c_1 +$

$c_3 g$. The interaction coefficient is the slope from a weighted linear fit of $b(Gap)$ on Gap, with positive weight $w(g) = \mathrm{Var}(A \mid Gap = g)$. Equivalently,

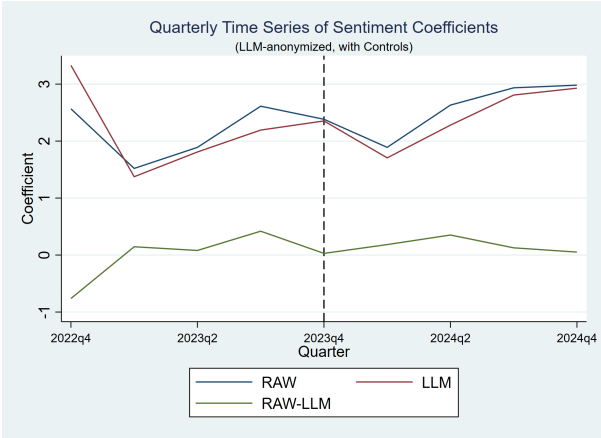
$$c_3 = \frac{\mathrm{Cov}_w(Gap, b(Gap))}{\mathrm{Var}_w(Gap)}.$$

Because $b(g)$ is decreasing, the weighted covariance in the numerator is negative, while the denominator is positive whenever the gap varies across documents. Thus $c_3 < 0$, which is the empirical prediction tested in the paper.

B Appendix



(a) Quarterly Sentiment Coefficients (RAW and TRF-anonymized)



(b) Quarterly Sentiment Coefficients (RAW and LLM-anonymized)

Figure B.1: **Quarterly Time Series of Sentiment Coefficients**

This figure plots the quarterly time series of coefficients on sentiment scores, estimated separately from regressions with controls following the specification in Panel B of Table 11. Sentiment scores are extracted using GPT-4o-mini from different versions of earnings call transcripts. Subfigure B.1a compares the coefficients derived from the RAW transcripts against those from the TRF-anonymized samples. Subfigure B.1b compares the coefficients from the RAW transcripts against those from the LLM-anonymized samples.

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

This table outlines the filtering procedures used to construct the final samples. Panel A details the selection process for the downgrade-prediction post-cutoff subsample, Panel B for earnings call transcripts, and Panel C for news headlines. Each row sequentially applies a filter, reporting the resulting sample size and the number of observations removed. For the downgrade subsample in Panel A, we additionally exclude utilities firms (SIC 4000–4999) and financial firms (SIC 6000–6299), following standard practice in the credit risk literature (Donovan et al., 2021), require non-missing controls, and require an S&P credit rating observed before the earnings call.

Panel A: Downgrade Prediction Subsample (Post-Cutoff)		
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
Drop utilities and financial firms (SIC 4000–4999, 6000–6299)	8,566	1,963
Require non-missing controls and an S&P credit rating observed before the earnings call	2,488	6,078
Panel B: 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 C: 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 B.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; the extraction prompts are reported in Appendix D. 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	
$\widehat{Downgrade}_{LLM}$	The predicted probability of an S&P credit rating downgrade within the 12 months following the earnings call, derived from LLM-anonymized earnings call transcripts.
$\widehat{Downgrade}_{RAW}$	The predicted probability of an S&P credit rating downgrade within the 12 months following the earnings call, derived from raw earnings call transcripts.
$\widehat{Downgrade}_{TRF}$	The predicted probability of an S&P credit rating downgrade within the 12 months following the earnings call, derived from TRF-anonymized earnings call transcripts.
$\widetilde{C}_j^v$	Winsorized and standardized version of the researcher-constructed S&P-aligned ten-factor composite C_j^v in equation (11), for transcript j and text version $v \in \{RAW, TRF, LLM\}$. The composite is the mean score across factors for which the model returns both a score and a supporting quote. It is not an official S&P assessment or rating.
Gap_v	Anonymization gap for method $v \in \{TRF, LLM\}$, defined as $Gap_v = \lvert A_v/\rho_v - R\rvert$, where $\rho_v = \mathrm{Corr}(R, A_v)$ is estimated in the pooled balanced sample. The gap is winsorized and standardized before entering the regressions.
SharedRaw	Mean raw factor score over shared reasons: factors for which the raw transcript and the anonymized transcript both return evidence and cite matched normalized quote anchors.
SharedAnon	Mean anonymized factor score over the same shared reasons. Comparing this measure with SharedRaw captures interpretive distortion.
RawOnly	Mean raw factor score over RAW-only reasons, comprising raw evidence that is absent after anonymization and the raw side of factors whose quote anchor changes.
AnonOnly	Mean anonymized factor score over anonymized-only reasons, defined symmetrically. Comparing this measure with RawOnly captures evidence displacement.

Continued on next page

Table B.2 – Continued from previous page

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.
$Investment_{RAW}$	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.
$Sentiment_{AGT}$	Sentiment of the transcript where only AGENTS entity types are preserved while all other entity categories are anonymized by the 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 preserved while all other entity categories are anonymized by the TRF model.
$Sentiment_{OTH}$	Sentiment of the transcript where only OTHERS entity types are preserved while all other entity categories are anonymized by the TRF model.
$Sentiment_{PLC}$	Sentiment of the transcript where only PLACES entity types are preserved while all other entity categories are anonymized by the TRF model.
$Sentiment_{RAW}$	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.
$Sentiment^{News}_{RAW}$	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.

Continued on next page

Table B.2 – Continued from previous page

Variable	Definition
$Sentiment_{TRF}$	Sentiment of the earnings call transcript text after anonymization using the spaCy `en_core_web_trf` (TRF) model.
$Sentiment_{TRF}^{News}$	Sentiment of the TRF-anonymized Yahoo news headlines.
$Uncertainty_{RAW}$	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.
Panel B: Dependent Variables	
$Capx_{t+2}$	Capital expenditure (CAPX) at the end of quarter $t+2$, scaled by total assets.
$DGTW_{i,t+1}$	DGTW-adjusted stock return that incorporates the information disclosed in the earnings call or news, defined following Daniel et al. (1997). Under the matching convention in the text, it is measured on the same trading day for events occurring before the U.S. market close (16:00 ET) and on the next trading day for events after the close.
$Downgrade_{i,t+1}$	An indicator variable equaling 1 if S&P downgrades the firm's credit rating within the following 12 months from the earnings call release, and 0 otherwise.
$SaleChange_{t+2}$	Defined as $\log(Sale_{t+2}/Sale_t)$; represents two-quarter-ahead sales growth.
$ValueAddChange_{t+2}$	Defined as $\log(ValueAdd_{t+2}/ValueAdd_t)$; represents two-quarter-ahead value-added growth.
Vol_{post}	Standard deviation of stock returns over the post-announcement window [0, 22], requiring at least 10 daily observations.
Panel C: Control Variables	
|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.
$Alpha_{pre}$	Alpha from a market model estimated over the window [-252, -6], requiring at least 60 daily observations.
BM	Book-to-market ratio at the end of June, calculated as book equity (fiscal year-end) divided by market equity (fiscal year-end).
$Capx_t$	Capital expenditure (CAPX) at the end of quarter t, scaled by total assets.
$DGTW_{i,t}$	DGTW-adjusted stock return on event day 0.
$DGTW_{i,t-1}$	DGTW-adjusted stock return on event day -1.
$DGTW_{i,t-2}$	DGTW-adjusted stock return on event day -2.
$DGTW_{i,t-21,t-3}$	DGTW-adjusted stock return over the event window [-21, -3].
$DGTW_{i,t-252,t-22}$	DGTW-adjusted stock return over the event window [-252, -22].

Continued on next page

Table B.2 – Continued from previous page

Variable	Definition
FE	Analyst forecast error, defined as Hirshleifer et al. (2009).
$FirmValue$	Market value of equity plus the book value of debt.
$Leverage_{Asset}$	Total debt scaled by total assets ((dlcq + dlttq)/atq).
$Leverage_{Market}$	Financial leverage, calculated as total debt (dlttq + dlcq) divided by total debt plus market value of equity.
MTB	Market value of equity scaled by book value of equity ((prccq * cshoq) / ceqq).
$Oscore$	O-Score measure of credit risk following Ohlson (1980).
$RevGrowth$	percentage changes in revenue (revtq) from the same quarter, prior year.
ROA	Income before extraordinary items scaled by total assets (ibq / atq).
$Size$	Market capitalization at the end of June, calculated as CRSP price times shares outstanding. This equity-size measure is used in the non-downgrade regressions.
$Std(Income)$	Standard deviation of quarterly income before extraordinary items measured over the previous twenty quarters.
$Std(Revenue)$	Standard deviation of total revenues measured over the previous twenty quarters.
$Tangibility$	Plant, Property, and Equipment (PPENTQ) scaled by total assets (atq) at the end of the quarter.
$Total\ Asset$	Logarithm of quarterly total assets from Compustat (atq).
$Turnover$	Share turnover at the end of June, calculated as prior year's trading volume divided by shares outstanding.
Vol_{pre}	Standard deviation of stock returns over the pre-announcement window [-252, -1], requiring at least 60 daily observations.
$Zscore$	Z-Score measure of bankruptcy risk following Altman (1968).

Table B.3: **Entity Categories**

This table presents the 18 mutually exclusive entity types identified by the `en_core_web_trf` NER model from Python's spaCy library. The entity types are grouped into four primary categories: NUMBERS, PLACES, AGENTS, and OTHERS. For each entity type, the table reports its description, an illustrative example, and its total frequency in the S&P-rated post-cutoff downgrade sample of 2,488 earnings call transcripts. Frequency per Transcript reports the mean number of occurrences in each category per transcript. The last two columns report the total frequency of each entity type in the 52,716 news headlines and the mean number of occurrences in each category 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	463,369	325.11	21862	0.678
	CARDINAL	Numerals that do not fall under another type	one	96,713		4670	
	MONEY	Monetary values, including unit	$69.7 billion	91,265		5419	
	PERCENT	Percentage, including "%"	6%	89,442		2322	
	ORDINAL	"first," "second," etc.	first	32,485		1214	
	TIME	Times smaller than a day	evening	29,063		120	
	QUANTITY	Measurements, as of weight or distance	100-foot	6,532		142	
PLACES	GPE	Countries, cities, states	Korea	47,021	30.74	6528	0.146
	LOC	Non-GPE locations, mountain ranges, bodies of water	the Middle East	24,996		773	
	FAC	Buildings, airports, highways, bridges, etc.	San Ciprián	4,468		384	
AGENTS	PERSON	People, including fictional	Tim Cook	277,546	197.36	5741	1.125
	ORG	Firms, agencies, institutions, etc.	Apple Inc.	207,753		52714	
	NORP	Nationalities or religious or political groups	Chinese	5,731		863	
OTHERS	PRODUCT	Objects, vehicles, foods, etc. (not services)	Watch Ultra	36,190	17.69	3100	0.084
	EVENT	Named hurricanes, battles, wars, sports events, etc.	Super Bowl	4,276		834	
	LAW	Named documents made into laws	the IRA Inflation Reduction Act	2,428		76	
	WORK_OF_ART	Titles of books, songs, etc.	America's Greatest Workplaces 2024	1,069		395	
	LANGUAGE	Any named language	English	40		7	

Table B.4: **Three-Month Forecast Horizon and the Predictive Power of Downgrade Estimates**

This table reproduces the knowledge-cutoff interaction analysis in Table 6 using a three-month rather than a twelve-month downgrade horizon. The dependent variable, $Downgrade^{3M}_{i,t+1}$, is an indicator equal to one if S&P downgrades the firm's credit rating within three months after the earnings call and zero otherwise. $\widehat{Downgrade}^{3M}_{RAW}$, $\widehat{Downgrade}^{3M}_{TRF}$, and $\widehat{Downgrade}^{3M}_{LLM}$ denote the predicted probabilities of a downgrade within the following three months, extracted from raw, TRF-anonymized, and LLM-anonymized transcripts, respectively. The scores are constructed using the same prompts as in the baseline analysis, except that the forecast horizon is changed from twelve months to three months. Pre is an indicator equal to one if the transcript falls in the pre-cutoff window (Nov. 2022–Oct. 2023) and zero if it falls in the post-cutoff window (Nov. 2023–Dec. 2024). The sample is the same balanced panel of firms used in Table 6. All variables are winsorized at the 1st and 99th percentiles, and all independent variables are standardized except for Pre. Control variables are identical to those in Panel B of Table 4. All regressions include SIC3 × Quarter fixed effects. T-statistics based on SIC3-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $Downgrade^{3M}_{i,t+1}$		
VARIABLES	(1)	(2)	(3)
$\widehat{Downgrade}^{3M}_{RAW}$	0.026***		
	(4.521)		
$\widehat{Downgrade}^{3M}_{RAW} \times Pre$	-0.006		
	(-0.684)		
$\widehat{Downgrade}^{3M}_{TRF}$		0.023***	
		(3.857)	
$\widehat{Downgrade}^{3M}_{TRF} \times Pre$		-0.010	
		(-1.299)	
$\widehat{Downgrade}^{3M}_{LLM}$			0.020***
			(3.712)
$\widehat{Downgrade}^{3M}_{LLM} \times Pre$			-0.002
			(-0.300)
Controls	Yes	Yes	Yes
SIC3 × Quarter FE	Yes	Yes	Yes
Observations	3,943	3,943	3,943
Adj. R^2	0.103	0.097	0.097

Table B.5: **Correlations of Sentiment Signals**

This table shows correlations among sentiment scores extracted using GPT-4o-mini. Panel A reports correlations among sentiment scores from raw transcripts (RAW) and the two fully anonymized transcripts (TRF and LLM). Panel B reports correlations among sentiment scores from raw transcripts and transcripts where only one category of entities is preserved while the other three categories are anonymized by the TRF model: $Sentiment_{NUM}$, $Sentiment_{PLC}$, $Sentiment_{AGT}$, and $Sentiment_{OTH}$ preserve only the NUMBERS, PLACES, AGENTS, or OTHERS category, respectively. All values are Pearson correlation coefficients.

Panel A: Raw and Fully Anonymized Sentiment

	$Sentiment_{RAW}$	$Sentiment_{TRF}$	$Sentiment_{LLM}$
$Sentiment_{RAW}$	1.000		
$Sentiment_{TRF}$	0.864	1.000	
$Sentiment_{LLM}$	0.858	0.859	1.000

Panel B: Raw and Single-Category-Preserved Sentiment

	$Sentiment_{RAW}$	$Sentiment_{NUM}$	$Sentiment_{PLC}$	$Sentiment_{AGT}$	$Sentiment_{OTH}$
$Sentiment_{RAW}$	1.000				
$Sentiment_{NUM}$	0.910	1.000			
$Sentiment_{PLC}$	0.872	0.878	1.000		
$Sentiment_{AGT}$	0.880	0.871	0.907	1.000	
$Sentiment_{OTH}$	0.872	0.876	0.943	0.905	1.000

Table B.6: **Comparing Explanatory Power Under Preserving Different Categories of Entities**

This table reports regressions of DGTW-adjusted stock returns ($DGTW_{i,t+1}$) on sentiment scores extracted from transcripts where only one of the four entity categories is preserved at a time, while the other three categories are anonymized using the TRF model. Following the matching convention in the text, the dependent variable is measured on the same trading day for calls before the U.S. market close (16:00 ET) and on the next trading day for calls after the close. The category definitions are provided in Appendix Table B.3. Specifically, $Sentiment_{NUM}$, $Sentiment_{PLC}$, $Sentiment_{AGT}$, and $Sentiment_{OTH}$ are extracted from transcripts where the TRF model preserves only the NUMBERS, PLACES, AGENTS, or OTHERS category, respectively. Columns (1)–(5) report standalone regressions for the RAW score and the four single-category-preserved variants; Columns (6)–(9) include the RAW score together with each single-category-preserved variant. Control variables are identical to those in Panel B of Table 11. All variables are winsorized at the 1st and 99th percentiles, and 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: $DGTW_{i,t+1}$								
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$Sentiment_{RAW}$	2.487***					1.627***	1.898***	1.669***	1.816***
	(18.832)					(6.313)	(9.067)	(7.554)	(8.301)
$Sentiment_{NUM}$		2.419***				0.950***			
		(17.743)				(3.628)			
$Sentiment_{PLC}$			2.315***				0.680***		
			(16.708)				(3.122)		
$Sentiment_{AGT}$				2.393***				0.936***	
				(17.295)				(4.077)	
$Sentiment_{OTH}$					2.341***				0.775***
					(17.423)				(3.540)
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	0.128	0.123	0.127	0.124	0.133	0.133	0.134	0.133

Table B.7: **Sample Period and Explanatory Power of Sentiment**

This table examines whether sentiment scores extracted from raw and anonymized earnings call transcripts differ in explanatory power between the pre- and post-knowledge-cutoff subsamples. The dependent variable is the DGTW-adjusted stock return ($DGTW_{i,t+1}$). Following the matching convention in the text, it is measured on the same trading day for calls before the U.S. market close (16:00 ET) and on the next trading day for calls after the close. $Sentiment_{TRF}$ and $Sentiment_{LLM}$ denote sentiment scores computed from TRF-anonymized and LLM-anonymized transcripts, respectively. Pre is an indicator variable equal to 1 for observations in the pre-cutoff subsample and 0 for those in the post-cutoff subsample. The three columns use sentiment scores extracted from raw, TRF-anonymized, and LLM-anonymized transcripts, respectively. All variables are winsorized at the 1st and 99th percentiles and standardized, except for the indicator variable Pre. All specifications include date fixed effects and the control variables used in the preceding tables. T-statistics based on time-clustered standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $DGTW_{i,t+1}$		
VARIABLES	(1)	(2)	(3)
$Sentiment_{RAW}$	2.506***		
	(19.564)		
$Sentiment_{RAW} \times Pre$	-0.553***		
	(-3.130)		
$Sentiment_{TRF}$		2.360***	
		(17.715)	
$Sentiment_{TRF} \times Pre$		-0.488***	
		(-2.728)	
$Sentiment_{LLM}$			2.397***
			(17.913)
$Sentiment_{LLM} \times Pre$			-0.515***
			(-2.900)
Controls	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Observations	12,772	12,772	12,772
Adj. R^2	0.132	0.125	0.126

Table B.8: **Anonymization Gap and Information Loss in Sentiment**

This table examines whether the attenuation-adjusted gap identifies where anonymization weakens the relation between earnings-call sentiment and DGTW-adjusted stock returns. The dependent variable is $DGTW_{i,t+1}$ in percentage points and follows the announcement-timing convention in the text. For each anonymization method, raw sentiment R and anonymized sentiment A are winsorized at the 1st and 99th percentiles and standardized over the pooled balanced sample. We estimate $\rho = \text{Corr}(R, A)$ on that sample and define $Gap = |A/\rho - R|$. The gap is then winsorized at the 1st and 99th percentiles and standardized. Each specification includes standardized anonymized sentiment, the standardized gap, and their product; the interaction itself is not re-standardized. The estimated values are $\rho_{TRF} = 0.8529$ and $\rho_{LLM} = 0.8510$, and the same pooled-sample estimate of ρ is used in all columns. Columns (1)–(2) use the full balanced sample, Columns (3)–(4) the pre-cutoff window (Nov. 2022–Oct. 2023), and Columns (5)–(6) the post-cutoff window (Nov. 2023–Dec. 2024). Controls are identical to Panel B of Table 11. All regressions include date fixed effects. T-statistics based on date-clustered standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	Dependent Variable: $DGTW_{i,t+1}$					
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
$Sentiment_{TRF}$	2.466***		2.139***		2.725***	
	(21.623)		(14.182)		(17.417)	
$Sentiment_{TRF} \times Gap_{TRF}$	-0.543***		-0.431***		-0.635***	
	(-9.734)		(-5.860)		(-7.752)	
Gap_{TRF}	-0.790***		-0.687***		-0.875***	
	(-6.725)		(-4.292)		(-5.095)	
$Sentiment_{LLM}$		2.507***		2.171***		2.800***
		(19.038)		(13.401)		(14.315)
$Sentiment_{LLM} \times Gap_{LLM}$		-0.541***		-0.401***		-0.666***
		(-7.863)		(-4.780)		(-5.982)
Gap_{LLM}		-0.865***		-0.671***		-1.003***
		(-7.801)		(-4.080)		(-6.781)
Sample	Full	Full	Pre	Pre	Post	Post
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,772	12,772	5,666	5,666	7,106	7,106
Adj. R^2	0.132	0.133	0.126	0.128	0.138	0.139

Table B.9: **Information Loss of Alternative LLMs**

This table compares sentiment scores generated by three distinct models: GPT-4o-mini, GPT-4o, and o3-mini. All regressions include the full set of control variables and use the DGTW-adjusted stock return ($DGTW_{i,t+1}$) as the dependent variable. Following the matching convention in the text, the dependent variable is measured on the same trading day for calls before the U.S. market close (16:00 ET) and on the next trading day for calls after the close. 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. Columns (7) through (9) conduct horse-race regressions for each LLM-extracted signal. Control variables are identical to those in Panel B of Table 11. 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: $DGTW_{i,t+1}$								
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

C Implementation Details for the S&P Ten-Factor Credit-Risk Audit

Section 5 explains the audit and reports its results. This appendix records the supplementary and implementation details needed to reproduce the results.

C.1 Factor Scope, Extraction Checks, and Samples

The ten factors are drawn from S&P Global Ratings' *Corporate Methodology*. Figure C.1 locates them in the broader criteria sequence, and Table C.1 gives the exact factor scopes supplied to the model. Group and government influence lie outside the audit because they are separate from the ten corporate-methodology dimensions and generally cannot be assessed from an earnings call.

C.2 Quote Matching and Evidence Decomposition

Using the notation introduced in Section 5, we compare the raw transcript separately with each anonymized version. Let $s^v_{jk} \in [0,1]$ denote the downgrade-risk score assigned to factor k for transcript j when the model reads text version $v \in \{RAW, TRF, LLM\}$. The score s^v_{jk} is included in the analysis only when the model returns both a non-null score and a non-null supporting quote for that factor.

When both the raw and anonymized versions contain valid evidence for a factor, we normalize their supporting quotes before matching. We remove named-entity spans identified by spaCy `en_core_web_trf` from the raw quote and remove explicit placeholders and tokens containing numbers from the anonymized quote. We then lowercase both quotes, remove punctuation and extra whitespace, and calculate a fuzzy token-sort ratio. Two quotes are treated as matched when this ratio is at least 75.

For method $a \in \{TRF, LLM\}$, define three sets. S^a_j contains factors with matched raw and anonymized quote anchors, R^a_j contains RAW-only reasons, and A^a_j contains anonymized-only reasons. Equivalently,

$$R^a_j = \mathcal{K}^{RAW}_j \setminus S^a_j, \qquad A^a_j = \mathcal{K}^a_j \setminus S^a_j.$$

If both versions populate a factor but cite different anchors, that factor belongs to both R^a_j and A^a_j because each version supplies a different reason.

Throughout, vertical bars denote set cardinality rather than absolute value. Thus, $|S^a_j|$ is the number of factors with valid score–quote pairs in both versions whose quote anchors match, while $|R^a_j|$ and $|A^a_j|$ count the valid RAW-only and anonymized-only reasons, respectively. For example, if both versions populate a factor but cite different quote anchors, that factor contributes one to $|R^a_j|$ and one to $|A^a_j|$, but contributes nothing to $|S^a_j|$.

The two means over shared reasons are

$$\begin{aligned} \text{SharedRaw}_{j,a} &= \frac{1}{|S_j^a|} \sum_{k \in S_j^a} s_{jk}^{RAW}, \\ \text{SharedAnon}_{j,a} &= \frac{1}{|S_j^a|} \sum_{k \in S_j^a} s_{jk}^{a}. \end{aligned} \tag{15}$$

The two means over version-specific reasons are

$$\begin{aligned} \text{RawOnly}_{j,a} &= \frac{1}{|R_j^a|} \sum_{k \in R_j^a} s_{jk}^{RAW}, \\ \text{AnonOnly}_{j,a} &= \frac{1}{|A_j^a|} \sum_{k \in A_j^a} s_{jk}^{a}. \end{aligned} \tag{16}$$

Each conditional mean is defined only when the corresponding set is nonempty. If a set is empty, its conditional mean is coded as missing rather than zero.
Because S_j^a and R_j^a partition the valid raw reasons, while S_j^a and A_j^a partition the valid anonymized reasons, the full raw and anonymized composites can be recovered as

$$\begin{aligned} C_j^{RAW} &= \frac{\sum_{k \in S_j^a} s_{jk}^{RAW} + \sum_{k \in R_j^a} s_{jk}^{RAW}}{|S_j^a| + |R_j^a|}, \\ C_j^{a} &= \frac{\sum_{k \in S_j^a} s_{jk}^{a} + \sum_{k \in A_j^a} s_{jk}^{a}}{|S_j^a| + |A_j^a|}. \end{aligned} \tag{17}$$

Equation (17) also provides a computational check on the decomposition. In each composite, the relevant factor scores are pooled before division by the total number of valid reasons; the separate conditional means are never added directly.

C.3 Shapley Decomposition

Equation (17) provides the two endpoint states used in the Shapley decomposition. Specifically, $C_{j,a}^{\text{Raw}} \equiv C_j^{RAW}$ is the original raw composite, while $C_{j,a}^{\text{Anon}} \equiv C_j^{a}$ is the fully anonymized composite. The two intermediate states are constructed directly from the underlying score sums and reason counts:

$$\begin{aligned} C_{j,a}^{\text{Replace}} &= \frac{\sum_{k \in S_j^a} s_{jk}^{RAW} + \sum_{k \in A_j^a} s_{jk}^{a}}{|S_j^a| + |A_j^a|}, \\ C_{j,a}^{\text{Reread}} &= \frac{\sum_{k \in S_j^a} s_{jk}^{a} + \sum_{k \in R_j^a} s_{jk}^{RAW}}{|S_j^a| + |R_j^a|}. \end{aligned} \tag{18}$$

The *Replace* state introduces evidence displacement alone by replacing the RAW-only reasons with the anonymized-only reasons while retaining the raw scores on shared reasons.

The *Reread* state introduces interpretive distortion alone by retaining the raw evidence set while replacing the raw scores on shared reasons with their anonymized counterparts. Each numerator pools the relevant factor scores before division; the conditional group means are never added directly.

Let

$$\mathcal{Q} = \{\text{Raw}, \text{Replace}, \text{Reread}, \text{Anon}\}$$

denote the four states. For each anonymization method a and each state $q \in \mathcal{Q}$, we estimate equation (12) separately by setting the composite score equal to $C^q_{j,a}$. Equivalently, each state is entered into the following single-score regression:

$$D_j = \alpha + \beta_q C^q_{j,a} + \gamma X_j + \delta_t + \varepsilon_j, \qquad q \in \mathcal{Q}. \tag{19}$$

Thus, the four regressions differ only in which state composite is used as the primary explanatory variable. The outcome, controls, SIC3 $\times$ Quarter fixed effects, and estimation sample are held fixed across the four regressions.

Before estimation, the four state composites are winsorized and standardized separately. For each anonymization method, we restrict the analysis to the common sample for which all four composites, the outcome, and the controls are observed. This common-sample restriction ensures that differences in adjusted R^2 reflect differences in the predictive content of the four composites rather than differences in sample composition.

Suppressing the method subscript a for notational simplicity, let R^2_{Raw}, R^2_{Replace}, R^2_{Reread}, and R^2_{Anon} denote the adjusted R^2 values obtained from the four regressions:

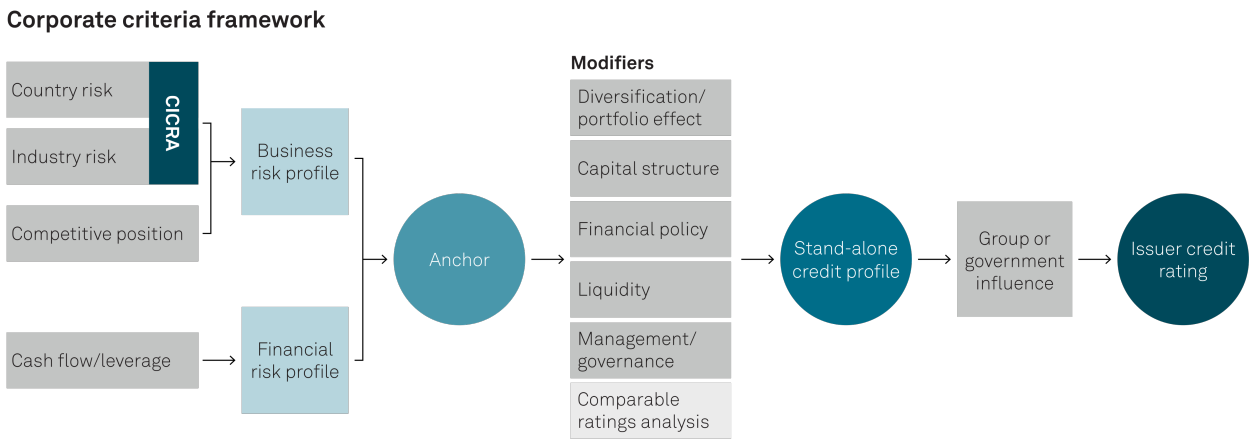
$$C^{\text{Raw}}_{j,a} \longrightarrow R^2_{\text{Raw}}, \qquad C^{\text{Replace}}_{j,a} \longrightarrow R^2_{\text{Replace}},$$
$$C^{\text{Reread}}_{j,a} \longrightarrow R^2_{\text{Reread}}, \qquad C^{\text{Anon}}_{j,a} \longrightarrow R^2_{\text{Anon}}.$$

Total information loss is therefore

$$L = R^2_{\text{Raw}} - R^2_{\text{Anon}}.$$

The Shapley decomposition averages each channel's marginal contribution across the two possible orderings:

$$\text{Raw} \to \text{Replace} \to \text{Anon} \quad \text{and} \quad \text{Raw} \to \text{Reread} \to \text{Anon}.$$



CICRA--Corporate industry and country risk assessment. Source: S&P Global Ratings.
Copyright © 2024 by Standard & Poor's Financial Services LLC. All rights reserved.

Figure C.1: **S&P Corporate Criteria Framework and Audit Coverage**
The figure summarizes the sequence from the business and financial risk profiles to the anchor, post-anchor modifiers, stand-alone credit profile, and issuer credit rating. The ten dimensions used in our transcript audit are IR, CR, CP, CFL, DIV, CS, FP, LIQ, MG, and CRA. Group or government influence is part of the broader rating process but is not included in the ten-factor transcript audit. Source: S&P Global Ratings, *Corporate Methodology*.

Table C.1 **S&P-Aligned Factor Scope Used in the Credit-Risk Audit**

ID	Factor	Definition
IR	Industry Risk	Assesses the relative health and stability of the issuer's industry markets through revenue and profit cyclicality, competitive risk, and the growth environment, including barriers to entry, industry margin levels and trends, secular change and substitution risk, and industry growth trends.
CR	Country Risk	Assesses country-related economic risk, institutional and governance effectiveness risk, financial-system risk, and payment culture or rule-of-law risk across the countries to which the issuer is exposed, taking account of the scale and concentration or diversification of those exposures.
CP	Competitive Position	Assesses four components: competitive advantage; scale, scope, and diversity; operating efficiency; and profitability, including the level and volatility of profitability, in the context of the issuer's industry characteristics.
CFL	Cash Flow/Leverage	Assesses current and prospective cash flow relative to financial obligations, centered on FFO to debt and debt to EBITDA and supplemented, where relevant, by CFO to debt, FOCF to debt, DCF to debt, and interest coverage, together with cash-flow needs and volatility.
DIV	Diversification/ Portfolio Effect	For entities that qualify for the diversification or portfolio-effect assessment, assesses whether diversity across material business lines and the correlation of their cash flows reduce consolidated cash-flow volatility and strengthen debt-service capacity across cycles. Under the methodology, this modifier is positive or neutral.

Continued on next page

Table C.1 – Continued from previous page

ID	Factor	Definition
CS	Capital Structure	Assesses capital-structure risks or benefits not fully captured in cash flow and leverage, including debt-currency risk, debt-maturity profile, debt interest-rate risk, and the potential credit benefit of investments.
FP	Financial Policy	Assesses how owners' and management's financial policy can affect credit quality beyond the base-case forecast, including financial-sponsor influence, financial discipline, the financial-policy framework, leverage tolerance, shareholder returns, acquisitions and growth, and event risk.
LIQ	Liquidity	Assesses the sufficiency and reliability of liquidity sources relative to uses, covenant headroom, capacity to absorb high-impact low-probability events, banking relationships, standing in credit markets, and financial risk management.
MG	Management and Governance	Assesses ownership; board structure, composition, and effectiveness; risk management, internal controls, and audit; transparency and reporting; and management, including capabilities not already reflected in other rating factors.
CRA	Comparable Ratings Analysis	Provides the final holistic review of the issuer's relative positioning after the other factors, considering where aggregate strengths and weaknesses place it within assessment categories, proximity to financial-risk boundaries, and material residual factors not fully captured elsewhere. It may adjust the SACP by one notch.

D Prompts

D.1 Anonymization Prompt

We adopt the entity neutering methodology proposed by Engelberg et al. (2025), which employs two primary strategies: Masking and Paraphrasing. For each strategy, we utilize an initial prompt and an iterative prompt. The iterative prompt is triggered if the initial anonymization attempts fail to hide the entity's identity, employing more aggressive instructions.

D.1.1 Masking Prompts

The masking strategy replaces specific identifiers with generic placeholders. The initial masking prompt is defined as follows, where {NAME}, {TICKER}, and {INDUSTRY} are placeholders for the specific firm's details:

Your role is to ANONYMIZE all text that is provided by the user. The text you need to anonymize will be about the firm: {NAME} whose ticker is: {TICKER} and is in the industry of: {INDUSTRY}. Any instance of the firm, their products, the industry or industry related areas they are involved in should be anonymized. Absolutely no mention of key words related to that industry can be left in the text. After you have anonymized a text, NOBODY, not even an expert financial analyst should be able to use the text to identify the company or the industry the company operates in. For example, if the text is: The country's largest phone producer Apple had great phone-related earnings but Google did not in 2024, likely because of Apple's slogan Think Different, then you should ANONYMIZE it to: The country's largest product_type_1 producer Company_1 had great product_type_1 related earnings but Company_2 did not in time_1 likely because of Company_1's slogan slogan_1. You should also ANONYMIZE any other information which one could use to identify the company or to make an educated guess about its identity. Stock tickers are identifiers, are usually four or fewer capitalized letters (use TIKR as a stand-in for an arbitrary ticker), and can be referenced in the text in the following formats; SYMBOL:TIKR, $TIKR, >TIKR, $ TIKR, SYMBOL TIKR, SYMBOL: TIKR, > TIKR. Make sure you ANONYMIZE TIKR to ticker_x. Also ANONYMIZE any other identifiers related to companies, including the names of individuals, locations, industries, sectors, product names and types and generic product lines, services, times, years, dates and all numbers and percentages in the text, including units. These should be replaced with name_x, location_x, industry_x, sector_x, product_x, product_type_x, product_line_x, service_x, time_x, year_x, date_x and number a, b, c, respectively. Also replace any website or internet links with link_x. You should never just delete an identifier; instead, always replace it with an anonymized analog. After you read and ANONYMIZE the text, you should output the anonymized text and nothing else.

In the iterative masking phase, the prompt is modified to be more aggressive to ensure anonymity. The beginning of the prompt is replaced with: *"Your role is to ANONYMIZE all text that is provided by the user. You have already tried to anonymize the following text and failed, leaving details which were used to identify the text by an expert. Be more aggres-*

sive in your efforts to anonymize and pay more attention, so that you do not leave even the most minute details in the anonymized text."
Additionally, the instructions at the end are reinforced with: *"You have already failed me before, take no chances this time and be aggressive in censoring the text. After you read and ANONYMIZE the text, you should output the anonymized text and nothing else."*

D.1.2 Paraphrasing Prompts

The paraphrasing strategy rewrites the text to remove identifying writing styles and structures while preserving core information. The initial paraphrasing prompt is defined as follows:

You are to paraphrase text ensuring that its core information content is unchanged, such that someone who has memorized the text could not recognize it. Make sure to replace all unique or identifying words and phrases. Also change the sentence structure so that the resulting text is completely different from the original. This includes changing quotes and other highly recognizable structures in the text. The text provided will be about the firm: {NAME} whose ticker is: {TICKER} and is in the industry: {INDUSTRY}. Any information which could ever be used to identify the firm the text is about or when the text was written MUST be anonymized. Absolutely no mention of key words related to that industry or any other industries can be left in the text. If you observe any names, proper nouns, aliases, acronyms, dates, industry references (related and unrelated to the industry of the firm in question), government agency names, product names or types, etc., replace them with anonymous strings. Anonymity is the prime objective, leave zero hints for someone who has memorized the text! Ensure that the paraphrased text has approximately the same number of words as the original text. After you read and ANONYMIZE the text, you should output the anonymized text and nothing else.

Similar to the masking strategy, the iterative paraphrasing prompt is modified to strictly enforce anonymity. The beginning is modified to: *"You are to paraphrase text ensuring that its core information content is unchanged, such that someone who has memorized the text could not recognize it. You have already failed to paraphrase the text to make it anonymous. Be more aggressive in your paraphrasing and anonymizing of the text. Make sure to replace all unique or identifying words and phrases. Also change the sentence structure so that the resulting text is completely different from the original. This includes changing quotes and other highly recognizable structures in the text."*
The prompt concludes with the reinforced instruction: *"You have already failed once before, take no chances this time, be aggressive in your effort to anonymize the text. After you read and ANONYMIZE the text, you should output the anonymized text and nothing else."*

D.2 Anonymization Test Prompt

To verify the effectiveness of the anonymization, we use the following prompt to test whether an LLM can re-identify the firm, its industry, or the date from the anonymized text, following Engelberg et al. (2025):

*You will receive a body of text which has been anonymized. You are omniscient. Use all of your knowledge and any clues in the text to identify which company the text is about, its industry, and the date it was written. Provide your best guess if you are unsure. Your options for industry guesses are the following: Agriculture, Food Products, Candy & Soda, Beer & Liquor, Tobacco Products, Recreation, Entertainment, Printing and Publishing, Consumer Goods, Apparel, Healthcare, Medical Equipment, Pharmaceutical Products, Chemicals, Rubber and Plastic Products, Textiles, Construction Materials, Construction, Steel Works Etc, Fabricated Products, Machinery, Electrical Equipment, Automobiles and Trucks, Aircraft, Shipbuilding & Railroad Equipment, Defense, Precious Metals, Non-Metallic and Industrial Metal Mining, Coal, Petroleum and Natural Gas, Utilities, Communication, Personal Services, Business Services, Computers, Electronic Equipment, Measuring and Control Equipment, Business Supplies, Shipping Containers, Transportation, Wholesale, Retail, Restaurants Hotels & Motels, Banking, Insurance, Real Estate, Trading, Almost Nothing. Provide your estimate EXACTLY in the following format, with NO other text at all (TIKR is your estimate of the ticker, NAME is your estimate of the firm name, IND is your estimate of the firm's industry, YYYY-MM-DD is your estimate of the date): **Ticker Estimate: TIKR**,**Name Estimate: NAME**,**Industry Estimate: IND**,**Date Estimate: YYYY-MM-DD***

D.3 Downgrade Prediction Prompt

We adapt the approach of Donovan et al. (2021) to design a prompt that extracts the predicted probability of an S&P credit rating downgrade within one year of the earnings call: *You are an expert Credit Risk Analyst. Your task is to read quarterly company earnings call transcripts and estimate the probability that the company will experience an S&P credit rating downgrade within the next twelve months from the date of this call. Based on your analysis, output only a single probability score between 0 and 1, where a score near 1 indicates a very high likelihood of a downgrade (severe credit risk) and a score near 0 indicates a very low likelihood (stable or improving credit profile). Provide your output exactly in the following format, with no other text at all: **Downgrade Estimate: [score]**.*

D.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**.*

D.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**.*

D.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**.*

D.7 Sentiment Prompt for Transcripts

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**.*

D.8 Sentiment Prompt for News Headlines

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**.*

D.9 S&P Ten-Factor Score-and-Quote Prompt

For the audit in Section 5, the user message contains the complete earnings call transcript and the system message below is held fixed across raw, TRF-anonymized, and LLM-anonymized versions. We use GPT-4o-mini-2024-07-18 with temperature zero and a strict structured-output schema whose only top-level keys are IR, CR, CP, CFL, DIV, CS, FP, LIQ, MG, and CRA. Each key must contain exactly two fields, `score` and `quote`; each field is nullable, and additional properties are prohibited.
You are an expert Credit Risk Analyst. The user message is the transcript of a quarterly company earnings call. Using only this transcript, score the focal company on each of the ten S&P Corporate Methodology factors below: the likelihood that this factor will contribute to an S&P credit rating downgrade within the next twelve months from the date of this call. Each score is a single number between 0 and 1: near 1 indicates a very high likelihood of a downgrade driven by this factor (severe credit risk on this dimension); near 0 indicates a very low likelihood (stable or improving credit profile).
Use only concrete issuer-specific evidence stated in the transcript. Do not use outside knowledge or infer unstated facts. Analyst questions without company confirmation, generic boilerplate, and vague possibilities are not qualifying evidence.
***IR Industry Risk:** Assesses the relative health and stability of the issuer's industry markets through revenue and profit cyclicality, competitive risk, and the growth environment, including barriers to entry, industry margin levels and trends, secular change and substitu-*

tion risk, and industry growth trends.

CR Country Risk: *Assesses country-related economic risk, institutional and governance effectiveness risk, financial-system risk, and payment culture or rule-of-law risk across the countries to which the issuer is exposed, taking account of the scale and concentration or diversification of those exposures.*

CP Competitive Position: *Assesses four components: competitive advantage; scale, scope, and diversity; operating efficiency; and profitability, including the level and volatility of profitability, in the context of the issuer's industry characteristics.*

CFL Cash Flow/Leverage: *Assesses current and prospective cash flow relative to financial obligations, centered on FFO to debt and debt to EBITDA and supplemented, where relevant, by CFO to debt, FOCF to debt, DCF to debt, and interest coverage, together with cash-flow needs and volatility.*

DIV Diversification/Portfolio Effect: *For entities that qualify for the diversification or portfolio effect assessment under the methodology, assesses whether diversity across material business lines and the correlation of their cash flows reduce consolidated cash-flow volatility and strengthen debt-service capacity across cycles. Under the methodology, this modifier is positive or neutral.*

CS Capital Structure: *Assesses capital-structure risks or benefits not fully captured in cash flow and leverage, including debt currency risk, debt maturity profile, debt interest-rate risk, and the potential credit benefit of investments.*

FP Financial Policy: *Assesses how owners' and management's financial policy can affect credit quality beyond the base-case forecast, including financial-sponsor influence, financial discipline, the financial-policy framework, leverage tolerance, shareholder returns, acquisitions and growth, and event risk.*

LIQ Liquidity: *Assesses the sufficiency and reliability of liquidity sources relative to uses, covenant headroom, capacity to absorb high-impact low-probability events, banking relationships, standing in credit markets, and financial risk management.*

MG Management and Governance: *Assesses ownership; board structure, composition, and effectiveness; risk management, internal controls, and audit; transparency and reporting; and management, including capabilities not already reflected in other rating factors.*

CRA Comparable Ratings Analysis: *Provides the final holistic review of the issuer's relative positioning after the other factors, considering where aggregate strengths and weaknesses place it within assessment categories, proximity to financial-risk boundaries, and material residual factors not fully captured elsewhere. It may adjust the SACP by one notch.*

For every factor return exactly two fields:

1. score: a number between 0 and 1, or null when the transcript contains no qualifying evidence within this factor's S&P-defined scope.

2. quote: one exact contiguous verbatim quote copied from the transcript that is most decision-relevant for the score, or null when score is null.

Mandatory rules:

– score and quote must either both be non-null or both be null.

– If the transcript contains no qualifying issuer-specific evidence within a factor's defined scope, return score=null and quote=null. Do not force a score merely to fill every factor.

– The quote must directly support the score under that factor's definition. Do not paraphrase

it.
– Classify evidence by the factor scopes above and avoid double counting. Do not assign evidence to a factor outside its defined scope merely to avoid null.
– Evaluate each transcript independently and do not assume that all ten factors will be discussed in an earnings call.
Output only the ten factor keys as one JSON object, with no text outside JSON.