Distortions in Financial Narratives: A ChatGPT Approach

In this study, we utilize ChatGPT to evaluate how financial content is distorted as it spreads across news outlets. We leverage a sample of exclusive news articles from the *Wall Street Journal* (WSJ) and track how these stories are retold in other news outlets. We aim to answer the following research questions: What news stories are retold (e.g., topics, companies referenced)? Who retells these stories (which news organizations retell original stories)? How are these original news stories retold (e.g., language style and details)? Can advanced AI models help us quantify how news is distorted as it is retold across news outlets? Does retellings of news and distortion have implications on asset prices and trading?

1. Data

1.1 Exclusive sample

Our study utilizes a sample of exclusive articles from WSJ. These articles represent exclusive reports to the WSJ on various issues, events, and developments, thereby serving as the primary source for subsequent news stories disseminated across other media outlets. By focusing on these original WSJ pieces, our research aims to trace and analyze the framing and narrative evolution as these stories are picked up, retold, and adapted by different publishers. We specifically focus only on WSJ articles that are labelled exclusive by Dow Jones (PMDM). We collect a sample of 29,702 exclusive WSJ articles that can potentially be retold by other news outlets.

1.2 Retellings sample

We compile a sample of articles specifically referencing the WSJ sourced from Factiva. We deliberately narrow our corpus to include only publications listed under “Newspapers: Top US newspapers”, while expressly excluding any articles published in “The Wall Street Journal” itself, as well as those associated with related tags, such as “WSJ Pro Private Equity”. To ensure consistency and relevance, we stipulate that all articles must be in English and contain explicit references to the WSJ within the full text. For a comprehensive enumeration of the references and tags considered, refer to Appendix A2. We collect a sample of 3,113 retellings between 2013 and 2022 across all the major newspapers in Factiva that can possibly match with exclusive WSJ articles from our sample.

1.3 Matching method

We perform document similarity matching between the exclusive WSJ articles and the Factiva sample of retellings using Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity. We combine the text of all articles from both groups for TF-IDF vectorization. This process transforms the articles into vectors, allowing the calculation of cosine similarities between each article in the exclusive WSJ articles group and all articles in the Factiva retellings group. Term Frequency (TF) measures how frequently a term appears in a document. It is calculated as:

$$TF(t,d) = \frac{\textit{Number of times term t appears in document d}}{\textit{Total number of terms in document d}}$$

Inverse Document Frequency (IDF) assesses the importance of a term *t* across all documents, formulated as:

$$IDF(t,D) = log\left(\frac{N}{|\{d \in D: t \in d\}|}\right)$$

where N is the total number of documents in the corpus D, and $(|\{d \in D: t \in d\}|)$ is the count of documents containing term t. The TF-IDF score combines these two measurements:

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

After transforming text into TF-IDF vectors, then we calculate the cosine similarity between vectors from the exclusive WSJ articles group and all vectors from Factiva retellings group. Cosine similarity is defined as:

$$\textit{cosine\_similarity}(E\ ,R\,) = \frac{E\ \cdot R}{||E|| \cdot ||R||}$$

Here, E and R are the TF-IDF vectors of the two documents, $(E\ \cdot R)$ is the dot product of the vectors, and $||E||$ and $||R||$ are the norms of the vectors E and R, respectively. For each article in the exclusive WSJ group, we identify the top five most similar articles from the Factiva retellings group (with replacement) within 14 days from the publication date of the exclusive WSJ article. We sort these articles by cosine similarity score. The potential matches are verified manually. We review the content of each WSJ article to grasp its primary focus and then evaluate the relatedness of each Factiva retellings article sequentially. A binary system is employed for assessment; 1 if a Factiva retellings group article directly referenced the content of the WSJ article in a meaningful way, such as mentioning a specific event or detail initially reported by the WSJ, and 0 for unrelated articles.

1.4 OpenAI ChatGPT variables

ChatGPT, developed by Open AI, is a large language model that can take sophisticated tasks and provide detailed and clear answers at a level similar to human experts. The model is based on the Generative Pre-trained Transformer (GPT) series of large language models. The GPT framework utilizes transformer architectures—advanced deep learning models adept at processing sequential data, notably text. Transformers are characterized by their self-attention mechanisms, enabling them to discern intricate word relationships within sentences. Pioneering this approach, Google's BERT (Bidirectional Encoder Representations from Transformers) emerged in 2018 as a foundational transformer-based model, garnering significant recognition. Subsequently, OpenAI's release of GPT-3 in June 2020, with its unprecedented 175 billion parameters and trained on 45TB of data, marked a significant advancement in the field. Building on this legacy, ChatGPT was introduced on November 30, 2022, astonishing the global community with its proficient generation of coherent and comprehensive responses across a multitude of knowledge areas. Continuing the evolution of transformer-based architectures, OpenAI introduced GPT-4, an even more advanced iteration of the Generative Pre-trained Transformer series. With a capacity exceeding its predecessor, GPT-4 is distinguished by its larger parameter count, enhanced training dataset, and improved fine-tuning techniques, resulting in superior understanding and generation of text.

We analyze how the style of language used in exclusive WSJ news and matched retellings along a wide range of dimensions. We utilize ChatGPT-4, to assess the text on the following five dimensions of language style and downstream consequence: Fact, Opinion, Negativity, Positivity, and Appeal. We utilize a questionnaire of 14 linguistic content and style elements to help us gauge the text on the five dimensions above. In each of the 14 questions, we use a scale of 1 = "not at all", and 7 = "very much so". The 14 elements in the questionnaire are motivated by Melumad, Meyer, and Kim (2021). The questionnaire is provided in Appendix A1. Next, we use the answers from the 14 elements to score the text on the five dimensions above as follows.

The first index captures the degree of specific factual details in the text. The first two elements in the questionnaire capture the degree of specific factual details and the degree of vagueness (reverse-coded, 7 = "not at all", and 1 = "very much so") in the text, respectively. The sum of the ratings on the first two elements is our *Fact* variable. We construct *Fact_A* and *Fact_Match* that correspond to the exclusive WSJ articles and the matched retellings (range between 2 and 14), respectively.

The second index captures the presence of opinions in the text. For this index, we sum ratings on elements 8, 11 (reverse-coded), and 14 (reverse-coded) as our *Opinion* variable. We construct *Opinion_A* and *Opinion_Match* that correspond to the exclusive WSJ articles and the matched retellings (range between 3 and 21), respectively.

The third and fourth indices capture negativity and positivity in the text. Negativity (positivity) is the degree of opposition (support) or disagreement (agreement) conveyed in the text. *Neg* is the sum of elements 6, 10, and 13. *Pos* is the sum of elements 5, 9, and 12. We construct *Neg_A* (*Pos_A*) and *Neg_Match* (*Pos_Match*) that correspond to the exclusive WSJ articles and the matched retellings (range between 3 and 21), respectively.

The final index is capturing the appeal of the text. In other words, how interesting the text is to read and the quality of the writing. We create *Appeal* as the sum of elements 4, 5, and 7. We construct *Appeal_A* and *Appeal_Match* that correspond to the exclusive WSJ articles and the matched retellings (range between 3 and 21), respectively.

1.5 Automated text variables

We compute various automated text variables to help us validate our responses from ChatGPT-4. To assess the level of complexity and specificity in the text, we compute the following variables. First, we compute Flesch Reading Ease (*Readability*), which is a readability test designed to assess the clarity of English writing. A higher score indicates easier readability, with scores typically ranging between 0 to 100, influenced by word and sentence length. Second, we count the numbers in the text (*numbers*) to assess the level of details and specificity in the text. Third, we calculate the percentage of words in the text that are complex (*complex*). We define complex words as those with two more syllables. To capture negativity/positivity, we use the 2018 version of the Loughran and McDonald (2011) lexicon. For each article, we count the number of positive words and negative words. We then scale these numbers by the total number of words in the document to get the percentage of positive words and the percentage of negative words (*positive*, *negative*).

1.6 CRSP variables

We collect returns and volume data from CRSP for companies mentioned by the exclusive WSJ articles and their retellings in our sample. We are able to identify 1,330unique stocks in our sample across 17,984 article-stock observations with return and price data available in CRSP. We calculate the market capitalization (*mve_m*) of the stock by multiplying the price of the share by the shares outstanding. We calculate the market beta (*BETA*) for the stock by estimating the Fama and MacBeth (1973) regressions using weekly returns and equal weighted market returns for three years. Idiosyncratic volatility (*idiovol*) is calculated following Ali, Hwang and Trombley (2003) by taking the standard deviation of residuals of weekly returns on weekly equal weighted market returns for three year window. We calculate the Amihud (2002) illiquidity measure (*ill*) by taking the average of daily absolute returns divided by dollar volume. We measure the leverage ratio (*lev*) following Bhandari (1988) by taking total liabilities and divide by the market capitalization. We measure the profitability of companies using the definition in Brown and Rowe (2007) by taking the annual earnings before interest and taxes net of non-operating income and scale by non-cash enterprise value. Finally, we capture market expectations by calculating book-to-market ratio following Rosenberg, Reid and Lanstein (1985) as the book value of equity divided by market capitalization. We calculate the abnormal returns for the referenced stocks using the Fama-French Plus Momentum model estimated using a 100-day trading window with a 50 day gap (70 day minimum window). *AbLogTurnovers* is the difference between log turnover on day t and the average log turnover from t -140 to t -20 trading days (6-month period, skipping most recent month).

2. Results:

2.1 *Descriptive statistics*

Table 1 presents data on exclusive news articles from the WSJ and their subsequent retellings by different news organizations. The dataset contains 1,940 observations of exclusive WSJ articles matched to a retelling. There are a total of 1,351 unique exclusive WSJ articles. The retellings are sourced from 18 different news organizations. The sample spans between 2013 and 2022, with a peak in 2014 and 2015 (174 articles each year) and the lowest in 2022 (107 articles). The organizations with the most retellings are 'INVD (Investor's Business Daily)' with 589 retellings, followed by 'NYPO (New York Post)' with 356, 'WPCO (Washington Post)' with 226, 'USAT (USA Today)' with 212, and 'NYDN (New York Daily News)' with 98.

In terms of the number of companies featured in each exclusive WSJ article, the average is 1.725 with a median of 1, a standard deviation of 2.167, and a range from 0 to a maximum of 26 companies. The average number of retellings per exclusive WSJ article is 1.404 with a standard deviation of 0.818, where the minimum number of retellings is 1 and the maximum is 5. The average number of days between the publication of an exclusive article and a retelling is 1.404, with a median of 1, a standard deviation of 3.38, and a range from 0 to a maximum of 14 days.

Table 2, Panel A, reports descriptive statistics on the ChatGPT indices of language style and downstream consequence. Panel B reports the mean difference between the exclusive WSJ sample and the matched Factiva retellings sample. We note that exclusive WSJ articles are deemed to contain more factual details (*Fact*) and are less vague than the matched retelling articles. Retelling articles are judged to be more opinionated (*Opinion*) than the exclusive WSJ articles. Moreover, retelling articles tend to have more negativity (*Neg*) and less positivity (*Pos*) than the exclusive WSJ articles. Finally, the exclusive WSJ articles have stronger appeal (*Appeal*) than the retelling articles. All the differences are statistically significant at the 1% level. The findings in Table 2 support the hypothesis that retellings go beyond relaying the key facts in the original story. Retellings attempt to provide guidance in a persuasive manner by becoming more opinionated and negative and less factual.

Table 3, Panel A, reports descriptive statistics of the automated text variables. Panel B compares the automated text variables between the exclusive WSJ group and the Factiva retellings group. Our results for the automated text variables help validate the responses from ChatGPT-4. If retellings become less factually detailed, we expect less specific and complex language. We compute the Flesch Reading Ease of the text to capture linguistic complexity (*Readability)*. Moreover, we look at the presence of numbers (*numbers*) and complex (*complex*) words in the text. Overall, we note that retelling articles tend to use less specific and complex language. We also examine how the tone varies between the original exclusive WSJ article and the retelling articles and note that the intensity of positive (*positive*) and negative (*negative*) tones both increases. In other words, the tone becomes more extreme in the retellings compared to the original story. Overall, the evidence from the automated text variables corroborates the results from the ChatGPT-4 indices.

Next, we report the descriptive statistics for the characteristics of the firms mentioned in the exclusive WSJ articles in Table 4, Panel A. We have 17,984 stock-WSJ exclusive articles or retelling article observations in our sample. 6.4% of those article-stock observations are retold by other newspapers. 8.9% represent the retellings by newspapers other than the WSJ. We note that the average market beta (*BETA*) of stocks in our sample is 1.110 and idiosyncratic volatility (*idiovol*) is 3.4%. The average market capitalization (*mve_m*) of companies in our sample is $143 billion. Firms in our sample have an average return on invested capital (*roic*) of 9.8% and book-to-market ratio (*bm*) of 0.474. The average leverage ratio (*lev*) is 2.738%. The average illiquidity using the Amihud (2002) measure is 0.002. The returns and abnormal returns corresponding to the date (t–1, 0) of the exclusive WSJ article or retelling article is 0.59 bps and 0.47 bps, respectively. In contrast, the monthly returns and abnormal returns after (t+1, t+21) the exclusive WSJ article or retelling article is 1.13% and –0.28 bps, respectively. We also report the average abnormal log turnover of stocks in our sample (*AbLogTurnovers)* during the time of the article and the month after, respectively.

2.2 *Disagreeable personalization*

Information from news can change when being retold by other news outlets. When news outlets perceive themselves as more knowledgeable about a specific topic, they are be more inclined to include their own opinions and interpretations. This personalization often leads to an emphasis on negative aspects of the content, driven by the reteller's desire to stand out and persuade the audience to value their guidance. As each recipient becomes a new source of the retold financial news, there is a compounding effect of opinionation and negativity. This means the final narrative received by the end audience might be significantly different from the original financial news, characterized by a pessimistic tone and possibly leading to misinformed or overly cautious decisions.

To examine how retellings of exclusive WSJ news might change details and language style, we regress the five ChatGPT-4 indices of language style and downstream consequence on an indicatory variable for whether the story is the original exclusive WSJ story or a retelling article (*retelling*). Each observation in our regression is an article (either an exclusive WSJ article or a retelling article). Moreover, we add 1) days between the retelling and the initial story (*log_retellings_same_day*); 2) how many tickers are mentioned in the exclusive article (*log_ticker_count*); and 3) number of other retellings the same day (*log_days_between*). We run the following regression with time and news outlet fixed effects and cluster standard errors by exclusive WSJ article id:

$$\begin{aligned} LanguageIndex_{i,t} \\ &= \alpha + \beta_1 \times retelling_{i,t} + \beta_2 \times log\_days\_between_{i,t} + \beta_3 \times log\_ticker\_count_{i,t} \\ &+ \beta_4 \times log\_retellings\_same\_day_{i,t} \end{aligned}$$

where *LanguageIndex* is one of the five ChatGPT-4 language style indices (*Fact*, *Opinion*, *Negativity*, *Positivity*, and *Appeal*) for article *i* at time *t*.

In the first model, where the dependent variable is *log_Fact*, a negative and statistically significant relationship is observed with *retelling*, indicating that retelling articles are deemed to have lower factual content than the exclusive WSJ articles. This finding suggests that as stories are retold, they become less factual, potentially due to alterations or exaggerations over time. In contrast, *log_Opinion* as a dependent variable in the second model shows a positive and significant relationship with *retelling*. In other words, retelling articles tend to be more opinionated compared to the original articles perhaps due to news outlets adding their interpretations or perspectives to the original story. For the negativity index (*log_Neg*), the third model shows a positive and significant relationship with *retelling* implying that retelling articles tend to have a more negative spin compared to the original story. Similarly, the fourth model with *log_Pos* as the dependent variable interestingly shows a negative significant relationship with *retelling*, suggesting that positive content diminishes in retellings compared to the original story. Finally, in the fifth model, *log_Appeal* has a negative significant relationship with *retelling*, indicating that retellings become less appealing (i.e., interesting, well-written) to read compared to the original story. All the coefficients on *retelling* are statistically significant at the 1% level.

Looking at the other independent variables, we note a positive and significant relation between *log_days_between* and *log_Opinion* and *log_Neg* suggesting that retellings that occur after more time passes since the original story tend to be more opinionated and negative. Moreover, stories with more tickers mentioned tend to be less subject to distortion when retold in terms of negativity and opinionisation. These results provide valuable insights into how news stories transform in their factual content, opinionation, tone, and appeal over time and through retellings.

2.3 *Asset pricing implications*

The retelling of financial news, especially when distorted or interpreted differently by various newspapers, can impact asset returns. When a financial story is retold with varying emphases or interpretations, it influences investors' perceptions and expectations, leading to divergent market reactions. For instance, an optimistic spin on economic data by one newspaper can foster positive investor sentiment, driving up asset prices. Conversely, a pessimistic interpretation by another source can incite risk aversion, precipitating a sell-off. This phenomenon underscores the role of media as a powerful agent in shaping market behavior, not merely through the dissemination of information but through its framing and interpretation. Such dynamics are critical in an academic context, as they offer insights into behavioral finance, highlighting how subjective interpretations of objective data can lead to market inefficiencies and volatility.

In this section, we explore how retellings of exclusive WSJ articles can impact asset returns. We regress abnormal returns on the date (t–1, 0) of the exclusive WSJ article or retelling article on indicators for whether the article has been retold (*retold*) or is a retelling of an exclusive WSJ article (*retelling*). Each observation in our regression is an article-firm (either an exclusive WSJ article or retelling article). Moreover, we add 1) *STORY* which is a vector of story characteristics including number of tickers mentioned, readability score, % of negative words in the text, count of numbers in the text, and % of complex words; and 2) *FIRM* which is a vector of observable firm characteristics including industry fixed effects and past returns. We cluster standard errors by firm.

$$AbRet_{i,j,t-1,t} = \alpha + \beta_1 \times retold_{i,j,t} + \beta_2 \times retelling_{i,t} + \beta_3 \times STORY_{i,t} + \beta_4 \times FIRM_{i,t-1}$$

where $AbRet_{i,t-1,t}$ is the compounded abnormal returns between t–1 and t adjusted using the Fama-French three factor model plus momentum. We use 100 days estimation window 50 days prior to time *t* and require a minimum of 70 valid returns to estimate the expected returns. We report the regression results in Table 6, Panel A. In specification (4), instead of *retelling* variable, we add more nuanced variables to capture specific characteristics of the retelling article including: 1) *log_Neg_per*; 2) *log_days_between*; and 3) *log_retellings_same_day*. These variables are 0 for all articles except those that are retelling articles of an exclusive WSJ article. *log_Neg_per* is the difference between the sum of the ChatGPT-4 language style indices between the original story and the retelling. A higher value indicates more negative personalization (more opinioned, more negative tone, less factual, and less appealing). To differentiate between retellings that occur soon after the initial report and those that come later, we include *log_days_between*, which is the log of the number of days between the retelling article and the exclusive WSJ article. Finally, we capture whether there are competing retelling articles about the a certain exclusive WSJ article on the same day using the *log_retellings_same_day*.

We find that exclusive WSJ articles that are retold are positively associated with contemporaneous abnormal returns. Moreover, if the retelling articles of exclusive WSJ articles are also positively associated with contemporaneous abnormal returns after controlling for past returns, story characteristics, and firm characteristics. In specification (4), we find a positive and significant relation between contemporaneous abnormal returns and negative personalization of articles. In other words, retelling articles with higher negative personalization (more opinioned, more negative tone, less factual, and less appealing), are associated with higher abnormal returns. In contrast, the longer the time between the retelling article and the exclusive WSJ article, the lower the contemporaneous abnormal returns.

In Panel B of Table 6, we examine how retelling articles can impact future abnormal returns. We run the following regression:

$$AbRet_{i,j,t+1,t+21} = \alpha + \beta_1 \times retold_{i,j,t} + \beta_2 \times retelling_{i,j,t} + \beta_3 \times STORY_{i,j,t} + \beta_4 \times FIRM_{i,j,t-1}$$

where $AbRet_{i,t,t+21}$ is the compounded abnormal returns between t and t+21 adjusted using the Fama-French three factor model plus momentum for stock *j* mentioned in article *i*.

Next, we examine how retellings can generate trading. In Table 7, Panel A, we run the following regression:

$$AbLogTurnovers_{i,j,t-1,t} = \alpha + \beta_1 \times retold_{i,j,t} + \beta_2 \times retelling_{i,t} + \beta_3 \times STORY_{i,t} + \beta_4 \times FIRM_{i,t-1}$$

where $AbLogTurnovers_{i,t-1,t}$ is the average of the difference between log turnover on day t–1 and t and the average log turnover from t -140 to t -20 trading days (6-month period, skipping most recent month). We also control for firm and story characteristics. If retelling articles are responsible for generating trading, we expect to find $\beta_2 > 0$. In specification (4), instead of *retelling* variable, we add more nuanced variables to capture specific characteristics of the retelling article including: 1) *log_Neg_per*; 2) *log_days_between*; and 3) *log_retellings_same_day*.

In Panel B of Table 7, we examine how retelling articles can generate future trading. We run the following regression:

$$AbLogTurnovers_{i,j,t+1,t+21} = \alpha + \beta_1 \times retold_{i,j,t} + \beta_2 \times retelling_{i,j,t} + \beta_3 \times STORY_{i,j,t} + \beta_4 \times FIRM_{i,j,t-1}$$

where $AbLogTurnovers_{i,t+1,t+21}$ is the average of the difference between log turnover on day t+1 and t+21 and the average log turnover from t -140 to t -20 trading days (6-month period, skipping most recent month).

2.4 Role of relative knowledge

The more knowledge the news outlet has about the subject, the more compelled they will be to add more opinions. Moreover, the news outlet might rely on negativity to increase attention and persuasion.

2.5 Retellings make the rumors true?

2.6 Is retelling good for society: price efficiency and noise

Appendix A1

You will be provided two related news articles. These two articles (delimited with XML tags, the first article has the <firstarticle> tag, and the second one has the <secondarticle>) come from different newspapers but are about the same story. Your task is to read and rate the two news articles. First, read the first article and rate it (on a 1: "not at all" to 7: "very much so" scale) based on how much you agree with the following statements related to the text:

1) This text is detailed;

2) This text is general / vague;

3) This text is interesting;

4) This text is well-written;

5) This text is positive about the subject matter;

6) This text is negative about the subject matter;

7) This text conveys interest in the subject matter;

8) This text is emotional about the subject matter;

9) In this text, the writer expressed support for the subject matter at hand;

10) In this text, the writer took an opposing stance towards the subject matter at hand;

11) In this text, the writer expressed no opinion about the subject matter itself;

12) In this text, the writer agreed with other people's opinions about the subject;

13) In this text, the writer disagreed with other people's opinions about the subject;

14) In this text, the writer expressed no opinion about other people's opinions.

Next, read the second article and rate it like above. However, this time when rating the second text, please think about how it compares to the first text (e.g., relative to the first article, the second article is more... or relative to the first article, in the second article the writer...:).

Your response should be in this format [first_article = {X, X, X, X, X, X, X, X, X, X, X, X, X, X}, second_article = {{X, X, X, X, X, X, X, X, X, X, X, X, X, X}], where X is a number between 1 and 7 that corresponds to the rating on the 14 elements above.

Appendix A2

Search phrases
According to the Wall Street Journal
appeared in The Wall Street Journal
article in The Wall Street Journal
by The Wall Street Journal
information of The Wall Street Journal
quoted in the Wall Street Journal
report from The Wall Street Journal
report in The Wall Street Journal
reported by the Wall Street Journal
reports the Wall Street Journal
Sources tell The Wall Street Journal
Speaking to the Wall Street Journal
story in The Wall Street Journal
The Wall Street Journal first reported
The Wall Street Journal had reported
The Wall Street Journal initially reported
The Wall Street Journal is reporting
The Wall Street Journal looks
The Wall Street Journal previously reported
The Wall Street Journal publishes
The Wall Street Journal say
The Wall Street Journal writes
told the Wall Street Journal
Wall Street Journal claimed
Wall Street Journal claims
Wall Street Journal described
Wall Street Journal describes
Wall Street Journal found
Wall Street Journal has reported
Wall Street Journal published
Wall Street Journal report
Wall Street Journal reported
Wall Street Journal reports
Wall Street Journal revealed
Wall Street Journal said
Wall Street Journal says

Table 1

Variable	N				
Exclusive news-match observations	1940				
Exclusive articles	1351				
News organizations	18				
Exclusive articles by year	2013: 100	2014: 174	2015: 174	2016: 121	2017: 143
	2018: 150	2019: 132	2020: 137	2021: 113	2022: 107
News organizations with most retellings	‘INVD’: 589	‘NYPO’: 356	‘WPCO’: 226	‘USAT’: 212	‘NYDN’: 98
	mean	median	std	max	min
# of companies reflected in each exclusive article	1.725	1	2.167	26	0
# of retellings per exclusive article	1.404	1	0.818	5	1
# of days between exclusive article and retelling	1.404	1	3.38	14	0

Table 2

Panel A: Distortion Variables (ChatGPT)

Variable	Number of Obs	Mean	Median	Std Dev	Max	Min
Fact_A	1941	11.915	12.000	1.141	14.000	3.000
Fact_Match	1941	10.458	11.000	2.696	14.000	2.000
Opinion_A	1941	8.732	8.000	3.133	20.000	3.000
Opinion_Match	1941	9.162	9.000	3.205	21.000	3.000
Neg_A	1941	6.441	6.000	2.466	21.000	3.000
Neg_Match	1941	7.105	7.000	3.097	21.000	3.000
Pos_A	1941	6.096	6.000	2.063	15.000	3.000
Pos_Match	1941	5.977	6.000	2.106	17.000	3.000
Appeal_A	1941	15.677	16.000	2.026	21.000	4.000
Appeal_Match	1941	14.808	15.000	2.990	21.000	4.000

Panel B: Means Difference

Variable	Diff (A-Match)	t-statistic	p-value
Fact	1.457	21.191	0.000
Opinion	-0.430	-7.544	0.000
Neg	-0.664	-12.256	0.000
Pos	0.118	3.372	0.001
Appeal	0.869	11.552	0.000

Table 3

Panel A: Automated Text Variables

Variable	Number of Obs	Mean	Median	Std Dev	Max	Min
Readability_A	1941	57.216	57.060	8.309	84.170	15.650
Readability_Match	1941	51.106	51.550	12.869	98.410	-9.820
numbers_A	1941	14.817	13.000	8.254	69.000	6.000
numbers_Match	1941	9.881	6.000	11.545	187.000	0.000
complex_A	1941	0.170	0.169	0.035	0.312	0.050
complex_Match	1941	0.162	0.162	0.037	0.321	0.039
positive_A	1941	0.612	0.529	0.555	3.846	0.000
positive_Match	1941	0.703	0.599	0.642	4.930	0.000
negative_A	1941	2.457	2.041	1.943	21.154	0.000
negative_Match	1941	2.566	2.216	2.071	14.894	0.000
Cosine_Similarity	1941	0.704	0.721	0.114	0.929	0.178

Panel B: Means Difference

Variable	Diff (A-Match)	t-statistic	p-value
Readability	6.110	19.346	0.000
numbers	4.936	16.117	0.000
complex	0.008	8.902	0.000
positive	-0.091	-5.517	0.000
negative	-0.109	-2.762	0.006

Table 4

Summary statistics for firm characteristics and market variables

Panel A: Firm Characteristics

Variable	Number of Obs	Mean	Median	Std Dev	Max	Min
retelling	17,984	0.089	0.000	0.285	1.000	0.000
retold	17,984	0.064	0.000	0.244	1.000	0.000
bm	17,984	0.474	0.329	0.467	4.008	-1.212
BETA	17,984	1.110	1.061	0.450	3.325	-0.124
idiovol	17,984	0.034	0.029	0.019	0.260	0.010
ill	17,984	0.002	0.000	0.036	2.205	0.000
lev	17,984	2.738	0.692	4.654	59.307	0.000
mve_m	17,984	143460.017	61594.095	252486.829	2902368.097	7.877
roic	17,984	0.098	0.083	0.289	1.090	-12.468

Panel B: Market variables

Variable	Number of Obs	Mean	Median	Std Dev	Max	Min
ret_-1_to_0	17,984	0.0059	0.0017	0.0936	7.8164	-0.6094
ret_1_to_21	17,984	0.0113	0.0110	0.1099	3.3916	-0.8504
abret_-1_to_0	17,984	0.0047	0.0001	0.0923	7.8079	-0.6428
abret_1_to_21	17,984	-0.0028	-0.0031	0.1042	2.8602	-0.9424
AbLogTurnovers_-1_to_0	17,984	0.0427	-0.0054	0.3597	4.6933	-1.5892
AbLogTurnovers_1_to_21	17,984	0.0163	-0.0056	0.2933	3.9652	-2.0869

Table 5

News organization and month-year fixed effects, cluster SEs by article ID

Dept Variable	(1) log_Fact	(2) log_Opinion	(3) log_Neg	(4) log_Pos	(5) log_Appeal
retelling	-0.1516***	0.0469***	0.0731***	-0.0197***	-0.0656***
	(-18.56)	(7.04)	(11.35)	(-3.80)	(-10.94)
log_days_between	0.0021	0.0200**	0.0477***	-0.0002	0.0056
	(0.48)	(2.17)	(4.82)	(-0.02)	(1.46)
log_ticker_count	0.0087	-0.0666***	-0.0793***	0.0176	-0.0039
	(1.37)	(-5.29)	(-6.29)	(1.49)	(-0.73)
log_retellings_same_day	0.0146	-0.0222	-0.0315	-0.0480**	0.0164*
	(1.31)	(-0.97)	(-1.29)	(-2.34)	(1.72)
_cons	2.5293***	2.2764***	2.0014***	1.9548***	2.7861***
	(193.21)	(86.21)	(73.38)	(78.17)	(245.71)
N	3794	3794	3794	3794	3794
adj. R-sq	0.171	0.085	0.132	0.063	0.092

Table 6

Industry fixed effects and SEs clustered by firm IDs

Panel A

	(1)	(2)	(3)	(4)
Dept Variable	abret_-1_to_0			
retold	0.0058**	0.0072***	0.0070**	0.0069**
	(2.04)	(2.59)	(2.48)	(2.49)
retelling	0.0056*	0.0077**	0.0070**	
	(1.84)	(2.47)	(2.04)	
log_Neg_per				0.0167***
				(2.92)
log_days_between				-0.0169***
				(-4.48)
log_retellings_same_day				0.0062
				(1.03)
log_ticker_count			-0.0041***	-0.0040***
			(-2.75)	(-2.69)
Readability			-0.0000	-0.0001
			(-0.27)	(-0.58)
negative			-0.0017***	-0.0016***
			(-3.36)	(-3.15)
numbers			-0.0000	-0.0001*
			(-1.32)	(-1.72)
complex_words			-0.0368	-0.0408
			(-0.94)	(-0.95)
BETA		-0.0008	-0.0006	-0.0000
		(-0.16)	(-0.12)	(-0.01)
idiovol		-0.3284	-0.3362	-0.3022
		(-1.03)	(-1.05)	(-0.92)
ill		0.2061**	0.2065**	0.2043**
		(2.47)	(2.48)	(2.46)
lev		0.0006	0.0006	0.0007
		(1.09)	(1.16)	(1.19)
roic		-0.0944	-0.0941	-0.0965
		(-1.44)	(-1.43)	(-1.44)
bm		-0.0113*	-0.0113*	-0.0118**
		(-1.93)	(-1.91)	(-2.00)
log_mv		-0.0056***	-0.0057***	-0.0056***
		(-4.47)	(-4.39)	(-4.19)
abret_-3_to_-2	-0.0069	-0.0079	-0.0089	-0.0101
	(-0.16)	(-0.21)	(-0.24)	(-0.27)
abret_-5_to_-4	0.0242	0.0152	0.0151	0.0086
	(0.33)	(0.24)	(0.24)	(0.14)
abret_-7_to_-6	-0.0647	-0.0619	-0.0626	-0.0633
	(-1.04)	(-1.11)	(-1.12)	(-1.07)
_cons	0.0039***	0.1277***	0.1451***	0.1458***
	(4.11)	(3.60)	(3.12)	(2.91)
N	17982	17890	17890	16823
adj. R-sq	0.007	0.103	0.104	0.109

Panel B

	(1)	(2)	(3)	(4)
Dept Variable	abret_1_to_20			
retold	**-0.0090****	**-0.0088****	**-0.0093****	**-0.0094****
	(-2.16)	**(-2.16)**	**(-2.25)**	**(-2.31)**
retelling	**-0.0099****	**-0.0094****	**-0.0108*****	
	(-2.55)	**(-2.49)**	**(-2.77)**	
log_Neg_per				**-0.0046**
				(-0.81)
log_days_between				**-0.0117****
				(-2.11)
log_retellings_same_day				**-0.0064**
				(-0.77)
log_ticker_count			-0.0024	-0.0033
			(-1.08)	(-1.55)
Readability			-0.0001	-0.0001
			(-1.16)	(-0.54)
negative			0.0001	0.0003
			(0.18)	(0.46)
numbers			-0.0001	-0.0001
			(-1.16)	(-0.87)
complex_words			0.0170	0.0252
			(0.59)	(0.80)
BETA		-0.0078	-0.0076	-0.0075
		(-1.28)	(-1.25)	(-1.22)
idiovol		-0.5499***	-0.5591***	-0.4946***
		(-3.60)	(-3.66)	(-3.10)
ill		-0.0015	-0.0024	-0.0032
		(-0.02)	(-0.03)	(-0.05)
lev		-0.0002	-0.0002	-0.0002
		(-0.36)	(-0.31)	(-0.38)
roic		-0.0066	-0.0065	-0.0047
		(-0.81)	(-0.80)	(-0.59)
bm		0.0055	0.0055	0.0050
		(0.69)	(0.69)	(0.60)
log_mv		-0.0048***	-0.0050***	-0.0048***
		(-4.22)	(-4.27)	(-4.24)
abret_-19_to_0	-0.0205	-0.0212	-0.0213	-0.0189
	(-0.89)	(-0.94)	(-0.94)	(-0.77)
_cons	-0.0011	0.1103***	0.1207***	0.1115***
	(-0.96)	(4.83)	(4.58)	(4.21)
N	17981	17889	17889	16822
adj. R-sq	0.027	0.036	0.036	0.036

Table 7

Industry fixed effects and SEs clustered by firm IDs

Panel A

	(1)	(2)	(3)	(4)
Dept Variable	AbLogTurnovers_-1_to_0			
retold	**0.1927***** **(3.60)**	**0.1986***** **(3.70)**	**0.1998***** **(3.72)**	**0.1975***** **(3.69)**
retelling	**0.2692***** **(4.74)**	**0.2793***** **(4.91)**	**0.2816***** **(4.88)**	
log_Neg_per				**0.1397***** **(2.89)**
log_days_between				**0.0001** **(0.00)**
log_retellings_same_day				**0.1633*** **(1.95)**
log_ticker_count			-0.0102 (-0.95)	-0.0004 (-0.06)
Readability			0.0001 (0.10)	0.0002 (0.30)
negative			0.0020 (0.75)	0.0022 (1.30)
numbers			0.0006** (2.00)	-0.0000 (-0.05)
complex_words			-0.2162 (-1.58)	-0.2429** (-2.12)
BETA		0.0366* (1.71)	0.0357* (1.66)	0.0171 (1.12)
idiovol		2.3552*** (4.90)	2.3480*** (4.88)	1.7878*** (4.85)
ill		-0.1982** (-2.01)	-0.1956** (-1.98)	-0.1238 (-1.44)
lev		0.0020 (0.87)	0.0020 (0.86)	0.0014 (0.88)
roic		-0.0153 (-0.90)	-0.0157 (-0.94)	-0.0195 (-1.36)
bm		-0.0406* (-1.83)	-0.0412* (-1.86)	-0.0223 (-1.31)
log_mv		-0.0274*** (-5.64)	-0.0273*** (-5.64)	-0.0192*** (-5.80)
abret_-3_to_-2	-0.0542 (-0.47)	-0.0532 (-0.53)	-0.0536 (-0.54)	-0.0597 (-0.62)
abret_-5_to_-4	0.4237*** (6.19)	0.3824*** (5.74)	0.3829*** (5.75)	0.4068*** (6.70)
abret_-7_to_-6	0.4592* (1.90)	0.3967* (1.83)	0.3956* (1.82)	0.3475* (1.70)
_cons	0.0053 (1.25)	0.3825*** (3.91)	0.4106*** (3.78)	0.3005*** (3.63)
N	17976	17884	17884	16818
adj. R-sq	0.118	0.165	0.166	0.162

Panel B

	(1)	(2)	(3)	(4)
Dept Variable	abret_1_to_20			
retold	0.1189**	0.1241**	0.1248**	0.1226**
	(2.39)	(2.50)	(2.51)	(2.47)
retelling	0.0945*	0.1033**	0.1010**	
	(1.96)	(2.15)	(2.10)	
log_Neg_per				0.1077***
				(2.64)
log_days_between				-0.0158
				(-0.31)
log_retellings_same_day				0.0195
				(0.31)
log_ticker_count			-0.0035	0.0029
			(-0.40)	(0.47)
Readability			-0.0005	-0.0001
			(-0.88)	(-0.25)
negative			0.0019	0.0021
			(0.89)	(1.34)
numbers			0.0004	0.0000
			(1.51)	(0.06)
complex_words			-0.1674	-0.1243
			(-1.27)	(-1.22)
BETA		0.0592***	0.0587***	0.0365***
		(3.08)	(3.04)	(2.68)
idiovol		2.0205***	2.0107***	1.4446***
		(4.30)	(4.29)	(3.93)
ill		-0.2018***	-0.2013***	-0.1508***
		(-3.02)	(-3.00)	(-2.76)
lev		0.0013	0.0013	0.0011
		(0.70)	(0.70)	(0.85)
roic		0.0055	0.0052	0.0012
		(0.29)	(0.27)	(0.07)
bm		-0.0336*	-0.0338*	-0.0206
		(-1.95)	(-1.96)	(-1.62)
log_mv		-0.0165***	-0.0167***	-0.0124***
		(-4.24)	(-4.27)	(-4.59)
abret_-19_to_0	0.1830***	0.1583***	0.1588***	0.1600***
	(2.74)	(2.69)	(2.69)	(3.21)
_cons	-0.0013	0.1663**	0.2189**	0.1527**
	(-0.33)	(2.08)	(2.28)	(2.14)
N	17973	17881	17881	16815
adj. R-sq	0.079	0.127	0.127	0.118