Finance-specific large language models: Advancing sentiment analysis and return prediction with LLaMA 2

I-Chan Chiu and Mao-Wei Hung

PUBLICATION RECORD

Pacific-Basin Finance Journal, Volume 90, Article 102632 (2025).

https://doi.org/10.1016/j.pacfin.2024.102632

Finance-Specific Large Language Models:

Advancing Sentiment Analysis and Return Prediction with LlaMA 2

by

I-Chan Chiu[a*], Mao-Wei Hung[b]

[a] Department of Finance, National Taiwan University. Address: No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan. E-mail: ichanchiu@gmail.com

[b] Department of International Business, National Taiwan University. Address: No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan. E-mail: mwhung@ntu.edu.tw

***Corresponding Author:** I-Chan Chiu

E-mail: ichanchiu@gmail.com

Phone: +886-2-3366-1096

Fax: +886-2-2366-0764

Postal Address: Department of Finance, National Taiwan University. Address: No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan.

Finance-Specific Large Language Models:

Advancing Sentiment Analysis and Return Prediction with LLaMA 2

Abstract

In this study, we develop an AI-driven summarization process for lengthy financial texts to improve our self-trained, finance-specific LLaMA-2 model. This approach allows for precise sentiment analysis, leading to more accurate return predictions on disclosures in the Management Discussion and Analysis sections of 10-K filings. Empirical results indicate that trading strategies based on LLaMA-2 sentiments produce significantly higher buy-and-hold returns (BHRs) compared to those derived from FinBERT (Huang et al., 2023) and traditional models. Furthermore, LLaMA-2 sentiment signals show a strong correlation with cumulative abnormal returns (CARs) and surpass traditional methods in predictive accuracy. The summarization process also enhances traditional models, generating significantly higher BHRs with summarized texts than with full texts. Both BHR and CAR results in our approach show robustness during periods of financial turbulence. These findings underscore the value of generative AI in finance and set a new standard for textual analysis.

Keywords: Generative AI, Large Language Model, AI-driven Summarization, Textual Analysis, Sentiment Analysis, Disclosures

JEL code: C40**,** C45, G11, G12, G17

1. Introduction

Textual analysis in finance requires navigating complex and diverse datasets. From SEC filings to earnings call transcripts, these sources are essential for understanding market dynamics and capturing investor sentiment. In this domain, the evolution of analytical methodologies has been noteworthy. The field has transitioned from dictionary-based methods (Bochkay et al., 2020; Loughran & McDonald, 2011) to machine learning models such as support vector machine (SVM) (Ren et al., 2018; Smailović et al., 2014), naive Bayesian classification (Azimi & Agrawal, 2021), and random forest (Frankel et al., 2022).

The incorporation of foundational deep learning architectures like convolutional neural networks (CNN), recurrent neural networks, and long short-term memory networks (Day & Lee, 2016; Devlin et al., 2018; Souma et al., 2019) has greatly enhanced financial text analysis. Moreover, recent studies have demonstrated the adaptability of these models in various financial applications. For example, Ghoshal & Roberts (2020) argue that traditional technical analysis lacks predictive rigor, and demonstrate that CNN-based deep learning offers a more reliable framework for financial forecasting by learning interpretable and effective features from data. Zhang et al. (2020) employ deep learning for ETF-based portfolio optimization, highlighting neural networks' potential in forming investment strategies.

Recently, large language models (LLMs) have gained widespread attention. ChatGPT, a version of the generative pre-trained transformer model designed specifically for chat, has been especially popular. Since ChatGPT's emergence, research into its financial applications has increased. Lopez-Lira & Tang (2023) show that ChatGPT can predict stock returns based on news headlines. Romanko et al. (2023) report that a ChatGPT-driven S&P 500 portfolio outperforms traditional strategies.

With the emergence of open-source LLMs like LLaMA (Touvron, Lavril, et al., 2023), LLaMA 2 (Touvron, Martin, et al., 2023), and Mixtral, self-training models has rapidly gained popularity. However, despite the existence of finance-specific models, most focus primarily on

task evaluation[1] rather than real-world empirical financial analysis, as seen in studies like Yang et al. (2023) and Zhu et al. (2024). A notable exception is FinBERT (Huang et al., 2023), a BERT-based LLM designed for sentiment analysis in finance. FinBERT has demonstrated superior empirical performance over the Loughran and McDonald (LM) dictionary and other machine learning algorithms in analyzing labeled sentences from analyst reports. However, as LLM technology continues to evolve, more advanced models may offer enhanced predictive capabilities, making them worthy of further investigation.

Sentiment analysis, foundational to financial forecasting, begin with Tetlock (2007)'s pioneering work, which demonstrates the predictive power of media sentiment on market prices. This shift broadens research beyond structured events, like earnings announcements (Beaver, 1968), to unstructured data. Researchers leverage various sentiment sources—social media, news articles, and more—to predict market behavior and uncover financial surprises (Bollen et al., 2011; Chen et al., 2014; Kelley & Tetlock, 2013; Malo et al., 2014; J. L. Zhang et al., 2016). Heston & Sinha (2017) have applied neural networks to assess how news aggregation affects stock return predictability. Frankel et al. (2022) demonstrate that machine learning models outperform dictionary-based methods in assessing sentiment around critical reporting periods.[2]

Building on these foundations, our study leverages advanced generative AI models to enhance sentiment analysis in finance. We employ Meta's LLaMA 2, a state-of-the-art LLM known for its sophisticated natural language understanding, and further fine-tune it with finance-specific sentiment datasets to embed domain-specific knowledge into the model. The

[1] Task evaluation typically involves benchmarking LLMs on structured tasks within predefined, simulated datasets, a common practice in finance to assess capabilities like question-answering, summarization, and classification. Representative studies have developed financial task benchmarks to measure model performance under these controlled settings (Islam et al., 2023; Xie et al., 2023). While these benchmarks provide insight into task-specific abilities, they may fall short of evaluating the model's performance on real-world financial analysis tasks, such as return prediction based on empirical market data.

[2] A broad spectrum of literature delves into the relationship between sentiment and financial variables, from asset return patterns to the predictability of returns (Bollen et al., 2011; Chang et al., 2019; Hiew et al., 2019; Kim et al., 2017; Lutz, 2016; Mishev et al., 2020; Ren et al., 2018; Renault, 2017; Schmeling, 2009; Souma et al., 2019; Stambaugh et al., 2012; Tetlock, 2010; Tetlock et al., 2008; Yu & Yuan, 2011).

target text data is on the Management Discussion and Analysis (MD&A) sections of 10-K filings, which provide critical insights into a company's financial health and outlook. Recognized for their strategic narratives and forward-looking disclosures, MD&A sections offer analysts valuable predictive information on future financial performance.[3] Our main objective is to accurately capture the price information from MD&A texts by our finance-specific LLaMA-2 model.

However, most LLMs, including LLaMA 2, face limitations in input length, computational demands, and insufficient tagged training data for extensive documents. These constraints pose challenges for effective sentiment analysis on lengthy texts. Financial document like MD&A in 10K filings, contain complex, nuanced information that standard LLMs often struggle to fully interpret. In our experiment, we find more than 93% of the outputs are erroneously generated by using the full-text inputs directly. To address this, we develop a two-step AI-driven summarization process specifically suited to the demands of financial text analysis. By leveraging Google's LongT5 (Guo et al., 2022), a generative AI model optimized for summarizing extensive documents, our method transforms lengthy financial texts into concise, information-dense summaries.

In our framework, the MD&A texts are summarized, and the sentiments are calculated by the LLMs, including fine-tuned LLaMA-2 variants and the FinBERT model for comparison. Additionally, traditional approaches, including machine learning methods and the LM dictionary, are also applied to the full texts and summarized texts for comparison. The flowchart is detailed in Figure 1. Our analyses demonstrate that summarized texts effectively retain essential price information for accurate sentiment analysis, benefiting both LLMs and traditional methods alike.

In our empirical studies, we focus primarily on buy-and-hold return (BHR) analysis and

[3] Several studies have analyzed the information contained in managerial disclosures from MD&A sections (Bochkay et al., 2023; Brown et al., 2023; Cole & Jones, 2004; Li, 2010; Muslu et al., 2015; Tavcar, 1998).

market-adjusted cumulative abnormal return (CAR) analysis. In the BHR analysis, we find that our best LLaMA-2 model achieves an average annual BHR of 12.37% based on the sentiment signals generated from summarized texts. This significantly outperforms FinBERT (5.98%) and traditional methods (best at 7.14%) on summarized texts, underscoring the effectiveness of the combination of the "summarization-first" and LLaMA-2 interpretation in return prediction. Additionally, our results indicate that, within traditional methodologies, focusing sentiment analysis on summarized texts yields significantly more profitable BHRs than using full texts. This suggests the critical role of summarization in preserving and distilling price information, even for traditional approaches.

For the filing event influences based on the CAR analysis, our results show that LLMs, including LLaMA-2 variants and FinBERT, effectively interpret summarized financial texts across all CAR windows. However, FinBERT's predictability fades in our robustness test under turbulent market conditions, while LLaMA-2 models remain robust. Traditional methods show no significance with summarized texts and only limited significance with full texts in CAR tests. This suggests that, for event-based price detection, traditional approaches to interpreting MD&A information—whether on full or summarized texts—are less accurate than LLM-based interpretation.

The contributions of our study are threefold. Firstly, we introduce a pioneering approach in finance by combining a finance-specific LLaMA-2 model with an AI-driven summarization process. By fine-tuning LLaMA 2 on finance-specific datasets and coupling it with LongT5-based summarization, we effectively capture price information embedded in extensive financial text. To our knowledge, we are the first to apply a generative AI-driven summarization as a preparatory step before analysis in financial text processing. This approach refines lengthy financial documents for subsequent LLM-based sentiment analysis, establishing a novel framework for extracting market-relevant insights and advancing NLP methodologies in finance.

Secondly, our study contributes to literature on sentiment analysis in finance by capturing directional investor sentiment with greater precision through our LLaMA-2 framework. Our approach highlights the increasing importance and enhanced accuracy of advanced models in sentiment analysis. Furthermore, our findings confirm a strong correlation between sentiment tones and asset prices, which demonstrates the significance of sentiment analysis in financial forecasting.

Finally, we extend the literature on disclosures in finance and accounting domain by focusing on MD&A sections of 10-K filings as a predictive source of market information. Our analysis shows that even when MD&A text is condensed, it still retains essential predictive content about future prices. This result affirms the value of MD&A disclosures for capturing managerial insights and strategic outlook. It also establishes our framework as a reliable method for extracting actionable data from corporate narratives. Together, these contributions chart new directions for integrating AI into financial disclosure analysis, sentiment interpretation, and return predictability.

This study is structured as follows. Section 2 describes our methodology of LLaMA-2 model training and AI-driven summarization process, Section 3 presents our data, Section 4 details our empirical findings, and Section 5 concludes with key insights and contributions.

2. Methodology

This section outlines our methodology. Section 2.1 details the fine-tuning process of the LLaMA-2 model and explains how it differentiates between sentiment classes. Section 2.2 discusses our AI-driven summarization process, highlighting its benefits and limitations for sentiment analysis.

2.1 Finance-specific LLaMA-2 Sentiment Model

Meta's LLaMA 2 is extensively pre-trained on a vast dataset of two trillion tokens, which

provides it with robust language knowledge and semantic understanding. Although LLaMA 2 is a powerful model and performs well in evaluation tasks[4], it lacks domain-specific knowledge. Therefore, fine-tuning LLaMA 2 with finance-specific datasets is essential to capture the distinct language and sentiment nuances found in financial texts.

For fine-tuning, we have trained the LLaMA-2 7B model[5] by four datasets: Stanford Alpaca (Taori et al., 2023), Financial PhraseBank (FPB) (Malo et al., 2014), Twitter Financial News Sentiment (TFNS)[6], and Climate Sentiment (CS) (Bingler et al., 2023). These datasets are selected to enhance the model's sentiment analysis capabilities across both finance and climate-related contexts. The details are provided in Appendix A, Table A1.

The fine-tuning process utilizes specific hyperparameters, aligned with Meta's training template and recommended hardware configurations. The fine-tuning process has been conducted on an RTX4090 GPU with 24GB RAM. Key settings include:

- **Low-Rank Adaptation (LoRA) Configuration:** LoRA parameters are set with a rank r=8, lora_alpha=32, and lora_dropout=0.05.
- **Learning Rate and Optimizer:** The model is fine-tuned using a learning rate of 10^{-5} with the AdamW optimizer.
- **Batch Size and Epochs:** Training is conducted over 5 epochs with a batch size of 4 and gradient accumulation set to 1.

These configurations ensure consistent performance, as reflected in the final training loss of 0.9372, detailed in Appendix A, Table A2 and Figure A1. The consistent downward trend in training loss indicates effective model learning while minimizing overfitting.

The sentiment classification approach is based on a combination of datasets representing

[4] Please refer to the website of LLaMA 2 (https://www.llama.com/llama2/).
[5] Among the available LLaMA-2 model sizes—7B, 13B, and 70B parameters—we select the 7B version to balance computational efficiency with high performance, ensuring suitability for our hardware environment without compromising analytical depth.
[6] "zeroshot/twitter-financial-news-sentiment" on Hugging Face (https://huggingface.co/).

two main domains: stock price-oriented datasets (FPB and TFNS) and climate risk-oriented datasets (CS). Each dataset follows distinct annotation standards that capture the sentiment characteristics relevant to its domain, enabling our model to generalize across both financial and environmental sentiment contexts.

- **Stock Price-Oriented Datasets:** The FPB dataset is widely used in financial sentiment analysis. It categorizes sentences as Positive, Negative, or Neutral based on their expected impact on stock prices from an investor's perspective. Labels in this dataset were assigned by financial experts, ensuring reliability. For the TFNS dataset, we adapt the original annotations to fit our framework, reclassifying "bullish" labels as Positive, "bearish" labels as Negative, and retaining "neutral" labels for content with no expected effect on stock prices. This dataset captures the market sentiment associated with "bullish" or "bearish" signals, helping to align our model's sentiment categories with real-world market interpretations.
- **Climate Risk-Oriented Dataset:** The CS dataset classifies text from financial reports based on the climate-related impacts on firms. Sentences that emphasize business opportunities or the positive effects of climate adaptation are originally labeled as "Opportunity," which we classify as Positive. Those discussing risks or adverse impacts associated with climate change are labeled as "Risk," which we classify as Negative. Statements without a clear positive or negative tone are assigned Neutral labels. This categorization ensures that our model can effectively interpret financial sentiment related to climate risks and opportunities.

Based on these standards, we define our three sentiment categories as follows:

- **Positive:** Text that conveys favorable information or an upward trend, such as anticipated growth, profit increases, or business opportunities, representing a positive signal for

investors.

- **Negative:** Text that includes risks or negative factors, such as losses, high costs, or business risks, signaling potential downsides from an investment perspective.
- **Neutral:** Descriptive text that does not explicitly influence stock prices, often providing context or background information without directional or price-related content.

The model classifies financial text sentiment based on predefined criteria, with Positive or Negative labels applied to texts containing clear directional cues, while Neutral is used for texts without such signals. To enhance transparency, Table B1 in Appendix B presents the 50 most frequent 1-grams and 2-grams for each sentiment category in our training data, highlighting the distribution of terms and unique characteristics of each dataset. For sentiment generation, the LLaMA-2 model is prompted with a financial text and a targeted question: "What is the sentiment of the following financial text? Positive, Negative, or Neutral?" This prompt guides the model's analysis and categorization, with its response formatted as "Sentiment: {Sentiment Tag}" to indicate the sentiment clearly.[7]

To learn more about our sentiment classification, we incorporate Bayesian estimation as a probabilistic framework to provide insights into the model's uncertainty across Positive, Negative, and Neutral categories. Using the English Finance News dataset[8], we manually adjust the dataset to ensure equal representation across each sentiment class, establishing a balanced prior expectation for sentiment distribution.

Because we recognize the limitations of LLMs in directly generating probability distributions for sentiment classes, we adopt an empirically driven Dirichlet Bayesian approach. This involves iteratively sampling and updating sentiment probabilities based on observed model predictions. For each batch, we draw 500 samples with replacement and prompt the model to assign a sentiment label to each text. The resulting proportions of Positive, Negative,

[7] {Sentiment Tag} = Positive, Negative or Neutral.
[8] "lukecarlate/english_finance_news" on Hugging Face (https://huggingface.co/).

and Neutral predictions in each batch act as empirical likelihoods, reflecting the model's observed sentiment distribution and guiding the Bayesian update process. With each batch, we combine the Dirichlet prior with the observed likelihoods to calculate the posterior distribution. This approach enables us to approximate the posterior distributions of sentiment probabilities for each category over multiple iterations and reveals the model's probabilistic inclinations.

Figure 2 presents the posterior distributions for each sentiment class and shows that "Neutral" is the most frequently selected category across all three classes. This distribution also indicates that the model tends to favor "Neutral" classifications when sentiment signals are unclear, which reflects a cautious approach in handling ambiguous cases. This behavior aligns with our objective of achieving balanced sentiment judgments, with the model reserving stronger Positive or Negative classifications for cases with more distinct sentiment cues. This cautious inclination is further supported by examples in Appendix B, where clear sentiment cases are shown in Table B2 and ambiguous cases in Table B3. The examples in Table B3 reveal that when the model encounters complex or mixed sentiment cues, it predominantly assigns a "Neutral" label.

Our adaptation of LLaMA 2 diverges from standard implementations by specifically targeting domain-specific language structures and terminology in financial contexts. Unlike typical sentiment models that focus on general emotional tones, our fine-tuning emphasizes capturing price directional information. Through training on finance-specific data, our model detects market sentiment nuances that are predictive of market movement, advancing beyond the scope of traditional sentiment analysis.

Additionally, prior machine learning and deep learning techniques in financial text analysis often use vector-based approaches, like *tf-idf* and *word2vec*, which deconstruct text by relying on isolated word vectors without contextual understanding. In contrast, LLaMA 2, with its billions of parameters, enables a more nuanced understanding of financial language. This model examines sequential and contextual relationships within text to capture sentiment with

greater depth and accuracy.

To broaden our research, we train four traditional machine learning models—SVM, logistic regression, random forest, and neural network—using the FPB, TFNS, and CS datasets for baseline comparison with LLaMA 2. The Stanford Alpaca dataset is excluded here, as it focuses more on general conversation than financial sentiment. We pre-process the data by removing special symbols and stop words, followed by *tf-idf* vectorization, ensuring clean, relevant input. Additionally, we include FinBERT for comparison with LLaMA 2, providing a robust framework to evaluate sentiment analysis performance across models.

2.2 AI-driven Summarization Process

Our goal is to develop a method that condenses extensive financial information into concise, manageable summaries while preserving its core message. Traditional single-step summarization methods often struggle to condense these texts accurately, as the volume of information can overwhelm model capacity and compromise interpretive depth. To address this, we have developed a structured two-step summarization process tailored to the specific demands of financial content. As outlined in Figure 3, this process begins with an iterative summarization of overlapping text segments, which preserves continuity across sections of the document. This is followed by an aggregative step that consolidates the summaries into a final, cohesive output. This method enables advanced models, like LLaMA 2, to interpret sentiment and predict returns from large-scale financial documents feasibly, establishing a new methodological standard for using LLMs in financial document analysis.

First, we assess the necessity of the "two-step" summarization process for effective sentiment analysis. We have conducted a 500-sample sentiment analysis experiment using human-tagged data as a subjective benchmark (see Appendix C) to validate our two-step summarization approach. Processing full-text inputs with LLaMA 2 resulted in a high error rate (more than 93%) due to input length limitations and the scarcity of training data for long-form

financial documents. A single-step summarization reduces text length but keeps the text overly lengthy and achieves only 33% accuracy against human-tagged labels. In contrast, our two-step summarization improves accuracy to 70.2%, which highlights the importance of this structured approach.

This accuracy metric primarily indicates the necessity and effectiveness of the two-step summarization. It shows improvement over the single-step process rather than directly evaluating LLaMA-2's sentiment classification ability. The true measure of LLaMA-2's capability in sentiment detection is more thoroughly demonstrated in our empirical analysis. Additionally, our two-step process improves computational efficiency by reducing processing time by over 95%—from 775.32 seconds of full texts to 36.93 seconds of summarized texts—making it both feasible and reliable for large-scale financial text analysis.

Our chosen generative AI model for summarization tasks, Google LongT5 (Guo et al., 2022), builds on the Google T5 (Raffel et al., 2020) framework and is designed to process extensive texts. It achieves state-of-the-art results in summarization tasks that require an understanding of longer contexts and multiple documents.[9] It excels in generating cohesive, information-rich summaries, crucial for applications where depth and clarity are paramount. In text summarization, two primary methodologies are widely used: extractive and abstractive.

Extractive summarization (e.g., BERT-based models) selects key phrases directly from the original text, often resulting in fragmented, keyword-focused summaries that lack fluidity and readability. For instance, from "Widget Corp's sales have skyrocketed due to the successful launch of their new Widget360," an extractive model might produce: "Widget Corp sales skyrocketed launch Widget360." Although effective for simpler documents, extractive summarization struggles with the depth and context embedded in complex financial texts like 10-K filings, where key insights are woven into lengthy sections.

[9] Please refer to the github website of LongT5 model (https://github.com/google-research/longt5).

Abstractive summarization, in contrast, restructures sentences and creates more coherent, reader-friendly summaries. Models like T5 and LongT5 excel at this approach, especially for financial contexts that require accuracy and readability. For example, LongT5 would generate: "Widget Corp's launch of the Widget360 has led to significant sales growth, surpassing market expectations." Despite its complexity, abstractive summarization is valuable for financial documents, where it captures higher-level insights across narratives, tables, and raw data.

Our two-step approach utilizes LongT5 specifically because it combines the advantages of large-scale text handling with the depth of abstractive summarization. Traditional T5 models are limited to 512 tokens, which is insufficient for the expansive MD&A sections of financial reports. LongT5, however, supports up to 16,384 tokens. This extended token capacity allows us to work with larger, context-rich "chunks," minimizing the loss of continuity commonly encountered in standard chunking methods. Unlike traditional chunk-based methods, which segment the text into small parts and risk creating disjointed summaries, our process retains more of the document's original flow, preserving nuanced financial insights.

This approach is strengthened by our choice of the "long-t5-tglobal-base-16384-book-summary" variant (Szemraj, 2022), fine-tuned on the *BookSum* dataset (Kryściński et al., 2022). Although *BookSum* is not customized for finance, its training on lengthy documents helps LongT5 create smooth, connected summaries. This reduces fragmentation in the text and keeps important connections across different parts of the content. By applying larger, overlapping windows, our iterative step captures broader context without sacrificing interpretability. This distinguishes our method from standard chunk-based approaches, enabling closer alignment with the original narrative.

To understand information retention in our summarization process, Table 1 presents metrics such as word count, compression rate, cosine similarity, and ROUGE scores for full-text, first-step, and second-step summarized texts. This breakdown highlights trade-offs in our approach. Final summarized texts reduce text length to an average of just 0.72% of the full

texts, with a ROUGE-1 score of 2.16%, which reflects significant compression. Although the average cosine similarity score of 0.4257 suggests partial retention of core content, it reflects a balanced approach aimed at maintaining key elements while achieving high brevity.

Figure 4 presents a Venn diagram showing key terms retained in both full texts and summarized texts. Additionally, Appendix D includes word frequency analysis and word clouds to illustrate content retention at each summarization stage. Table D2 in Appendix D shows that 31 of the top 50 words (1-grams and 2-grams) remain consistent across summarization stages. Figure 4 highlights key terms crucial to business (e.g., Company, Customer, Product, Service, Market), financial reporting (e.g., Revenue, Income, Expense, Cost), performance (e.g., Increase, Loss, Result), and operations (e.g., Operating, Credit). This demonstrates that important nuances for pricing information are effectively preserved. However, forward-looking or potentially negative terms, like "Future" and "Decrease," appear less frequently, underscoring a limitation in capturing the full spectrum of sentiment nuances.

Appendix E provides examples illustrating the process's effectiveness, showing that most summarized texts retain meaningful content. The AI's capability to filter out less relevant details while preserving the core narrative is evident. For instance, an MD&A section (Accession Num: 0000950170-22-002517), initially 37,469 words, was condensed to 301 words while maintaining key information and insights. This example, along with others in the appendix, highlights the model's ability to produce summarized texts that are suitable for further analysis.

Our summarization strategy delivers manageable, interpretable data without overloading the model, enhancing both computational efficiency and predictive accuracy. However, it still presents certain limitations and biases that warrant further attention.

LongT5 presents challenges for capturing domain-specific details in financial contexts. Since *BookSum* was trained on book-length narratives, it lacks financial specificity, which can lead to potential misinterpretations of specialized terminology. For instance, terms like "interest"

and “line” may default to common meanings like curiosity or a physical boundary rather than the financial meanings “interest rates” or “in line with expectations.” These subtle shifts may introduce biases by prioritizing general interpretations over context-specific financial meanings. However, given the limited availability of finance-specific, long-form summarization data, this *BookSum*-finetuned model remains a practical choice for capturing the essence of extensive texts.

Additionally, as previously discussed, our analysis in Table 1 provides insight into the trade-offs between similarity and compression in our two-step approach. Specifically, the final second-step summarized texts reach a cosine similarity of 0.4257 with the full text, compared to 0.5572 for the first-step summarized texts, indicating a degree of deviation from the initial content. While this lower similarity is a necessary compromise to achieve conciseness, it suggests that, even as key points are retained, the second-step summarization may result in a partial loss of the full document’s depth and nuance.

Moreover, Figure 4 and Appendix D illustrate that “decrease” appears less frequently in the second-step summarized texts. This finding suggests a tendency toward a neutral or positive tone, potentially impacting downstream sentiment analysis. Appendix C’s sentiment analysis experiment supports this observation, showing an accuracy increase with the second-step summarized texts but also revealing a shift in sentiment distribution, with fewer negative sentiments and more positive ones in the second-step summarized texts (Figure C2). This bias may lead to overly optimistic sentiment classifications, indicating that some nuanced aspects of the original text may be lost.

While our two-step summarization approach improves computational feasibility and model performance, it entails trade-offs in interpretive accuracy and informational depth. Despite the improvements we have implemented, a certain degree of information loss and tonal bias is unavoidable in achieving a manageable, effective summary suitable for sentiment and predictive analysis on complex financial texts. From a technical standpoint, our summarization

processes have been executed in a modest computing environment, utilizing a single-core V100 GPU with 16GB RAM.

3. Data

Our dataset spans 1994 to 2022. Stock price data is sourced from the Center for Research in Security Prices and supplemented by Compustat annual financial reports. Our analysis focuses on firms listed on the NYSE, Amex, and Nasdaq (EXCHCD 1, 2, and 3), limited to ordinary common shares (SHRCD 10 and 11) and excluding financial firms (SICCD 6000–6999).

We use the "EDGAR" package in R (Lonare et al., 2021)[10] to retrieve isolated MD&A texts for sentiment analysis with LLaMA 2, FinBERT, and traditional machine learning models. We have also customized the package's *getSentiment* function to calculate LM sentiment scores from the original MD&A texts. For texts beyond the package's scope, we calculate LM sentiment scores according to the protocols outlined by the Software Repository for Accounting and Finance.[11]

Table 2 presents descriptive statistics across three panels. Panel A includes return metrics such as daily stock returns, risk-free rates, and market returns, alongside BHRs over market-adjusted CARs. It also lists control variables including turnover rate, market value, book-to-market ratio, return on assets, volatility, leverage ratio, Tobin's q, financial text word count, NASDAQ listing indicator, sales and investment growth rates, gross margin, and cumulative returns. Panel B presents sentiment analysis statistics and compares sentiment measures extracted by LLMs, including LLaMA-2 variants and FinBERT, alongside traditional methods. Panel C displays the correlations between these sentiment measures.

[10] This package was developed by Lonare, Patil, and Raut (2021). The details of usage methodology and funtionalities can be checked on the website (https://cran.r-project.org/web/packages/edgar/index.html).
[11] SRAF (https://sraf.nd.edu/) is a website designed to provide a central repository for programs and data used in accounting and finance research, initially proposed in (Loughran & McDonald, 2016).

4. Empirical Findings

This section presents the empirical analysis of how sentiment influences financial market performance under various conditions and through different methodologies. Section 4.1 provides an analysis of BHRs across models and compares the performance of LLaMA 2, FinBERT, and traditional approaches. Section 4.2 explores the relationship between sentiment tones and CARs and presents results from both LLMs and traditional methods. Section 4.3 assesses the robustness of LLaMA 2 during the volatile period from 2000 to 2009, covering major events such as the dotcom bubble and the 2008 financial crisis.

4.1 Results of Buy-and-Hold Strategies

This subsection presents an empirical analysis of average BHRs influenced by sentiment tones, as shown in Table 3, highlighting the practical impact of sentiment on investment decisions. Panel A examines BHRs based on the sentiment tone at day t, with trades executed at the close of day $t+1$. The strategy involves buying after a positive sentiment or short selling following a negative sentiment, with positions closed after varying holding periods.

The table summarizes the average BHRs for these strategies across holding durations of one week, two weeks, one month, three months, six months, and one year. The results illustrate how sentiment tones impact performance. These tones are derived from summarized texts processed by various LLMs, including LLaMA-2 model variants across five training epochs, FinBERT, and traditional methodologies like the LM dictionary and machine learning models. The analysis also includes full texts assessed with traditional methods for comparison.

Panel A of Table 3 presents the average BHR results across various sentiment models. Among the LLMs, all LLaMA-2 variants trained over five epochs achieve strong one-year BHRs above 10%, with the 5-epoch model (*LLAMA2_TONE_E5*) reaching the highest at 12.37%. This significantly outperforms FinBERT, which achieves a one-year BHR of 5.98%

on summarized texts. Traditional methods applied to summarized texts show one-year BHRs ranging from -2.55% to 7.14%, also underperforming compared to LLaMA-2 counterparts. These results underscore the superior capability of LLaMA 2 in sentiment-driven investment strategies compared to other models.

The results for traditional methods applied to full texts reveal substantial underperformance, with one-year BHRs ranging from -1.27% to -15.01%. The LM dictionary model fares the worst, likely due to an overemphasis on negative cases, which could stem from excessive pessimism in full texts. This result contrasts with the slight optimistic bias observed in summarized texts, suggesting that full text analysis may contain overly pessimistic noise affecting sentiment accuracy.

Panel B examines the significance of differences in BHRs between summarized and full texts across various traditional methodologies. This panel tests how summarized texts impact the performance of traditional models by comparing BHRs for each model with 1% significance in most cases. These results highlight that the summarization process enhances sentiment-based predictions, refining information in a way that benefits traditional methods. Despite this improvement, LLaMA-2 variants remains the strongest in leveraging summarized texts, achieving the most robust results across all models.

4.2 Results of Cumulative Abnormal Returns (CARs)

This section analyzes the effect of sentiment derived from both LLMs and traditional methods on market-adjusted CARs surrounding filing events. By examining CARs across multiple timeframes from $CAR[0,1]$ to $CAR[0,5]$, we evaluate the efficacy of each sentiment analysis approach in detecting abnormal returns indicative of market reactions to disclosed information.

Firstly, we assess the performance of LLMs. Table 4 presents the regression results, with Panel A detailing the outcomes for *LLAMA2_TONE_E5*, and Panel B offering a comparative analysis across additional LLaMA-2 variants and FinBERT. The results show that all LLMs

yield statistically significant positive CARs across the examined time windows ($CAR[0,1]$ to $CAR[0,5]$). For instance, *LLAMA2_TONE_E5* achieves a t-stat of 1.99 at $CAR[0,1]$ to of 2.67 at $CAR[0,5]$. These positive abnormal returns underscore each model's interpretative capability to extract meaningful sentiment signals from summarized texts and capture market reactions to filing events.

This section shifts focus to traditional methodologies and their efficacy in detecting CARs based on sentiment derived from both full and summarized texts, as presented in Table 5. Panel A examines sentiment analysis using traditional methods applied to full texts, specifically the LM dictionary and machine learning models. Our findings show that traditional sentiment analysis methods—particularly the LM dictionary—exhibit some correlation with CARs, especially over longer timeframes (a t-stat of 0.78 at $CAR[0,1]$ to a t-stat of 2.92 at $CAR[0,5]$). This suggest some predictive power in capturing price-relevant sentiment from MD&A sections, though less accurately than LLMs.

Panel B explores traditional methodologies applied to sentiment derived from summarized texts. Here, sentiment scores and tones generated by traditional methods lack statistical significance across all CAR windows, highlighting their limitations in interpreting sentiment from summarized texts. Traditional models appear to struggle with capturing the nuanced sentiment signals that LLMs effectively discern in summarized content.

All findings in Table 4 and Table 5 highlight the superior precision of LLMs over traditional models in capturing market reactions through event-driven sentiment analysis, especially with summarized texts.

4.3 Results during the Financial Turbulence Times

In this subsection, we examine BHR and CAR analyses during the turbulent period from 2000 to 2009, marked by events like the dotcom bubble burst and the 2008 financial crisis. Focusing on this volatile era allows us to test the robustness of our methodology and other models under

extreme market conditions.

Table 6 reinforces the findings from Table 3, showing that our BHR analyses remain strong even during market instability. The LLaMA-2 models consistently perform better than other models, highlighting their strength across evaluations. For traditional methodologies, BHR differences between summarized and full texts become less significant over longer periods, especially in models like neural networks, logistic regression, and SVM. Apart from these few exceptions, the results consistently show strong significance. These results highlight our framework's robustness in BHR analysis across varying market conditions.

Table 7 presents the CAR analysis during periods of financial turbulence, assessing how sentiment indicators respond under extreme market conditions. The *LLAMA2_TONE_E5* model displays consistently strong performance across all CAR windows, particularly evident in short- and long-term timeframes. For example, *LLAMA2_TONE_E5* achieves a t-stat of 1.95 at $CAR[0,1]$ and 3.28 at $CAR[0,5]$, underscoring its robust predictive ability in both immediate and extended event windows. In contrast, FinBERT shows a weaker response, with a t-stat of 0.89 at $CAR[0,1]$ to of 2.31 at $CAR[0,5]$, suggesting a relatively diminished effectiveness compared to *LLAMA2_TONE_E5* in capturing nuanced sentiment during volatile periods.

These results highlight LLaMA-2's adaptability and superior sentiment-capturing capability, which prove advantageous in navigating complex market reactions tied to significant financial events. While the LM dictionary applied to full texts maintains some predictive power in line with past analyses, it remains less effective than the LLaMA-2 models. The strong performance of *LLAMA2_TONE_E5* reinforces the effectiveness of advanced LLMs in extracting meaningful insights from summarized texts during turbulent periods, making it a highly reliable tool for event-driven sentiment analysis.

5. Conclusion

In this study, we introduce an innovative framework that combines the LLaMA-2 model with an AI-driven summarization process, enhancing textual analysis in finance. By fine-tuning LLaMA 2 on finance-specific datasets and employing a "summarization-first" approach with LongT5, we effectively condense the lengthy texts, filter out noise and capture nuanced sentiment cues from MD&A sections of 10-K filings. Empirical results demonstrate that LLaMA 2 consistently outperforms other analytical methods in both BHR and CAR analyses.

Our research contributes to applications of generative AI in textual analysis, sentiment analysis, and disclosure by providing a scalable framework that enables deeper, contextually aware sentiment detection. The ability of LLaMA 2 to capture price-sensitive insights and predict future market responses underscores the potential of advanced AI models to interpret complex financial narratives more accurately.

Looking forward, our framework opens promising avenues for future research. Potential directions include refining AI interpretability in finance, enhancing real-time analysis capabilities, and extending these methods across sectors for cross-market insights. This study sets a foundation for AI-enhanced financial analysis, where event-specific sentiment detection can support data-driven decision-making and embed AI insights into financial strategy and operations.

Figure 1. Research Process Flowchart

This figure illustrates the research methodology that underpins our study. On the left, we outline the sentiment recognition model training process, which starts with the merging of datasets and extends to the training of both LLMs and traditional machine learning models. The right segment of the figure delineates the core process, commencing with data collection, processing, and integration. Here, MD&A texts from the merged dataset undergo an AI-driven summarization process, resulting in concise, summarized data. These summarized texts, alongside the original MD&A texts, are then analyzed using the previously trained models to produce the final dataset. The process culminates in an empirical analysis of this refined data, drawing insights and conclusions from the sentiment-laden financial narratives.

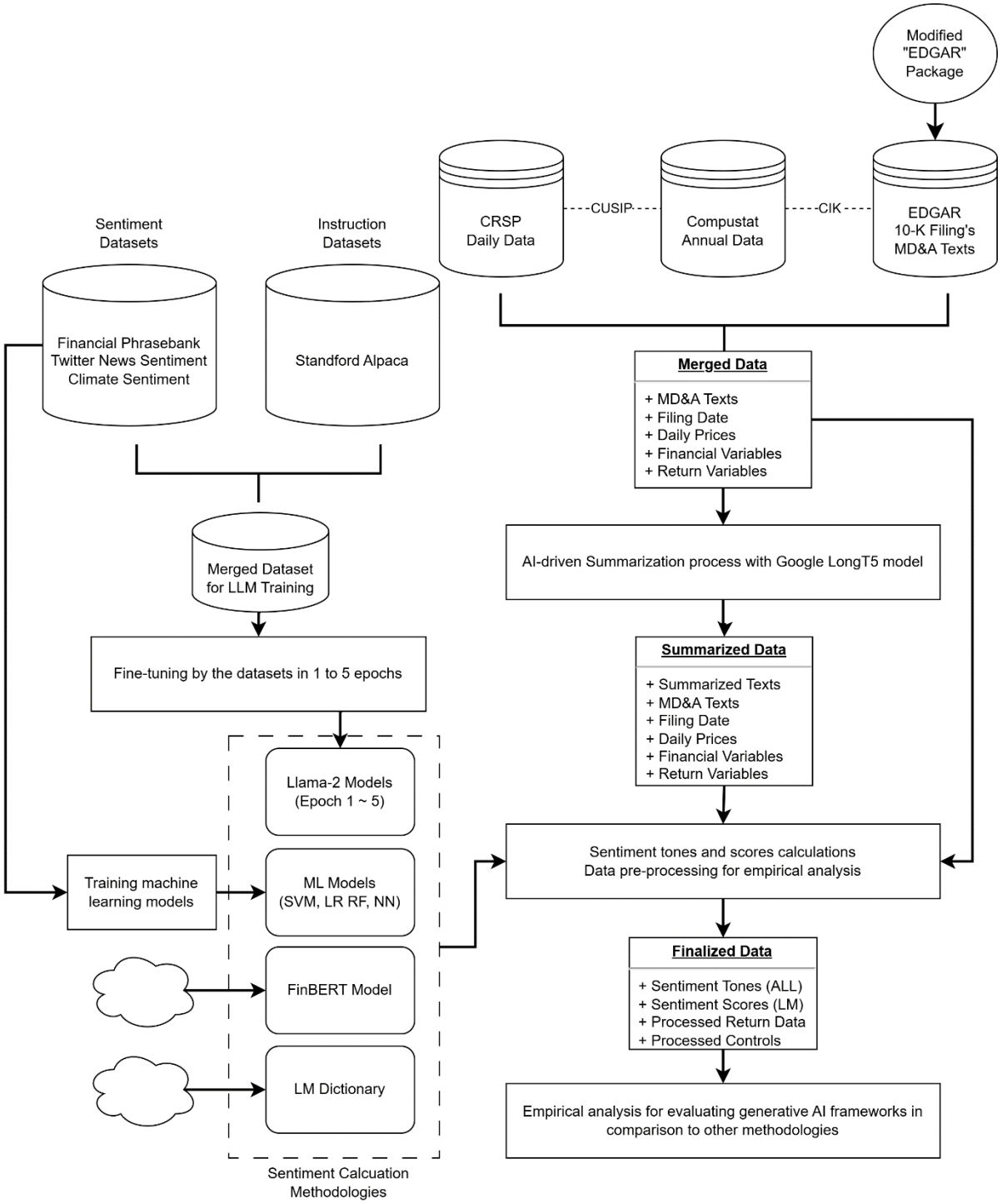


Figure 2. Bayesian Posterior Probability Distributions of LLaMA-2 Sentiment Model
This figure presents the posterior probability distributions for each sentiment class (Positive, Negative, and Neutral) in the LLaMA-2 model, updated with a Dirichlet prior. The distributions illustrate the model's estimated probability range for each sentiment category, providing insight into the model's probabilistic assessment across different sentiment classes.

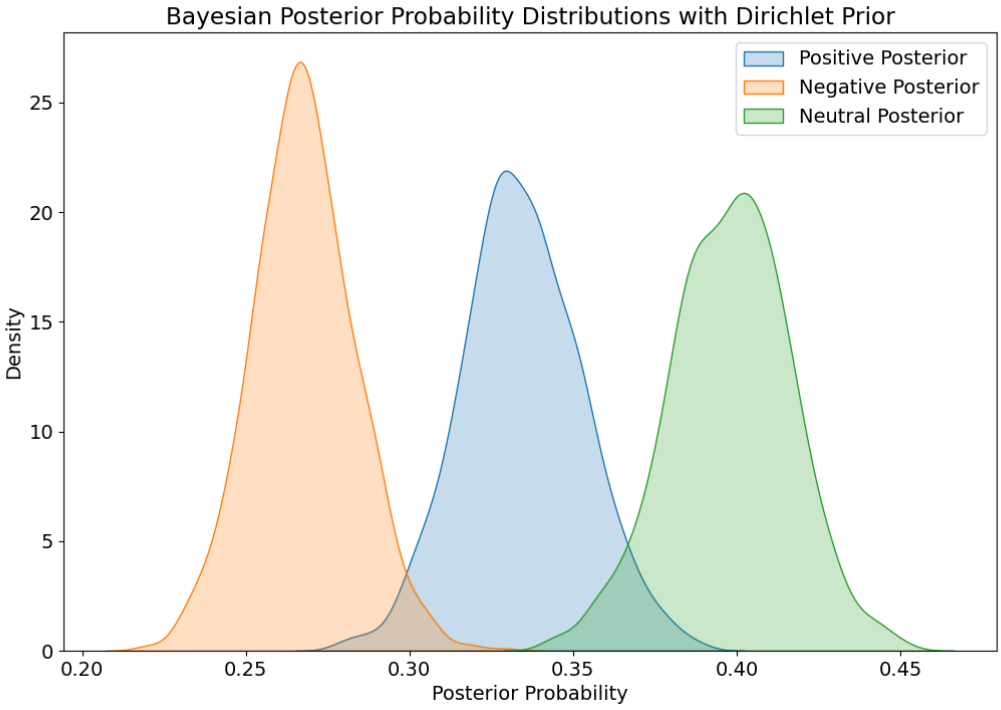


Figure 3. Summarization Process for MD&A Texts

This figure depicts the summarization process with the AI model applied to the MD&A sections of 10-K filings. In the first step, a sliding window moves through the document to perform sequential summarizations. In the second step, the summarized segments are concatenated and then subjected to an additional summarization step, resulting in a consolidated summary. This extracted summary is rich in information, free of noises, and focuses on the essential points, achieving a significant reduction in length while preserving the critical content.

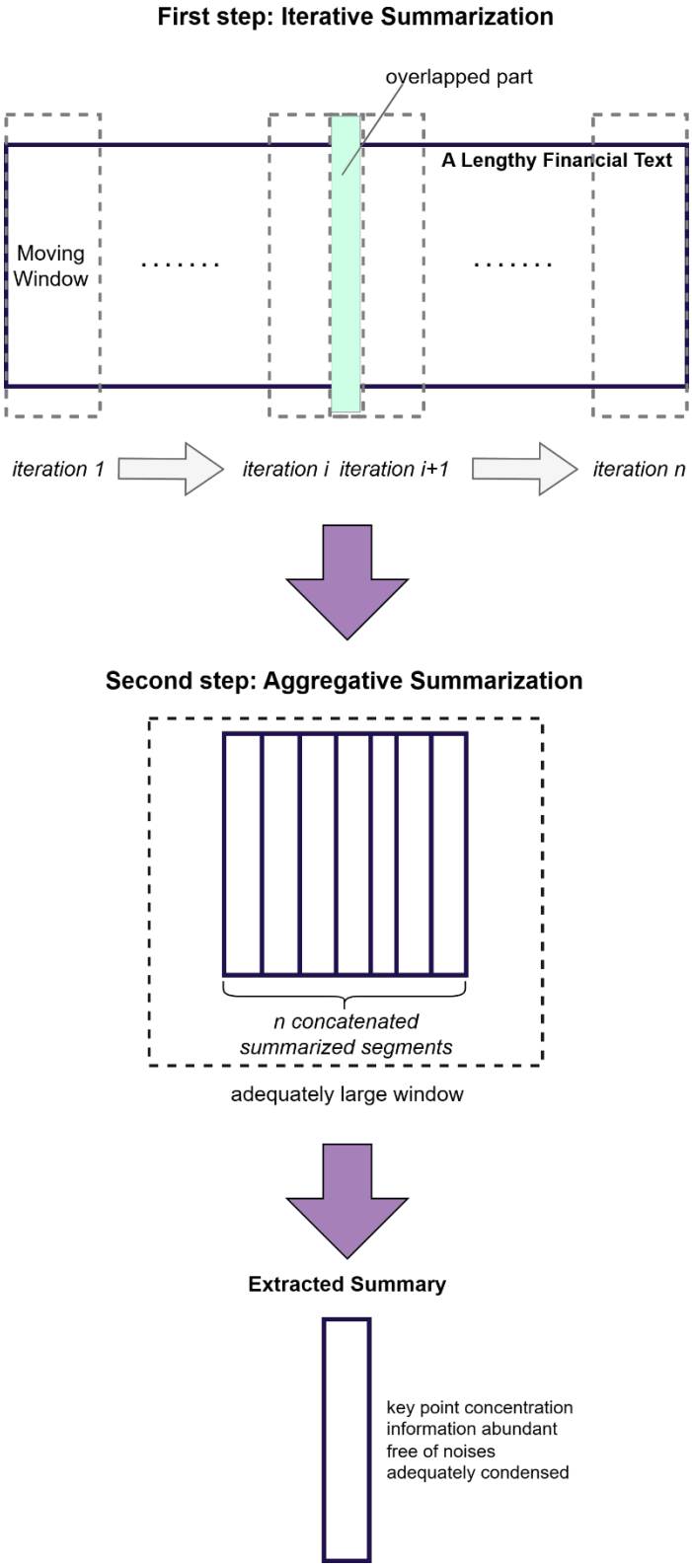


Figure 4. Key Terms Shared and Unique Among Full Texts and Summarized Texts
This Venn diagram shows the main shared and unique terms among the full texts, first-step summarized texts, and second-step summarized texts. Terms in overlapping sections are common across multiple text types, while terms in non-overlapping areas are unique to each specific text type.

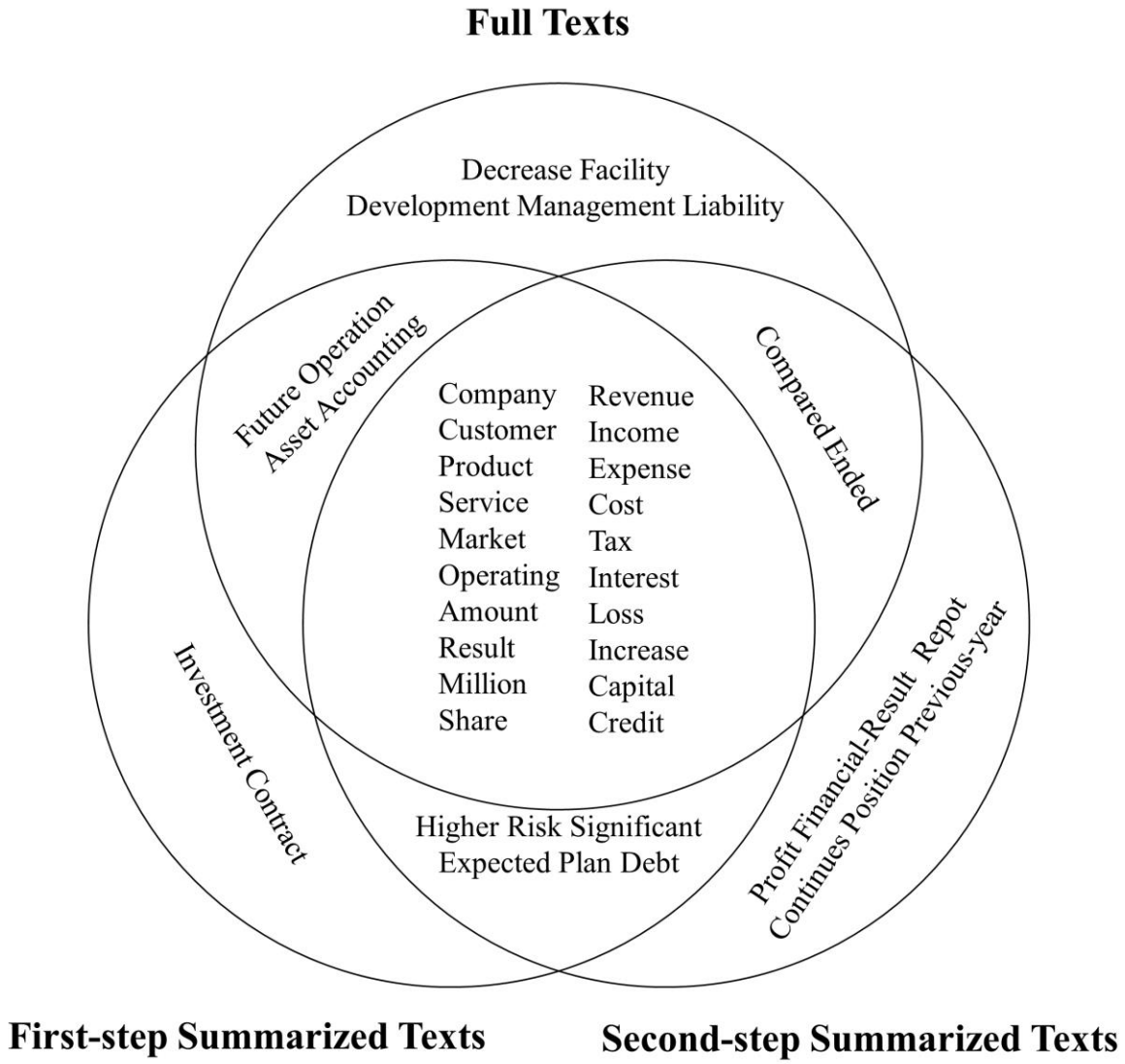


Table 1. Comparison of Full Text and Summarized Text Characteristics and Trade-offs

This table provides a comprehensive overview of full text (FT), first-step summarized text (ST_1), and second-step summarized text (ST_2) characteristics, emphasizing word count statistics and trade-off measures after each summarization stage. Panel A presents summary statistics for word counts (WC) across these three types of texts, including mean, standard deviation (SD), percentiles (P25, P50, P75), minimum (MIN), and maximum (MAX) values over the period from 1994 to 2022. Panel B highlights the average trade-off metrics by year, examining Word Count (WC), Compression Percentage (CP) — defined as the word count proportion of summarized texts relative to the full texts— along with Cosine Similarity, ROUGE-1, and ROUGE-L scores between each pair of text types (FT, ST_1, ST_2). These metrics provide insights into the balance between compression and textual fidelity achieved through summarization.

Panel A. Summary Statistics of Word Counts for Full Text and Summarized Texts						
Text Type	**MEAN**	**SD**	**P25**	**P50**	**P75**	**MIN**	**MAX**
Full-text Wordcount ($\boldsymbol{WC(FT)}$)	8972.10	7710.89	4137	7419	11482	101	253907
First-step Summarized-text Wordcount ($\boldsymbol{WC(ST_1)}$)	479.81	438.51	205	377	618	3	13310
Second-step Summarized-text Wordcount ($\boldsymbol{WC(ST_2)}$)	64.59	46.06	33	52	84	3	446

Panel B. Yearly Average Trade-off Metrics between Full Text and Summarized Texts

Year	$\boldsymbol{WC(FT)}$	$\boldsymbol{WC(ST_1)}$	$\boldsymbol{WC(ST_2)}$	$\boldsymbol{CP(ST_1)}$	$\boldsymbol{CP(ST_2)}$	$\boldsymbol{CosSim}$ $\boldsymbol{(FT, ST_1)}$	$\boldsymbol{CosSim}$ $\boldsymbol{(FT, ST_2)}$	$\boldsymbol{CosSim}$ $\boldsymbol{(ST_1, ST_2)}$	$\boldsymbol{ROUGE\text{-}1}$ $\boldsymbol{(FT, ST_1)}$	$\boldsymbol{ROUGE\text{-}1}$ $\boldsymbol{(FT, ST_2)}$	$\boldsymbol{ROUGE\text{-}1}$ $\boldsymbol{(ST_1, ST_2)}$	$\boldsymbol{ROUGE\text{-}L}$ $\boldsymbol{(FT, ST_1)}$	$\boldsymbol{ROUGE\text{-}L}$ $\boldsymbol{(FT, ST_2)}$	$\boldsymbol{ROUGE\text{-}L}$ $\boldsymbol{(ST_1, ST_2)}$	**Samples**
1994	3926.95	213.73	48.35	5.44%	1.23%	0.5940	0.4855	0.7271	9.87%	4.15%	46.95%	7.10%	3.22%	42.27%	488
1995	3325.73	182.69	49.20	5.49%	1.48%	0.5829	0.4788	0.7614	9.84%	4.39%	50.78%	7.01%	3.32%	46.32%	1064
1996	3197.80	176.30	48.59	5.51%	1.52%	0.5894	0.4787	0.7534	10.07%	4.44%	50.46%	7.24%	3.35%	45.92%	2371
1997	3440.65	189.76	47.60	5.52%	1.38%	0.5919	0.4805	0.7247	9.99%	3.99%	46.14%	7.17%	3.03%	41.22%	4042
1998	4033.44	225.11	47.32	5.58%	1.17%	0.5824	0.4610	0.6948	10.08%	3.41%	39.90%	7.17%	2.61%	34.52%	4259
1999	4726.31	264.62	47.52	5.60%	1.01%	0.5748	0.4411	0.6697	10.13%	2.87%	33.93%	7.17%	2.20%	28.23%	4249
2000	4682.91	273.78	49.50	5.85%	1.06%	0.5776	0.4477	0.6803	10.40%	3.03%	35.48%	7.30%	2.29%	29.91%	4353
2001	5332.75	312.34	53.73	5.86%	1.01%	0.5708	0.4493	0.6785	10.40%	2.98%	34.99%	7.25%	2.24%	29.41%	4253
2002	7434.03	422.60	61.57	5.68%	0.83%	0.5614	0.4361	0.6571	10.24%	2.50%	29.73%	7.01%	1.88%	23.85%	4071
2003	9173.17	510.48	66.34	5.56%	0.72%	0.5608	0.4283	0.6403	10.08%	2.15%	26.22%	6.88%	1.63%	20.50%	3926
2004	9917.51	543.96	69.19	5.48%	0.70%	0.5594	0.4267	0.6385	9.86%	1.94%	25.11%	6.71%	1.47%	19.29%	3841
2005	10510.10	579.95	72.32	5.52%	0.69%	0.5634	0.4242	0.6355	9.90%	1.82%	24.20%	6.70%	1.38%	18.33%	3850
2006	9384.93	507.95	65.82	5.41%	0.70%	0.5602	0.4163	0.6352	9.71%	1.83%	24.19%	6.62%	1.40%	18.37%	3794
2007	9851.34	526.06	67.77	5.34%	0.69%	0.5575	0.4160	0.6329	9.51%	1.71%	23.67%	6.47%	1.30%	17.78%	3747
2008	11413.45	599.15	71.89	5.25%	0.63%	0.5527	0.4105	0.6289	9.33%	1.57%	22.76%	6.36%	1.21%	16.94%	3610
2009	11015.13	577.56	71.96	5.24%	0.65%	0.5521	0.4117	0.6249	9.32%	1.54%	22.63%	6.32%	1.17%	16.72%	3415
2010	11074.57	568.99	68.45	5.14%	0.62%	0.5478	0.4067	0.6271	9.20%	1.48%	22.02%	6.27%	1.15%	16.29%	3218
2011	10873.82	554.97	68.02	5.10%	0.63%	0.5482	0.4048	0.6268	9.08%	1.44%	22.22%	6.18%	1.11%	16.48%	3154
2012	11132.45	567.74	69.17	5.10%	0.62%	0.5431	0.4007	0.6236	9.02%	1.41%	22.27%	6.13%	1.09%	16.49%	3077
2013	11250.41	583.67	71.42	5.19%	0.63%	0.5482	0.4046	0.6268	9.16%	1.41%	22.29%	6.21%	1.08%	16.46%	3016
2014	11324.67	590.23	71.84	5.21%	0.63%	0.5444	0.4065	0.6251	9.23%	1.41%	22.17%	6.27%	1.09%	16.29%	3041
2015	11572.02	608.16	72.77	5.26%	0.63%	0.5460	0.4071	0.6232	9.27%	1.38%	21.56%	6.30%	1.06%	15.67%	3054
2016	11910.10	625.14	73.75	5.25%	0.62%	0.5474	0.4106	0.6280	9.27%	1.38%	21.38%	6.27%	1.06%	15.54%	3019
2017	11894.89	619.77	74.55	5.21%	0.63%	0.5460	0.4067	0.6280	9.17%	1.38%	22.00%	6.22%	1.06%	16.10%	2954
2018	12118.21	630.95	76.06	5.21%	0.63%	0.5436	0.4068	0.6293	9.20%	1.42%	22.25%	6.24%	1.09%	16.42%	2920
2019	11709.58	605.61	73.94	5.17%	0.63%	0.5407	0.4022	0.6203	9.14%	1.40%	22.15%	6.21%	1.08%	16.25%	2942
2020	11018.20	574.95	70.92	5.22%	0.64%	0.5317	0.4037	0.6234	9.17%	1.44%	22.57%	6.20%	1.10%	16.62%	2935
2021	10891.37	583.86	71.40	5.36%	0.66%	0.5173	0.3936	0.6191	9.40%	1.42%	21.98%	6.34%	1.09%	16.00%	3097
2022	10498.76	566.88	69.60	5.40%	0.66%	0.5224	0.4004	0.6230	9.44%	1.38%	21.47%	6.35%	1.06%	15.31%	3236
Total	8972.10	479.81	64.59	5.35%	0.72%	0.5572	0.4257	0.6520	9.60%	2.16%	28.40%	6.61%	1.65%	22.74%	94996

Table 2. Descriptive Statistics

This table compiles comprehensive descriptive statistics from 1994 to 2022. Panel A outlines various return metrics, such as daily stock (R_i), risk-free (R_f), and market returns (R_m), along with buy-and-hold returns spanning from one week to one year ($BHR_{1w} \sim BHR_{1y}$) and market-adjusted cumulative returns for firms across time windows from [0,1] to [0,5] ($CAR[0,1] \sim CAR[0,5]$). Additionally, Panel A covers control variables like turnover rate (*TURNOVER*), market value (*lnME*), book-to-market ratio (*BM*), return on assets (*ROA*), volatility (*VOLATILITY*), leverage ratio (*LEVERAGE*), Tobin's Q (*TOBIN-Q*), word count of financial texts (*TEXT_COUNT*), NASDAQ listing indicator (*IsNasdaq*), sales growth rates ($\triangle SALE$), investment growth rates($\triangle INVESTMENT$), gross margin (*GROSSMARGIN*), and cumulative returns over 1 month ($CRET_{1m}$) and 2 to 12 months ($CRET_{2m\to12m}$). Panel B explores the descriptive statistics of sentiment indicators derived from financial texts. It encompasses generated tones of summarized texts conducted by LLMs, including various LLaMA-2 model variants with regards to training epochs (*LLAMA2_TONE_E1 ~ LLAMA2_TONE_E5*) and the FinBERT mode l(*FINBERT_TONE*). The panel also examines sentiment scores and tones from both full and summarized texts, as identified by the LM dictionary (*LM_TEXT_TAG*) and traditional machine learning models such as logistic regression (*LR_TEXT_TONE*), support vector machine (*SVM_TEXT_TONE*), random forest (*RF_TEXT_TONE*), and neural network (*NN_TEXT_TONE*). Note that *TEXT* has 2 types: summarized texts (*SUM*) and full texts (*FULL*); and *TAG* has two types: *SCORE* and *TONE*. Panel C examines the correlations among the different sentiment scores and tones. Definitions for all variables are provided in Appendix F for reference. The data spans from 1994 to 2022.

Panel A. Descriptive Statistics of Returns and Control Variables

	Mean	SD	P25	P50	P75	Skew	Kurt
RETURNS							
R_i	-0.000079	0.059857	-0.018386	0.000000	0.016575	6.425275	269.7871
R_f	0.000082	0.000081	0.000000	0.000060	0.000180	0.502461	1.685292
R_m	-0.000670	0.014318	-0.006400	0.000040	0.006060	-0.927011	15.32288
BHR_{1w}	0.001441	0.098495	-0.036036	0.000000	0.033900	6.631193	335.8007
BHR_{2w}	0.003457	0.135765	-0.048076	0.000000	0.050000	3.944050	120.0385
BHR_{1m}	0.019081	0.199117	-0.063153	0.008475	0.082564	5.320485	178.0489
BHR_{3m}	0.054492	0.373924	-0.116684	0.015741	0.161791	5.901946	118.9059
BHR_{6m}	0.081198	0.582727	-0.181475	0.014925	0.222222	8.762281	221.8851
BHR_{1y}	0.181801	1.089018	-0.247656	0.044078	0.363683	21.56043	1267.626
$CAR[0,1]$	-0.000403	0.079603	-0.024323	-0.000867	0.022125	4.930795	165.9012
$CAR[0,2]$	-0.000414	0.090282	-0.029227	-0.001066	0.026866	3.293812	91.20033
$CAR[0,3]$	-0.000507	0.099227	-0.033243	-0.001232	0.030270	3.623802	110.3710
$CAR[0,4]$	0.000228	0.106049	-0.036712	-0.001070	0.034350	3.011503	80.15916
$CAR[0,5]$	0.000693	0.112610	-0.039587	-0.000799	0.037654	2.797406	69.15068
CONTROLS							
TURNOVER	0.012118	0.043298	0.002180	0.005681	0.012251	43.55164	2884.638
lnME	12.90359	2.157159	11.34735	12.85160	14.35496	0.179287	2.794696
BM	0.000669	0.001945	0.000175	0.000402	0.000781	14.40300	860.6180
ROA	-0.075463	1.229822	-0.055154	0.027650	0.071995	-80.85300	8217.202
VOLATILITY	0.037657	0.025771	0.021802	0.031579	0.046358	6.372288	153.5546
LEVERAGE	2.019734	311.6210	0.004863	0.284307	0.871239	231.7540	56307.18
TOBIN-Q	0.998556	0.842549	1.000000	1.000046	1.024554	-153.6260	29250.42
TEXT_COUNT	9283.654	6999.398	4820.000	8048.000	11780.00	3.742705	50.33055
IsNasdaq	0.608965	0.487986	0.000000	1.000000	1.000000	-0.446594	1.199446
$\triangle SALE$	1.163712	0.423575	0.982085	1.082287	1.229481	3.513682	22.31521
$\triangle INVESTMENT$	1.363655	1.192562	0.757807	1.077935	1.525246	3.544897	20.42153
GROSSMARGIN	0.337557	0.260330	0.179478	0.312602	0.479525	0.166074	4.194095
$CRET_{1m}$	0.012871	0.137518	-0.062765	0.008037	0.079790	0.445602	4.657292
$CRET_{2m\to12m}$	0.133669	0.522285	-0.176991	0.061125	0.332518	1.556979	7.443157

Table 2. (Continue.)

Panel B. Descriptive Statistics of Sentiment Tones and Scores

	Mean	SD	P25	P50	P75	Positive	Neutral	Negative
Summarized Texts								
LLMs								
LLAMA2_TONE_E1	0.606799	0.732005	1	1	1	47,440	6,027	9,333
LLAMA2_TONE_E2	0.539554	0.641681	0	1	1	39,012	18,660	5,128
LLAMA2_TONE_E3	0.557293	0.676356	0	1	1	41,615	14,568	6,617
LLAMA2_TONE_E4	0.510048	0.626512	0	1	1	36,509	21,813	4,478
LLAMA2_TONE_E5	0.538551	0.627997	0	1	1	38,401	19,819	4,580
FINBERT_TONE	0.302892	0.764826	0	0	1	30,760	20,303	11,737
TRADITIONAL								
LM_SUM_SCORE	0.001363	0.049862	-0.000000	0.000000	0.012658	16,055	29,292	17,453
LM_SUM_TONE	-0.022261	0.730123	-1	0	1	16,055	29,292	17,453
LR_SUM_TONE	0.205573	0.612456	0	0	1	19,560	36,590	6,650
SVM_SUM_TONE	0.149252	0.538522	0	0	0	14,492	43,189	5,119
RF_SUM_TONE	0.092787	0.613940	0	0	0	15,019	38,589	9,192
NN_SUM_TONE	0.064761	0.597846	0	0	0	13,888	40,091	9,321
Full Texts								
LM_FULL_SCORE	-0.007846	0.007538	-0.000000	-0.007198	-0.002903	7,432	769	54,599
LM_FULL_TONE	-0.751067	0.650892	-1	-1	-1	7,432	769	54,599
LR_FULL_TONE	0.153217	0.914015	-1	1	1	31,780	8,862	22,158
SVM_FULL_TONE	0.124268	0.866586	-1	0	1	27,967	14,670	20,163
RF_FULL_TONE	-0.428137	0.663350	-1	-1	0	6,129	23,655	33,016
NN_FULL_TONE	0.030350	0.401714	0	0	0	6,049	52,608	4,143

Panel C. Correlations

	A.	*B.*	*C.*	*D.*	*E.*	*F.*	*G.*	*H.*	*I.*	*J.*	*K.*	*L.*	*M.*	*N.*	*O.*	*P.*
A. LLAMA2_TONE_E1	1.00	0.80	0.85	0.74	0.77	0.54	0.35	0.24	0.23	0.18	0.20	0.20	0.15	0.16	0.08	0.07
B. LLAMA2_TONE_E2	0.80	1.00	0.92	0.93	0.95	0.61	0.39	0.28	0.26	0.24	0.23	0.20	0.16	0.17	0.08	0.08
C. LLAMA2_TONE_E3	0.85	0.92	1.00	0.87	0.9	0.59	0.38	0.27	0.26	0.23	0.22	0.20	0.16	0.17	0.08	0.07
D. LLAMA2_TONE_E4	0.74	0.93	0.87	1.00	0.96	0.61	0.39	0.29	0.27	0.25	0.23	0.21	0.16	0.17	0.09	0.08
E. LLAMA2_TONE_E5	0.77	0.95	0.9	0.96	1.00	0.61	0.39	0.28	0.27	0.24	0.22	0.20	0.16	0.17	0.09	0.08
F. FINBERT_TONE	0.54	0.61	0.59	0.61	0.61	1.00	0.44	0.32	0.31	0.28	0.25	0.20	0.15	0.17	0.08	0.08
G. LM_SUM_SCORE	0.35	0.39	0.38	0.39	0.39	0.44	1.00	0.27	0.26	0.27	0.27	0.20	0.11	0.12	0.07	0.06
H. LR_SUM_TONE	0.24	0.28	0.27	0.29	0.28	0.32	0.27	1.00	0.78	0.43	0.42	0.15	0.15	0.15	0.07	0.07
I. SVM_SUM_TONE	0.23	0.26	0.26	0.27	0.27	0.31	0.26	0.78	1.00	0.47	0.39	0.14	0.14	0.15	0.08	0.07
J. RF_SUM_TONE	0.18	0.24	0.23	0.25	0.24	0.28	0.27	0.43	0.47	1.00	0.23	0.15	0.13	0.14	0.11	0.05
K. NN_SUM_TONE	0.20	0.23	0.22	0.23	0.22	0.25	0.27	0.42	0.39	0.23	1.00	0.11	0.09	0.09	0.04	0.06
L. LM_FULL_SCORE	0.20	0.20	0.20	0.21	0.20	0.20	0.20	0.15	0.14	0.15	0.11	1.00	0.39	0.37	0.28	0.15
M. LR_FULL_TONE	0.15	0.16	0.16	0.16	0.16	0.15	0.11	0.15	0.14	0.13	0.09	0.39	1.00	0.82	0.26	0.24
N. SVM_FULL_TONE	0.16	0.17	0.17	0.17	0.17	0.17	0.12	0.15	0.15	0.14	0.09	0.37	0.82	1.00	0.27	0.23
O. RF_FULL_TONE	0.08	0.08	0.08	0.09	0.09	0.08	0.07	0.07	0.08	0.11	0.04	0.28	0.26	0.27	1.00	0.11
P. NN_FULL_TONE	0.07	0.08	0.07	0.08	0.08	0.08	0.06	0.07	0.07	0.05	0.06	0.15	0.24	0.23	0.11	1.00

Table 3. Average Buy-and-Hold Return by Sentiment Tone

This table illustrates the average buy-and-hold returns based on the sentiment tone of investor signals. Specifically, it shows returns when an investor buys a stock at the last price of day $t+1$ following a positive sentiment at time t, or shorts the stock based on a negative sentiment at time t at $t+1$. Panel A details the results for various durations, including 1 week (BHR_{1w}), 2 weeks (BHR_{2w}), 1 month (BHR_{1m}), 3 months (BHR_{3m}), 6 months (BHR_{6m}), and 1 year (BHR_{1y}), alongside the number of positive (Pos), negative (Neg), and total (Total) samples. The buy-sell sentiment indicators comprise sentiment tones derived from summarized texts. These include tones generated by LLMs, featuring five LLaMA-2 model variants and the FinBERT model, alongside traditional methodologies such as the LM dictionary (LM), logistic regression (LR), support vector machine (SVM), random forest (RF), and neural network (NN) models. Panel B compares the buy-and-hold returns between full texts and summarized texts in traditional methods, as predicted by the same model. Given the variance in sample sizes, a different-sample t-test is used to assess the significance of these differences. The data covers the period from 1994 to 2022.
*Significance at the 10% level; **significance at the 5% level; ***significance at the 1% level.

Panel A. Average Buy-and-hold Returns of Different Sentiment Indicators

VARIABLES	BHR_{1w}	BHR_{2w}	BHR_{1m}	BHR_{3m}	BHR_{6m}	BHR_{1y}	Pos	Neg	Total
Summarized Texts									
LLMs									
LLAMA2_TONE_E5	0.11%	0.23%	1.39%	3.91%	5.83%	12.37%	37,959	4,499	42,435
LLAMA2_TONE_E4	0.11%	0.21%	1.33%	3.81%	5.66%	12.01%	36,083	4,398	40,460
LLAMA2_TONE_E3	0.06%	0.15%	1.19%	3.57%	5.27%	11.32%	41,143	6,504	47,619
LLAMA2_TONE_E2	0.13%	0.21%	1.30%	3.76%	5.61%	11.82%	38,608	5,048	43,630
LLAMA2_TONE_E1	0.07%	0.18%	1.13%	3.30%	4.91%	10.60%	46,915	9,182	56,063
FINBERT_TONE	0.10%	0.16%	0.74%	2.17%	3.30%	5.98%	30,413	11,563	41,950
TRADITIONAL									
LM_SUM_TONE	-0.01%	0.00%	-0.11%	-0.55%	-0.92%	-2.55%	15,879	17,266	33,124
LR_SUM_TONE	0.10%	0.16%	0.85%	2.33%	3.02%	7.14%	19,315	6,582	25,880
SVM_SUM_TONE	0.13%	0.22%	0.90%	2.23%	2.83%	6.82%	14,302	5,068	19,357
RF_SUM_TONE	0.10%	0.19%	0.40%	0.68%	0.60%	2.36%	14,809	9,106	23,898
NN_SUM_TONE	-0.01%	0.01%	0.25%	0.73%	0.74%	2.19%	13,216	9,191	22,391
Full Texts									
TRADITIONAL									
LM_FULL_TONE	-0.16%	-0.37%	-1.70%	-4.61%	-6.75%	-15.01%	7,318	53,992	61,270
LR_FULL_TONE	-0.19%	-0.39%	-0.46%	-0.65%	-1.11%	-1.27%	31,417	21,963	53,346
SVM_FULL_TONE	-0.16%	-0.35%	-0.44%	-0.86%	-1.48%	-2.71%	27,673	19,357	47,634
RF_FULL_TONE	-0.09%	-0.28%	-1.46%	-4.16%	-6.65%	-14.34%	6,044	32,734	38,750
NN_FULL_TONE	-0.11%	-0.22%	-0.55%	-0.98%	-1.75%	-2.04%	5,976	4,070	10,039

Panel B. Differences between Summarized Text and Full Texts in Traditional Methods

VARIABLES	BHR_{1w}	BHR_{2w}	BHR_{1m}	BHR_{3m}	BHR_{6m}	BHR_{1y}
Diff_LM	0.17%**	0.37%***	1.59%***	4.06%***	5.83%***	12.46%***
	(2.5617)	(4.0034)	(11.4341)	(15.2561)	(14.1128)	(17.0891)
Diff_LR	0.29%***	0.55%**	1.31%***	2.98%***	4.13%***	8.42%***
	(4.2049)	(5.6816)	(9.1746)	(10.4786)	(9.0277)	(9.8764)
Diff_SVM	0.29%***	0.57%***	1.34%***	3.09%***	4.31%***	9.53%***
	(3.8426)	(5.2792)	(8.5320)	(10.0418)	(8.6982)	(9.7533)
Diff_RF	0.19%**	0.47%***	1.86%***	4.84%***	7.25%***	16.70%***
	(2.5612)	(4.5330)	(12.1919)	(16.3838)	(15.5103)	(21.1557)
Diff_NN	0.10%	0.23%	0.79%***	1.71%***	2.49%***	4.23%***
	(0.9805)	(1.6734)	(3.7720)	(4.1565)	(3.8030)	(3.5363)

Table 4. LLM Analysis of Summarized Text Sentiments and Cumulative Abnormal Returns (CARs)
This table showcases the efficacy of LLMs in interpreting the sentiment of summarized texts and their correlation with market-adjusted CARs from $CAR[0,1]$ to $CAR[0,5]$. Panel A presents the regression outcomes using the sentiment tone of well-trained LLaMA-2 model (*LLAMA2_TONE_E5*) over five epochs, demonstrating its prowess in sentiment analysis. Panel B extends the analysis to include other four variants of the LLaMA 2 model, trained across 1 to 4 epochs (*LLAMA2_TONE_E1 ~ LLAMA2_TONE_E4*), and incorporates comparisons with the FinBERT model (*FINBERT_TONE*). The regression tests incorporate control variables such as market value (*lnME*), book-to-market ratio (*BM*), and turnover rate(*TURNOVER*), alongside NASDAQ listing status (*IsNasdaq*), textual word count (*TEXT_COUNT*), leverage ratio (*LEVERAGE*), return on assets (*ROA*), volatility (*VOLATILITY*), and Tobin's Q (*TOBIN-Q*). The models are adjusted for year and industry effects, encapsulating data from 1994 to 2022. The regression function can be represented as: $CAR[0,T]_{i,t} = \alpha + \beta \times LLMTONE_{i,t} + \gamma' Controls_{i,t} + \varepsilon_{i,t}$.
*Significance at the 10% level; **significance at the 5% level; ***significance at the 1% level.

Panel A. A well-trained LLaMA 2 model results

VARIABLES	$CAR[0,1]$	$CAR[0,2]$	$CAR[0,3]$	$CAR[0,4]$	$CAR[0,5]$
LLAMA2_TONE_E5	0.0010**	0.0014**	0.0015**	0.0014**	0.0020***
	(1.9916)	(2.3645)	(2.3432)	(2.0661)	(2.6697)
lnME	0.0005**	0.0004	0.0003	0.0003	0.0002
	(2.4159)	(1.6039)	(1.2706)	(1.1476)	(0.7782)
BM	-0.5045***	0.0687	0.6018***	1.6424***	2.5582***
	(-2.9426)	(0.3533)	(2.8096)	(7.1734)	(10.5216)
TURNOVER	0.1148***	0.0982***	0.0897***	0.0866***	0.0883***
	(15.3270)	(11.5540)	(9.6129)	(8.6849)	(8.3385)
IsNasdaq	-0.0006	-0.0012	-0.0003	-0.0005	-0.0011
	(-0.6929)	(-1.2650)	(-0.2525)	(-0.4945)	(-0.9234)
TEXT_COUNT	-0.0000**	-0.0000**	-0.0000*	-0.0000*	-0.0000*
	(-2.4734)	(-2.4876)	(-1.8447)	(-1.6696)	(-1.8266)
LEVERAGE	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.5407)	(0.3369)	(0.4655)	(0.7333)	(0.7180)
ROA	-0.0000	-0.0002	-0.0003	-0.0001	-0.0006
	(-0.0839)	(-0.7763)	(-0.8936)	(-0.3959)	(-1.5130)
VOLATILITY	0.0045	0.0382**	0.0704***	0.0826***	0.0810***
	(0.2781)	(2.0715)	(3.4792)	(3.8132)	(3.5223)
TOBIN-Q	0.0005	0.0007	0.0005	0.0003	-0.0000
	(1.2049)	(1.5206)	(1.0581)	(0.5296)	(-0.0393)
INTERCEPT	-0.0148*	-0.0153	-0.0189*	-0.0261**	-0.0210*
	(-1.7802)	(-1.6187)	(-1.8163)	(-2.3568)	(-1.7772)
Adjusted R-squared	0.00764	0.00758	0.00998	0.00972	0.00965
Observations	62,755	62,727	62,709	62,695	62,675
Year and Industry FE	YES	YES	YES	YES	YES

Panel B. Results of FinBert Model and Other LLaMA 2 Variants

VARIABLES	$CAR[0,1]$	$CAR[0,2]$	$CAR[0,3]$	$CAR[0,4]$	$CAR[0,5]$
LLAMA2_TONE_E4	0.0010*	0.0013**	0.0014**	0.0013*	0.0017**
	(1.8339)	(2.1855)	(2.2192)	(1.8596)	(2.3483)
Adjusted R-squared	0.00763	0.00757	0.00997	0.00971	0.00963
LLAMA2_TONE_E3	0.0011**	0.0013**	0.0014**	0.0014**	0.0017**
	(2.3592)	(2.4680)	(2.2872)	(2.2131)	(2.4468)
Adjusted R-squared	0.00766	0.00759	0.00997	0.00973	0.00964
LLAMA2_TONE_E2	0.0010*	0.0013**	0.0014**	0.0015**	0.0020***
	(1.9545)	(2.2359)	(2.2296)	(2.1715)	(2.7360)
Adjusted R-squared	0.00764	0.00757	0.00997	0.00973	0.00966
LLAMA2_TONE_E1	0.0011**	0.0013***	0.0013**	0.0014**	0.0017***
	(2.4148)	(2.5922)	(2.4279)	(2.2981)	(2.7401)
Adjusted R-squared	0.00767	0.00760	0.00998	0.00974	0.00966
Observations	62,755	62,727	62,709	62,695	62,675
FINBERT_TONE	0.0008**	0.0012**	0.0013**	0.0014**	0.0016***
	(1.9813)	(2.5371)	(2.5258)	(2.5485)	(2.7263)
Adjusted R-squared	0.00764	0.00759	0.00999	0.00976	0.00966
Observations	62,753	62,725	62,707	62,693	62,673
Year and Industry FE	YES	YES	YES	YES	YES

Table 5. Sentiment Analysis on CARs for Full and Summarized Texts Using Traditional Methods

This table features regression analyses that investigate the relationship between market-adjusted CARs, ranging from $CAR[0,1]$ to $CAR[0,5]$, and the sentiment derived from both full (Panel A) and summarized texts (Panel B) using traditional methodologies. These texts are evaluated with the LM dictionary (*LM_TEXT_SCORE*) and through sentiment predictions made by four machine learning models: logistic regression (*LR_TEXT_TONE*), support vector machine (*LR_TEXT_TONE*), random forest (*LR_TEXT_TONE*), and neural network (*LR_TEXT_TONE*), where *TEXT* has two types: full texts (*FULL*) and summarized texts (*SUM*). The regression tests incorporate control variables such as market value, book-to-market ratio, and turnover rate, alongside NASDAQ listing status, textual word count, leverage ratio, return on assets, volatility, and Tobin's Q. The models are adjusted for year and industry effects, encapsulating data from 1994 to 2022. The regression function can be represented as: $CAR[0,T]_{i,t} = \alpha + \beta \times SENTIMENT_{i,t} + \gamma' Controls_{i,t} + \varepsilon_{i,t}$.

*Significance at the 10% level; **significance at the 5% level; ***significance at the 1% level.

Panel A. Full Texts

VARIABLES	$\boldsymbol{CAR[0,1]}$	$\boldsymbol{CAR[0,2]}$	$\boldsymbol{CAR[0,3]}$	$\boldsymbol{CAR[0,4]}$	$\boldsymbol{CAR[0,5]}$
LM_FULL_SCORE	0.0374	0.1279**	0.1306**	0.1665***	0.1979***
	(0.7797)	(2.3530)	(2.1873)	(2.6090)	(2.9208)
Adjusted R-squared	0.00758	0.00758	0.00996	0.00976	0.00968
LR_FULL_TONE	0.0001	0.0002	0.0001	-0.0001	-0.0001
	(0.2605)	(0.4000)	(0.1280)	(-0.1178)	(-0.2037)
Adjusted R-squared	0.00758	0.00749	0.00989	0.00965	0.00954
SVM_FULL_TONE	0.0001	0.0005	0.0006	0.0003	0.0001
	(0.2949)	(0.9809)	(1.1320)	(0.4852)	(0.1953)
Adjusted R-squared	0.00758	0.00751	0.00991	0.00965	0.00954
RF_FULL_TONE	0.0004	0.0011*	0.0012*	0.0014**	0.0014*
	(0.7753)	(1.8625)	(1.8788)	(2.0564)	(1.8782)
Adjusted R-squared	0.00758	0.00755	0.00994	0.00972	0.00960
NN_FULL_TONE	-0.0006	-0.0006	-0.0001	-0.0004	0.0002
	(-0.6898)	(-0.6247)	(-0.0616)	(-0.3584)	(0.1431)
Adjusted R-squared	0.00758	0.00750	0.00989	0.00965	0.00954
Observations	62,843	62,814	62,796	62,781	62,762
Year and Industry FE	YES	YES	YES	YES	YES

Panel B. Summarized Texts

VARIABLES	$\boldsymbol{CAR[0,1]}$	$\boldsymbol{CAR[0,2]}$	$\boldsymbol{CAR[0,3]}$	$\boldsymbol{CAR[0,4]}$	$\boldsymbol{CAR[0,5]}$
LM_SUM_SCORE	0.0085	0.0092	0.0121	0.0109	0.0128
	(1.3124)	(1.2483)	(1.5073)	(1.2706)	(1.3977)
Adjusted R-squared	0.00760	0.00751	0.00992	0.00968	0.00957
LR_SUM_TONE	-0.0000	0.0002	0.0001	0.0002	0.0006
	(-0.0010)	(0.2993)	(0.2280)	(0.2826)	(0.7546)
Adjusted R-squared	0.00757	0.00749	0.00989	0.00965	0.00955
SVM_SUM_TONE	0.0001	0.0001	0.0005	0.0006	0.0009
	(0.1328)	(0.1863)	(0.6549)	(0.7455)	(1.0778)
Adjusted R-squared	0.00757	0.00749	0.00989	0.00966	0.00956
RF_SUM_TONE	0.0003	0.0005	0.0008	0.0008	0.0006
	(0.5119)	(0.8723)	(1.1330)	(1.0631)	(0.8216)
Adjusted R-squared	0.00758	0.00750	0.00991	0.00967	0.00955
NN_SUM_TONE	-0.0001	0.0004	-0.0003	-0.0007	-0.0005
	(-0.1331)	(0.6707)	(-0.4120)	(-0.9568)	(-0.6436)
Adjusted R-squared	0.00757	0.00750	0.00989	0.00967	0.00955
Observations	62,843	62,814	62,796	62,781	62,762
Year and Industry FE	YES	YES	YES	YES	YES

Table 6. Average Buy-and-hold Return by Sentiment Tone During Financial Turbulence

This table illustrates the average buy-and-hold returns based on the sentiment tone of investor signals. Specifically, it shows returns when an investor buys a stock at the last price of day $t+1$ following a positive sentiment at time t, or shorts the stock based on a negative sentiment at time t at $t+1$. Panel A details the results for various durations, including 1 week (BHR_{1w}), 2 weeks (BHR_{2w}), 1 month (BHR_{1m}), 3 months (BHR_{3m}), 6 months (BHR_{6m}), and 1 year (BHR_{1y}), alongside the number of positive (Pos), negative (Neg), and total (Total) samples. The buy-sell sentiment indicators comprise sentiment tones derived from summarized texts. These include tones generated by LLMs, featuring five LLaMA-2 model variants and the FinBERT model, alongside traditional methodologies such as the LM dictionary (LM), logistic regression (LR), support vector machine (SVM), random forest (RF), and neural network (NN) models. Panel B compares the buy-and-hold returns between full texts and summarized texts in traditional methods, as predicted by the same model. Given the variance in sample sizes, a different-sample t-test is used to assess the significance of these differences.. The data covers the period from 2000 to 2009, which encompasses two significant financial crises.

*Significance at the 10% level; **significance at the 5% level; ***significance at the 1% level.

Panel A. Average Buy-and-hold Returns of Different Sentiment Indicators

VARIABLES	BHR_{1w}	BHR_{2w}	BHR_{1m}	BHR_{3m}	BHR_{6m}	BHR_{1y}	Pos	Neg	Total
Summarized Texts									
LLMs									
LLAMA2_TONE_E5	0.07%	0.06%	0.35%	2.07%	4.75%	13.35%	17,980	1,777	19,748
LLAMA2_TONE_E4	0.09%	0.07%	0.32%	2.04%	4.73%	13.12%	17,017	1,729	18,738
LLAMA2_TONE_E3	0.03%	0.00%	0.24%	1.89%	4.38%	12.29%	19,630	2,638	22,257
LLAMA2_TONE_E2	0.06%	0.03%	0.29%	2.02%	4.64%	12.90%	18,362	2,009	20,360
LLAMA2_TONE_E1	0.04%	0.03%	0.22%	1.72%	4.15%	11.43%	22,460	3,842	26,290
FINBERT_TONE	0.13%	0.15%	0.33%	1.36%	2.97%	7.68%	14,234	4,872	19,095
TRADITIONAL									
LM_SUM_TONE	0.00%	0.01%	0.18%	-0.17%	-0.22%	-1.01%	7,595	8,431	16,017
LR_SUM_TONE	0.15%	0.18%	0.37%	1.50%	2.97%	7.86%	8,555	3,145	11,695
SVM_SUM_TONE	0.17%	0.25%	0.43%	1.42%	2.81%	7.61%	6,217	2,483	8,695
RF_SUM_TONE	0.05%	0.19%	0.31%	0.23%	0.74%	2.56%	6,458	4,711	11,164
NN_SUM_TONE	-0.02%	0.10%	0.14%	0.77%	1.22%	3.80%	5,629	3,791	9,413
Full Texts									
TRADITIONAL									
LM_FULL_TONE	-0.21%	-0.28%	-0.68%	-2.44%	-4.58%	-13.00%	3,025	25,705	28,714
LR_FULL_TONE	-0.23%	-0.36%	-0.51%	0.37%	2.26%	5.17%	15,412	10,117	25,515
SVM_FULL_TONE	-0.24%	-0.37%	-0.47%	0.27%	0.20%	0.44%	22,928	8,700	28,903
RF_FULL_TONE	-0.19%	-0.22%	-0.52%	-1.63%	-3.79%	-10.15%	2,346	16,900	19,234
NN_FULL_TONE	-0.26%	-0.44%	-0.75%	-0.01%	1.55%	3.89%	2,921	1,026	3,944

Panel B. Differences between Summarized Text and Full Texts in Traditional Methods

VARIABLES	BHR_{1w}	BHR_{2w}	BHR_{1m}	BHR_{3m}	BHR_{6m}	BHR_{1y}
Diff_LM	0.21%***	0.35%***	0.87%***	2.26%***	4.36%***	11.99%***
	(2.6021)	(3.1428)	(4.8853)	(7.3378)	(9.2289)	(11.6672)
Diff_LR	0.38%***	0.54%**	0.87%***	1.13%***	0.71%	2.69%**
	(4.3277)	(4.3535)	(5.0909)	(3.4593)	(1.3154)	(1.9968)
Diff_SVM	0.41%***	0.61%***	0.90%***	0.55%***	0.85%	3.17%**
	(4.3307)	(4.5165)	(4.7329)	(3.3994)	(1.5624)	(2.0091)
Diff_RF	0.24%***	0.40%***	0.84%***	1.86%***	4.53%***	12.71%***
	(2.6482)	(3.1386)	(4.6659)	(5.7385)	(9.0913)	(12.6976)
Diff_NN	0.24%*	0.53%***	0.89%***	0.85%***	-0.33%	-0.09%
	(1.7551)	(2.9845)	(3.4179)	(1.8416)	(-0.4834)	(-0.0501)

Table 7. Sentiment Analysis on CARs During Financial Turbulence

This table presents test combining insights from previous analyses, examining the impact of sentiment indicators on market-adjusted CARs from $CAR[0,1]$ to $CAR[0,5]$ during the volatile period of 2000 to 2009, which encompasses two significant financial crises. This table brings together sentiment analysis results from both full and summarized texts, applying an array of methodologies that include LLMs and traditional techniques. The analysis encompasses sentiments scored using the LM dictionary as well as predictions made by logistic regression, support vector machine, random forest, and neural network models. The regression tests incorporate control variables such as market value, book-to-market ratio, and turnover rate, alongside NASDAQ listing status, textual word count, leverage ratio, return on assets, volatility, and Tobin's Q. The models are adjusted for year and industry effects. The regression function can be represented as: $CAR[0,T]_{i,t} = \alpha + \beta \times SENTIMENT\_TURB_{i,t} + \gamma' Controls_{i,t} + \varepsilon_{i,t}$, where $SENTIMENT\_TURB$ denotes the sentiment indicators generated by LLMs or traditional methodologies by either summarized texts or full texts in financial turbulence time.

*Significance at the 10% level; **significance at the 5% level; ***significance at the 1% level.

VARIABLES	$CAR[0,1]$	$CAR[0,2]$	$CAR[0,3]$	$CAR[0,4]$	$CAR[0,5]$
LLAMA2_TONE_E5	0.0016*	0.0028***	0.0034***	0.0030**	0.0042***
	(1.9499)	(2.9024)	(3.1251)	(2.5279)	(3.2796)
Adjusted R-squared	0.00642	0.0125	0.0159	0.0171	0.0182
LLAMA2_TONE_E4	0.0014	0.0021**	0.0027**	0.0021*	0.0032**
	(1.6023)	(2.1990)	(2.4028)	(1.8043)	(2.4822)
Adjusted R-squared	0.00637	0.0123	0.0157	0.0170	0.0180
LLAMA2_TONE_E3	0.0017**	0.0023***	0.0027***	0.0026**	0.0034***
	(2.1864)	(2.6071)	(2.6482)	(2.3560)	(2.8902)
Adjusted R-squared	0.00647	0.0124	0.0158	0.0171	0.0181
LLAMA2_TONE_E2	0.0017**	0.0026***	0.0033***	0.0031***	0.0041***
	(2.0403)	(2.7346)	(3.0561)	(2.6598)	(3.3165)
Adjusted R-squared	0.00644	0.0125	0.0159	0.0172	0.0182
LLAMA2_TONE_E1	0.0016**	0.0022***	0.0026***	0.0027***	0.0036***
	(2.2597)	(2.7433)	(2.8452)	(2.7266)	(3.4055)
Adjusted R-squared	0.00648	0.0125	0.0158	0.0172	0.0182
FINBERT_TONE	0.0006	0.0013	0.0015*	0.0020**	0.0024**
	(0.8860)	(1.5902)	(1.7222)	(2.0613)	(2.3117)
Adjusted R-squared	0.00629	0.0122	0.0156	0.0170	0.0180
LM_SUM_SCORE	0.0130	0.0134	0.0238*	0.0278*	0.0279*
	(1.2524)	(1.1287)	(1.7553)	(1.9076)	(1.7864)
Adjusted R-squared	0.00633	0.0122	0.0156	0.0170	0.0179
LR_SUM_TONE	-0.0005	-0.0002	-0.0002	-0.0001	0.0003
	(-0.6333)	(-0.1962)	(-0.1888)	(-0.0576)	(0.2552)
Adjusted R-squared	0.00628	0.0121	0.0155	0.0169	0.0178
SVM_SUM_TONE	-0.0005	-0.0003	0.0001	0.0004	0.0009
	(-0.5320)	(-0.2942)	(0.0757)	(0.2583)	(0.5865)
Adjusted R-squared	0.00627	0.0121	0.0155	0.0169	0.0178
RF_SUM_TONE	0.0005	0.0003	0.0006	0.0009	0.0005
	(0.5251)	(0.3218)	(0.4809)	(0.7096)	(0.4104)
Adjusted R-squared	0.00627	0.0121	0.0155	0.0169	0.0178
NN_SUM_TONE	-0.0008	-0.0000	-0.0007	-0.0017	-0.0014
	(-0.8950)	(-0.0460)	(-0.5814)	(-1.4133)	(-1.0776)
Adjusted R-squared	0.00630	0.0121	0.0155	0.0169	0.0178
LM_FULL_SCORE	0.0472	0.1691**	0.1893*	0.2514**	0.3747***
	(0.6282)	(1.9668)	(1.9296)	(2.3879)	(3.3167)
Adjusted R-squared	0.00628	0.0123	0.0157	0.0171	0.0182
LR_FULL_TONE	0.0001	0.0004	0.0010	0.0009	0.0008
	(0.1065)	(0.5687)	(1.2181)	(0.9714)	(0.8808)
Adjusted R-squared	0.00626	0.0122	0.0156	0.0169	0.0178
SVM_FULL_TONE	0.0003	0.0010	0.0019**	0.0018*	0.0019*
	(0.4129)	(1.2472)	(2.2024)	(1.9311)	(1.8935)
Adjusted R-squared	0.00627	0.0122	0.0157	0.0170	0.0179
RF_FULL_TONE	0.0002	0.0010	0.0011	0.0016	0.0018
	(0.2218)	(1.0243)	(1.0397)	(1.4117)	(1.4528)
Adjusted R-squared	0.00626	0.0122	0.0155	0.0169	0.0178
NN_FULL_TONE	0.0005	0.0011	0.0026	0.0022	0.0030
	(0.4246)	(0.7969)	(1.5914)	(1.2471)	(1.6076)
Adjusted R-squared	0.00627	0.0122	0.0156	0.0169	0.0179
Observations	23,903	23,892	23,885	23,882	23,876
Year and Industry FE	YES	YES	YES	YES	YES

References

Azimi, M., & Agrawal, A. (2021). Is positive sentiment in corporate annual reports informative? Evidence from deep learning. *The Review of Asset Pricing Studies*, *11*(4), 762–805.

Bingler, J., Kraus, M., Leippold, M., & Webersinke, N. (2023). *How Cheap Talk in Climate Disclosures relates to Climate Initiatives, Corporate Emissions, and Reputation Risk* (SSRN Scholarly Paper 4000708).

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2023). Textual analysis in accounting: What's next? *Contemporary Accounting Research*, *40*(2), 765–805.

Bochkay, K., Hales, J., & Chava, S. (2020). Hyperbole or reality? Investor response to extreme language in earnings conference calls. *The Accounting Review*, *95*(2), 31–60.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Brown, S. V., Hinson, L. A., & Tucker, J. W. (2023). *Financial Statement Adequacy and Firms' MD&A Disclosures* (SSRN Scholarly Paper 3891572).

Chang, E. C., Lin, T.-C., Luo, Y., & Ren, J. (2019). Ex-Day Returns of Stock Distributions: An Anchoring Explanation. *Management Science*, *65*(3), 1076–1095.

Chen, H., De, P., Hu, Y. (Jeffrey), & Hwang, B.-H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *The Review of Financial Studies*, *27*(5), 1367–1403.

Cole, C. J., & Jones, C. L. (2004). The Usefulness of MD&A Disclosures in the Retail Industry. *Journal of Accounting, Auditing & Finance*, *19*(4), 361–388.

Day, M.-Y., & Lee, C.-C. (2016). *Deep learning for financial sentiment analysis on finance news providers*. 1127–1134.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*.

Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: Machine learning vs. Dictionary methods. *Management Science*, *68*(7), 5514–5532.

Ghoshal, S., & Roberts, S. (2020). Thresholded ConvNet ensembles: Neural networks for technical forecasting. *Neural Computing and Applications*, *32*(18), 15249–15262.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H., & Yang, Y. (2022). *LongT5: Efficient Text-To-Text Transformer for Long Sequences* (arXiv:2112.07916). arXiv.

Heston, S. L., & Sinha, N. R. (2017). News vs. Sentiment: Predicting Stock Returns from News Stories. *Financial Analysts Journal*, *73*(3), 67–83.

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., & Xu, Y. (2019). BERT-based financial sentiment index and LSTM-based stock return predictability. *arXiv Preprint arXiv:1906.09024*.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text*. *Contemporary Accounting Research*, *40*(2), 806–841.

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). *FinanceBench: A New Benchmark for Financial Question Answering* (arXiv:2311.11944). arXiv.

Kelley, E. K., & Tetlock, P. C. (2013). How Wise Are Crowds? Insights from Retail Orders and Stock Returns. *The Journal of Finance*, *68*(3), 1229–1265.

Kim, J. S., Kim, D.-H., & Seo, S. W. (2017). Investor Sentiment and Return Predictability of the Option to Stock Volume Ratio. *Financial Management*, *46*(3), 767–796.

Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., & Radev, D. (2022). *BookSum: A Collection of Datasets for Long-form Narrative Summarization* (arXiv:2105.08209). arXiv.

Li, F. (2010). Textual Analysis of Corporate Disclosures: A Survey of the Literature. *Journal of Accounting Literature*, *29*, 143–165.

Lonare, G., Patil, B., & Raut, N. (2021). edgar: An R package for the U.S. SEC EDGAR retrieval and parsing of corporate filings. *SoftwareX*, *16*, 100865.

Lopez-Lira, A., & Tang, Y. (2023). *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models* (SSRN Scholarly Paper 4412788).

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

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Lutz, C. (2016). THE ASYMMETRIC EFFECTS OF INVESTOR SENTIMENT. *Macroeconomic Dynamics*, *20*(6), 1477–1503.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, *8*, 131662–131682.

Muslu, V., Radhakrishnan, S., Subramanyam, K. R., & Lim, D. (2015). Forward-Looking MD&A Disclosures and the Information Environment. *Management Science*, *61*(5), 931–948.

Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, *13*(1), 760–770.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, *84*, 25–40.

Romanko, O., Narayan, A., & Kwon, R. H. (2023). ChatGPT-Based Investment Portfolio Selection. *Operations Research Forum*, *4*(4), 91.

Schmeling, M. (2009). Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance*, *16*(3), 394–408.

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, *285*, 181–203.

Souma, W., Vodenska, I., & Aoyama, H. (2019). Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, *2*(1), 33–46.

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, *104*(2), 288–302.

Szemraj, P. (2022). Long-t5-tglobal-base-16384-book-summary (Revision 4b12bce). *Hugging Face*.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... & Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, *3*(6), 7.

Tavcar, L. R. (1998). Make the MD&A more readable. *The CPA Journal*, *68*(1), 10.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168.

Tetlock, P. C. (2010). Does Public Financial News Resolve Asymmetric Information? *The Review of Financial Studies*, *23*(9), 3520–3557.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, *63*(3), 1437–1467.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., & Azhar, F. (2023). Llama: Open and efficient foundation language models. *arXiv Preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., & Bhosale, S. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv Preprint arXiv:2307.09288*.

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2023). *PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance* (arXiv:2306.05443). arXiv.

Yang, Y., Tang, Y., & Tam, K. Y. (2023). *InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning* (arXiv:2309.13064). arXiv.

Yu, J., & Yuan, Y. (2011). Investor sentiment and the mean–variance relation. *Journal of Financial Economics*, *100*(2), 367–381.

Zhang, J. L., Härdle, W. K., Chen, C. Y., & Bommes, E. (2016). Distillation of News Flow into Analysis of Stock Reactions. *Journal of Business & Economic Statistics*, *34*(4), 547–563.

Zhang, Z., Zohren, S., & Roberts, S. (2020). Deep Learning for Portfolio Optimization. *The Journal of Financial Data Science*, *2*(4), 8–20.

Zhu, F., Liu, Z., Feng, F., Wang, C., Li, M., & Chua, T.-S. (2024). *TAT-LLM: A Specialized Language Model for Discrete Reasoning over Tabular and Textual Data* (arXiv:2401.13223). arXiv.

Appendices

Appendix A. Fine-tuning Details of LLaMA-2 Model

This appendix provides a detailed summary of the fine-tuning process for the LLaMA-2 7B model, including the datasets, hyperparameters, and training performance metrics. Training data specifics are outlined in Table A1. The model's hyperparameters were configured based on Meta's guidelines for LLaMA-2 and adjusted to fit our hardware constraints.

Key hyperparameter settings include LoRA parameters (r=8, lora_alpha=32, lora_dropout=0.05), a learning rate of 10^{-5}, the AdamW optimizer with weight decay, and a batch size of 4 with gradient accumulation steps of 1. These settings were chosen to ensure stable training, as demonstrated by the training loss trend across five epochs.

Table A2 provides statistical details of the training loss per epoch, including the mean, standard deviation, and 25th, 50th, and 75th percentiles (P25, P50, and P75). The average training loss over five epochs was recorded at 0.9566, indicating consistent model convergence. Figure A1 illustrates the downward trend in training loss, highlighting the effectiveness of the selected configurations.

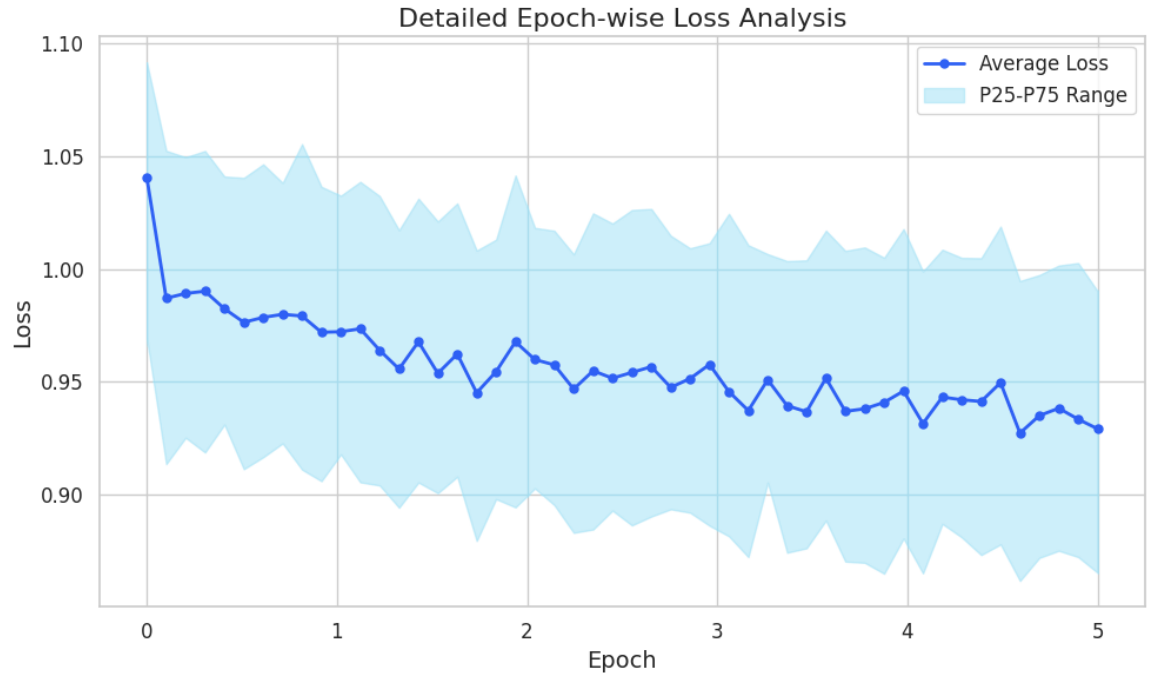
Table A1. Training Data Summary

Dataset	Purpose	Source	Sentiment			Description
			Pos	Neg	Neu	
Financial Phrasebank	Sentiment	Malo et al. (2014)	1363	604	2879	Polar sentiment dataset of sentences from financial news.
Twitter Financial News Sentiment	Sentiment	Hugging Face	2398	1789	7744	An English-language dataset containing an annotated corpus of finance-related tweets.
Climate Sentiment	Sentiment	Bingler et al. (2023)	301	448	571	An expert-annotated dataset for climate-related sentiment in corporate disclosures.
Standford Alpaca	Instruction	Taori et al. (2023)				A dataset of 52,000 instructions and demonstrations generated by OpenAI's text-davinci-003 engine.

Table A2. Epoch-wise Training Analysis

Epoch Training Loss	Mean	SD	P25	P50	P75
Loss of epoch 0 to 1	0.9875	0.1030	0.9227	0.9873	1.0517
Loss of epoch 1 to 2	0.9617	0.0969	0.9003	0.9632	1.0281
Loss of epoch 2 to 3	0.9539	0.0940	0.8904	0.9552	1.0189
Loss of epoch 3 to 4	0.9424	0.1008	0.8762	0.9464	1.0106
Loss of epoch 4 to 5	0.9372	0.0977	0.8729	0.9377	1.0028
Loss of Total 5 Epochs	0.9566	0.1001	0.8901	0.9591	1.0236

Figure A1. Epoch-wise Training Analysis



Appendix B. LLaMA-2 Model Sentiment Survey

This appendix provides an overview of the sentiment classification approach and model outputs for the fine-tuned LLaMA-2 model, focusing on the sentiment categories of Positive, Negative, and Neutral. It includes a survey of high-frequency terms used in training data for each sentiment class, as well as example texts that illustrate how the model handles clear and ambiguous sentiment cases. Table B1 presents the top 50 most frequent 1-grams and 2-grams for each sentiment type in the training data, providing insight into the typical language patterns associated with each category. Tables B2 and B3 display example sentences used to assess the model's performance, with B2 covering clear examples for each sentiment type and B3 highlighting texts with mixed or ambiguous sentiments, illustrating the model's tendency towards neutrality when sentiment direction is unclear.

Table B1. High-Frequency Terms by Sentiment Category

	Positive		Negative		Neutral	
1	eur	1.05%	risk	1.89%	company	0.92%
2	company	0.82%	climate	0.95%	market	0.56%
3	market	0.73%	eur	0.81%	stock	0.55%
4	share	0.64%	impact	0.78%	business	0.52%
5	energy	0.60%	change	0.75%	report	0.51%
6	beat	0.56%	climate change	0.66%	result	0.48%
7	mn	0.55%	company	0.58%	group	0.48%
8	profit	0.54%	business	0.58%	marketscreener	0.46%
9	net	0.49%	mn	0.56%	share	0.45%
10	target	0.49%	market	0.53%	emission	0.38%
11	finnish	0.45%	profit	0.47%	energy	0.37%
12	business	0.43%	loss	0.46%	earnings	0.36%
13	investment	0.42%	cost	0.41%	stock marketscreener	0.35%
14	group	0.40%	oil	0.40%	financial	0.33%
15	project	0.39%	share	0.40%	climate	0.32%
16	revenue	0.39%	financial	0.38%	dividend	0.31%
17	fund	0.38%	group	0.38%	investment	0.31%
18	service	0.38%	result	0.38%	service	0.31%
19	period	0.37%	cut	0.36%	fund	0.30%
20	growth	0.33%	physical	0.34%	bank	0.29%
21	mln	0.32%	event	0.33%	inc	0.28%
22	eps	0.32%	energy	0.33%	risk	0.28%
23	operating	0.31%	environmental	0.32%	economy	0.27%
24	climate	0.30%	economy	0.30%	ceo	0.26%
25	economy	0.28%	miss	0.30%	buy	0.24%
26	raised	0.27%	operating	0.30%	global	0.24%
27	increase	0.27%	weather	0.29%	eur	0.24%
28	solution	0.26%	increase	0.29%	update	0.23%
29	rise	0.25%	investment	0.29%	target	0.23%
30	technology	0.25%	due	0.28%	management	0.22%
31	higher	0.24%	net	0.27%	declares	0.22%
32	increased	0.24%	including	0.26%	finnish	0.22%
33	today	0.24%	increased	0.25%	project	0.21%
34	buy	0.24%	operation	0.25%	industry	0.21%
35	operating profit	0.24%	revenue	0.25%	environmental	0.21%
36	bank	0.24%	finnish	0.25%	investor	0.21%
37	product	0.24%	global	0.24%	oil	0.21%
38	gain	0.24%	potential	0.24%	operation	0.21%
39	trade	0.24%	bank	0.24%	change	0.21%
40	carbon	0.23%	period	0.24%	fed	0.20%
41	global	0.23%	product	0.24%	product	0.20%
42	oil	0.23%	report	0.24%	capital	0.20%
43	target raised	0.23%	transition	0.23%	conference	0.20%
44	earnings	0.22%	coronavirus	0.22%	plan	0.20%
45	sustainable	0.22%	emission	0.22%	trump	0.20%
46	finance	0.22%	management	0.22%	carbon	0.19%
47	oyj	0.22%	carbon	0.22%	corporation	0.19%
48	year	0.22%	asset	0.21%	presentation	0.18%
49	order	0.21%	future	0.21%	technology	0.18%
50	renewable	0.21%	gas	0.21%	data	0.18%

Table B2. Certain Examples of LLaMA-2 Sentiment

LLaMA-2 Sentiment	Financial Text Examples
Positive	The company's revenue grew by 15% this quarter, exceeding analyst expectations and indicating strong financial performance.
Positive	Management announced an increase in dividends, signaling confidence in future cash flows.
Positive	The company reported record-breaking sales for the third consecutive quarter, surpassing analyst expectations by 15%. This growth was largely driven by the success of its new product line, which has gained significant market share in North America and Europe. Management has also announced plans to expand into the Asian market, further fueling optimism among investors.
Negative	The firm's operating costs have surged due to rising commodity prices, cutting into profit margins.
Negative	The company missed its quarterly earnings target, and its stock dropped by 8% in after-hours trading.
Negative	The company's profits fell sharply, declining by 20% year-over-year, mainly due to rising raw material costs and weakening consumer demand. The CFO warned that, with inflation pressures expected to continue, the firm's profitability could remain under stress. Additionally, recent layoffs indicate a struggle to balance costs, leaving investors concerned about the company's financial health.
Neutral	The company is scheduled to release its quarterly report next Tuesday, providing an update on its recent financial performance.
Neutral	The company recently held its annual shareholder meeting, where it reviewed past performance and provided guidance for the upcoming fiscal year. Management discussed macroeconomic challenges, such as fluctuating exchange rates, but refrained from making specific predictions about the company's growth. The tone remained cautiously optimistic, with no major announcements or unexpected updates.

Table B3. Ambiguous Examples of LLaMA-2 Sentiment

LLaMA-2 Sentiment	Financial Text Examples
Negative	While the company posted strong revenue growth, the increasing debt levels are raising concerns among investors.
Positive	The acquisition of a competitor is expected to increase market share, though it may take time for the integration costs to pay off.
Neutral	Despite launching a new product line, the firm is still struggling with supply chain issues that could impact future profitability.
Neutral	Although profits are up compared to last year, regulatory investigations could potentially harm the company's reputation.
Neutral	The company achieved record-breaking sales, yet its heavy reliance on debt financing has analysts worried about long-term sustainability, especially if interest rates continue to rise.
Neutral	While the stock has seen steady gains over the past few months, lingering concerns about market volatility, regulatory risks, and high valuations keep some investors on the sidelines.
Neutral	The firm's expansion into emerging markets shows potential for significant revenue, though political instability, currency fluctuations, and supply chain challenges could hinder results.
Neutral	The firm's expansion into emerging markets shows potential for significant revenue, though political instability, currency fluctuations, and supply chain challenges could hinder results.
Neutral	The company announced a 30% increase in quarterly revenue, driven by strong demand in emerging markets. However, the CFO warned that escalating production costs and supply chain disruptions could offset these gains in the upcoming quarters. Investors are left wondering if this growth is sustainable under current conditions.
Positive	The board announced a significant dividend increase, which delighted shareholders and sent the stock up 5%. Yet, at the same time, the company is laying off a large portion of its workforce due to cost-cutting measures. This dual move reflects the complex trade-off between maintaining investor confidence and managing operational expenses.
Neutral	Despite reporting a loss for the fiscal year, management highlighted promising early results from its new product line, suggesting a potential turnaround. Analysts, however, are cautious, noting that high competition and ongoing debt obligations could undermine these initial gains. The market reaction has been mixed, with some seeing potential and others remaining skeptical.
Neutral	In a bold move, the company announced plans to expand into renewable energy, which aligns with global trends and could enhance its public image. However, the upfront costs are expected to be high, and some investors worry about the impact on near-term profitability. The stock responded with volatility, reflecting the divided opinions on this ambitious shift.

Appendix C. Comparative Sentiment Analysis of Full Texts and Summarized Texts using Fine-tuned LLaMA-2 Model

This appendix provides a detailed analysis of sentiment classification differences among full text, first-step summarized text, and second-step summarized text, using 500 randomly sampled text passages and a fine-tuned LLaMA-2 model. This study aims to investigate the effectiveness and trade-offs of single-step versus two-step summarization in sentiment analysis tasks.

Table C1 presents the distribution of predicted sentiment labels (including Unknown labels indicating misclassification), average word count, and execution time across the three types of texts. Table C2 provides a classification report comparing the sentiment predictions of first-step and second-step summarized texts against human-tagged labels, focusing on performance metrics such as Precision, Recall, and F1-Score. Here, the human-tagged sentiment serves as the true label, while first-step and second-step predictions are treated as model outputs. This table facilitates an assessment of the trade-offs between the two summarization approaches.

Figure C1 visualizes the confusion matrices for first-step and second-step summarized text sentiment predictions, using the human-tagged sentiment as the reference. This analysis aids in understanding areas where the two summarization steps differ in accuracy. Figure C2 delves deeper into the relationship between first-step and second-step summarizations for the human-tagged sample, examining sentiment distribution, agreement rate, and the sentiment-specific agreement in predictions.

Finally, Figure C3 presents the word count distributions across the 500-sample texts and the training data used to fine-tune LLaMA-2. This comparison sheds light on the variations in text length across different text types and their alignment with the training data length distribution, offering insight into the model's familiarity with varying text lengths.

Table C1. Distribution of Sentiment Tags, Word Count, and Execution Time Across Text Types

	Sentiment Tags					
Text Type	Unknown	Positive	Negative	Neutral	Avg Wordcount	Execution Time
Full Text	468	22	1	9	9122.33	775.32 sec
1st Step Summarized Text	4	116	8	372	487.12	79.74 sec
2nd Step Summarized Text	0	282	42	176	62.92	36.93 sec

Table C2. Classification Report of First-step and Second-step Summarized Texts Using Human-Tagged Labels

	First-step Summarized Text				Second-step Summarized Text			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
Positive	0.6466	0.3488	0.4532	215	0.7730	0.9008	0.8321	242
Negative	1.0000	0.0457	0.0874	175	0.9048	0.3423	0.4967	111
Neutral	0.2231	0.7830	0.3473	106	0.5398	0.6463	0.5882	147
Accuracy			0.3347	496			0.7020	500
Micro Avg	0.6232	0.3925	0.2960	496	0.7392	0.6298	0.6390	500
Weighted Avg	0.6808	0.3347	0.3015	496	0.7337	0.7020	0.6859	500

Figure C1. Confusion Matrices for Sentiment Prediction on 500-Sample Experiment (First-step and Second-step Summarized Texts)

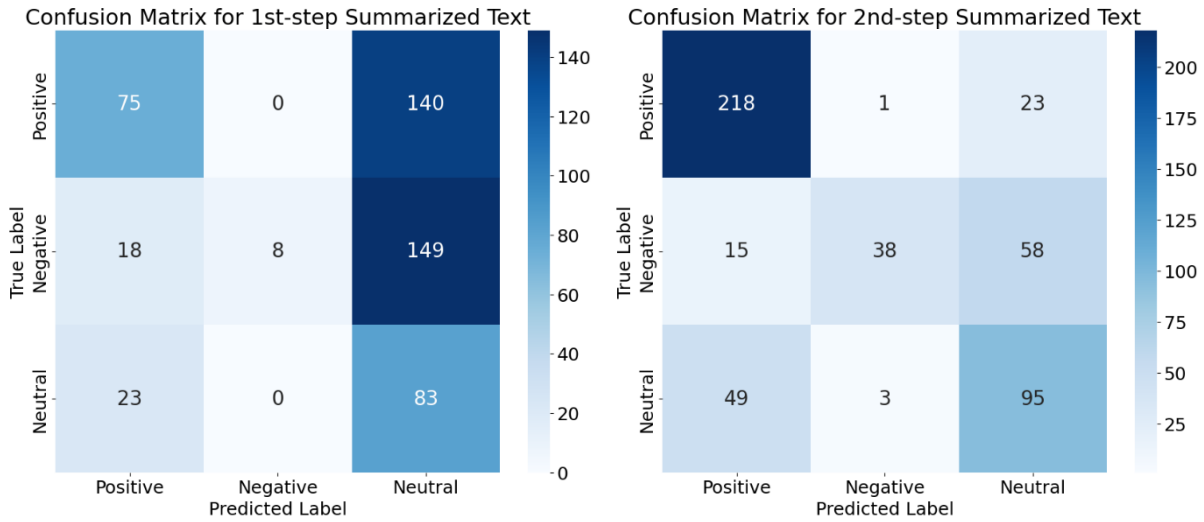


Figure C2. Comparison of First-step and Second-step Summarized Sentiments in Human-Tagged Texts: Distribution and Agreement Rate

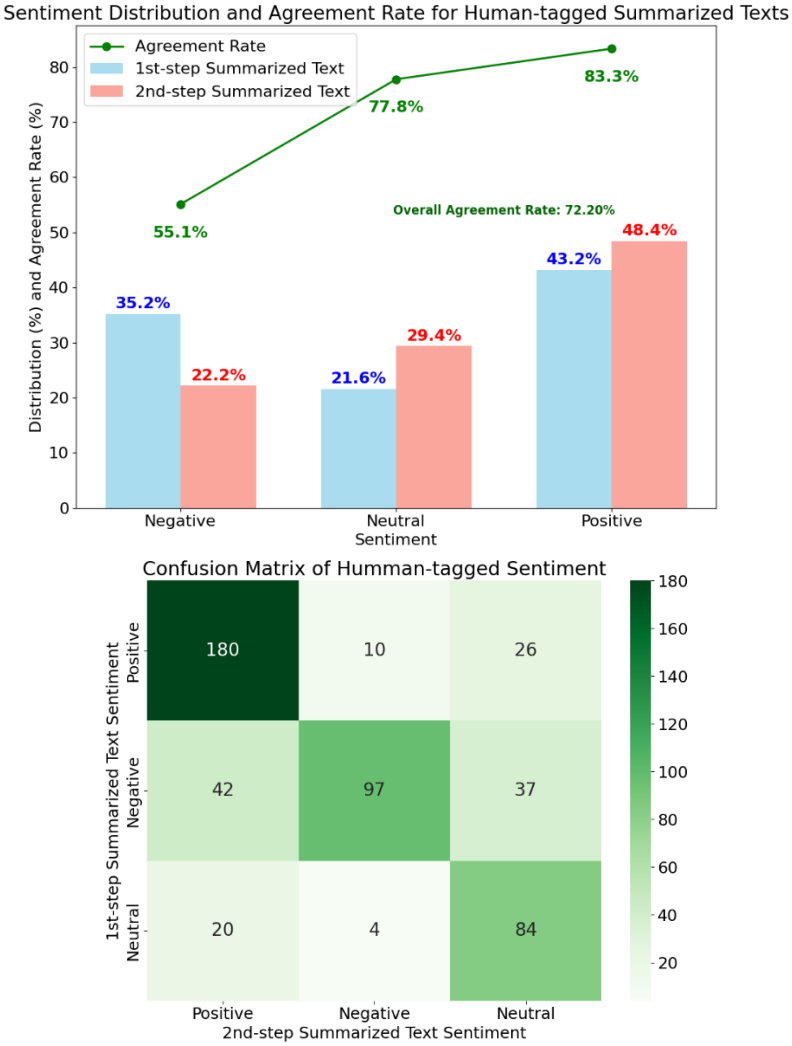
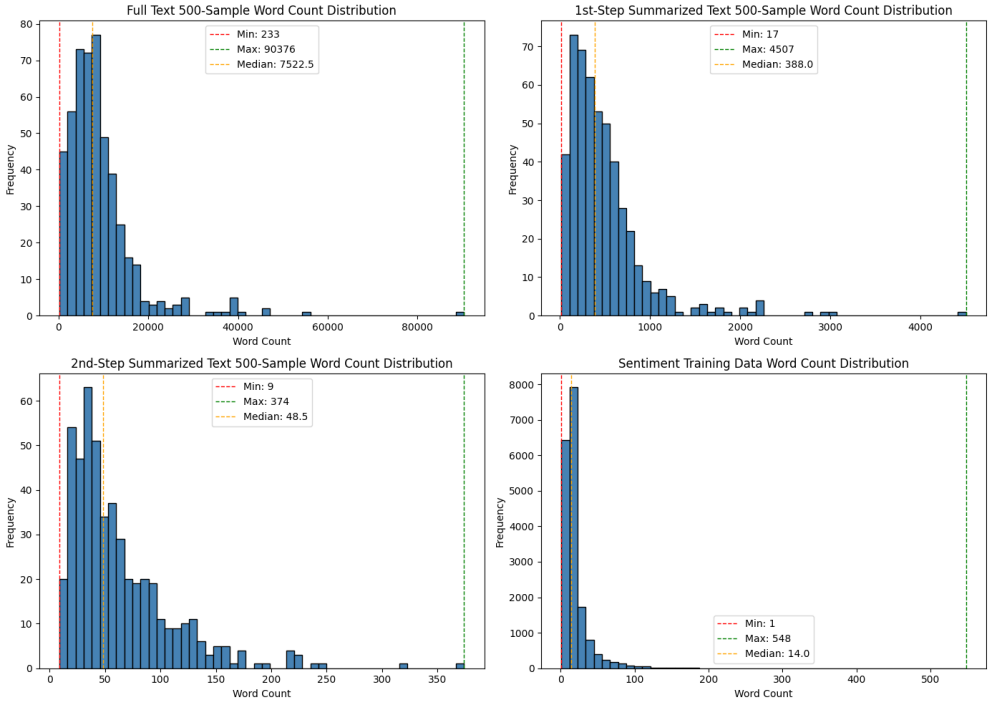


Figure C3. Word Count Distribution Across Full Text, Summarized Texts, and Fine-tuning Training Data



Appendix D. Word Analysis of The Three Types of Texts

This appendix provides an in-depth examination of the top words and linguistic structure within Full Text, first-step Summarized Text, and second-step summarized text versions. Table D1 presents the top 50 most common words (including 1-grams and 2-grams) within each text type, along with their percentage representation of the overall word count. Table D2 further breaks down the intersection of these words across the three text versions, detailing the words common to all sets, those unique to each set, and specific combinations of overlap. Figure D1 visualizes the frequency and emphasis of these common words within each text type using word clouds, offering a visual comparison of word prominence and distribution across different summarization stages.

Table D1. Top 50 Words (1-grams and 2-grams)

	Full Text		**First-step Summarized Text**		**Second-step Summarized Text**	
1	company	1.06%	company	2.99%	company	5.05%
2	expense	0.81%	business	1.14%	business	1.99%
3	revenue	0.80%	financial	0.97%	financial	1.14%
4	cost	0.79%	expense	0.95%	result	1.02%
5	net	0.73%	year	0.87%	year	0.93%
6	income	0.64%	product	0.77%	expense	0.87%
7	result	0.60%	cost	0.75%	product	0.81%
8	increase	0.60%	result	0.69%	increase	0.74%
9	tax	0.60%	due	0.68%	due	0.73%
10	product	0.57%	increase	0.68%	significant	0.73%
11	fiscal	0.57%	revenue	0.67%	revenue	0.69%
12	financial	0.56%	net	0.59%	report	0.67%
13	asset	0.54%	tax	0.58%	net	0.67%
14	ended	0.52%	significant	0.57%	cost	0.63%
15	operation	0.51%	interest	0.50%	continues	0.62%
16	operating	0.48%	service	0.48%	service	0.49%
17	due	0.46%	rate	0.44%	higher	0.44%
18	interest	0.44%	amount	0.43%	interest	0.43%
19	rate	0.44%	operating	0.40%	tax	0.42%
20	million	0.44%	income	0.40%	customer	0.38%
21	service	0.43%	customer	0.39%	risk	0.37%
22	related	0.42%	statement	0.39%	financial result	0.37%
23	increased	0.41%	market	0.38%	debt	0.37%
24	primarily	0.41%	debt	0.37%	market	0.36%
25	statement	0.41%	increased	0.37%	operating	0.36%
26	business	0.40%	million	0.36%	make	0.35%
27	loss	0.39%	asset	0.35%	statement	0.34%
28	compared	0.38%	risk	0.35%	increased	0.34%
29	period	0.35%	credit	0.35%	ended	0.34%
30	year	0.35%	higher	0.34%	previous	0.33%
31	amount	0.35%	agreement	0.33%	profit	0.31%
32	credit	0.34%	operation	0.32%	rate	0.30%
33	market	0.33%	share	0.32%	loss	0.29%
34	note	0.32%	example	0.31%	income	0.29%
35	customer	0.32%	accounting	0.31%	million	0.29%
36	total	0.31%	capital	0.31%	amount	0.28%
37	capital	0.31%	loss	0.31%	pay	0.28%
38	approximately	0.31%	term	0.30%	share	0.28%
39	future	0.30%	purchase	0.30%	plan	0.27%
40	agreement	0.30%	activity	0.30%	fy	0.26%
41	share	0.30%	future	0.29%	line	0.26%
42	facility	0.29%	addition	0.29%	order	0.25%
43	activity	0.27%	order	0.29%	purchase	0.25%
44	including	0.27%	however	0.28%	expected	0.25%
45	decrease	0.26%	contract	0.28%	capital	0.25%
46	accounting	0.26%	plan	0.27%	compared	0.24%
47	development	0.25%	previous	0.27%	previous year	0.24%
48	consolidated	0.25%	expected	0.27%	activity	0.24%
49	management	0.25%	primarily	0.27%	credit	0.23%
50	liability	0.24%	investment	0.26%	position	0.23%

Table D2. Word Overlap and Unique Words Across Three Types of Texts

Words Common to Three Sets (Common Words: 29)
Core Business and Market Terms: company, business, customer, product, service, financial, market *Financial Indicators and Resources:* revenue, income, expense, cost, tax, interest, net, rate *Performance Metrics and Results:* loss, increase, increased, result, statement *Operational and Activity Indicators:* operating, activity, capital, credit *Quantity and Temporal Terms:* year, due, amount, million, share
Words common to full texts and first-step summarized texts but not in all three sets (Common Words: 35)
future, operation, accounting, asset, agreement, primarily
Words common to full texts and second-step summarized texts but not in all three sets (Common words: 31)
compared, ended
Words common to first-step and second-step summarized texts but not in all three sets (Common words: 38)
debt, expected, higher, risk, significant, order, plan, previous, purchasec
Unique words in full texts only (13 words)
decrease, facility, development, management, liability, approximately, consolidated, fiscal, including, note, period, related, total
Unique words in first-step summarized texts only (6 words)
investment, contract, example, addition, however, term
Unique words in second-step summarized texts only (10 words)
profit, continues, position, previous year, financial result, report , fy, make, line, pay

Figure D1. Word Cloud of Three Types of Texts

Panel A. Word Cloud of Full Texts

Panel B. Word Cloud of First-step Summarized Texts

Panel C. Word Cloud of Second-step Summarized Texts

Appendix E. Summarized Text Examples

This appendix showcases contextual summarization examples produced by the Long-T5 model. The process involves a two-step summarization process.

Access Num	Company (CIK)	Filing Date	Original Word Counts	1st summarized Word Counts	2nd summarized Word Counts
0000950170-22-002517	873303	2022-03-01	37,469	2,139	301
Summary: This annual report describes the financial condition of the company and its business activities. Although the company has sufficient funds to support its mission, it faces significant risks because of the continuing use of the duchenne disease in the treatment of patients with confirmed Mutinogen gene defects. The cost-of-virgil disease is a significant risk to the companys ability to commercialize and/or market their products. These risks include delays in obtaining approval for product candidates from the federal authorities, delays in conducting pre-clinical or early-stage studies, and delays in receiving final approval from the Food and drug Regulatory Council. It also depends on third parties to conduct some of their earlystage or pre-clinical development, but they must comply with stricter regulations than those of other drugs approved by the West. Finally, our dependence on third party distributors for manufacturing and selling our products may cause delays in the distribution of our products. We must maintain strict control over the manufacturing process; we may have to perform additional laboratory trials or trials to prove that our product's efficacy is as good as or superior to any other product available on the marketplace. Other risks related to intellectual property relate to the health laws and regulation in the U.S. As well as the financial situation, there are numerous other risks that can affect the business. For instance, the certificate of Incorpation, the Delaware law, and certain business combination provisions may make it harder for stockholders to gain control or effect change in the board. Certain business combination with stocksowners who hold more than 15 Percent of their shares could result in serious difficulties for these potential acquirers to get into control of us. Furthermore, foreign laws may lead to civil or criminal penalties, such as delays in approval nor rejection of an approved product.					
0001558370-20-001127	1374310	2020-02-21	11,938	500	60
Summary: In this report, the Company compares its financial results to that of the previous year in order to better understand how it manages its business. The company's total operating expenses are lower than expected due to a number of factors, including a significant decline in the price of certain services and an increase in the cost of certain business-critical offices.					
0001047469-15-001069	1082554	2015-02-24	37,428	2,573	131
Summary: Fargo Bank's financial position continues to improve as the company continues to develop new drugs for patients. The FDA extends the review process for new drug approvals for three months and allows manufacturers to submit new applications without having to undergo pre-approved studies. Other laws in the United States and foreign countries govern the manufacture and distribution of medicinal products, including biologics. These laws also require that companies submit similar samples to FDA before approving or withdrawing a new product. In particular, there are provisions in the Patient Care Reform Act that require companies to disclose certain information about safety and efficacy of their products. This law may lead to increased costs for the company, but it will not be completely affected until all of these provisions are passed into effect.					
0001445305-13-000433	882104	2013-03-01	12,180	589	214
Summary: PDL is not a public company, but it does have a significant amount of intellectual property to invest in. PDL has a board of directors set the annual dividend each year according to generally accepted accounting principle. The company hedges certain foreign currency risks with forward contracts so that they do not have to pay back its debts. In December 31, PDL issues a warrant for facet's redwood city based leases as long as FACET does not default on the contract. For the rest of the year, PDL expects to have enough cash to support its business. Wellstat borrows money from PDL under an agreement whereby PDL pays back the loan when it is due. Wellsta agrees to pay PDL at different interest and royalty levels. As previously mentioned, Wellsta had breached PDL by sending a default note on February 22, 2013. PDL uses another remedy to borrow money off Wellsta for the remaining part of the loan because Wellsta doesn't have sufficient funds to repay the loan. On March 29, 2013, PDL enters into a buy and sale agreement with Hyperion whereby Hippobe sells all PDL shares in trade for two milestone payment that are due April and March. These notes are subject to conversion or cancellation at the discretion of the noteowners.					
0001206774-08-001977	854860	2008-12-11	5,875	154	31
Summary: The Company reports that for the third year in a row, its gross profit has declined as a result of an increase in the tax rate and an excisable income tax.					
0000950152-99-002941	880456	1999-03-31	6,761	305	58
Summary: The company's financial results are affected by a number of business factors, such as downsizing, restructuring, and a default on its senior secured loan. After the purchase of Nextran Company, the company takes an additional charge of $300,000 during the second quarter of 2000. Other income for the year is primarily due to interest and foreign exchange gains.					

Appendix F. Variable Definitions

This appendix contains a table that lists the variables used in this study, providing a clear reference for understanding the various elements analyzed throughout the research.

VARIABLE	DEFINITION	SOURCE
RETURNS		
R_i	Daily return of stock i.	CRSP
R_f	Daily return of risk-free rate.	French's website
R_m	Daily return of market.	French's website
BHR_{1w}	The average buy-and-hold return of long positive tone and short negative tone stocks in one week.	CRSP
BHR_{2w}	The average buy-and-hold return of long positive tone and short negative tone stocks in two week.	CRSP
BHR_{1m}	The average buy-and-hold return of long positive tone and short negative tone stocks in one month.	CRSP
BHR_{3m}	The average buy-and-hold return of long positive tone and short negative tone stocks in three months.	CRSP
BHR_{6m}	The average buy-and-hold return of long positive tone and short negative tone stocks in six months.	CRSP
BHR_{1y}	The average buy-and-hold return of long positive tone and short negative tone stocks in one year.	CRSP
$CAR[0,1]$	Market-adjusted cumulative abnormal return with a window of [0,1]. The market beta is estimated in one year window.	CRSP & French's website
$CAR[0,2]$	Market-adjusted cumulative abnormal return with a window of [0,2]. The market beta is estimated in one year window.	CRSP & French's website
$CAR[0,3]$	Market-adjusted cumulative abnormal return with a window of [0,3]. The market beta is estimated in one year window.	CRSP & French's website
$CAR[0,4]$	Market-adjusted cumulative abnormal return with a window of [0,4]. The market beta is estimated in one year window.	CRSP & French's website
$CAR[0,5]$	Market-adjusted cumulative abnormal return with a window of [0,5]. The market beta is estimated in one year window.	CRSP & French's website
CONTROLS		
TURNOVER	Trading volumn / (shareout x 1000)	Compustat
lnME	The longarithm of market value of the firm.	CRSP
BM	Book to market ratio = BE / ME	CRSP & Compustat
ROA	Return on asset = NI / AT	Compustat
VOLATILITY	Standard deviation of the daily return in a year window.	CRSP
LEVERAGE	Debt to equity = (DLCC+DLC)/SEQ	Compustat
TOBIN-Q	(AT+BE-CEQ) / AT	Compustat
TEXT_COUNT	Total words in the MD&A texts	EDGAR
IsNasdaq	The dummy to represent nasdaq stocks. (EXCHCD = 3)	CRSP
$\triangle SALE$	Sale growth rate	Compustat
$\triangle INVESTMENT$	Investment growth rate (growth rate of CAPX)	Compustat
GROSSMARGIN	Gross Profit / Asset = GP / AT	Compustat
$CRET_{1m}$	Cumulative return of last month	CRSP
$CRET_{2m\to12m}$	Cumulative return of last 2 to 12 months	CRSP

LLM TONES		
LLAMA2_TONE_E1	Prompted tone of LLaMA-2 1-epoch model from summarized texts	EDGAR, META
LLAMA2_TONE_E2	Prompted tone of LLaMA-2 2-epoch model from summarized texts	EDGAR, META
LLAMA2_TONE_E3	Prompted tone of LLaMA-2 3-epoch model from summarized texts	EDGAR, META
LLAMA2_TONE_E4	Prompted tone of LLaMA-2 4-epoch model from summarized texts	EDGAR, META
LLAMA2_TONE_E5	Prompted tone of LLaMA-2 5-epoch model from summarized texts	EDGAR, META
FINBERT_TONE	Prompted tone of FinBert model from summarized texts	EDGAR, Huang et al. (2023)
SUMMARY TEXTS TONES & SCORES		
LM_SUM_SCORE	Calculated score by LM dictionary from summarized texts	EDGAR, LM Dictionary
LM_SUM_TONE	Calculated tone by LM dictionary from summarized texts	EDGAR, LM Dictionary
LR_SUM_TONE	Predicted tone by Logistic Regression Model from summarized texts	EDGAR
SVM_SUM_TONE	Predicted tone by Support Vector Machine from summarized texts	EDGAR
RF_SUM_TONE	Predicted tone by Random Forest Model from summarized texts	EDGAR
NN_SUM_TONE	Predicted tone by Neural Network Model from summarized texts	EDGAR
FULL TEXTS TONES & SCORES		
LM_FULL_SCORE	Calculated score by LM dictionary from full texts	EDGAR, LM Dictionary
LM_FULL_TONE	Calculated tone by LM dictionary from full texts	EDGAR, LM Dictionary
LR_FULL_TONE	Predicted tone by Logistic Regression Model from full texts	EDGAR
SVM_FULL_TONE	Predicted tone by Support Vector Machine from full texts	EDGAR
RF_FULL_TONE	Predicted tone by Random Forest Model from full texts	EDGAR
NN_FULL_TONE	Predicted tone by Neural Network Model from full texts	EDGAR