Is Small Really Beautiful for Central Bank Communication? Evaluating Language Models for Finance: Llama-3-70B, GPT-4, FinBERT-FOMC, FinBERT, and VADER

Wonseong Kim
Institute of Economics and Statistics, Korea University
KR
wonseongkim@korea.ac.kr

Jan Spörer
Institute of Computer Science, Chair of Natural Language Processing, University of St. Gallen
CH
jan.spoerer@unisg.ch

Choong Lyol Lee
Institute of Economics and Statistics, Korea University
KR
cllee@korea.ac.kr

Siegfried Handschuh
Institute of Computer Science, Chair of Natural Language Processing, University of St.Gallen
CH
Siegfried.Handschuh@unisg.ch

Abstract

This study compares the sentiment detection capabilities of language models for the domain of central bank communication, particularly the official statements released by the U.S. Federal Open Market Committee (FOMC). While previous studies have explored FOMC communication, this work is one of the few studies that use a natural language processing-based approach. The analysis employs VADER, FinBERT, a fine-tuned FinBERT model (FinBERT-FOMC), GPT-4, and Llama-3-70B.

Within the scope of our labeled dataset on FOMC minutes, Llama 3 is the most accurate model, followed by GPT-4, FinBERT-FOMC, FinBERT, and VADER. The FinBERT-FOMC model, which was fine-tuned on central bank communication and utilizes a text simplification pipeline, performs better than the original FinBERT model. Llama 3 and GPT-4 outperform at the expense of large model sizes. Unlike GPT-4, FinBERT and FinBERT-FOMC are open-source and can be deployed on consumer-grade hardware. Llama 3 requires substantial hardware investments to deploy.

The work thus finds that there is still a between model size and performance, and that the notion that "small is beautiful" can still hold for use cases where maximum accuracy is a lesser concern than inference speed and cost.

Human performance is still significantly above all models, indicating that further improvements in language models and FOMC-specific prompting are possible. The labeled dataset for central bank communication we present in this paper is thus a challenging benchmark for future research.

The dataset is freely available for download.[1]

[1] huggingface.co/datasets/janspoerer/fomc.

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698675

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Information systems** → **Sentiment analysis**.

Keywords

Federal Open Market Committee (FOMC), natural language processing, sentiment analysis, financial text

ACM Reference Format:
Wonseong Kim, Jan Spörer, Choong Lyol Lee, and Siegfried Handschuh. 2024. Is Small Really Beautiful for Central Bank Communication? Evaluating Language Models for Finance: Llama-3-70B, GPT-4, FinBERT-FOMC, FinBERT, and VADER. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3677052.3698675

1 Introduction

1.1 Introduction to the FOMC

The Federal Open Market Committee (FOMC) is part of the United States Federal Reserve and plays a significant role in the country's central banking system. It is responsible for directing monetary policy in the United States, which has far-reaching effects on the global economy. As the core policymaking body within the Federal Reserve System, the FOMC's decisions and communication impact various economic aspects, such as financial markets, investment strategies, and the general stability of the economy.

The FOMC convenes eight times a year and the official statements it releases after these meetings are of paramount importance. The statements reflect the Committee's current economic assessments and policy decisions, shaping the expectations of various economic stakeholders [35]. The statements' impact on market sentiment is substantial, leading to immediate and noticeable shifts in financial behavior and investment climate.

The official statements are scrutinized by economists, financial analysts, and investors, providing them with critical insights into the Federal Reserve's views on economic conditions, inflationary pressures, and future policy directions. Analyzing these statements is not just about decoding the immediate market reactions but also

understanding their longer-term implications. This understanding is vital for anyone engaged in the financial sector or interested in the health and trajectory of the U.S. and global economies.

1.2 Policy Transparency

Central banks worldwide have recently increased their emphasis on transparency and communication as essential tools for effective monetary policy [2, 34]. The FOMC's official statements are a prime example of this communication strategy, serving as a critical channel for conveying the Committee's policy stance and forward guidance. Despite evolving content, these statements maintain their relevance and impact on financial markets.

Market participants, policymakers, and researchers alike inspect the language and tone of FOMC statements for signals that may guide their decision-making and forecasting processes. Investors and banks use natural language processing to analyze the Federal Reserve's FOMC Minutes. The FOMC Minutes reflect the opinions of the most relevant monetary policy decision-making body in the U.S. on inflation, the state of the economy, and labor markets [37].

1.3 Contributions

This study has two primary contributions. Firstly, it presents an open FOMC minutes dataset that can be used as a foundational resource for sentiment analysis in central bank communication. The dataset captures the complexity of sentiments in financial texts, where a single sentence may contain multiple aspects. Secondly, this research article compares the performance of language models for the sentiment classification of complex financial texts. FOMC minute texts are subject to the ongoing academic debate surrounding the role of central bank communication in influencing financial markets and the economy. While numerous studies explore FOMC communication, only some cover the linguistic aspects of FOMC transcripts. Our work is a milestone in this domain, focusing on a linguistic approach to understand these documents better. We employ a multi-method approach, integrating qualitative content analysis and quantitative techniques. Our analysis encompasses the period since the inception of the FOMC's meeting minutes on January 3rd, 2006.

1.4 Overview of the Study's Structure

The remainder of this article is organized as follows: Section 2 reviews the literature on FOMC communication. Section 3 describes our data, including the data sources and findings from natural language processing (NLP) techniques. Section 4 presents our methods, highlighting sentiment prediction in FOMC meeting minutes. Finally, sections 5 and 6 present the predicted results and a discussion of the study's implications, limitations, and avenues for future research.

2 Related Work

2.1 Related Work on FOMC Communication

Language and communication have been increasingly recognized as essential factors in economic policy and decision-making, particularly in the context of central bank communications. [37] finds that the FED funds futures rate and the exchange rate of the U.S. dollar are partly reflected in FOMC transcript sentiment, which emphasizes the role that FOMC sentiment may play for market participants. [43] argues that increased central bank transparency and willingness to share assumptions about future policy have improved policy predictability and effectiveness.

[5] showed that text analysis techniques such as Latent Semantic Analysis can provide valuable insights into market participants' reactions to FOMC minutes. The work by [11] on transparency in central bank communication is another example of the growing use of text analysis in economics. They develop a new measure of central bank transparency based on NLP, which is used to assess the effectiveness of transparency policies in various countries. Ongoing research in this area aims to develop more accurate and efficient methods for analyzing financial text sentiment, potentially improving financial decision-making and forecasting.

More recently, [8] used natural language processing to identify the tone of FOMC post-meeting statements, highlighting the importance of qualitative statement language in policy decision-making. While some state-of-the-art methods are used in text classification in the financial domain [6, 23], the current linguistic analysis of FOMC minutes is still limited to vector language models. It has yet to incorporate more advanced models, such as transformers. Improving linguistic analysis techniques could provide even more accurate and efficient methods for analyzing central bank communication and informing economic policy.

2.2 Related Work on Methods for Sentiment Classification

There are multiple approaches to solving sentiment classification tasks, and [26] provide a summary of pre-LLM methods. Dedicated classification models were the first methods. Those models include VADER [10] and FinBERT [4, 15], which we also use in this study. But the task can also be solved using general or fine-tuned language models. This is a paradigm shift outlined by [36, pp. 173–174]. Examples of such general-purpose models used for sentiment classification include encoder-only transformers such as BERT [7] and its variants [12, 18, 20], encoder-decoder models such as BART [17], and decoder-only transformers such as the GPT family [1, 9, 29, 30]. Model engineers may also reformulate sequence-to-sequence (seq2seq) models to perform sentiment classification tasks [36, 44, p. 177]. Language models can also be fine-tuned for (sentiment) classification tasks, as [25, pp. 8–9] summarized in a recent survey.

[19] reveal that GPT-4 can solve a range of financial tasks with state-of-the-art performance. Their study finds that the sentiment classification performance of GPT-4 is comparable to FinBERT (pp. 410–411). We confirm their findings in section 5. [19] also report that GPT-4 has the highest performance on financial sentiment tasks among the language models under consideration. It performs better than other language models on the Financial PhraseBank task [24] and on the FiQA sentiment task [22].

The "small is beautiful" principle finds support through knowledge distillation, as demonstrated by DistilBERT [33]. This involves training a smaller "student" model to replicate the behavior of a larger "teacher" model. By leveraging the teacher model's outputs, the student model learns the task and absorbs the nuances of the

larger model, retaining up to 97% of the performance while requiring fewer parameters and running faster. This illustrates that smaller, fine-tuned models can be just as effective as larger ones for downstream tasks.

[27] use a fine-tuned RoBERTa model to achieve better performance on the TweetFinSent dataset than GPT-4 in a few-shot setting [19, p. 411]. This shows how the performance of sentiment models is influenced by the type of text that the model was trained on. While GPT-4 is comparable to fine-tuned smaller models on formal text, it underperformed in the context of X (formerly Twitter) texts, which are usually short and sometimes written in a casual style.

Prompt engineering aims to improve the capabilities of language models without requiring changes to the model [31, 41]. [42] propose to not only prompt models with question-answer prompts (few-shot prompting), but to provide question-reasoning-answer triplets in the few-shot prompt. They find that eight question-reasoning-answer triplets significantly improve the reasoning abilities of language models.

Flow engineering [32] is a concept that emphasizes the benefits of gradual knowledge extraction and reasoning from language models. It is a shift from one-off prompts to a conversational, multi-prompt approach. While the authors work with a programming use case, the concept was proposed only recently and will likely be generalized by other researchers to more domains.

3 Data

3.1 Overview

The dataset is freely available for download. [2] Our study employs language modeling techniques to 1,065 randomly selected medium-length sentences from the FOMC since January 3rd, 2006, covering 131 FOMC minute transcripts (32,034 sentences) over 18 years. We chose to analyze medium-length sentences from FOMC communications, given their complexity. These communications often cover intricate economic situations and future actions, and medium-length sentences are better suited for capturing multiple aspects of a discussion. Shorter sentences may not fully convey the intended sentiment in context. To ensure accuracy, we specifically curated the FOMC dataset to include sentences of medium length.

Existing research on sentiment analysis in FOMC minutes has typically utilized word-level analysis, even when dealing with longer paragraphs [14]. Word-level approaches fail to detect contextualized nuances. Alternative, contextualized approaches such as BERT [7] and ELMo [28] can overcome the limitations of word-level models. Consequently, this study explores the benefits of sentence-level sentiment analysis to gain a contextual understanding of FOMC minutes.

The preprocessing steps included four stages, but not all were performed for all models: splitting, removing stopwords, removing short sentences, and manual inspection. The first step involved splitting paragraphs into individual sentences and removing headlines. The FinBERT-FOMC model post-processes the sentences using a text simplification pipeline described by [6]. Next, and only for the FinBERT model, we removed stopwords, which are words that do not carry significant meaning. Then, for all models, we filtered out sentences too short to provide meaningful insights, which we set at fewer than eight words or fewer than 43 characters.

[2] huggingface.co/datasets/janspoerer/fomc.

3.2 Labeling

To assess the precision of the language model, qualified annotators labeled a total of 1,065 sentences. The research design accounted for the possibility that one person might exaggerate aspects and sentiments in complex text. A voting system was implemented to mitigate this issue. If one annotator failed to choose the correct sentiment, but the others selected the correct sentiment, the label would be revised to reflect the majority. Thus, the final label required at least two matching labels as the ground truth. A significant distinction of our financial dataset is the rule for perfect information in labeling. Annotators were permitted to search for unknown words and concepts; otherwise, the dataset for FinBERT was to disregard external knowledge in detecting market sentiment. This approach ensures the dataset's labels are highly accurate, as our objective is to have a solid ground truth. By leveraging online information, annotators achieved near-perfect information circumstances for precise labeling.

Of the 32,034 FOMC sentences we had extracted from the FOMC transcripts, we chose 1,065 sentences of medium length. This decision is based on the rationale that short sentences often do not contain multiple statements about the same sentiment aspect and often only contain statements about a single aspect. This makes short sentences too easy to classify. Long sentences, however, often explain general guiding principles of the FOMC and do not express any sentiment. Medium sentences were thus the most difficult to classify, allowing for the best estimation of a model's capabilities.

All annotators had at least an master level education in business, finance, or economics degree. The authors provided about half of the labels. None of the sentences were only annotated by the authors. None of the annotators were incentivized to label sentences quickly.

When discussing discrepancies between annotators, we established the following non-trivial definitions of FOMC sentiment across the three aspects. Foreign positive growth is positive for the growth sentiment, even if no direct link to growth in the U.S. can be established. Decreases in government spending are negative for growth. A minor change in an indirectly growth-related indicator, such as domestic debt, is neutral.

Below is an example of a hard sentence (ID 12087 of the dataset) that led to disagreements between the models: "Recent data suggested that growth rates of household spending and business fixed investment had moderated from their strong fourth-quarter readings." VADER: positive, FinBERT: negative, FinBERT-FOMC: negative, GPT-4: neutral, Llama 3: negative. The human annotation of this sentence is "neutral" because a moderation implies a decrease (negative), but from a high level, indicating that growth is still higher than zero, albeit not high. This sentence is one of the most debatable sentences, showing the complexity of the underlying language and the need for high-quality annotations. We believe that our dataset is an important contribution to the scientific community, as it is a useful and hard benchmark component for future research.

4 Methods

4.1 VADER

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a sentiment analysis tool that operates on a rule-based framework to evaluate the sentiment expressed in each text [10]. It analyzes sentiments conveyed through short social media posts (microblogs), reviews, and forum discussions. VADER employs a lexicon of sentiment-related words and a set of rules to determine the polarity (positive, negative, or neutral) and intensity of sentiment in the text. The VADER score is normalized to a range of -1 to 1, where a score close to 0 indicates neutral sentiment. We chose -0.05 and 0.05 as the cutoffs for negative and positive sentiment, respectively.

4.2 FinBERT

FinBERT was proposed by [4] and is a pre-trained transformer language model [39] based on the BERT architecture [7]. The model is designed to capture the nuances and complexities of financial language and terminology, making it suitable for various financial text analysis tasks [21]. The model has been fine-tuned on a large corpus of financial text data, including news articles, corporate reports, and regulatory filings. In this study, a fine-tuned version of the FinBERT language model was utilized, based on [40]. The sentiment score of the FinBERT model is normalized to a range of -1 to 1.

The dataset used to train FinBERT includes Financial PhraseBank [24] and FiQA Task 1 sentiment scoring [22]. While the model was trained on existing financial datasets, including a high agreement level dataset that was further fine-tuned by [40] to reduce noise in the labeled data, the dataset may not be well-suited for analyzing sentiment in FOMC minutes. This is because the FOMC often mentions multiple sentiment aspects in a single sentence, which may have different valences. Furthermore, FOMC sentences may contain opposing statements about a single sentiment.

4.3 FinBERT-FOMC

[6] describes a version of FinBERT that is fine-tuned on 1,375[3] FOMC sentences. In addition to being fine-tuned, FinBERT-FOMC preprocesses the sentences with the Sentiment Focus method (p. 361), which uses conjunctions to detect phrases that are irrelevant to the sentences' sentiment. The results demonstrated an overall improvement of 5% in accuracy over the baseline FinBERT model. In cases of complex sentences containing conjunctions like "but," "while," and "though" with contradicting sentiments, the fine-tuned model outperformed the original FinBERT by a margin of 17.4%.

4.4 GPT-4 and Task-Specific Prompt Engineering

GPT-4 is a transformer-based language model and is considered one of the most capable models to date [1, pp. 7–8]. [19] show that GPT-4 delivers the best language model performance for some financial sentiment analysis benchmarks, as we summarize in section 2.2. Due to being closed-source, details about the model's architecture are not known. We use GPT-4 in this study due to its current status as a state-of-the-art language model despite it being closed-source.

When experimenting with GPT-4 for aspect-based FOMC sentiment analysis, we studied how to ensure that the model output, which can be unstructured text of varying length, can reliably be parsed by our evaluation system. As GPT-4 is a text-to-text model, receiving an unstructured text and returning a structured text requires a prompt that enforces structure on the output. We experimented with several prompts, including the following:

- An instruction to provide the aspect-based sentiment in JSON format with the three keys growth, employment, and inflation. No further explanation or background about FOMC sentiment was provided.
- A description with the information provided in the prompt above, in addition to 14 examples in human-readable form, such as "Lower petroleum prices are good for inflation." This prompt was similar to the instructions the human annotators received (see subsection 3.2).
- No verbal explanation, just eight example sentences followed by the desired JSON output (see B for the full prompt).

When testing the first prompt, we received unparsable responses for the growth sentiment in 11.25% of cases; with the second prompt, the rate dropped to 4.13%; with the third prompt, the rate dropped to 0.0%. We, therefore, used the third prompt for all GPT-4 and Llama 3 experiments in this study.

We also tested the prompt suffix "Let's think step by step" proposed by [16] by allowing the model to output its reasoning in a JSON field called "explanation" to the beginning of the model output. This suffix reduced the GPT-4's ability to reliably provide parsable JSONs with the correct keys. As a result, we did not use this suffix in our experiments despite its merits in other domains (pp. 5–9).

On a side note, GPT-4 and Llama 3 allowed us to calculate the sentiments for the aspects of employment and inflation without much additional effort. Employment and inflation were not predicted by the other models.

4.5 Llama 3 and Task-Specific Prompt Engineering

Llama 3 is a decoder-only transformer language model. Unlike GPT-4, Llama 3 is open-source, making its architecture transparent and making the model available for deployment by anybody with appropriate hardware. Compared to previous versions of Llama [38], Llama 3 has an extended vocabulary with 128'000 tokens and it uses recent architectural advancements such as Grouped Query Attention [3].

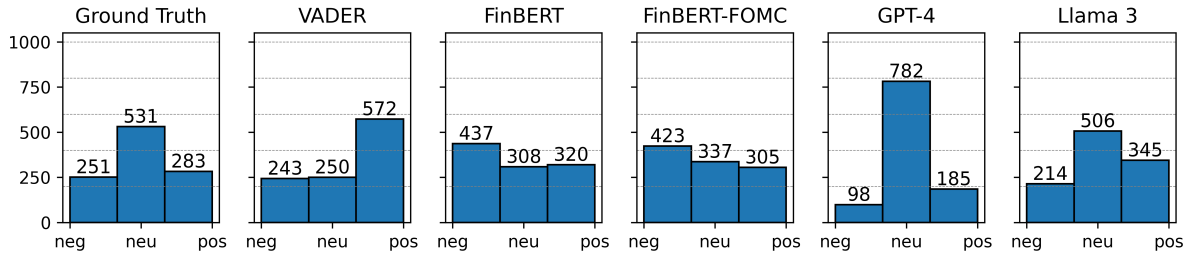
Llama 3 training sequences were 8,192 tokens long, which is much longer than necessary for the task at hand.

We deployed the Llama-3-70B version of the model on a professional-grade NVIDIA DGX-2 GPU cluster. Llama-3-70B suffers from significant degradation from quantization [13], hence we chose the original 16-bit model.

We used the same prompt for Llama 3 as for GPT-4.

[3]The original dataset was in its early stages and labeled by two researchers, and in this paper, we build upon an initial FOMC dataset, extending it to encompass 1,065 sentences.

Figure 1: Frequencies of sentiment labels of the ground truth and all models.



5 Results

5.1 Confusion Matrix

Metric	VADER	FinBERT	FinBERT-FOMC	GPT-4	Llama 3
F1 (Neg)	0.437	0.602	0.620	0.504	**0.744**
F1 (Neu)	0.376	0.574	0.634	0.748	**0.815**
F1 (Pos)	0.508	0.624	0.663	0.628	**0.794**
F1 (Avg)	0.440	0.600	0.639	0.627	**0.784**
Accuracy	0.443	0.597	0.638	0.682	**0.793**

Table 1: Performance Comparison of Language Models.

5.1.1 VADER. VADER is designed to analyze online communication, which often exhibits strong emotions. In contrast to this, the FOMC statements have an unemotional tone, and, unsurprisingly, FOMC statements were predominantly classified as neutral by VADER, with more than 30% of sentences being classified as such.

VADER tends to produce a centered distribution of sentiment predictions, which prompted us to allocate a neutral sentiment interval between -0.05 and 0.05. This adjustment was made to enhance the classification power of VADER and to better align with the *actual* sentiment distribution. After adjusting the interval for the neutral rating, VADER demonstrated a skewed sentiment distribution in favor of more positive predictions.

As can be derived from the confusion matrix in table 5.1.6, VADER achieved a high recall of 0.767 on positive sentences. Overall, however, VADER has a weak performance with an accuracy of only 44.3% and an F1 score of 0.44.

VADER is the weakest model. This is likely due to the emotional text from X (formerly Twitter) that VADER was designed for [10, p. 222]. Simple models have the potential to be competitive in financial classification tasks, as the strong performance of other small models [27] shows when compared to GPT-4 [19, p. 411]. But as VADER is a rule-based model, it is not surprising that VADER performs poorly on the formally written FOMC dataset.

5.1.2 FinBERT. The confusion matrix for FinBERT indicates that the model better detected positive and negative sentiments than neutral sentiments. The model had the highest accuracy in detecting negative sentiments, with 205 (81.7%) true negatives, and the accuracy for positive sentiments was also high, with 233 (82.3%) true positives. However, the model had a low recall in detecting neutral sentiments, only detecting 157 (29.6%) out of 531 neutral sentences.

In the sentence below, which states that historically low unemployment insurance benefits can be a positive signal for the economy, FinBERT classified the sentiment as negative: "The four-week moving average of initial claims for unemployment insurance benefits through mid-October remained near historically low levels." Moreover, the assigned sentiment score of -0.964 by FinBERT in this sentence is not only extreme but also in the opposite direction of the actual sentiment, which is concerning. This finding highlights a potential limitation of FinBERT in accurately capturing the sentiment of financial text data, especially in the context of FOMC minutes. Furthermore, compared to VADER, the FinBERT model determines sentiment more confidently and sometimes in the wrong direction.

Overall, the FinBERT model performed better than VADER in detecting positive and negative sentiments, but its recall could be improved for neutral sentiments.

5.1.3 FinBERT-FOMC. The fine-tuned FinBERT-FOMC model improves in this area, as figure 5.1.6. It is able to detect neutral sentiment more accurately than the original FinBERT model. Its negative sentence accuracy is also better, increasing from 44.1% (FinBERT's negative-sentiment accuracy) to 49.4% (FinBERT-FOMC's negative-sentiment accuracy). Likewise, the recall on negative-sentiment sentences increased from 81.7% to 83.3%. FinBERT-FOMC loses positive sentence recall against FinBERT. The recall drops from 82.3% to 68.9%. Nevertheless, the positive-sentiment prediction accuracy of FinBERT-FOMC (63.9%) is higher than that of FinBERT (54.7%).

Overall, FinBERT-FOMC is superior to FinBERT, with higher F1 scores on negative (from 0.57 to 0.62) and neutral (from 0.45 to 0.63) sentiments and an unchanged F1 score for positive (0.66) sentiment.

5.1.4 GPT-4. While GPT-4's accuracy is the second-highest with 68.2%, it reports neutral sentiment too frequently. GPT-4 only classifies sentences as negative or positive if they have a clear sentiment, leading to good precision scores for negative and positive sentiments. However, GPT-4 is too cautious in its predictions, as it has a low recall for negative and positive sentiments.

As a result, GPT-4 has a lower F1 score for negative sentiments than FinBERT-FOMC (0.50 vs. 0.62). GPT-4, unlike FinBERT-FOMC, reports too many neutral sentiments, leading to a high recall of

neutral sentiment but also to too many false neutral predictions (and simultaneously low recall for positive and negative sentiment). Table 5.1.6 shows that the model reports negative and positive sentiments insufficiently often.

The average F1 score of FinBERT (0.64) is slightly higher than GPT-4's F1 score (0.63), indicating that they have a comparable performance on the FOMC dataset.

Our results are in line with prior comparative research by [19, pp. 410–411], who showed that GPT-4 overall has comparable sentiment analysis capabilities to FinBERT [4] on tasks such as the Financial PhraseBank [24] and the FiQA sentiment task [22].

5.1.5 Llama 3. Llama 3's accuracy is the highest with 79.3%. Llama 3 tends to under-detect negative sentiment. It's prediction distribution closely resembles that of the ground truth, as figure 1 shows.

The average F1 score of Llama 3 is the highest of all models, with 0.78. With an F1 score of 0.65 for negative sentiment, the negative sentiment is Llama 3's weakest sentiment. It has an F1 score of 0.82 for neutral and 0.79 for positive sentiments.

Upon closer inspection of the responses of GPT-4 and Llama 3, we found that Llama 3 is able to provide reasoning for its predictions, while maintaining the ability to follow the output format that we had requested. This is a significant advantage over GPT-4, which often failed to provide reasoning for its predictions when prompted to explain its predictions. For most FOMC sentences, Llama 3 produced an explanation before and after the JSON-like output, which likely gave Llama 3 the edge needed to perform better than GPT-4 [16].

5.1.6 Comparison Between Models. In Summary, VADER achieves an accuracy of 44.3%, FinBERT 59.7%, FinBERT-FOMC 63.8%, GPT-4 68.2%, and Llama 3 79.3%. FinBERT-FOMC's areas under the curve (AUCs) outperform GPT-4's AUCs. Llama 3 clearly stands out as the best-performing model The results show that fine-tuning of a moderately small language model can be valuable even compared to a large foundation language model such as GPT-4, but that there are significant differences among large language models.

The result of VADER, based on a lexicon and rules, had an accuracy of 44.3%, serving as a baseline. Naive guessing would have an expected accuracy of only 33.3%. FinBERT, a model pre-trained on financial data, achieved higher accuracy at 59.7%, indicating the benefits of domain-specific pre-training. Further fine-tuning of FinBERT on FOMC texts resulted in FinBERT-FOMC, which showed a significant increase in accuracy to 63.8%. The much larger GPT-4 and Llama 3 models had accuracies of 68.2% and 79.34%, respectively. However, GPT-4's AUC scores were not superior to those of FinBERT-FOMC, suggesting that larger models do not necessarily perform better in sentiment analysis. Our findings add to the ongoing debate about the trade-offs between large foundation models and smaller, finely-tuned counterparts. Moderate-sized models like FinBERT-FOMC, when optimally adjusted for specific text types, can get close to some large language models.

Human annotators demonstrated superior accuracy compared to machine analysis, ranging from 87.5%–89.6% for the growth sentiment, surpassing the best AI model (Llama 3 with an accuracy of 79.3% and an average F1 score of 0.78). The discrepancy between human and model performance highlights the difficulty presented by the FOMC dataset, which can be attributed to the subtleties of economic language.

		VADER			
		negative (44.4%)	neutral (58.8%)	positive (37.9%)	Total
Truth	negative (43.0%)	**108**	56	87	251
	neutral (27.7%)	116	**147**	268	531
	positive (76.7%)	19	47	**217**	283
	Total	243	250	572	1065

		FinBERT			
		negative (47.4%)	neutral (78.2%)	positive (58.8%)	Total
Truth	negative (82.5%)	**207**	25	19	251
	neutral (45.4%)	177	**241**	113	531
	positive (66.4%)	53	42	**188**	283
	Total	437	308	320	1065

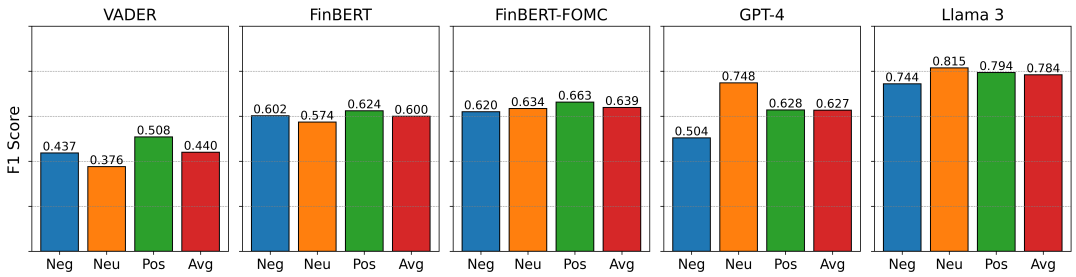
		FinBERT-FOMC			
		negative (49.4%)	neutral (81.6%)	positive (63.9%)	Total
Truth	negative (83.2%)	**209**	29	13	251
	neutral (51.8%)	159	**275**	97	531
	positive (68.9%)	55	33	**195**	283
	Total	423	337	305	1065

		GPT-4			
		negative (89.8%)	neutral (62.8%)	positive (79.5%)	Total
Truth	negative (35.1%)	**88**	159	4	251
	neutral (92.5%)	6	**491**	34	531
	positive (51.9%)	4	132	**147**	283
	Total	98	782	185	1065

		Llama 3			
		negative (80.8%)	neutral (83.3%)	positive (72.5%)	Total
Truth	negative (68.9%)	**173**	61	17	251
	neutral (79.8%)	30	**424**	77	531
	positive (87.6%)	11	24	**248**	283
	Total	214	509	342	1065

Table 2: Confusion matrices for all models. The numbers in parentheses show the recall and the precision.

Figure 2: F1 Scores by Model and Sentiment for FOMC Sentiment Analysis



6 Conclusion

Key observations from the F1 scores of different models for FOMC sentiment analysis indicate that Llama 3 is the most effective, achieving the highest average F1 score of 0.784 with strong performance across negative, neutral, and positive sentiments. GPT-4 also performs well, particularly excelling in negative sentiment with an F1 score of 0.748, though its performance in neutral sentiment is relatively lower at 0.504. Both FinBERT and FinBERT-FOMC show similar and competitive results, with FinBERT-FOMC slightly ahead in the average score (0.639 versus 0.600).

FinBERT-FOMC, in particular, has shown enhanced capability due to its fine-tuning process, outperforming the base FinBERT model. Although it does not reach the accuracy level of Llama 3, FinBERT-FOMC offers a good size-performance trade-off. With its relatively compact architecture, FinBERT-FOMC stands out as a promising candidate for specialized sentiment analysis in the context of FOMC communications compared to GPT-4.

In summary, while Llama 3 outperforms other models, adhering to the principle of 'small is beautiful,' FinBERT-FOMC demonstrates that fine-tuning can yield efficient and cost-effective solutions. It remains beneficial for researchers or institutions to create customized language models tailored to their specific requirements, leveraging the versatility and efficiency of models like FinBERT-FOMC.

7 Limitations and Future Work

The primary limitation is that the approach did not consider inter-sentence contexts due to our methodology of isolating sentences. More attention is required on the richness of contextual information, especially when interpreting statements about economic changes that are often inherently relative. While the FOMC's systematic style of communication somewhat mitigates this issue, future research must develop an approach that takes into account the entirety of the statements, considering the broader economic discourse, to enhance sentiment prediction accuracy.

In future research, constructing more sophisticated FOMC sentiment classifiers should be contemplated. Such models could incorporate an ability to discern the relevance of statements to the domestic economy, distinguishing between comments on internal affairs and those about international contexts. This would necessitate an analytical framework capable of situating individual sentences within the larger economic narrative, recognizing the implications for the U.S. economy.

Recent advances in prompt engineering, especially research on the paradigm of flow engineering [32], are worth being investigated in future research. Additionally, future research should work to overcome existing limitations of chain-of-thought (CoT) prompting concerning reliably parsable output formats (see subsection 4.4 for documentation of our experiments that informed us not to use CoT in this study) and make use of the improvements made possible by CoT [42]. To overcome the limitations of sentence-by-sentence sentiment labels, contextualized approaches to aspect-based sentiment prediction are a promising open field for research on FOMC minutes.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, and Diogo Almeida. 2023. GPT-4 Technical Report.
[2] Miguel Acosta. 2023. A New Measure of Central Bank Transparency and Implications for the Effectiveness of Monetary Policy. *International Journal of Central Banking* 19, 3 (2023), 49–97.
[3] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints.
[4] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv* (2019).
[5] Ellyn Boukus and Joshua V. Rosenberg. 2006. The Information Content of FOMC Minutes. *SSRN* (2006).
[6] Ziwei Chen, Sandro Gössi, Wonseong Kim, Bernhard Bermeitinger, and Siegfried Handschuh. 2023. FinBERT-FOMC: Fine-Tuned FinBERT Model with Sentiment Focus Method for Enhancing Sentiment Analysis of FOMC Minutes. *ACM International Conference on AI in Finance* (2023), 357–364.
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1, Long and Short Papers (2019), 4171–4186.
[8] Taeyoung Doh, Sungil Kim, and Shu-Kuei Yang. 2021. How You Say It Matters: Text Analysis of FOMC Statements Using Natural Language Processing. *Economic Review-Federal Reserve Bank of Kansas City* 106, 1 (2021), 25–40.
[9] Tom Brown et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
[10] Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *International Conference on Weblogs and Social Media (ICWSM-14)*, Vol. 8. 216–225.
[11] Stephen Hansen, Michael McMahon, and Matthew Tong. 2019. The Long-Run Information Effect of Central Bank Communication. *Journal of Monetary Economics* 108 (2019), 185–202.
[12] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-Based Sentiment Analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Mareike Hartmann and Barbara Plank (Eds.). 187–196.
[13] Wei Huang, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. How Good Are Low-bit Quantized LLaMA3 Models? An Empirical Study.
[14] Yu-Lieh Huang and Chung-Ming Kuan. 2021. Economic Prediction With the FOMC Minutes: An Application of Text Mining. *International Review of Economics & Finance* 71 (2021), 751–761.

[15] Jinhyoung Kim and Wonseong Kim. 2024. Advanced Natural Language Processing Analysis on Cross-Border Media Sentiment from China and South Korea. *International Area Studies Review* 27, 1 (2024), 43–56.
[16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems* 35 (2022), 22199–22213.
[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
[18] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. *arXiv* (2019).
[19] Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 408–422.
[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* (2019).
[21] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. FinBERT: A Pre-Trained Financial Language Representation Model for Financial Text Mining. In *International Conference on International Joint Conferences on Artificial Intelligence*. 4513–4519.
[22] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. *Companion Proceedings of the Web Conference* (2018), 1941–1942.
[23] Macedo Maia, Juliano Efson Sales, André Freitas, Siegfried Handschuh, and Markus Endres. 2021. A Comparative Study of Deep Neural Network Models on Multi-Label Text Classification in Finance. In *International Conference on Semantic Computing (ICSC)*. IEEE, 183–190.
[24] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796.
[25] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *Comput. Surveys* 56, 2 (2023), 1–40.
[26] Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir Chitkushev, and Dimitar Trajanov. 2020. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access* 8 (2020), 131662–131682.
[27] Yulong Pei, Amarachi Mbakwe, Akshat Gupta, Salwa Alamir, Hanxuan Lin, Xiaomo Liu, and Sameena Shah. 2022. TweetFinSent: A Dataset of Stock Sentiments on Twitter. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. 37–47.
[28] Matthew Peters, Mark Neumann, Mohit Iyyer†, Matt Gardner†, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2018), 2227–2237.
[29] Alex Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI* (2018).
[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* (2019).
[31] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–7.
[32] Tal Ridnik, Dedy Kredo, and Itamar Friedman. 2024. Code Generation with AlphaCodium: From Prompt Engineering to Flow Engineering. *arXiv* (2024).
[33] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
[34] Ruttachai Seelajaroen, Pornanong Budsaratragoon, and Boonlert Jitmaneeroj. 2020. Do Monetary Policy Transparency and Central Bank Communication Reduce Interest Rate Disagreement? *Journal of Forecasting* 39, 3 (2020), 368–393.
[35] Adam Hale Shapiro, Moritz Sudhof, and Daniel Wilson. 2022. Measuring News Sentiment. *Journal of Econometrics* 228, 2 (2022), 221–243.
[36] Tian-Xiang Sun, Xiang-Yang Liu, Xi-Peng Qiu, and Xuan-Jing Huang. 2022. Paradigm shift in Natural Language Processing. *Machine Intelligence Research* 19, 3 (2022), 169–183.
[37] Raul Cruz Tadle. 2022. FOMC Minutes Sentiments and Their Impact on Financial Markets. *Journal of Economics and Business* 118 (2022), 106021.
[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv* (2023).
[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).
[40] Yifei Wang. 2021. Aspect-Based Sentiment Analysis in Document-FOMC Meeting Minutes on Economic Projection. *arXiv* (2021).
[41] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.
[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
[43] Michael Woodford. 2005. Central Bank Communication and Policy Effectiveness. *National Bureau of Economic Research Cambridge, Mass., USA* (2005).
[44] Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A Unified Generative Framework for Aspect-Based Sentiment Analysis. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2416–2429.

A Disclosure of Direct Costs for the Study

The GPT-4 costs were no higher than $101.98. The costs for the GPT-4 (8k context window) API were $6 per 100k tokens at the time of the experiment (lower today). In addition to the main run, we also ran test experiments to try different prompt strategies. We estimate the total direct costs of the GPT-4 study to be below $300.

Furthermore, we performed inference with FinBERT and VADER on a local machine. The costs for these runs are negligible and are mainly due to consumer-grade hardware (GPU) usage and electricity costs from running the GPU for several hours. We ran Llama-3-70bn on our institute's NVIDIA DGX-2 GPU cluster.

B GPT-4 & Llama3 Prompt

```
Sentence: Imports rose in December, with an increased volume of petroleum imports, but declined in January, driven by lower prices and volumes for petroleum.
Output: { "growth_sentiment": "neutral", "employment_sentiment": "neutral", "inflation_sentiment": "positive" }

Sentence: Commercial and industrial loans on banks' books continued to expand strongly, reportedly in part to fund increased merger and acquisition activity.
Output: { "growth_sentiment": "positive", "employment_sentiment": "neutral", "inflation_sentiment": "neutral" }

Sentence: Available indicators of drilling activity, such as counts of rigs in operation, suggested spending would decline less rapidly in the third quarter.
Output: { "growth_sentiment": "negative",  "employment": "neutral", "inflation": "neutral" }

Sentence: Almost all participants judged that the surprisingly weak May employment report increased their uncertainty about the outlook for the labor market.
Output: { "growth_sentiment": "neutral", "employment_sentiment": "negative", "inflation_sentiment": "neutral" }

Sentence: Although payrolls for state and local governments expanded in July and August, nominal construction spending by these governments declined in July.
Output: { "growth_sentiment": "negative", "employment_sentiment": "positive", "inflation_sentiment": "neutral" }

Sentence: And, even if nominal wages should accelerate somewhat, relatively wide profit margins could buffer the effect on prices of final goods and services.
Output: { "growth_sentiment": "neutral", "employment_sentiment": "positive", "inflation_sentiment": "positive" }

Sentence: The pass-through of the substantial rise in energy prices could account for a considerable part of the step-up in core inflation in recent quarters.
Output: { "growth_sentiment": "neutral", "employment_sentiment": "neutral", "inflation_sentiment": "negative" }

Sentence: Growth of nonfinancial domestic debt was estimated to have slowed a little in the third quarter from the average pace in the first half of the year.
Output: { "growth_sentiment": "neutral", "employment_sentiment": "neutral", "inflation_sentiment": "neutral" }
```