Contextualized Sentiment Analysis using Large Language Models

Christian Breitung*, Garvin Kruthof[−] and Sebastian Müller[+]

January 15, 2025

Abstract

This study explores the capabilities of large language models (LLMs) in employing economic reasoning to predict industry-specific news impacts. Using a large dataset of commodity news headlines, we find that LLMs are able to derive industry-specific sentiments, highlighting their economic reasoning capabilities. Motivated by these findings, we apply GPT-4o-mini to assess the macroeconomic information in monetary policy reports. We find that a portfolio that is long (short) in the best (worst) industry generates a highly significantly positive alpha of up to 110 bps, which may not be explained by a potential look-ahead bias.

Keywords: Contextualized Sentiment, Macroeconomic News, Large Language Models, GPT-4

We want to thank Jochen Hartmann, Christoph Georg Schmidt and David Wuttke for their valuable comments and feedback.

*Technical University of Munich, TUM School of Management, Campus Heilbronn, Center for Digital Transformation, Am Bildungscampus 9, 74076 Heilbronn, Germany (christian.breitung@tum.de, Phone: +49 7131 264 18 822),

[−]Technical University of Munich, TUM School of Management, Campus Heilbronn, Center for Digital Transformation, Am Bildungscampus 9, 74076 Heilbronn, Germany (garvin.kruthof@tum.de),

[+]Technical University of Munich, TUM School of Management, Campus Heilbronn, Center for Digital Transformation, Am Bildungscampus 9, 74076 Heilbronn, Germany (sebastian.mueller.hn@tum.de, Phone.: +49 713 126418806)

1. Introduction

Efficient information processing is essential in various areas like marketing, supply chain management, finance, and accounting. For instance, social media and news analysis can be utilized for studying brand perception in marketing (Hansen et al., 2018; Hartmann et al., 2021), demand forecasting and risk management in supply chain management (Lau et al., 2018; Swain and Cao, 2019), as well as understanding how efficiently markets price new information in finance and accounting (Cohen et al., 2020).

To accomplish this, researchers and practitioners embraced Natural Language Processing (NLP) techniques like sentiment analysis to quantify text information. Early research in finance employed general-purpose dictionaries in this regard (Tetlock, 2007), but Loughran and McDonald (2011) strongly argued for using domain-specific dictionaries to deal with the unique jargon of the corresponding discipline.

While recent applications of fine-tuned machine learning (ML) models for assigning sentiment to news headlines show improved performance over traditional dictionary approaches (Araci, 2019; Yang et al., 2020), both methods share the same simplification: they assume sentiment is homogeneous across all industries within an economy. It is not difficult to find examples where this assumption seems far stretched. For instance, a news headline portraying an oil price increase should positively affect firms in the oil refining industry due to the expected revenue increase but adversely impact airlines as their operating costs increase. Current approaches to sentiment measurement fail to capture these important variations.

In this study, we propose a novel approach to sentiment analysis by leveraging the economic reasoning capabilities of large language models (LLMs). Specifically, we explore how these capabilities can be utilized to derive contextualized sentiments for news headlines. Testing the contextualization ability of various LLMs on a dataset of commodity-related news headlines, we find that the models correctly assess news impacts in over 84% of cases. We then apply GPT-4o-mini to a comprehensive dataset of news headlines covering commodities, currencies, interest rates, and inflation. The results reveal a highly significant correlation between changes in industry-specific sentiment and industry returns, highlighting the effectiveness of GPT-4o-mini in capturing meaningful economic relationships. Moreover, we demonstrate that GPT-4o-mini can be employed to construct portfolios based on industry-specific sentiments derived from monetary policy reports. These portfolios generate significantly positive alphas, showcasing the potential of LLMs to enhance investment strategies through sentiment-driven insights.

Assessing the ability of LLMs to contextualize news sentiment is essential, before using these models for investment strategies. On a firm-specific level, Lopez-Lira and Tang (2023) show that sentiments derived from LLMs may predict future stock returns for individual stocks. The ability of large language models to interpret macroeconomic news, however, is relatively unexplored. Given that macroeconomic news are an essential part of the overall news that may affect stock markets, there is an urgent need to fill this research gap.

We, therefore, assess the accuracy of three smaller LLMs (*falcon-7b, llama-7b, mistral-7b*) and three larger ones (*GPT-4, GPT-4o-mini and GPT-4o*) on a dataset of 1600 news headlines reflecting commodity price increases and decreases. For each of the 40 commodities under study, we create 40 headlines that are linguistically aligned with real-world data through manual oversight, 20 conveying a price increase and 20 conveying a decrease.

The ground truth against which the LLMs' predictions are evaluated is established based on well-founded economic principles. For example, in the context of rising oil prices, it is widely accepted that this scenario exerts pressure on the airline industry due to increased operational costs, particularly kerosene, a substantial component of expenses. Conversely, the same circumstance is favorable for companies in the oil sector, as their profit margins expand with rising prices. We use this dichotomy as a benchmark for assessing the accuracy of the LLMs' economic reasoning. We classify impacts as positive, negative, or neutral. Thereby, we aim to ensure that the assessment of LLMs' performance is based on clear, objective criteria rather than speculative gradations of impact.

Contrary to Zhang et al. (2023), our study aims to evaluate the intrinsic economic reasoning capabilities of LLMs in their *vanilla* state, without resorting to the *few-shot* learning technique or any form of tailored tutoring about economic dynamics. This approach is based on the notion that, having been trained on extensively diverse and rich datasets, contemporary LLMs might already possess a substantive foundation of economic understanding.

Using our generated news dataset, we quantitatively evaluate the capabilities of current LLMs (proprietary models from OpenAI and smaller open-source variants) with respect to deriving industry-specific sentiments. We prompt the language models to provide industry-specific sentiments and then extract the predicted sentiment labels from their responses. Finally, we compare these labels with the previously determined ground truth using performance metrics such as accuracy, precision, and recall scores to gain insights into the strengths and limitations of existing models on generating contextualized sentiments.

We investigate three different research questions. First, we test whether LLMs can contextualize the sentiment of news headlines. Second, we evaluate if the ability of language models differs across topics, which are different commodities in our context. Finally, we explore whether investors efficiently price industry sentiment changes by testing whether there is a significant relation between industry sentiment change and industry stock returns.

Concerning our first research question, we find that all models surpass mono-directional sentiment prediction models. The accuracy of the smaller models (7 billion parameters) ranges between 38.92% and 63.46%, larger than the 33.33% a random allocation would yield on expectation.

Next to these rather small LLMs, larger models like GPT-4 and its variants GPT-4o and GPT-4o-mini perform significantly better. For instance, GPT-4 generates an accuracy of 77.31%. GPT-4o, a newer version performs even better with 84.3%, followed by its smaller variant GPT-4o-mini with 75.74%. These high accuracies suggest that state-of-the-art LLMs are indeed able to contextualize.

Regarding our second hypothesis, we compare the accuracy of LLMs across different commodities. Overall, we find substantial differences across commodities, which aligns with our idea that the knowledge depth might vary across different topics. While we observe an accuracy of up to 95.83% for news headlines portraying price changes in *wheat* if we use GPT-4o, the best value we receive for liquified petroleum gas is 64.17%, as obtained via mistral-7b, the only case where one of the smaller models under evaluation achieves the highest accuracy. Incorporating this observation into the broader understanding of LLMs, it is plausible to posit that less-covered industries or commodities might be underrepresented in the training datasets of these models.

However, while GPT-4o performs best in most cases, there are some commodities where the accuracy is surprisingly low, for instance in case of *crude oil* and *natural gas*. Given that these commodities are one of the most discussed commodities, a lack of training data is unlikely to explain this pattern. An alternative explanation could be that larger models might apply a more complex reasoning, incorporating indirect effects, while smaller models like *mistral-7b* focus more on the direct effect.

Regarding our third hypothesis, we find strong evidence that changes in industry-specific sentiments correlate with industry returns. By running a Fama MacBeth regression, we observe that the difference between the industry sentiment of day t and the average industry sentiment

in the previous week is highly correlated with the stock return on day t. The industry-specific sentiments obtained via GPT-4o-mini thus seem to accurately measure shifts in industry sentiment. These results hold once controlling for overall market sentiment and past stock return changes. We do not find any significant relation between lagged sentiment change and industry returns, suggesting that investors do not overlook important industry-specific effects.

In our final test, we investigate the ability to identify industry-specific impacts using macroeconomic information from the Federal Reserve's Monetary Policy Reports. Employing GPT-4o-mini, we generate scores for various macroeconomic dimensions and evaluate their ability to inform long-short investment strategies. The results reveal that long-short portfolios based on the overall score provided by GPT-4o-mini generate significant five-factor alphas, with monthly five-factor alphas of up to 110 basis points (bps) that are significant at the 1% level.

Robustness tests, including the removal of temporal information and the use of report summaries, suggest that the significantly positive alphas are not driven by look-ahead bias. Further analysis reveals that GPT-4o-mini assigns varying weights to individual macroeconomic dimensions to determine the overall score, with dimensions like consumer sentiment, financial conditions, and governmental fiscal policy receiving the highest weights. These weights exhibit intuitive variations across industries, highlighting the model's ability to consider industry-specific characteristics.

We contribute to the literature of advanced NLP tools in management science in two main dimensions. On the one hand, we shed light on the economic reasoning capabilities of current general-purpose LLMs. On the other hand, we illustrate how LLMs can be instructed to generate industry-specific sentiments for news and macroeconomic reports.

The remainder of this paper is structured as follows. First, we differentiate contextualized sentiment analysis from traditional, coarse-grained sentiment analysis in section 2. Second, we discuss our methodology in section 3 and data in section 4.1. In section 5, we use industry-specific sentiments to construct investment portfolios. We conclude our findings in section 6.

2. Sentiment prediction methods

With the growing amount of textual data, researchers started to embrace sentiment classification methods to study economic questions using textual data. According to Hirschberg and Manning (2015, p. 265), sentiment analysis represents the "identification of positive or

negative orientation of textual language". News headlines that appear to have a positive tone are thus associated with a positive effect on the market according to this definition, whereas negatively toned news are believed to have an adverse effect.

2.1. Traditional sentiment analysis

Traditionally, researchers determined news sentiment using a dictionary approach, where text is classified as positive, neutral, or negative based on the presence of certain words. Early research employed general-purpose dictionaries in this regard (Tetlock, 2007), but Loughran and McDonald (2011) strongly argued for using domain-specific dictionaries to deal with the unique financial jargon.

Despite the straightforward nature and respectable performance, dictionary-based strategies have inherent limitations. Among these is the failure to consider the context of words, leading to possible misunderstandings due to negations or confusion regarding the subject to which a positive or negative term pertains.

Machine learning models offer a promising solution to these issues by considering word context to accurately decipher textual sentiment. For instance, Frankel et al. (2022) illustrate that machine learning models surpass traditional domain-specific dictionaries, a success attributable to these models' nuanced comprehension of word interactions and relations.

Researchers have also examined the performance of more sophisticated deep neural networks such as BERT (Bidirectional Transformers for Language Understanding). This pre-trained model learns bidirectional representations from unlabelled text using an 'attention mechanism'. Its high-dimensional vector output can serve as the foundation for tasks ranging from sentiment prediction to text classification (Devlin et al., 2018). Among others, Araci (2019) investigate to what extent BERT may be used to predict financial news sentiment and find that fine-tuning the model on finance data yields a highly accurate sentiment prediction model that surpasses existing models. Yang et al. (2020) suggest an alternative version of *FinBERT* that is pre-trained on financial communication data before fine-tuning on the financial phrasebank dataset and receive similar results.

Despite their ability to consider the context of words to determine the textual sentiment, also ML-based model can not be used to derive sentiments for different contexts. For instance, positive news for a specific firm may be positive, neutral, or even negative for its competitors, depending on the news content. While a decrease in commodity prices might benefit a company and its competitors, an increase in the market share of one company ultimately leads to a market

share decrease of its competitors.

2.2. Contextualized Sentiment Analysis

We introduce Contextualized Sentiment Analysis (CSA) as the methodology for attributing context-specific sentiments to news articles. We posit that Generative AI (GenAI) models, when provided with a specific context, can discern context-specific sentiments to both firm-specific and macroeconomic news.

The reason is that the underlying architecture of GenAI models is designed to generate contextually coherent and semantically relevant text. Fundamentally, these models are trained on voluminous datasets to predict subsequent words in a given sequence based on the preceding context. Through specialized tasks like masked language modeling, these models iteratively fine-tune their internal parameters, thereby acquiring the ability to identify linguistic patterns (Brown et al., 2020). This supervised learning approach involves input and expected output and facilitates the model's continuous performance enhancement.

In theory, we could also fine-tune models such as those discussed by Frankel et al. (2022) to provide the news sentiment for a specific context, e.g. an industry. However, such an approach is less practicable. The reason is that it not only requires adequately labeled datasets that are costly to obtain, but it also does not scale well, given that we need one model for each context we aim to consider.

Despite the theoretical advantages of GenAI models in contextualizing information, its *CSA* capabilities are yet to be explored. While Lopez-Lira and Tang (2023) offer empirical evidence suggesting that LLMs possess contextualization abilities by correlating stock returns with firm-specific sentiments generated by GPT-3 and GPT-4, the exact mechanisms contributing to enhanced return predictability are not fully understood yet.

Furthermore, LLMs might also excel in contextualizing macroeconomic news next to firm-specific news. For instance, commodity price fluctuations can have heterogeneous impacts on firms, contingent on their business models. We therefore require a more rigorous analysis of the economic reasoning capabilities of LLMs to understand whether they may be used to generate contextualized sentiments.

3. Methodology

3.1. The challenge of identifying the ground truth

Assessing the ability of LLMs to interpret economic news is a complex task. The main challenge arises from the absence of a definitive ground truth, making objective evaluation difficult. In this paper, we tackle this issue from two perspectives.

First, we adopt a direct evaluation approach to assess the contextualized sentiment analysis (CSA) capabilities of large language models (LLMs), utilizing economic reasoning to assess the impact of individual commodity news headlines on different industries. For example, consider a scenario in which aluminum prices surge. At first glance, this may indicate rising production costs for the automotive sector, given the heavy use of aluminum in vehicle manufacturing. In contrast, the agricultural sector appears to be unaffected, as aluminum has no direct connection to crop production.

Nevertheless, a more comprehensive analysis reveals broader interdependencies. For instance, the automotive industry might respond to increased costs by passing them on to consumers or by innovating in response to price pressures, such as adopting alternative materials to reduce production expenses. While aluminum is not directly used in crop production, the agricultural sector depends on machinery, such as harvesters, that incorporate aluminum, suggesting potential indirect cost effects.

Although accounting for indirect effects would ideally enhance the robustness of our analysis, it may introduce additional noise and biases into our ground truth assessment. Thus, our experiments are confined to direct effects. Similarly, we categorize impacts as positive, neutral, or negative, rather than employing a continuous impact measure, to mitigate inconsistencies and avoid introducing unnecessary complexity.

3.2. Hypotheses

Extant research indicates that as large language models (LLMs) are scaled and trained with more data, their comprehension and reasoning capabilities improve significantly (Talmor et al., 2020). Their performance in tasks that evaluate relational understanding and common-sense reasoning reflects the extensive world knowledge embedded within these models, which is derived from their exposure to vast datasets (Petroni et al., 2019; Bosselut et al., 2019). The use of pretrained masking techniques further enhances the contextual understanding of LLMs by enabling them to predict masked tokens based on surrounding context (Devlin et al., 2018).

Additionally, when fine-tuned on domain-specific content, LLMs demonstrate strong proficiency in understanding specialized topics, suggesting their potential applicability in interpreting economic news (Gururangan et al., 2020). These findings collectively suggest that LLMs may possess the contextual processing capabilities to effectively interpret commodity price news in an economic context.

We therefore derive the first hypothesis as:

Hypothesis 1: *Language models can effectively contextualize commodity price news headlines.*

We hypothesize that the contextualization capabilities of language models may vary across different commodities, reflecting disparities in the availability and representation of training data. While large-scale pre-training equips LLMs with a broad base of general knowledge, there may be considerable heterogeneity in the amount of text related to specific topics. Indeed, Rogers et al. (2021) document that language models exhibit uneven performance across diverse language understanding tasks, indicating that they assimilate certain types of knowledge more effectively than others. This finding suggests the existence of potential knowledge gaps in specialized areas with limited data coverage. Similarly, Gehrmann et al. (2021) show that LLM accuracy varies significantly across subjects when evaluated against a comprehensive benchmark, which may be attributable to inconsistent coverage in pre-training datasets. Collectively, this evidence suggests that pre-training corpora may be biased, offering dense coverage of some subjects while providing limited exposure to niche fields, such as those involving rare commodities. Such imbalances could result in knowledge deficiencies that hinder the accurate contextualization of less-represented domains. Accordingly, we anticipate heterogeneity in LLM interpretive performance across different commodities.

Our second hypothesis thus reads as follows:

Hypothesis 2: *The ability of language models to accurately contextualize commodity price news likely varies across different topical areas and domains.*

Under the assumption of market efficiency and given that the contextualization capabilities of current LLMs are sufficient to accurately assess the impact of news headlines on industry performance, we would expect a strong contemporaneous correlation between changes in industry sentiment and industry returns. Specifically, an increase in industry sentiment would indicate favorable news, which should prompt investors to purchase stocks within the corresponding industry, thereby resulting in an increase in industry returns.

Our third hypothesis thus reads as follows:

Hypothesis 3: *There is a significant relation between industry sentiment change and industry stock market returns.*

4. Data

We obtain macroeconomic news headlines[1] from Ravenpack news headlines, encompassing the period from 2000 to December 2022, with access provided through the Wharton Research Data Services (WRDS) platform. We then apply a named entity recognition model [2] to discern and eliminate remaining firm-specific news headlines. Based on this dataset, we then construct two datasets.

4.1. Commodity news dataset

On the one hand, a dataset of commodity price news headlines to directly evaluate the contextualization abilities of LLMs in the context of commodity price changes. On the other hand, we use more than 600,000 macroeconomic news headlines to estimate the correlation between industry-specific sentiments and stock market returns.

Regarding the commodity news headline dataset, we apply filters to narrow our selection to commodity news using the *entity types* column. From this subset, we further refine our collection by discarding any headlines with a relevance score below 95 out of 100. This metric signifies the extent to which the headline is pertinent to that specific entity. We also exclude news that lack significant content, as indicated by values below 0.05 for the Composite Sentiment Scores (CSS) or the News Impact Projections (NIP). We further refine the selection based on headlines suggesting a commodity price change, employing keywords indicative of a price increase or decrease.

Finally, for every news headline, we generate two analogously styled news headlines that convey a commodity price increase or decrease respectively using the open-source *Falcon-7B* GenAI model.[3]

The reason we do not use original news headlines is twofold. On the one hand, we cannot share Ravenpack news headlines with external providers since Ravenpack does not allow sharing

[1] We do not use entire news articles because they regularly contain multiple price changes, which complicates determining a clear ground truth.

[2] We utilize the *en_core_web_trf* named entity recognition model (NER) provided by the Python package *SpaCy*.

[3] We select this model because it is the least resource-intensive open-source GenAI model we could access.

data with third party providers. Thus, we could not apply proprietary LLMs operated by third parties (e.g., GPT-4) to these news headlines. On the other hand, some of the news headlines do not necessarily convey a clear direction of the commodity price change, which complicates the determination of ground truth. Our procedure instead ensures that we steer the direction of the commodity price change, allowing us to determine the ground truth. Furthermore, it ensures that the generated news headlines align with real-world data by language models from external data providers.[4]

To ensure a diverse and equally distributed number of headlines for each commodity, we randomly select 40 news headlines for each commodity, equally split across commodity price increases and decreases. Commodities with less than 40 unique news headlines are omitted from the analysis. In total, our dataset comprises 1600 news headlines for 40 commodities. To better understand our dataset, we provide an overview of exemplary commodity news headlines in Table 1.

[Table 1 about here.]

To determine the true industry-specific sentiments of our news headlines, we consider that industries involved in the production of the commodity should benefit from a price increase, and industries that use it as a preliminary product should be adversely affected due to an increase in production costs. Industries that neither produce nor use the commodity or are exposed to strong indirect effects should be unaffected. Following this argumentation, we manually assign industry names[5] to each commodity and derive the true industry-specific sentiment accordingly. We provide an overview of all commodities and associated industries in Table 2.

[Table 2 about here.]

4.2. Macroeconomic news dataset

We filter the initially collected macroeconomic news headlines to include only those related to commodities, the US Dollar, or the United States more broadly. Additionally, we incorporate news headlines that mention "inflation" or "interest rate." To exclude headlines that merely

[4]Note that we manually omit those news headlines that fail to convey the designated commodity price change. Neglecting to do so could lead to a downward bias in our performance metrics if the created news does not align with the assigned label.

[5]Note that the assigned industry names do not necessarily coincide with known NAIC or SIC industry classifications.

report new information (e.g., a new copper price) without providing additional context, we refine the dataset to include only those containing at least one verb or adjective, leveraging SpaCy's natural language processing capabilities. Each news headline is considered only once; any repeated headline appearing later in the sample period is excluded to prevent incorporating stale information that has likely already been priced into the market. This filtering process results in a final sample of 621,132 unique news headlines.

To mitigate potential look-ahead bias, we shift any news released after the New York Stock Exchange (NYSE) closing time to the subsequent trading day. The same procedure is applied to headlines released on weekends or national holidays to ensure that no information is omitted when calculating sentiments.

Finally, we paraphrase the Ravenpack headlines using a locally hosted language model (Mistral-7B) prior to obtaining sentiment scores to address licensing restrictions that prevent sharing the original data with third parties. After careful review, we find that the paraphrased headlines maintain high quality and accurately capture the meaning of the original content. While minor deviations in meaning may occur, this would primarily introduce noise, leading to more conservative estimates of the correlation between industry sentiment and industry returns.

5. Contextualization Abilities of LLMs

5.1. Anecdotal Evidence

Commodity news headlines news may have substantially different implications across industries. For instance, some industries might be involved in mining, drilling, or growing commodities and, therefore, profit from increased commodity prices. On the other hand, some industries heavily rely on commodities as preliminary products so any price increase will have a negative effect. If a model can identify these patterns, we may conclude that it possesses sufficient economic understanding and therefore may be used to interpret commodity price news.

[Table 3 about here.]

Let us illustrate this argument by considering some exemplary news headlines as well as the implications for the industries "technology", "automotive", and "renewable energy" in Table 3. The first sentence, "oil prices plummet amidst oversupply", should have a positive direct effect on the automotive industry, as customers might be more willing to buy a car if oil prices are low, given the lower operational costs. Depending on the reason for the oversupply, we might

also observe indirect effects. For instance, an excess supply of oil may be indicative of a macroeconomic recession, which in turn could dampen consumer demand for automobiles. However, given the absence of specific information pertaining to the causal factors of the oversupply, such indirect effects are not incorporated into our present analysis.

In contrast, the renewable energy industry is likely to be negatively affected, as lower oil prices make the transition from fossil fuels to renewable energy less attractive. Simultaneously, there should be no direct effect on the technology industry. There are also news headlines that do not directly affect any of the three industries, for instance, a decrease in coffee prices.

However, traditional sentiment prediction models, such as a dictionary approach, may not be used to approximate these industry-specific effects.

5.2. Commodity News Analysis

5.2.1. Model selection and evaluation metrics

We prompt various GenAI models to assess the impact of a news headline concerning different industries. More specifically, we evaluate the three 7 billion parameter models Falcon-7b (Penedo et al., 2023), LLaMA-2-7b (Touvron et al., 2023) and mistral-7b (Jiang et al., 2023). In addition, we evaluate three large models from OpenAI, GPT-4 (OpenAI, 2023) and its more recently introduced variants GPT-4o and GPT-4o-mini. We use a direct prompting strategy that reads as follows:

How does the news headline affect the profit of firms within the {*industry*} *industry? Answer with 'increase' if profits increase, 'decrease' if profits decrease or 'unaffected' if profits are unaffected. News headline: '*{*head*}*'*

Since some language models do not return a clear label in rare cases, even when instructed likewise, we remove certain parts from the model's responses before extracting the label. First, we remove the original headline if it is present in the response. Second, we delete any statements repeating the commodity price change.

We then extract the sentiment from the models' responses and compare the industry-specific sentiment to the labels assigned during the construction of the dataset. In this context, we calculate accuracy, precision, and recall scores. Accuracy refers to the overall proportion of correct predictions, precision measures the correctness of positive predictions, and recall evaluates the model's ability to detect all positive instances. This process allows us to quantitatively assess each model's capability in inferring industry-specific impacts. The results will highlight

the strengths and limitations of current LLMs concerning nuanced financial sentiment analysis tasks.

To identify whether performance differences are significant, we first calculate the accuracy on the commodity level and then compare the distributions of these forty accuracies across models or inferences.

5.2.2. Contextualization ability

We calculate multiple evaluation metrics using our benchmark query for the six models (falcon-7b, llama-2-7b, Mistral-7b, GPT-4, GPT-4o-mini, and GPT-4o). Next to the accuracy across all three classes of positive, negative, and unaffected industries, we also calculate class-wise precision, recall, and f1 scores. By doing so, we should be able to identify key strengths of the different models.

[Table 4 about here.]

According to Panel A of Table 4, all tested language models achieve an average accuracy significantly exceeding 33.33%, which serves as the benchmark for context-unaware sentiment prediction models.[6] However, there exists a notable performance disparity among the smaller models. For example, Falcon-7B attains an accuracy of 38.92%, whereas the similarly sized Mistral-7B achieves 63.46%. This illustrates the rapid advancements in the field of natural language processing (NLP), considering that Falcon-7B was released on June 20th, followed by Llama-2-7B on July 18th, and Mistral-7B on September 27th.

All GPT-4 variants, however, significantly outperform even the best-performing smaller models. We observe accuracies of approximately 75% for GPT-4o-mini and up to 84% for GPT-4o.

To assess whether the models are more accurate in identifying news headlines with positive, negative, or neutral effects, we present the precision, recall, and F1 scores across the different classes in Panel B. For Mistral-7B, the precision scores are 56.49% for the positive impact class and 63.68% for the negative impact class. In contrast, GPT-4o-mini exhibits higher accuracies for the positive class and lower for the negative class. GPT-4o performs best, achieving approximately 82% precision in each class. Notably, all models except Falcon-7B attain their highest precision scores for the unaffected class.

[6] Given that one industry is positively affected, one is negatively affected, and another is unaffected by construction, any prediction will be correct precisely one out of three times.

Recall scores further underscore the models' effectiveness in classifying the neutral category, with GPT-4 leading at 97.29%. However, analyzing precision and recall scores in isolation may not provide a complete picture. Therefore, we also compute the F1 scores (Panel C), which offer a balanced assessment of precision and recall. Here, GPT-4 achieves the highest F1 score for the unaffected class at 95.48%, while GPT-4o attains the highest F1 scores for positive and negative news.

Overall, these results support our first hypothesis that large language models (LLMs) can be effectively utilized to determine the impact of commodity price news, particularly in the case of the largest LLMs.

5.2.3. Differences across commodities

We evaluate our second hypothesis that the performance of different language models might vary across different commodities. We therefore calculate the accuracy scores for the different models for each individual commodity.

[Table 5 about here.]

According to Table 5, we indeed observe substantial differences in the accuracy across the different commodities. On the one hand, we observe accuracies up to 97.5% for news related to *wool*. On the other hand, the best score we observe for *crude oil* is 66.67%. This is surprising given that oil prices are among the most discussed commodities in the financial market and therefore not in line with our initial hypothesis that LLMs perform better on news dealing with content that is more frequently covered in the training data.

Instead, we propose that the discrepancies in accuracy across various commodities may be attributable to indirect effects that language models consider when evaluating the sentiment of news headlines. Specifically, fluctuations in the price of *crude oil* exert an impact across multiple industries, whereas variations in *wool* prices predominantly influence the food processing sector. This hypothesis is supported by the high accuracy rates observed for commodities such as *sugar and cotton*, contrasted with the relatively low accuracy rates in commodities like *coal*.

5.3. Capital Market Analyses

5.3.1. News headlines

We investigate the efficiency of financial markets by applying contextualized sentiment analysis to a dataset of paraphrased news headlines spanning from 2000 to 2022.

We instruct the language model GPT-4o-mini[7] hosted by OpenAI to generate industry-specific sentiments between -5 and 5, where larger (smaller) values indicate more positive (negative) industry sentiment, and 0 if the model believes the news headline does not affect the stock market at all. The exact prompt template is as follows:

Imagine you are a stock market analyst. Rate the impact of the attached hypothetical news headline on the $\{impact - entity\}$ *industry from -5 to 5. Negative values indicate a negative impact, positive values indicate a positive impact, and zero indicates no significant impact. An impact is not significant if it is expected to have no immediate effect on the stock prices. Only provide the score, do not add any comments. Headline:* $\{headline\}$

We classify the news headlines as "hypothetical" within the prompt for three primary reasons. First, this approach prevents the model from declining to respond to the query due to internal content moderation guidelines, a behavior observed in certain language models. Second, it mitigates the potential for the model to rely on internal knowledge regarding the publication date of the headline, thereby reducing the risk of introducing look-ahead bias into our predictions. Third, the news headlines are paraphrased versions of original texts and may not correspond exactly to publicly available headlines. By framing the headlines as hypothetical, we ensure that the model treats each headline as an independent instance, isolated from any broader contextual or temporal information it might possess.

We adopt the industry classifications from the 48-sector categorization proposed by Fama and French, excluding the residual category ("Other"), which encompasses all remaining stocks. As a result, we obtain 47 distinct industry sentiments for each news headline in our dataset. For each industry, we remove any news headline that receives a sentiment score of zero, as such headlines are unlikely to produce a significant impact on market valuations. To demonstrate the frequency distribution of each sentiment class across industries, we provide a corresponding boxplot in Figure 1.

[Figure 1 about here.]

As we can see, extreme values -5 and 5 are least often allocated, which is in line with the expectation. However, it seems that there are more extremely negative news headlines in the dataset than positive ones. Another observation we make is that most news headlines are

[7] Although GPT-4 overall performed best, we decided to apply the substantially cheaper model GPT-4o-mini, given the large number of industry sentiments we needed to obtain.

classified as having a moderately negative impact (class -2). Roughly 125,000 news headlines out of the more than 600,000 industry sentiments are allocated to this class.

To obtain a market sentiment, we apply a highly similar prompt that looks as follows:

Imagine you are a stock market analyst. Rate the impact of the attached hypothetical news headline on the stock market from -5 to 5. Negative values indicate a negative impact, positive values indicate a positive impact, and zero indicates no significant impact. An impact is not significant if it is expected to have no immediate effect on the stock prices. Only provide the score, do not add any comments. Headline: {*headline*}

To determine whether industry-specific sentiments can explain industry returns, we run a Fama-MacBeth regression. We regress equally (value) weighted industry returns from Fama & French on day t against the change in industry sentiment between time t and *t-1*. We also implement two additional definitions where we compare sentiment to the average sentiment over the previous five (twenty) days to reduce noise in the benchmark sentiment. If the calculated average industry sentiments correctly capture the news content available on a given day, we should observe a significantly positive relationship between changes in industry sentiment and industry returns.

[Table 6 about here.]

According to column (1) in Table 6, we observe a highly significant relationship between sentiment change and industry return. The coefficient is 0.04, with a t-statistic of 5.10. A similar result is found when using value-weighted returns (0.03), as shown in column (4). We also observe a small positive effect for general market sentiment, derived by instructing GPT-4-mini to assess the impact of the same news headlines on the overall stock market. However, the effect is economically smaller, with a coefficient of 0.01. There also appears to be a positive correlation with lagged stock returns, although it is only slightly statistically significant.

The effects are more pronounced when sentiment change is calculated as the difference between today's sentiment and the average over the previous five days, where we observe a coefficient of 0.08 (0.06) with a t-statistic of 7.3 (5.16) for equally (value) weighted returns. This makes intuitive sense, as the measure is overall more robust. Further increasing the rolling window for sentiment comparison does not seem to increase the coefficients. In fact, we observe slightly smaller coefficients when using the average over a 20-day period as the reference. Overall, these findings suggest that the obtained industry-sentiments are accurate, given our initial hypothesis that a high correlation between industry returns and sentiments.

Industry-specific sentiments thus appear to accurately capture the overall news impact for individual industries. A question that arises is whether investors fully price the news on the first day or whether lagged sentiment changes can still predict stock returns. Therefore, we repeat the analysis using lagged industry sentiment changes to predict the next-day industry stock return.

[Table 7 about here.]

According to Table 7, we do not find evidence that investors systematically overlook value-relevant industry information. If we consider the lagged industry sentiment, we do not find a significant relation with the equally or value weighted stock return. It seems that investors overall efficiently price the information provided.

5.3.2. Macroeconomic reports

The Federal Reserve Board regularly publishes a *Monetary Policy Report*, which provides the U.S. Congress with an in-depth overview of the Federal Reserve's monetary policy decisions, economic conditions, and future economic outlook. The report is released semiannually and a dataset of past reports is publicly available, reaching back to 1996. In contrast to the dataset of news headlines, it contains an in-depth analyses of various macroeconomic factors, from inflation to consumer sentiment. Using their economic reasoning capabilities, LLMs should be able to assess the impact of this macroeconomic information on individual industries.

To evaluate the ability of large language models (LLMs) to infer industry-specific effects from lengthy macroeconomic reports, we analyze 56 reports published between 1996 and 2024. After downloading the reports in PDF format, we automatically extract their text and instruct GPT-4o-mini to assess the industry-specific impacts. In addition to providing an overall score, the models are tasked with generating scores for specific topics, including inflation, commodity prices, employment, consumer spending, housing prices, financial conditions, fiscal policy, and external trade.[8]

Using the scores generated by the LLMs, we construct a portfolio strategy that takes a long position in industries identified as most positively affected and a short position in industries deemed most negatively affected. The investment begins one day after the publication of each report and is adjusted one day after the next report's release. The performance of this strategy

[8]The detailed prompt used for this analysis is provided in the appendix.

is evaluated using the Fama-French five-factor model (Fama and French, 2015) to assess risk-adjusted returns.

To account for potential randomness in the LLMs' responses, we repeat this process ten times. For each iteration, the portfolio comprises long and short positions based on the industries selected as best (long) and worst (short). The final portfolio returns are averaged across these ten iterations.

To mitigate the risk of look-ahead bias when evaluating the original monetary policy reports using GPT-4o-mini, given its knowledge cutoff in December 2023, we employ a two-step approach to generate alternative versions of the reports. These alternatives aim to convey the key economic insights without including temporal information.

On the one hand, we instruct GPT-4o-mini to mask and exclude all relevant temporal markers in the reports. Specifically, the model replaces dates, years, and other time indicators with placeholders and removes explicit references to events or policies tied to specific timeframes, ensuring that the core economic insights remain intact.[9]

On the other hand, we direct the model to generate a concise summary of the reports, capped at 5,000 tokens, while also masking all temporal information. This summarization process is designed to distill the primary economic messages of the reports without introducing temporal dependencies.

Portfolios constructed based on the evaluations of these alternative, temporally-neutral reports should avoid the risk of look-ahead bias, provided that the model consistently and effectively eliminates all temporal markers. This approach ensures that the insights derived from the reports are independent of any future knowledge embedded in the model, thereby maintaining the integrity of the evaluation process.

[Table 8 about here.]

Panel A in Table 8 shows the five-factor alphas of long-short portfolios on the basis of the original monetary policy reports. A portfolio that is long (short) in the best (worst) industry according to the overall score allocated by GPT-4o-mini (SCORE) yields a positive monthly five-factor alpha of 110 basis points (bps) which is significant at the 1% level. Expanding the strategy to include the top three (five, ten) industries reduces the alpha to 79 (68, 49) bps per month but maintains highly statistical significance at the 1% level. It thus seems as

[9]The detailed prompt for this masking process is provided in the appendix.

if GPT-4o-mini is able to assess the impact of the current macroeconomic situation on the individual industries. When considering only individual dimensions to identify the best and worst industries, we also observe significantly positive alphas for some dimensions, although they are typically smaller. While long-short portfolios based on individual dimensions such as the housing market do not generate significant alpha, there are also some scores that can be used as a sorting variable to construct löng-short portfolios that generate significant alpha. For instance, the governmental fiscal policy (GOVT) and the external trade (TRADE) dimensions tend to produce portfolios with alphas up to 102 bps that are significant at the 1% level.

To ensure that the results are not driven by a look-ahead bias, we calculate alphas for long-short portfolio based on scores obtained for reports where temporal information has been removed in Panel B. A portfolio that is long (short) in the best (worst) industry generates a monthly five-factor alpha of 124 bps, which is highly significant at the 1% level. Investing in the best (worst) three (five, ten) industries instead generate alphas of 60 (44, 38) bps monthly five-factor alpha, which are highly significant at the 1% level. These results suggest that a look-ahead bias is unlikely to fully explain the significant alpha, as we should observe lower alphas when deleting information that may help the model to effectively identify the time horizon the report was published. This time, external trade, inflation and financial conditions seem to be most relevant, as they generate highly significant alphas for different numbers of industries.

Panel C shows the alphas for long-short investments based on report summaries without temporal information. Overall, the observed alphas are relatively similar to those obtained when using the masked report. The observed alphas range from 118 bps to 35 bps for investments into the top one and ten industries. This is in line with the idea that the language model effectively extracts the most relevant information from the monetary policy report. Again, portfolios constructed on the basis of individual scores underperform portfolios based on the overall score, highlighting the importance of considering various macroeconomic dimensions.

A question that arises is how GPT-4o-mini weights the individual dimensions to form its overall score. To investigate this, we run Non-Negative Least Squares (NNLS) regressions of the overall score on the eight individual scores and normalize the weights. By doing so, we can estimate the weights the model assigns to the individual subcomponents when constructing the overall score.

[Figure 2 about here.]

Figure 2 presents the results of the NNLS regressions for each industry in the form of a

heatmap. Notably, some dimensions seem to receive higher weights than others. The first row of the heatmap, which displays the estimated weights for the eight subscores, reveals that housing prices are largely disregarded when the model constructs the overall score. In contrast, consumer sentiment, financial conditions, and governmental fiscal policy receive the highest weights.

Substantial differences in weighting are evident across industries. For example, in industries such as *Food products, Personal Services, Restaurant and Retail*, the overall score is strongly correlated with consumer sentiment scores. Conversely, consumer sentiment plays a minimal role in the defense industry, aligning with intuitive expectations. Similarly, in the trading sector, the overall score exhibits a high correlation with financial conditions.

While the assigned weights align with intuition, it remains unclear whether GPT-4o-mini's weighting of the individual dimensions is optimal. To address this, we compare the alpha derived from these weights with alphas from portfolios constructed using randomly initialized weights for the subscores.

[Figure 3 about here.]

Figure 3 depicts the distribution of five-factor alphas for long-short portfolios constructed using scores generated by randomly weighting the individual subscores (INFL, COMM, EMPL, CONSSENT, HOUSE, FIN, GOVT, and TRADE) obtained from GPT-4o-mini. The analysis includes two variants of the monetary policy reports: the original version and a modified version with masked temporal information.

A common feature across the distributions is that their mean alpha is consistently greater than zero. This indicates that the individual subscores contain some degree of informative value, aligning with our earlier findings. Additionally, the alphas obtained using the overall score provided by GPT-4o-mini tend to be larger than the majority of those derived from randomly weighted scores. This observation suggests that GPT-4o-mini's weighting of the subscores may outperform random weighting, although the difference is not statistically significant.

6. Conclusion

This study explores the potential of large language models to derive the implications of commodity price news on different industries.

In a first step, we construct a dataset of commodity news headlines and assign industry-specific sentiments using economic reasoning. We then evaluate the performance of smaller open-source and larger proprietary LLMs on this dataset.

Most importantly, we observe that all language models under study seem to be able to derive industry-specific impacts from commodity price news to some extent. We observe the highest performance for the largest models. GPT-4o achieves the highest average accuracy of 84.30% across all commodities. These findings suggest that an extensive knowledge base enables a model to correctly interpret the impacts of commodity price news on other companies. Furthermore, it seems that the models are best in identifying industries that are unaffected by certain commodity news headlines. Using GPT-4, we achieve a f1 score of 95.48% in this regard. We also evaluate how strongly sentiment predictions vary across different topics and obtain substantially lower accuracy for a few commodities.

Motivated by these findings, we apply the LLMs to capital market data. On the one hand, we find that the difference between industry sentiment on day t and the sentiment change over the previous week is strongly associated with stock returns on day t. The industry-specific sentiments, derived from GPT-4o-mini, appear to effectively capture shifts in industry sentiment. On the other hand, we apply GPT-4o-mini to monetary policy reports from the FED to obtain industry specific assessments. Going long (short) in the industries with the best (worst) score generates highly significant five-factor alpha of 110 bps. We do not find evidence that this effect could be driven by a look-ahead bias, as dropping temporal information from the report does not lead to significant reductions in alpha.

Overall, our findings suggest that LLMs possess economic reasoning capabilities that can be leveraged to automatically derive contextualized-sentiments.

References

Araci, D., 2019. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 .

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y., 2019. Comet: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317 .

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Cohen, L., Malloy, C., Nguyen, Q., 2020. Lazy prices. The Journal of Finance 75, 1371–1415.

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 .

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Frankel, R., Jennings, J., Lee, J., 2022. Disclosure sentiment: Machine learning vs. dictionary methods. Management Science 68, 5514–5532.

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P.S., Anuoluwapo, A., Bosselut, A., Chandu, K.R., Clinciu, M., Das, D., Dhole, K.D., et al., 2021. The gem benchmark: Natural language generation, its evaluation and metrics. arXiv preprint arXiv:2102.01672 .

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 .

Hansen, N., Kupfer, A.K., Hennig-Thurau, T., 2018. Brand crises in the digital age: The short-and long-term effects of social media firestorms on consumers and brands. International Journal of Research in Marketing 35, 557–574.

Hartmann, J., Heitmann, M., Schamp, C., Netzer, O., 2021. The power of brand selfies. Journal of Marketing Research 58, 1159–1177.

Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349, 261–266.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023. Mistral 7b. arXiv preprint arXiv:2310.06825 .

Lau, R.Y.K., Zhang, W., Xu, W., 2018. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. Production and Operations Management 27, 1775–1794.

Lopez-Lira, A., Tang, Y., 2023. Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619 .

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

OpenAI, 2023. Gpt-4 technical report.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J., 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S., 2019. Language models as knowledge bases? arXiv preprint arXiv:1909.01066 .

Rogers, A., Kovaleva, O., Rumshisky, A., 2021. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics 8, 842–866.

Swain, A.K., Cao, R.Q., 2019. Using sentiment analysis to improve supply chain intelligence. Information Systems Frontiers 21, 469–484.

Talmor, A., Elazar, Y., Goldberg, Y., Berant, J., 2020. olmpics-on what language model pretraining captures. Transactions of the Association for Computational Linguistics 8, 743–758.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62, 1139–1168.

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

Yang, Y., Uy, M.C.S., Huang, A., 2020. Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097 .

Zhang, B., Yang, H., Liu, X.Y., 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. arXiv preprint arXiv:2306.12659 .

7. Appendix

Monetary Policy Report Prompt You are an expert in economics and industry analysis. Your task is to analyze a macroeconomic report and assess the implications of its findings on various industries. Do not provide comments or explanations, simply return the json file as indicated below.

Input: 1. The macroeconomic report (text content provided): report 2. A list of industry names: ['Agriculture','Aircraft','Apparel','Automobiles and Trucks','Banking','Beer & Liquor','Business Services','Business Supplies','Candy & Soda','Chemicals','Coal','Communication','C Materials','Construction','Consumer Goods','Defense','Electrical Equipment','Electronic Equipment','Entertainment','Fabricated Products','Food Products','Healthcare','Insurance','Machine and Control Equipment','Medical Equipment','Non-Metallic and Industrial Metal Mining','Personal Services','Petroleum and Natural Gas','Pharmaceutical Products','Precious Metals','Printing and Publishing','Real Estate','Recreation','Restaurants, Hotels, Motels','Retail','Ru and Plastic Products','Shipbuilding, Railroad Equipment','Shipping Containers','Textiles','Tobacco Products','Trading','Transportation','Utilities','Wholesale']

Output: Generate a JSON file where: - Each industry is a key. - For each industry, provide: - An overall score from -1 (very negative) to +1 (very positive), representing the general impact of the macroeconomic conditions on the industry. - Scores for the following economic subtopics, using the same scale (-1 to +1): - Inflation Trends: General trends in inflation, excluding food and energy. - Commodity Prices: Focus on energy, oil, and agricultural products. - Employment/Unemployment: Labor market trends, wage growth, and participation rates. - Consumer Spending and Sentiment: Trends in durable and nondurable goods consumption. - Housing Market: Mortgage rates, housing starts, and real estate trends. - Financial Conditions: Interest rates, credit availability, and debt servicing. - Government Fiscal Policy: Federal and state budget impacts. - External Trade: Trade deficits and international demand for goods and services. - A justification for each score in 1-2 sentences, explaining how the subtopic affects the industry. - No further comments or explanations, simply return the json file.

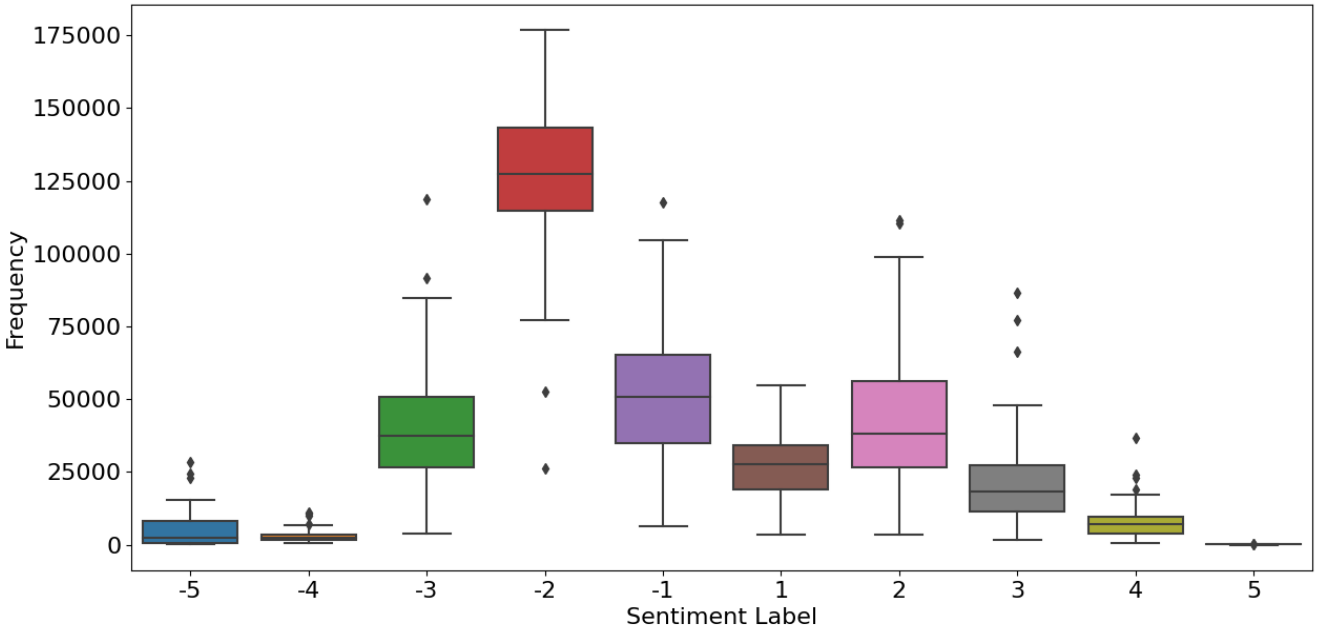
Removing Temporal Information You are an expert in text preprocessing and temporal bias removal. Your task is to modify the provided text of a macroeconomic report to remove all temporal references or phrases that could reveal the time of its origin or the

historical context it describes.

Follow these steps: 1. Replace all explicit dates, years, and time markers with a placeholder. 2. Remove all names with a placeholder. 3. Replace relative temporal indicators with neutral terms (e.g., replace ”recent,” ”current,” ”next year,” or ”last quarter” with ”period under analysis”). 4. Eliminate any mentions of events or policies, clearly tied to a specific timeframe (e.g., ”Full Employment and Balanced Growth Act of 1978” should become ”the relevant legislative act”). 5. Avoid removing substantive economic insights; focus on neutralizing temporal context only. 6. Ensure that the text remains coherent and meaningful for analysis.

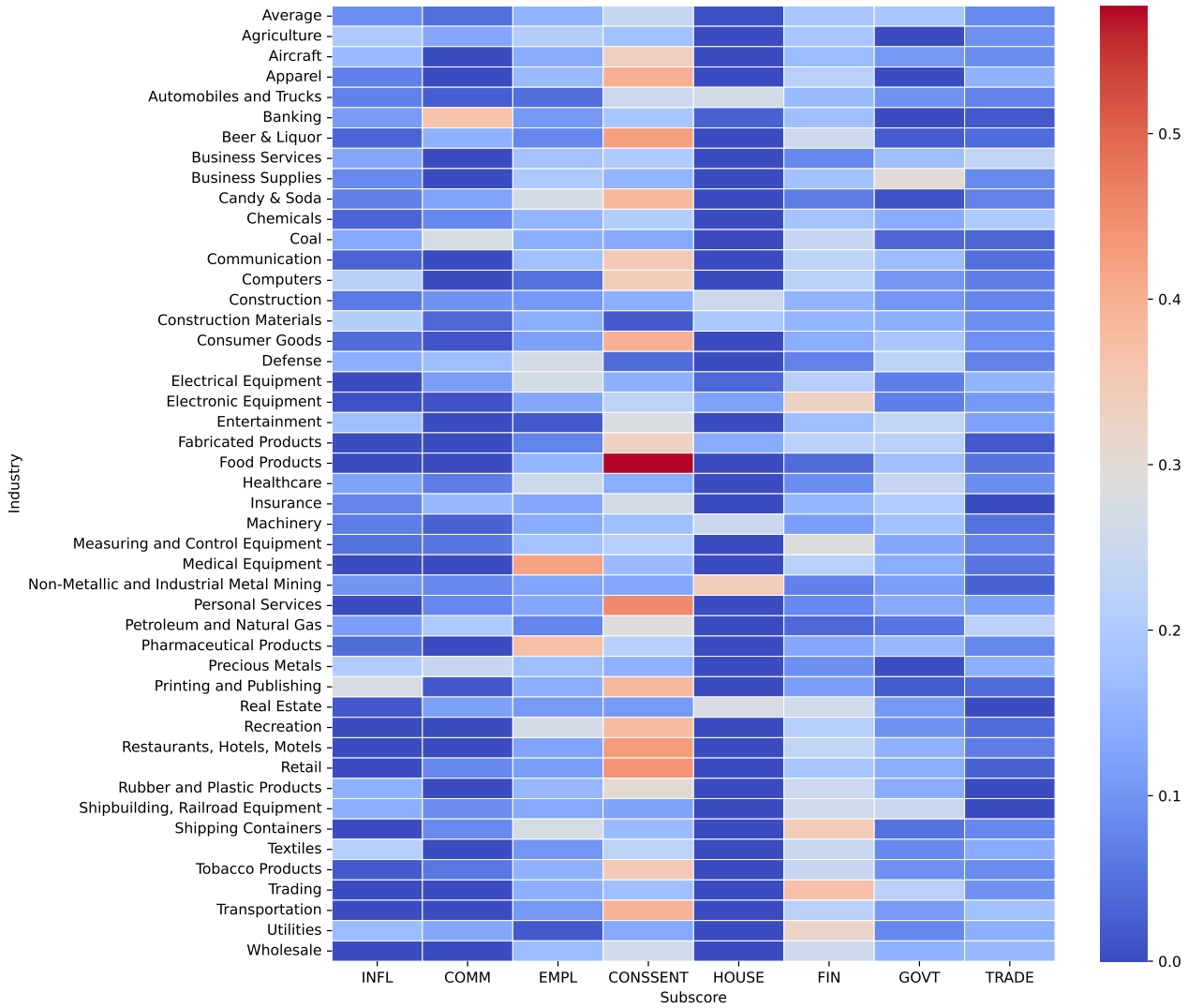
Output: Return the modified text with all potential look-ahead biases removed, ensuring the processed content is neutral with respect to its temporal origin. Just return the text, do not add any comments.

Figure 1: Industry-specific sentiment distribution



This figure contains a boxplot for each of the sentiment labels allocated by GPT-4o-mini. It thus provides an overview how often each industry label is allocated to any of the news headlines. Each boxplot is based on 47 frequency values, one for each industry.

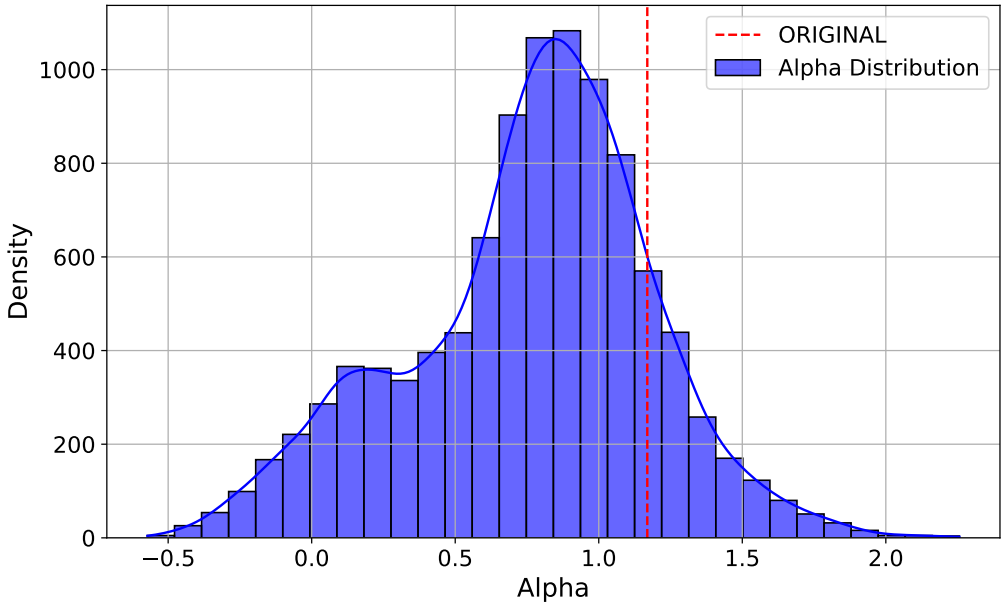
Figure 2: Estimated relevance of subtopics for different industries according to GPT-4o-mini



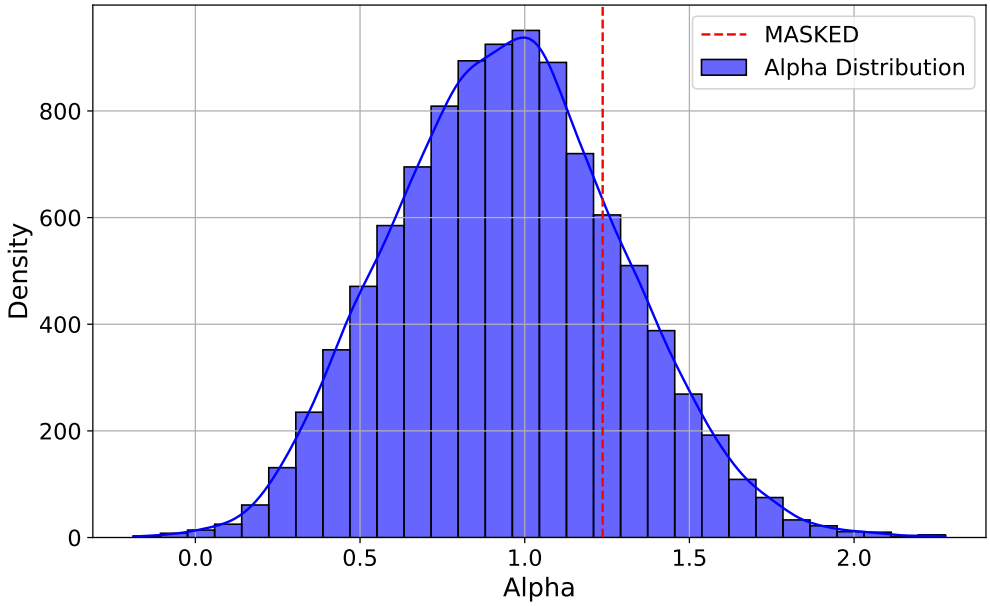
This figure presents a heatmap illustrating the estimated weights derived from non-negative least squares regressions, where the overall score is regressed on the individual dimension scores provided by GPT-4o-mini. The first row shows the estimated weights over all industries, whereas all other rows show the estimated weights for individual industries.

Figure 3: Alpha distribution for alternative industry-specific impact scores

Original Report:



Masked Report (No temporal information):



The two figures illustrate the distribution of five-factor alphas derived from long-short investment strategies. These strategies are based on scores generated by randomly weighting the individual subscores (INFL, COMM, EMPL, CONSSENT, HOUSE, FIN, GOVT, and TRADE). The alphas of the random portfolios are compared to those obtained from long-short investments based on the overall score provided by the LLM. The first figure refers to the alphas derived from the original monetary policy reports, whereas the second one refers to the variant where temporal information is masked.

Table 1: **Exemplary news headline per commodity**

Commodity	Headline
Aluminium	aluminum prices surge as lme copper rebounds from 3-month low
Arabica coffee	arabica coffee prices skyrocket in northwest cameroon due to overproduction
Beef	dj texas beef prices soar amidst drought and canada's mad cow disease
Coal	new hope for coal industry as output jumps amidst rising prices
Cocoa	nigeria's cocoa price inches up in cross river and edo states
Coffee	coffee prices rise as sao paulo and london markets open lower
Copper	copper prices surge higher as supply concerns mount
Corn	corn prices rise on strong demand and limited supply
Cotton	cotton prices surge in india due to reduced supply and strong demand
Crude oil	breaking news: nymex crude oil futures soar past $90/bbl, up by $1.99
Electricity	dj german press: electricity prices set to increase by 3.5% in 2004
Frozen orange juice	frozen orange juice prices skyrocket amid shortage
Fuel oil	fuel oil prices up slightly amid market demand
Gasoil	oil prices rise as market anticipates opec+ meeting
Gasoline	gasoline prices surge 5.8% on weekly basis
Gold	gold prices see steady increase, set for weekly gain
Iron ore	iron ore prices surge 6% in weekly trade amid china demand
Lead	lead price up 4.8% in shanghai as chinese demand continues to strengthen
Lean hogs	lean hog prices surge amid low supplies and strong demand
Liquified petroleum gas	propane prices surge as weekly stocks decline
Lumber	lumber prices rise as supply concerns mount
Naphtha	asian naphtha prices up 2.6% amid crude rally
Natural diamond	sanctions bite: natural diamond prices to surge 15% in 2022
Natural gas	natural gas prices increase as power demand eases
Nickel	lme metals in asia mixed; nickel prices may soar past $10,000
Palm oil	palm oil prices set to increase as crude futures drop
Platinum	platinum prices surge as johnson matthey confirms higher base prices
Pulp	wood pulp mills in the us to experience steady price increase in 2013, says report
Rough rice	japan's consumer price index (cpi) jumps 3.1% as high-grade rice prices soar
Rubber	rubber prices rise amidst uncertainty in thailand, indonesia, and malaysia
Silver	silver prices surge as cpm group warns of possible $37/oz decline
Soy oil	us soy soars above $13 bushel, reaching highest levels since sept 2008
Soybeans	breaking news: soybean prices surge 1.3% on cbot as supply concerns mount
Steel	steel prices up in germany following chinese imports
Sugar	sugar prices surge in india, affecting exports
Tin	tin prices soar over 9% as supply concerns mount
Uranium	uranium price up as amer lake main zone uranium deposit offers potential for ...
Wheat	wheat prices surge on crop damage concerns
Wool	australian wool prices up on pre-christmas demand
Zinc	zinc prices surge 7% as supply concerns mount

This table presents exemplary news headlines from our news dataset. We select one news headline portraying a price increase for each commodity included in our dataset.

Table 2: **Commodity-industry mapping**

commodity	POS	NEUT	NEG
Aluminium	Aluminium production	Coffee shops	Airplane manufacturing
Arabica coffee	Arabica coffee farming	Steel mills	Coffee shops
Beef	Beef cattle farming	Information technology	Meat retailer
Coal	Coal mining	Apparel manufacturing	Fossil electric utilities
Cocoa	Cocoa farming	Petrochemical	Chocolate manufacturing
Coffee	Coffee farming	Electrical wiring	Coffee shops
Copper	Copper mining	Chocolate manufacturing	Electrical wiring
Corn	Corn farming	Jewelry manufacturing	Animal feed
Cotton	Cotton farming	Uranium mining	Apparel manufacturing
Crude oil	Crude oil extraction	Coffee shops	Petrochemical
Electricity	Electricity generation	Beef cattle farming	Information technology
Frozen orange juice	Orange farming	Tin mining	Beverage
Fuel oil	Fuel oil refining	Corn farming	Shipping
Gasoil	Gasoil refining	Wheat farming	Trucking
Gasoline	Gasoline refining	Rubber production	Automotive
Gold	Gold mining	Soybean farming	Jewelry manufacturing
Iron ore	Iron ore mining	Soybean farming	Steel mills
Lead	Lead mining	Frozen orange juice	Car battery manufacturing
Lean hogs	Lean hog farming	Silver mining	Meat processing
Liquified petroleum gas	Liquified petroleum gas	Wool production	Petrochemical
Lumber	Lumber production	Silver mining	Home building
Naphtha	Naphtha production	Frozen orange juice	Petrochemicals
Natural diamond	Natural diamond mining	Gasoline refining	Jewelry manufacturing
Natural gas	Natural gas extraction	Frozen orange juice	Electric utilities
Nickel	Nickel mining	Sugar farming	Stainless steel manufacturing
Palm oil	Palm oil production	Crude oil extraction	Processed foods
Platinum	Platinum mining	Lean hog farming	Microchip manufacturing
Pulp	Pulp production	Gold mining	Paper mills
Rough rice	Rough rice farming	Nickel mining	Rice processing
Rubber	Rubber production	Iron ore mining	Tire manufacturing
Silver	Silver mining	Lumber production	Jewelry manufacturing
Soy oil	Soy oil production	Platinum mining	Processed foods
Soybeans	Soy farming	Natural gas extraction	Animal feed
Steel	Steel mining	Cocoa farming	Arms
Sugar	Sugar farming	Copper mining	Confectionery
Tin	Tin mining	Frozen orange juice	Electronics manufacturing
Uranium	Uranium mining	Cotton farming	Electric utilities
Wheat	Wheat farming	Fuel oil refining	Flour mills
Wool	Wool production	Liquified petroleum gas	Apparel manufacturing
Zinc	Zinc mining	Natural diamond mining	Galvanized steel manufacturing

This table presents the allocated industries that experience a positive, neutral or negative effect in case of a commodity price increase.

Table 3: **Industry-specific impacts of synthetic news headlines**

Headline	**Tech**	**Auto**	**RE**
Oil Prices Plummet Amidst Oversupply	—	↑	↓
U.S. Shale Boom Lowers Natural Gas Prices	—	↑	↓
China Dominates Rare Earth Metals Market Amidst Trade Tensions	↓	↓	↓
Gold Prices Skyrocket as Investors Seek Safe Havens	—	—	—
Copper Price Surges as Chilean Production Slows	↓	↓	↓
Uranium Prices Rise as Nuclear Energy Demand Increases	—	—	↑
Global Steel Shortage Impacts Car Manufacturers	—	↓	—
Cobalt Prices Soar Due to High Demand for Lithium-Ion Batteries	↓	↓	—
Silicon Shortage Disrupts Solar Panel Production	—	—	↓
Global Coffee Prices Jump Amid Drought Concerns	—	—	—
Aluminium Prices Surge Amid Russian Sanctions	↓	↓	—
Silver Demand Increases with Solar Power Popularity	—	—	↑
Platinum Price Drops Amid Decreased Demand in Auto Industry	—	↓	—
Nickel Prices Soar as Indonesia Imposes Export Ban	↓	↓	—
Corn Prices Rise as Drought Continues	—	—	—
Zinc Prices Spike Due to Mine Shutdowns	↓	↓	—
Rubber Prices Plunge as Demand Drops	—	↑	—
Iron Ore Prices Soar Amid China's Infrastructure Boom	—	↓	—
Coffee Prices Fall as Brazil Crop Fears Ease	—	—	—
Palm Oil Prices Surge Amid Concerns Over El Nino	—	—	—

This table provides an overview on twenty synthetic datasets as well as the industry-specific implications for the technology (*Tech*), the automotive (*Auto*), and the renewable energy (*RE*) industry. A negative impact is indicated by a "↓", a positive impact is symbolized by a "↑". "–" indicates that there is no significant effect on an industry.

Table 4: **Performance evaluation: Precision, Recall and F1**

	Smaller			Larger		
	Falcon-7b	LLama-2-7b	Mistral-7b	GPT-4	GPT-4o-mini	GPT-4o
Panel A: Accuracy						
accuracy	38.92	40.78*	63.46***	77.31***	75.74	84.3***
		(1.67)	(26.61)	(9.55)	(-0.62)	(2.82)
Panel B: Precision						
positive	36.56	40.82***	56.49***	66.78***	72.84*	82.34**
		(3.5)	(8.11)	(4.07)	(1.85)	(2.56)
unaffected	41.81	74.42***	72.24	94.56***	95.88	94.57
		(8.16)	(-0.55)	(7.96)	(0.54)	(-0.57)
negative	49.26	35.45***	63.68***	73.6***	66.94*	81.96***
		(-4.28)	(10.81)	(2.74)	(-1.87)	(4.32)
Panel C: Recall						
positive	57.9	53.42**	46.56	56.52***	70.93	72.77
		(-2.11)	(-0.17)	(9.19)	(-0.18)	(1.63)
unaffected	39.1	28.57***	56.69***	97.29	92.26	81.84
		(-3.02)	(12.78)	(-0.09)	(-1.2)	(-0.01)
negative	19.77	39.47***	49.14***	53.8***	68.24	79.86
		(10.04)	(-2.82)	(6.16)	(1.53)	(0.75)
Panel D: F1						
positive	44.41	45.89	50.6***	60.94***	71.3	76.45**
		(1.4)	(9.15)	(7.22)	(0.85)	(2.24)
unaffected	38.35	38.05	62.86***	95.48***	92.25	83.54
		(-0.28)	(10.48)	(6.74)	(-0.91)	(-0.1)
negative	25.88	37.1***	54.64***	61.66***	66.93	80.14***
		(5.77)	(12.4)	(4.98)	(-0.28)	(2.96)

The table presents a comprehensive performance evaluation of six distinct LLMs (Falcon-7B, Llama-2-7B, Mistral-7B, GPT-4, GPT-4o-mini, and GPT-4o) applied to a dataset of 1,600 news headlines across 40 commodities. The performance metrics reported include Accuracy, defined as the proportion of correctly classified instances across all classes; Precision, calculated as the ratio of correctly classified instances for each class relative to the total instances classified as that class; and Recall, defined as the ratio of correctly classified instances for each class over the total actual instances of that class. The F1 Score, representing the harmonic mean of Precision and Recall, is computed for three sentiment categories: Positive, Unaffected, and Negative. All metrics are reported in percentage points. Statistical significance, denoted by *, **, or ***, is assigned to model-prompt pairs that demonstrate a statistically significant improvement over the preceding model.

Table 5: **Accuracy of industry-specific sentiment across commodities**

commodity	falcon-7b	llama-2-7b	mistral-7b	gpt-4	gpt-4o-mini	gpt-4o
Aluminium	46.67	35.83	62.50	87.50	89.17	**90.83**
Arabica coffee	38.33	50.00	61.67	81.67	85.00	**89.17**
Beef	48.33	36.67	65.83	67.50	69.17	**77.50**
Coal	35.00	38.33	61.67	**72.50**	59.17	65.83
Cocoa	40.00	40.00	63.33	80.83	81.67	**94.17**
Coffee	44.17	45.83	62.50	84.17	**90.00**	**90.00**
Copper	29.17	38.33	59.17	80.00	80.00	**87.50**
Corn	44.17	45.00	65.00	86.67	**91.67**	**91.67**
Cotton	39.17	45.00	65.00	90.83	88.33	**95.83**
Crude oil	45.83	37.50	60.00	**66.67**	45.00	43.33
Electricity	38.33	49.17	62.50	63.33	64.17	**65.00**
Frozen orange juice	44.44	45.24	64.29	69.05	67.46	**90.48**
Fuel oil	41.67	40.00	65.00	**79.17**	58.33	68.33
Gasoil	42.50	40.00	55.00	**77.50**	68.33	64.17
Gasoline	37.50	41.67	61.67	**78.33**	60.00	65.00
Gold	28.33	37.50	63.33	70.83	70.83	**97.50**
Iron ore	33.33	36.67	61.67	73.33	78.33	**91.67**
Lead	34.17	35.83	58.33	83.33	83.33	**91.67**
Lean hogs	32.50	39.17	65.00	66.67	65.83	**79.17**
Liquified petroleum gas	33.33	36.67	**64.17**	63.33	49.17	50.83
Lumber	43.33	38.33	65.83	92.50	89.17	**95.00**
Naphtha	40.00	41.67	65.83	66.67	60.83	**85.00**
Natural diamond	27.50	33.33	65.83	66.67	69.17	**90.00**
Natural gas	44.17	40.83	65.83	**78.33**	58.33	62.50
Nickel	35.83	35.83	62.50	72.50	80.00	**88.33**
Palm oil	35.83	42.50	63.33	76.67	**87.50**	82.50
Platinum	40.00	39.17	61.67	73.33	71.67	**89.17**
Pulp	39.17	38.33	64.17	75.83	63.33	**88.33**
Rough rice	40.83	53.33	63.33	**67.50**	66.67	**67.50**
Rubber	46.67	45.83	65.00	90.00	86.67	**95.83**
Silver	32.50	37.50	64.17	78.33	93.33	**98.33**
Soy oil	35.71	40.48	65.08	85.71	90.48	**96.83**
Soybeans	35.00	47.50	61.67	81.67	90.83	**96.67**
Steel	36.67	35.00	64.17	67.50	70.00	**81.67**
Sugar	45.00	47.50	68.33	93.33	**94.17**	**94.17**
Tin	40.00	35.83	60.83	87.50	80.00	**92.50**
Uranium	44.17	40.83	65.83	71.67	83.33	**94.17**
Wheat	42.50	46.67	66.67	94.17	**95.83**	**95.83**
Wool	36.67	38.33	65.83	77.50	70.83	**97.50**
Zinc	40.00	36.67	65.00	70.83	83.33	**91.67**

This table provides the accuracy scores of industry-specific sentiments across different commodities.

Table 6: **Industry-specific sentiment and contemporaneous industry returns**

	(1)	(2)	(3)	(4)	(5)	(6)
	Equally weighted			Value weighted		
ind_sent_diff_1d	0.04***			0.03***		
	(5.10)			(3.80)		
ind_sent_diff_5d		0.08***			0.06***	
		(7.30)			(5.16)	
ind_sent_diff_20d			0.07***			0.04***
			(6.41)			(3.83)
market_sent_diff_1d	0.01***			0.01***		
	(3.13)			(3.65)		
market_sent_diff_5d		-0.47			-0.15	
		(-0.42)			(-0.16)	
market_sent_diff_20d			-0.06			0.51
			(-0.09)			(0.91)
ew_ret_l1	0.01*	0.01*	0.01*			
	(1.78)	(1.92)	(1.76)			
ew_ret_l5	0.01**	0.01**	0.01**			
	(2.14)	(2.15)	(1.97)			
vw_ret_l1				0.01*	0.01**	0.01**
				(1.94)	(1.96)	(1.97)
vw_ret_l5				-0.00	-0.00	-0.00
				(-0.79)	(-1.06)	(-0.70)
cons	0.06***	-0.00	-0.09	0.04***	-0.44	-0.21
	(3.87)	(-0.01)	(-0.15)	(3.18)	(-0.61)	(-0.39)
r2	0.18	0.21	0.21	0.19	0.21	0.22
N	271942	271754	271049	271942	271754	271049

This table presents the results from Fama-MacBeth regressions to explain equally (columns 1-3) and value weighted US industry returns (columns 4-6). The main variables of interest are the difference of the sentiment of day t minus the average sentiment in the previous one, two and three weeks. Additionally, we include the sentiment differences of the general market and the returns of the same day. We use Newey-West standard errors with five lags and denote the corresponding t-statistics in parentheses. *, ** and *** indicates significance at the 10%, 5%, and 1% levels, respectively.

Table 7: **Industry-specific sentiment and future industry returns**

	(1)	(2)	(3)	(4)	(5)	(6)
	Equally weighted			Value weighted		
ind_sent_diff_5d_l1	0.00	-0.00	-0.01	-0.00	0.00	0.00
	(0.02)	(-0.10)	(-0.38)	(-0.15)	(0.06)	(0.00)
market_sent_diff_5d_l1	-1.08	-0.89	-2.50	-0.69	0.21	-0.19
	(-0.87)	(-0.45)	(-1.05)	(-1.00)	(0.16)	(-0.15)
ew_ret_l1	0.01**	0.02***	0.02***			
	(2.04)	(2.88)	(2.74)			
ew_ret_l5	0.01**	0.01*	0.01**			
	(2.29)	(1.91)	(2.01)			
vw_ret_l1				0.01**	0.01	0.00
				(2.25)	(1.09)	(0.41)
vw_ret_l5				-0.00	-0.01	-0.01
				(-0.88)	(-1.55)	(-0.76)
cons	1.16	2.43	3.15	-0.05	0.87	1.10
	(0.85)	(1.10)	(1.26)	(-0.15)	(1.15)	(1.07)
r2	0.21	0.21	0.20	0.21	0.21	0.21
N	271707	271707	271660	271707	271707	271660

This table presents the results from Fama-MacBeth regressions to explain equally (columns 1-3) and value weighted (columns 4-6) US industry returns. The main variables of interest are the difference of the sentiment of day t-1 minus the average sentiment in the previous one, two and three weeks. Additionally, we include the sentiment differences of the general market and the returns of the same day. We use Newey-West standard errors with five lags and denote the corresponding t-statistics in parentheses. *, ** and *** indicates significance at the 10%, 5%, and 1% levels, respectively.

Table 8: **Long-short portfolio based on monetary policy reports**

Industries	INFL	COMM	EMPL	CONSSENT	HOUSE	FIN	GOVT	TRADE	SCORE
Panel A: Original									
1	0.64***	0.57*	0.5	0.6**	0.2	0.66**	1.02***	1.01***	1.1***
	(2.77)	(1.74)	(1.45)	(2.11)	(0.86)	(2.48)	(3.08)	(3.69)	(2.9)
3	0.44***	0.43*	0.55**	0.52**	0.1	0.66***	0.76***	0.54***	0.79***
	(2.71)	(1.86)	(2.49)	(2.57)	(0.67)	(3.35)	(3.5)	(3.06)	(3.19)
5	0.38***	0.25	0.42**	0.38**	0.06	0.52***	0.58***	0.48***	0.68***
	(2.93)	(1.48)	(2.3)	(2.15)	(0.49)	(3.14)	(3.5)	(3.38)	(3.47)
10	0.32***	0.16	0.28**	0.28**	-0.01	0.38***	0.33***	0.34***	0.49***
	(3.36)	(1.57)	(2.05)	(2.08)	(-0.05)	(3.15)	(3.25)	(3.65)	(3.59)
Panel B: Masked									
1	0.41*	0.23	0.47	0.67**	-0.0	0.52**	0.73**	0.82***	1.24***
	(1.75)	(0.65)	(1.19)	(2.2)	(-0.02)	(2.06)	(2.02)	(3.04)	(2.93)
3	0.37*	0.26	0.33	0.28	-0.07	0.27	0.46**	0.52***	0.6**
	(1.96)	(1.09)	(1.32)	(1.25)	(-0.43)	(1.3)	(1.98)	(2.64)	(2.24)
5	0.34**	0.21	0.27	0.16	-0.02	0.24	0.37**	0.32**	0.44**
	(2.27)	(1.16)	(1.38)	(0.88)	(-0.16)	(1.38)	(2.16)	(2.08)	(2.08)
10	0.25**	0.08	0.26*	0.18	-0.08	0.2	0.29***	0.27***	0.38***
	(2.36)	(0.75)	(1.84)	(1.27)	(-0.68)	(1.51)	(2.61)	(2.68)	(2.61)
Panel C: Summary									
1	0.22	0.09	0.79*	0.4	0.01	0.45	0.84**	0.65**	1.18***
	(0.91)	(0.25)	(1.95)	(1.29)	(0.05)	(1.64)	(2.25)	(2.35)	(2.59)
3	0.37**	0.2	0.26	0.27	-0.13	0.41*	0.59**	0.41**	0.68**
	(2.05)	(0.8)	(1.01)	(1.13)	(-0.77)	(1.79)	(2.4)	(2.09)	(2.41)
5	0.24	0.21	0.22	0.25	-0.03	0.26	0.41**	0.21	0.48**
	(1.6)	(1.13)	(1.08)	(1.22)	(-0.21)	(1.39)	(2.28)	(1.33)	(2.17)
10	0.24**	0.11	0.23	0.17	-0.08	0.24*	0.27**	0.18*	0.35**
	(2.11)	(0.93)	(1.6)	(1.15)	(-0.66)	(1.81)	(2.43)	(1.75)	(2.27)

This table reports the five-factor alphas for portfolios constructed by taking equally weighted long (short) positions in the top 1 (bottom 3, 5 or 10) industries, as determined by overall scores and scores for individual dimensions obtained from GPT-4o-mini when provided with the monetary policy reports of the Federal Reserve Board. The individual dimensions include inflation, commodities, employment, consumer sentiment, housing market, financial services, governmental fiscal policy, and external trade. Panel A shows the results obtained when analyzing the original monetary policy reports, whereas Panel B shows the results obtained for modified versions where temporal information has been removed with GPT-4o-mini to reduce the possibility of a look-ahead bias. Panel C shows the results obtained for report summaries without temporal information, generated with GPT-4o-mini. We use robust standard errors and denote the corresponding t-statistics in parentheses. *, ** and *** indicates significance at the 10%, 5%, and 1% levels, respectively.