Quantifying Market Sentiment: A Multi-Model Analysis of Wall Street Reports

Martin Li[1], Amy Tran[1], Holden Carrillo[1], Amber Yu[1] **Mentor: Ethan Yen[1,2]**

[1]University of California, Berkeley[1]; [2]Gilead Sciences[2]

Overview

Wall Street analyst reports contain detailed narrative insights that are difficult to quantify, yet they play a major role in shaping market expectations. Existing sentiment tools often miss the nuance in long-form equity research, especially when reports include mixed commentary, technical language, or firm-specific context.

We sought to answer the question: Can large language models accurately measure sentiment in Wall Street analyst reports, and how do design choices affect their performance?

We built a pipeline to measure sentiment in analyst reports using LLMs (Claude, Llama, and FinBERT) and test how different design choices, such as chunk size, prompting, and model selection, affect performance. We extract and structure report text, run standardized prompts in Databricks, and compare model outputs to human ratings collected on a 1–7 scale.

Preliminary Methods

Text Extraction

In Phase 1, we developed a pipeline to convert analyst report PDFs into a structured, sentence-level dataset. We extracted raw text using PyPDF, split it into full sentences, and assigned each sentence a unique identifier before exporting it as a CSV. This dataset serves as the foundation for chunking, prompt construction, and model inference in later stages of the project.

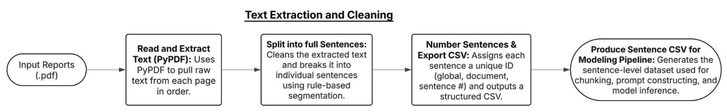

Figure 1: Workflow used to convert analyst report PDFs into a structured sentence-level dataset. The resulting dataset serves as the standardized input for downstream chunking, prompt construction, and model inference.

Manual Scoring

We scored a stratified sample of chunks among reports on a 1–7 Likert scale. Aggregated ratings served as ground truth for evaluating model accuracy. Models that score similarly to manual scores are considered more accurate, which is key to verify when comparing models and testing different strategies.

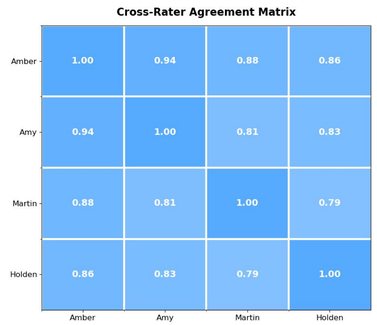

Figure 2: Manual Validation Comparison. Rater agreement was strong, indicating reliable and consistent human judgment across raters.

Model Testing

We evaluated multiple LLMs on the processed text. Each chunk was evaluated using Claude, Llama, and FinBERT through Databricks. We standardized outputs into 1–7 Likert scores, recorded non-commentary flags, and logged experimental settings to enable direct comparison across models and prompt configurations.

After producing the sentence dataset, we tested different strategies to potentially make the models more accurate at analyzing sentiment similarly to our manual scoring.

Window Size

We tested the performance of the evaluation of chunks of different sizes, including 1, 3, 5, and 7 sentences. A 5-line window yielded the best outcomes when compared to our manual scores.

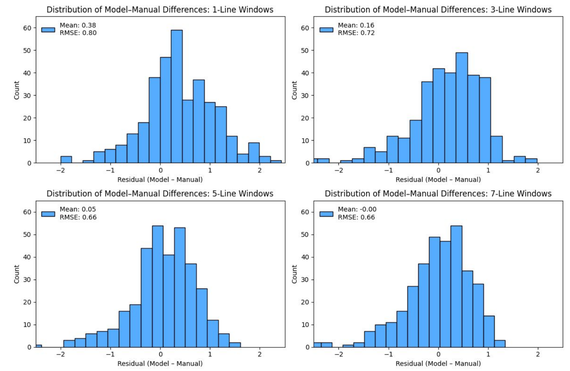

Figure 3: Distribution of Model - Manual Differences of 1-, 3-, 5-, and 7-Line Windows. The RMSE of the residuals decrease as window size increases, but becomes diminishing after 5 lines.

Few-Shot Prompting

We tested few-shot prompting the models with examples to see if accuracy would increase. More examples added noise instead of helping.

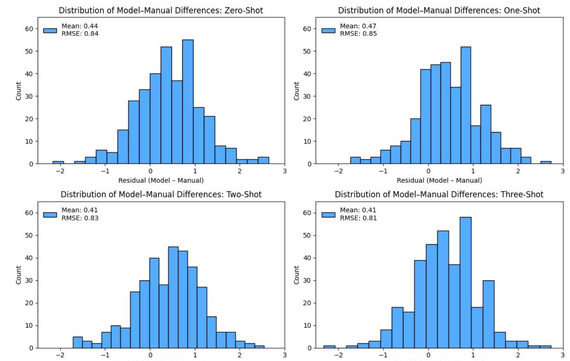

Figure 4: Distribution of Model - Manual Differences of 0-, 1-, 2, and 3-Shot Prompts. Zero-shot prompting aligns best with manual validation, while additional examples had negligible impacts.

Results

Across the full dataset, firms showed clear differences and similarities in the sentiment distribution of their analyst reports, both between each other and the market as a whole.

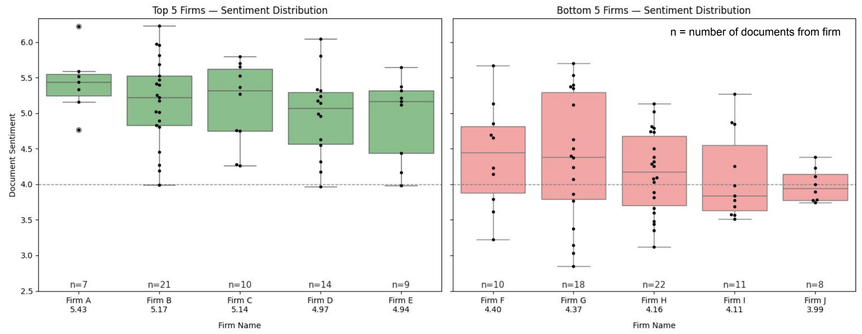

Figure 5: Top and Bottom 5 Firms in Average Sentiment. When ranking all firms by their average sentiment score, the five most positive firms showed median scores around 5.0–5.5, while the five most negative firms clustered around 4.0–4.4.

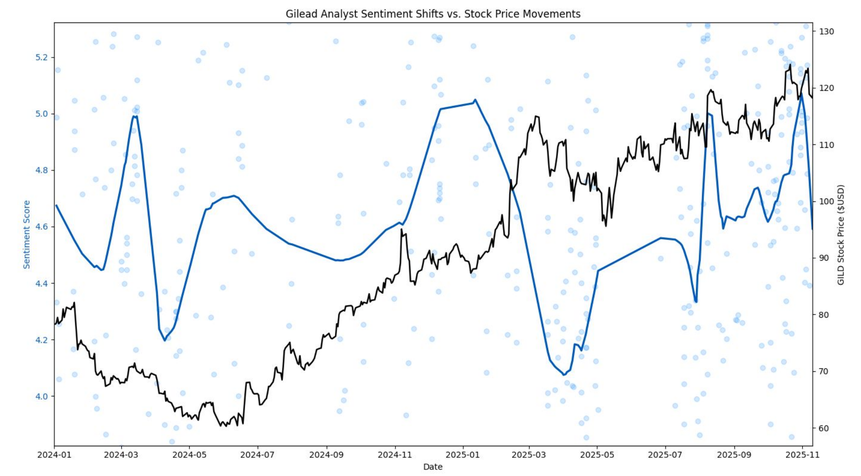

Figure 6: Gilead's Moving Average Sentiment Score vs. Gilead's Stock Price Over Time. While both sentiment and stock price fluctuate, the overall pattern indicates that analyst opinion contains useful information about market trends, particularly around inflection points.

Future Steps

Next semester, we will build long- and short-term trading portfolio simulations for each firm using our sentiment scores to test whether increases or decreases in analyst tone can generate useful trading signals. These simulations will help determine firm accuracy and whether sentiment signals can guide basic investment decision-making.

Acknowledgements

We thank Ethan Yen for his guidance and mentorship throughout this project, and we gratefully thank the Gilead Sciences leadership team for their support and feedback.