# Quantifying the Effects of Central Bank Communication on Asset Prices

# (Can ChatGPT understand the FED?)

Cesare Papa Malatesta

Master of Finance Thesis

31/08/2023

# Table of Contents

## Abstract

This paper investigates a novel approach to classifying a database of text data with the use of ChatGPT. I construct a database consisting of speeches, statements, and other official communications made by FED officials since 2005 and classify the sentiment of each communication using an automatic process involving ChatGPT. I also fully automated the database building process creating a tool which extracts all communications including relevant labels such as date, author, title of speech. The sentiment classification made with ChatGPT scores each communication on a range from -1 (dovish) to 1 (hawkish). This provides a chronological sentiment landscape of FED communications. The score holds significant relationships with changes in treasury rates across different maturities and time horizons. Notably, the score captures the difference between long- and short-term maturities (10Y – 2Y) which is a closely observed variable by FED officials and investors alike. The paper attempts to add to the literature a methodology which uses ChatGPT with few shot learning and answer extraction techniques. This methodology can be applied to a much wider range of classification tasks by financial researchers and analysts to rapidly and without much prior knowledge classify large amounts of data without the need to manually classify a dataset beforehand.

## 1. Introduction

The main objective of this paper is to explore the capabilities of ChatGPT in classifying economic text data without the use of a human labelled dataset. To assess the model's classification, I am going to measure Federal Reserve communications through time. To test the relevancy of the classification, asset prices will be considered along with each classification to assess their relationship. In this introduction I will focus on relevant definitions, the motivation for this sort of analysis, where academic literature currently is on similar topics and finally briefly explain monetary policy communication of the FED.

It is important to start defining what ChatGPT is and how it produces its results. To make the discussion more interesting we should start by looking at the definition of what ChatGPT does which will come up when you first start diving into how the model works and how the words appear on your screen. Taking the words from Wolfram Alpha: "what ChatGPT is always fundamentally trying to do is to produce a "reasonable continuation" of whatever text it's got so far, where by "reasonable" we mean "what one might expect someone to write after seeing what people have written on billions of webpages, etc." (Wolfram). Although this sort of innovation and way of thinking on how to create a machine that produces results like ChatGPT seems fairly recent, it is relevant to see where the ideas have originated. For this we have to start looking at the academics in the field of Information Theory. One of its founders and most prominent figures, Claude Shannon, in his 1948 paper "A Mathematical Theory of Communication", proposes a methodology for a machine to create phrases in the English language: "by giving a machine certain statistics of a language, the probabilities of finding a particular letter or group of l, or 2, or 3, or n letters, and by giving the machine an ability equivalent to picking a ball from a hat, flipping a coin, or choosing a random number, we could make the machine produce a close approximation to English text or to text in some other language. The more complete information we gave the machine, the more closely would its product resemble English or other text, both in its statistical structure and to the human" (Pierce) ; (Roberts). Shannon creates what he calls n-th order approximations of the English language which is quite similar to what the current Large Language Models (LLMs), such as ChatGPT, are doing. Noam Chomsky, American linguist, rejected the "finite-state machine as either a possible or a proper, model of grammatical structure. Chomsky points out that there are many rules for constructing sequences of characters which cannot be embodied in a finite-state machine" (Pierce). This brief explanation of these discussions which date between the

1940s and 1970s show the importance of these modern LLMs and put into context for how long the problems they are solving have been studied and have captivated academia.
In finance in particular, machine learning has been heavily focused on numerical models and their ability to predict future states of say, asset prices or volatility. This only seems a natural development as financial data is always regarded as mostly numerical. In addition, machine learning in commercial applications seems to always have been more oriented towards predicting numerical values such as customer spending power or loan default probabilities. Interest towards Natural Language Processing (NLP) has started to pick up in the last decade with academia and the industry finding new ways to analyze text and pickup information such as main topics or sentiment. The financial analyst, with the use of NLP, can quickly analyze large amounts of text data. But this analysis is not particularly quick if the type of text analysis to make is new or if she is not particularly knowledgeable with NLP models and how to use them. These obstacles are well highlighted by looking at the problem this paper is focused on, analyzing central bank communication to determine its effects on asset prices. Looking at the existing literature, there are several papers which first assess sentiment of each central bank governor speech, and then check how closely communication follows assets prices. This sort of analysis has reached its pinnacle with the use of Google's embedding model: BERT. A typical workflow for producing this sort of analysis is to manually classify each speech (or sentence within each speech) as say hawkish or dovish on some scale. The classification would then be used to fine tune the pre-trained model BERT. Model fine-tuning is a process used in machine learning where a pre-trained model is further trained or "tuned" on a specific task or dataset. This process allows the model to adapt its previous knowledge from the large dataset (on which it was originally trained) to a more specific or smaller dataset related to a unique task. The purpose of fine-tuning is to leverage the broad learning of the pre-trained model and apply it to more specialized tasks. If the analysts would then decide to start analyzing company earning calls, they would then have to start from scratch and manually classify again an number of earnings calls to fine tune again BERT. This approach is the current gold standard and has made NLP ever more relevant. But the potential innovation brought by ChatGPT and the main motivation for this paper is its ability to create a classification and garner an "understanding" of a piece of text with no need for fine tuning. This would open many possibilities to modern financial analysts and researchers alike which could quickly classify text databases without the need for hours of manual labor. It can reveal useful also for other fields of the social sciences in which open ended questionnaires can be

classified rapidly without the need of human labelling. In addition, machine bias in classification could be easier to determine ex post than human bias might be.

### 1.1. Forward Guidance as a Monetary Policy Instrument of the FED

The American FED tries to innovate its monetary policy toolkit through the years to improve its efficacy in maintaining financial stability. Innovations are always backed by extensive research and gradually experimented and yet present its challenges and unknowns. After the 2008 crisis, the FED has adopted the new tool of Quantitative Easing (QE) and increased its use of providing Forward Guidance.

Forward guidance is communication by central bankers on what their expectations are regarding the evolution of economic conditions and monetary policy. This communication involves many sources of publications such as the FED's Monetary Policy Report to Congress. The most effective and most looked at communication by investors are the federal open market committee (FOMC) post meeting statements. These are then elaborated to the public through the press conference after the statements' release. This practice of the press conference was introduced by Chairman Ben Bernanke around 2011. The economic concept behind forward guidance is that financial conditions depend not only on the short-term rates but also on market expectations of future rates. As forward guidance can influence market participant's expectations, it has become an additional policy tool. Not only does central bank communication affect sophisticated financial market investors, it also plays an important role for all households and firms which can understand better current and future economic conditions (Bernanke). Additionally, it serves as an instrument for increased transparency and democratic accountability. It is important not to forget central banks are independent of politics and only through this independence can they have a more long-term course of action. Only in the mid 1990s FED officials have started to issue post FOMC statements. These only occurred if an interest rate change occurred. Moreover, this change happened with Chairman Alan Greenspan, known for its qualitative and indirect comments and outlooks about monetary policy and economic conditions. Bernanke, in his book "21$^{st}$ Century Monetary Policy" points out that especially after 2008, the importance of forward guidance has been highlighted. When rates are at the lower bound, there is the necessity to further affect expectations without the ability of further lowering the rates. An interesting paper by Chicago FED President Charles Evans and other Chicago FED Economists, "Macroeconomic Effects of Federal Reserve Forward Guidance", attempts at dividing forward guidance into two types: Delphic and Odyssean Guidance. Delphic Guidance, named after the oracle of Delphi, is

aimed solely at informing. This is done by providing policy explanations and forecasts of future states of the economy. This sort of guidance is data driven and it is meant to make market participants understand what factors shaping monetary policy are being observed by policy makers.

Odyssean guidance, named after Odysseus binding himself to the mast of his ship to avoid the temptations of the sirens, is aimed at stating a commitment to which central bankers bind themselves (or attempt to) to conduct policy in a specific way in the future (Campbell, Fisher e Evans). This sort of guidance is useful when short term rates are at the lower bound and central bankers want to increase the downward pressure on longer term rates by convincing investors they will keep short term rates lower for an extended period of time. With Odyssean guidance policymakers face a tradeoff between flexibility of action and commitment. Making a specific commitment leads to a clearer message to the markets. Diverging from these commitments becomes more apparent, making it harder for central bankers to change course. This affects their flexibility to react to unexpected situations with new solutions. This sort of guidance thus comes with "escape clauses" (Bernanke) within statements, aimed at allowing wiggle room for policymakers to change course in case of rare events. It is clear that an important factor for the effectiveness of odyssean guidance is credibility of the institution and the policymakers. The reputation of central bankers thus is very important to conduct forward guidance.

## 2. Methodology

The methodology of this analysis is divided into two distinct parts. Firstly, the method used for the classification of the Federal Reserve speeches. Secondly, the regression ran onto the asset prices to determines the impact of communication on them.

### 2.1 Classification Methodology – Data Preparation of Speeches

Data from the American FED is accessible online. Any speech made by officials of the FED is transcribed and accessible from the relevant section within their website. With the use of the Beautiful Soup library in python, I have scraped all the speeches made by any official of the FED since 1998 which is present on the website. In order to do this 2 points have to be highlighted. The website is split into speeches before and after 2006, with websites which have two different structures, hence the scraping method was different for the two cases. Secondly, for both cases there is a main page which links to each page dedicated to the speech. This final page contains the following data which I have taken and inserted into a convenient dataframe in which each row is a speech: Title, Governor, Date, Speech.

Additionally due to the structure of the web page, each speech contains its references as well. To make the data clean, I removed any form of text which came after the word “References” in each speech. Please note that the speeches vary greatly in nature, ranging from commencement addresses in universities to discussions at industry conferences. In figure 1 we can see how many speeches were made each year.

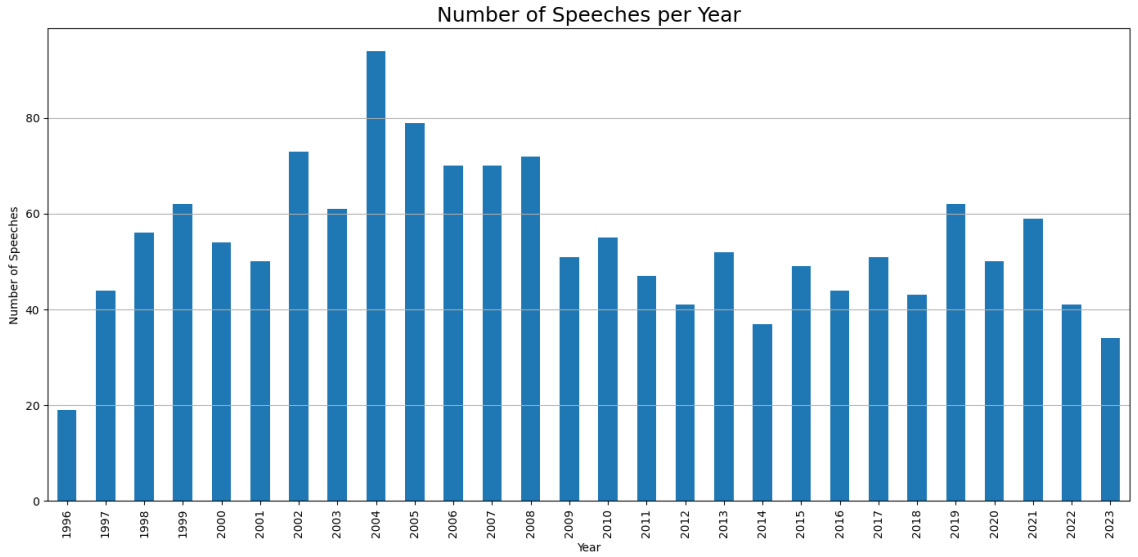


*Figure 1: Each bar represents the number of speeches made each year by officials of the FED*

From the data we observe there are some years in which there are far more speeches than others. In particular 2004/2005 seem to be the years with the most speeches.

Another important aspect which is relevant about the speech data is the tokens in each speech. LLMs like ChatGPT do not use regular words instead see words as tokens, different from words. Tokens are important because they are the basic units that the model uses to understand and generate text. The model learns to predict the next token in a sequence given the previous tokens. In particular, ChatGPT uses BPE (Byte pair encoding) and offers a python library (tiktoken) to calculate the amount of tokens a text contains. Encoding can prove useful as: “It attempts to let the model see common subwords. For instance, "ing" is a common subword in English, so BPE encodings will often split "encoding" into tokens like "encod" and "ing" (instead of e.g. "enc" and "oding"). Because the model will then see the "ing" token again and again in different contexts, it helps models generalise and better understand grammar.” (OpenAI). As the ChatGPT model measures the number of tokens of input it is important to count the number of tokens in each speech to make considerations on how to best use the speeches within the API. Two factors render important the number of tokens per speech.

1. The model has a limited amount of tokens it can consider in its prompt (request to the API). The GPT-3 model used within this analysis has a maximum context (length of the prompt) of 4,096 tokens. Weeks after I had completed my first classification a new model came out which allowed for a 16,000 token context. The model comes with a steeper price point and not all speeches would fit within the 16k context limit.
2. The API has rate limits which restrict the number of tokens and requests per minute. Using it to classify every speech in the database has the major issue that performing every API request one at a time would take days for a larger database. Hence, concurrent API calls must be made in parallel. These concurrent requests need to take into consideration how many speeches are processed per minute and how long these speeches are.

In conclusion, counting the number of tokens in each speech is important both for the model itself which has a limited token context and for operating the model through the API conveniently. Figure 2 shows the amount of total tokens per year.

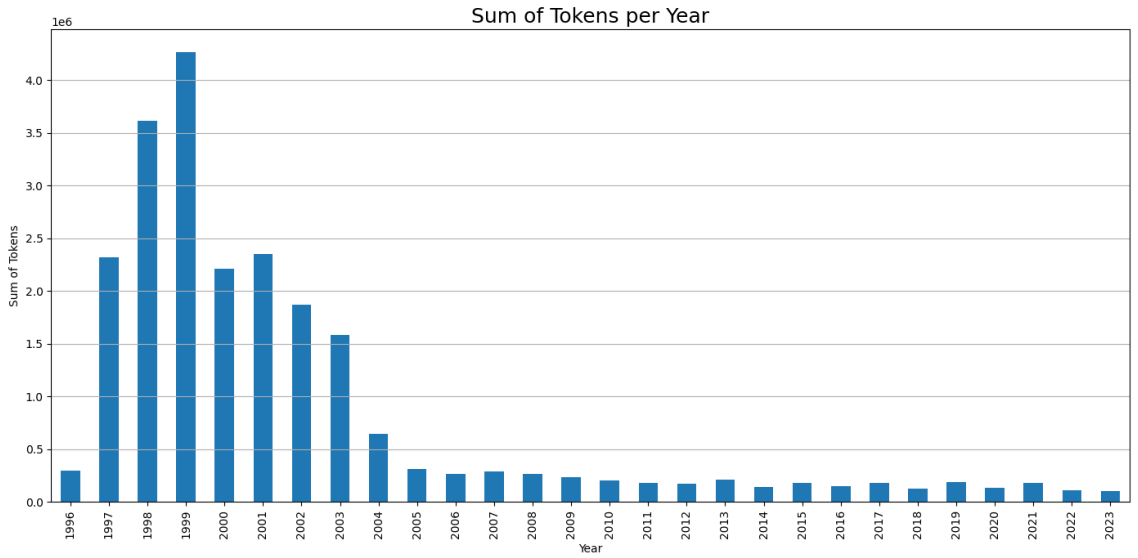


*Figure 2: Each bar represents the sum of the tokens of all the speeches in each year.*

At a first glance of Figure 2 we notice an imbalance in the data. In particular, there are much more tokens before 2004. To analyze this further and understand the issue at hand let's look at Figure 3.

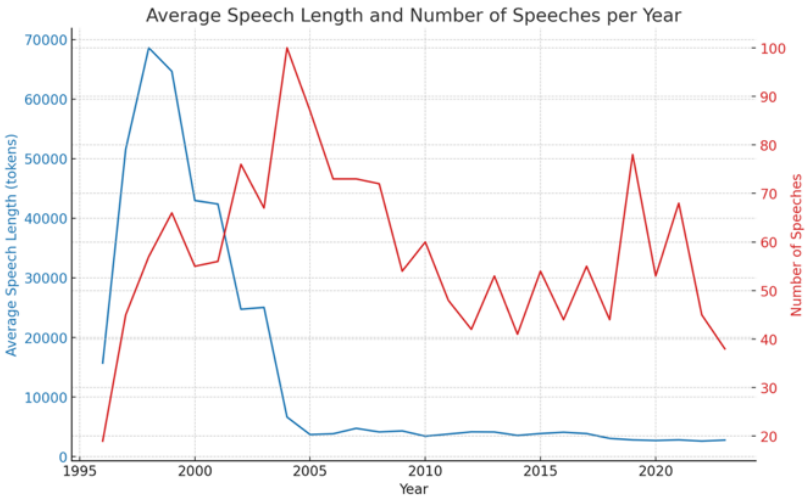


*Figure 3: The blue axis and line represent average token count of each speech throughout the years. The red axis and line is to compare the number of speeches in each year.*

In Figure 3 we get a more granular understanding of these imbalances within our speeches. We notice that speeches before 2005 seem to have significantly lengthier content. In addition, the number of speeches in each year peaks in 2005. An explanation for this pattern is the kind of speech made by the officials differs among mandates, thus speeches during the time of Chairman Greenspan might have been longer or in a different format. Taking an example of this, press conferences after the FOMC meetings only start occurring when Chairman Bernanke introduced the practice in 2011 (FED). This detail is important, as the opening remarks of these press conferences have been proven to deliver a high informational content as highlighted in (Pooter).

To solve the token number problem imposed by context and API limits, the speeches are split into parts of maximum 1000 tokens to avoid rate limits. This division of the speeches into smaller parts as input for the model deserves a note. Literature mostly divides speeches into sentences and classifies the sentences 1 by 1 (Hansen e Kazinnik, Can ChatGPT Decipher Fedspeak?). In addition, newspapers would suggest that internal banking models which are already using ChatGPT for these classification tasks also use sentence by sentence classification (Elder). These seem different from the use case within this paper, thus rendering the simple division of speeches in 1000 token parts sensible. The discussion on using the pieces of speeches instead of a sentence by sentence approach are picked up again in section 5.1.1.

Lastly, due to the imbalanced nature of the speech data, which presents a larger quantity and length of speeches before 2005, this part of the dataset has been removed. This makes our

data look much better in terms of being balanced each year which is a relevant detail at least to start off the analysis with. Below the new graph:

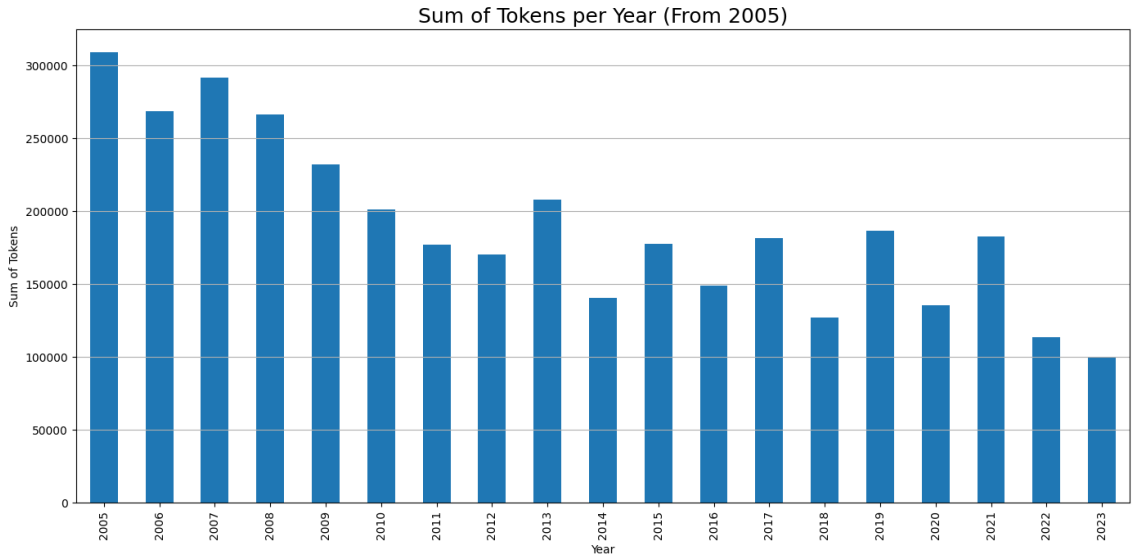


*Figure 4: Shows the sum of all tokens in each year starting from 2005. From this year on, accounted communication from FED officials is shorter.*

In addition to every speech made by governors of the FED, every statement issued by the board of governors of the FED is included within the statement (FOMC Statements). Each speech was also entirely downloaded using web scraping. The website of the FED contains a yearly directory which includes all the link to the statements. In each page, the date is also present. Differently from the speeches of the governors, these are on average shorter, with the mean length of being of around 350 tokens and the longest being of around 800 tokens. Thus, these did not require being divided into smaller chunks to input within the model. As the rest of the speeches include a "governor" and "title" column these we standardized to governor being FOMC for each record and title being the number of the statement starting from 2005. Other relevant elements, proven in the literature (Pooter) to have informational content, are the contents of the press conferences held after each FOMC meeting. Transcripts are released two weeks after the live press conference and are also proven to have an impact on the markets upon their release. Obtaining the data is not as straightforward as for the rest of the speeches. The press conference transcripts are not included within an html page put are available for download as pdfs which present a structure which is not immediately machine readable. Additionally, they not only present the opening remarks made by the chair of the committee, but also all the Q&A with journalists after the opening remarks. While the structure is easily readable by a human, the change in person which speaks within the pdf files is not easy to structure with a dataset with python. This is acknowledged within the paper

(Shah, Paturi e Chava) in which thankfully they make available all the datasets created by them on Github. They state that the task required ample manual input. Their transcript dataset includes every opening remark and Q&A phase in each press conference, including the speaker for each statement. The data ends in September 2022. This dataset was used to build this part of the dataset used within this paper. All the transcripts were aggregated and the date of the conferences inferred by the filename within the directory. All the sentences which were not made by the chair of the committee at the time of the press conference were removed (to stay in line with all the other data which only comprises communication controlled by the FED) and each sentence of the chain aggregated into one row. In this way, a clean dataset with all the press conference from 2011 to 2022 was created.

In summary, a new dataset including all the speeches, statements and press conference transcripts of the FED was created for this paper. To achieve this, two different web scraping tools were created: one to find all the speeches of FED officials, one to find all statements of the FOMC. All include relevant time series indexing, speakers and titles (if any).

### 2.2 Classification Methodology – Prompting

Prompting is one of the most interesting aspects of this analysis and it is only fair that some space is dedicated to topic. We will first look at the possibilities of prompting that the OpenAI API offers among its models and the parameters, relevant literature which has been used to create relevant prompts, issues found during the classification phase and ultimately the prompts used for the classification.

OpenAI offers two solutions for classification tasks which are suggested within their documentation: Completion and Chat Models (OpenAI). The text completion model (Text-Da-Vinci-003) was their most expensive model and not recommended within the documentation due to its steeper price point, additionally it runs the older GPT-3 Model. The documentation points at how the completion model could be better suited for (as the name suggests) completing text following the prompt. On the other hand, since the GPT-4 model has been released to the public within the API, the Text-Da-Vinci-003 completion model was announced to be deprecated after July 2023. For the analysis in this paper, the model used was the GPT-3.5 Turbo with the ChatCompletion endpoint. This is the most efficient model in terms of price point and speed according to the documentation (OpenAI).

To use the API, important input parameters which are relevant for this paper are:

Messages: within this parameter of the API users can insert their prompts assuming 3 different roles (this is limited to one role for the Da-Vinci model). It offers the possibility of

adding a “system” role which is hidden from the final user. This system prompt is particularly useful to provide context for the subsequent chat and steer the model towards a desired goal throughout the whole chat. The “user” role is the space with which users interact with ChatGPT on their website. Lastly, the “assistant” role can help simulate answers of the model and also provide input space in which the output of the API should focus on.

Temperature: This is a more complex concept common to LLMs. As previously said, language models determine the next token given the previous tokens within the limit of their context. They do not consider all possibilities given the previous tokens but limit their next choice based on a probability distribution. One might think they should always pick the next token with the highest probability, but this is not necessarily the case. The temperature setting allows to tweak which token to pick as the next one among the tokens with the highest probabilities to be picked next. In Figure 5 an example of this process taken from (Wolfram):

*The best thing about AI is its ability to*

| | |
|---|---|
| learn | 4.5% |
| predict | 3.5% |
| make | 3.2% |
| understand | 3.1% |
| do | 2.9% |

*Figure 5: An excerpt from the short book: "What is ChatGPT doing, and why does it work?". The snippet confers the reader a good sense of how the next words are chosen as output from the model*

From Figure 5 we notice the relevancy of the temperature parameter which make the model seem less deterministic and appear more random when picking the next word. After having experimented with adjusting this parameter, setting the temperature of the initial tasks of finding the topics, answers tend to be more in line with “No Topic” as the category the model decides. Setting a temperature closer to 1, leads to the model offering more topics mentioned within the speech.

Max_Tokens: this parameter is very straightforward. It limits the API on the maximum number of tokens the model can output. Note that this does not mean the model will try to limit its answers to the selected amount of tokens. It simply means the output will be cutoff, this is mostly used to regulate the answer to avoid lengthy responses from increasing the cost of the API. To regulate the number of tokens the model should use as output, inserting this request within the prompt itself can lead to desired results (OpenAI).

### 2.2.1. Choosing the prompts for the tasks at hand

In this next section I will discuss the process through which the final prompts for the classification tasks in this paper were determined.

The classification task is twofold. Firstly, I want to assess whether the topics of Inflation, Unemploymnent, GDP, Interest Rates and Inflation are mentioned within the speech. Secondly, I want a measure of whether the speech is hawkish or dovish within a scale, in order to assess the intensity of the sentiment. To test each of the prompts I used 100 speeches to determine whether they were providing a valid output or they had to be changed. OpenAI does offer a "Playground" function within their website which allows to test prompts but ultimately many instances of responses need to be requested in order to have a clear picture of whether the prompts have the desired output. There were 3 main issues encountered. Firstly, it would be ideal to aggregate all the classification questions to submit to the API into one prompt. This would make the classification task cheaper and faster. Separating the questions entails resubmitting every speech for each question which leads to multiplied costs. At the time of writing, OpenAI did offer a guide which offered a tutorial on classifying multiple tweets at once using only one API call. This has been removed from their website for a good reason. Figure 6 is the first prompt example I used to classify each speech:

| |
|---|
| "role": "system", "content": "You are a speech classifier. Answer each of the 3 questions.<br>\n 1. Is inflation, GDP, unemployment, interest rates mentioned yes or no?<br>\n 2. Categorize the input speech in one of these 3 categories: hawkish dovish, neutral. Only use category as answer.<br>\n 3. on a 1 to 5 scale, 1 being hawkish, 5 being dovish how would you classify the speech? only number as answer" |
| "role": "user", "content": "<br>Here is the Speech: \n\n {INSERT SPEECH HERE}" |

*Figure 6: An example of one of the first prompts designed for the classification.*

This prompt follows some of the main guidelines highlighted by Openai such as: providing delimiting characters to indicate different section of the prompt (/n in the figure), indicating a role for the assistant and providing the acceptable answers to choose from. This sort of prompt does not give the expected output. Even by providing relevant examples of how the answer should be structured the results were always too messy to be properly used. This means that answers were not properly separated among them in a consistent manner and often mixed together. A solution to this was passing again the answer to the API to discern the 3 answers given the prompt. This strategy was also unsuccessful and the model did not deal well with separating the answers into 3 parts consistently enough to be able to generate clean

records from the queries. Thus, the classification tasks need to be split into multiple parts in order to have a clearer output.

The second issue which affected the choice of prompts relates to the kind of classification which should result in the output. For the first classification task (determining whether the speech is relevant or not for monetary policy) asking the model to provide relevant topics among a list (including a No Topic category) was the most successful strategy. This follows from other papers which assess main topics within the FED speeches (Gnan, Schmeling e Schleritzko). Not only this offers the desired results of assessing whether the speech is relevant for monetary policy but also provides main topics in each speech which can provide further insights for our analysis. In addition, starting the classification with this "relevancy" prompt renders the process more efficient as the second classification can only be run on the "relevant" speeches. Additionally, as the speeches were divided into 1000 token parts, a lot of token savings come from irrelevant parts of the speeches being removed from further classifications. It is important to keep in mind that many of these speeches should be completely irrelevant for monetary policy as they include university speeches and personal interviews of FED officials as well. Another limitation and issue faced even when having separated the prompts is that results are still not always the desired ones, of 6000 rows analyzed in Table 1 are the 10 most common categories.

| **Topics** | **Count** |
|---|---|
| No Topic | 2846 |
| Category: None | 251 |
| Interest Rates | 232 |
| Inflation | 210 |
| Employment | 108 |
| Category: Interest Rates | 60 |
| Category (Inflation, Economic Growth, Employment, | 30 |
| Inflation, Interest Rates | 28 |
| Category: Inflation | 25 |
| Economic Growth | 23 |

*Table 1: These are the unique value counts of the categories as raw output from the model. These are not all the proper labels and they require cleaning through Answer Extraction.*

These are very good results already as the categories seem to be somewhat homogeneous yet require cleaning.

The third issue encountered is with the second prompt where a sentiment direction and measure have to be given by the model. Initially, as seen in Figure 6, the idea was to place the speech on a numerical scale. This method did not provide the expected results. The output became very mixed between words used for classification and numbers. Additionally, trying to give a prompt which paired the category (hawkish) and a respective intensity for this category also failed. This last attempt led to numbers which were inconsistent with the category assigned with the word. Using Figure 6 as an example this means the model would output Dovish with associated numeric scale value of 2, which should be Hawkish. Ultimately, the decision was to follow what was done in (Hansen e Kazinnik, Can ChatGPT Decipher Fedspeak?) of providing 5 categories for the model to choose from: Hawkish, Mostly Hawkish, Neutral, Mostly Dovish, Dovish. In this way the prompt does not have to include any instructions on converting words to numbers for the classification and the categories can be easily converted into numbers for purposes of the regression afterwards.
In order to make an informed decision on how to design the most effective prompt, there are multiple resources from which one can understand important dynamics that improve the output of the model to our objectives. In particular, the papers (Brown, Mann e Ryder) and (Kojima, Gu e Reid) explain in detail different strategies for designing a prompt and its effectiveness particularly for problem solving tasks. The first paper is written by members of the OpenAI team themselves and the second is among the recommended resources for prompting from the OpenAI website. The papers offers insightful diagrams which clearly explain different strategies for prompting, Figure 7.


*Figure 7: Different prompting methods explained by the team at OpenAI*

In Figure 6, the first prompt used, which included multiple classification tasks, was using the technique of zero shot prompting. The combination of multiple queries with only one prompt and not providing any examples has resulted in unsuccessful results. Using the few-shot learning technique for the prompts has yielded the results which were most aligned with the required output. Interestingly enough, even with few-shot learning, the model had difficulties giving as outputs numbers.

The two final prompts used for the classification have been constructed following the same procedure. Below the prompts for the classification tasks. Marked in red are the system inputs, blue the user inputs and green the assistant inputs. The black text in each box is the input to the model. Each table represents an entire prompt.

**For the topics classification task:**

| |
|---|
| "system": "You are a knowledgeable assistant specialized in identifying economic aspects mentioned in given texts. Provide the topic mentioned among these: Inflation, Economic Growth, Employment, Interest Rates." |
| "user": "Speech: 'The current fiscal policy has led to unprecedented growth. GDP has risen 3% in the last quarter alone, indicating our economy is on a solid trajectory.' \n Topic (Inflation, Economic Growth, Employment, Interest Rates, No Topic):" |
| "assistant": "Economic Growth" |
| "user": "Speech: 'Our employment rates have never been better. More people are joining the workforce every day, contributing to the nation's prosperity' \n Topic (Inflation, Economic Growth, Employment, Interest Rates, No Topic):" |
| "assistant": "Employment" |
| "user": "Speech: 'There has been a lot of discussion recently regarding new technology trends shaping the banking industry.' \n Topic (Inflation, Economic Growth, Employment, Interest Rates, No Topic):" |
| "assistant": "No Topic" |
| "user": "Speech: 'The Federal Reserve has decided to increase the interest rates by 0.25% in an effort to curb the rising inflation and stabilize the economy.'\n Topic (Inflation, Economic Growth, Employment, Interest Rates, No Topic):" |
| "assistant": "Inflation, Interest Rates" |
| "user": "Speech: "{INSERT SPEECH TO ANALYZE}" \n Topic (Inflation, Economic Growth, Employment, Interest Rates, No Topic):" |

**Answer Extraction for the classification task:**

| |
|---|
| ""system":  "You are performing a data cleaning task.<br>Please clean the input and use only these 5 labels, use comma as separator:<br>Inflation, Economic Growth, Employment, Interest Rates, No Topic" |

| |
|---|
| "user": " 'the mentioned categories are inflation, and rates'\n Labels:" |
| "assistant" : "Inflation, Interest Rates" |
| "user": “Labour market, Economy growing'\n Labels:" |
| "assistant": "Employment, Economic Growth" |
| "user": "'Inflation'\n Labels:" |
| "assistant": "Inflation" |
| "user": "'{INSERT SPEECH TO ANALYZE}' \n Labels:" |

**For the sentiment classification task:**

| |
|---|
| "system": "You are a knowledgeable assistant specialized in identifying sentiment in Federal Reserve Speeches. Please label each speech only with these 5 labels: Dovish, Mostly Dovish, Neutral, Mostly Hawkish, Hawkish." |
| "user": "'We believe inflation is temporary and we will keep interest rates low to support economic recovery.'\n Sentiment (Dovish, Mostly Dovish, Neutral, Mostly Hawkish, Hawkish):" |
| "assistant": "Dovish" |
| "user": "'The current economic situation may require us to maintain a lower interest rate for a while.'\n Sentiment (Dovish, Mostly Dovish, Neutral, Mostly Hawkish, Hawkish):" |
| "assistant": "Mostly Dovish" |
| "user": "'We're monitoring the situation carefully and will adjust monetary policy as needed.'\n Sentiment (Dovish, Mostly Dovish, Neutral, Mostly Hawkish, Hawkish):” |
| "assistant": "Neutral" |
| "user": "'Given the positive economic indicators, it may be necessary to start considering raising interest rates.'\n Sentiment (Dovish, Mostly Dovish, Neutral, Mostly Hawkish, Hawkish):" |
| "assistant": "Mostly Hawkish" |
| "user": "'The risk of inflation is high, it's critical to increase interest rates immediately.'\n Sentiment (Dovish, Mostly Dovish, Neutral, Mostly Hawkish, Hawkish):" |
| "assistant": "Hawkish" |
| "user": "Speech: "{ INSERT SPEECH TO ANALYZE }" \n Sentiment (Dovish, Mostly Dovish, Neutral, Mostly Hawkish, Hawkish):" |

Please note how all prompts:

- Make use of separators such as \n
- Contain a system message which defines the role of the model and the provides the context for the classification task
- All the examples of user messages for the prompts are taken by querying ChatGPT4 through the browser interface.
- Contain examples for every category (few shot learning). These are the exchanges between the user and assistant.
- Repeat the categories as possible answers
- Provide for each category an example for what the answer should look like
- Each prompt ends with a final user message in which the desired speech is inserted. The API will then output the assistant message which should follow.

Using these techniques, as suggested by the documentation, the prompts enabled the model to act as a classifier. Furthermore, the use of few-shot learning with examples artificially generated by the model itself opens up the concept of creating synthetic data to perform model fine tuning.

### 2.3 Classification Methodology – Classifying the relevance of the speeches

The classification task is divided into two parts. Firstly, I classify the relevance of the speech to assess future monetary policy decisions. I use the simple approach of determining whether the speech mentions: employment, economic growth, interest rates and inflation. Employment and Inflation are relevant as they are the mandate of the FED, representing the main focus of the monetary policy. Interest Rates has been taken as it is likely to appear in statements that determine future rates and thus should be relevant to include. Lastly, Economic Growth is a relevant factor to determine monetary policy as it is a metric the FOMC evaluates to make hints at future policy decisions. Another logical approach would have been to use a more direct prompt for the classification task such as: “Is the text relevant to infer future monetary policy decisions”. Although this would have been an interesting scenario one factor was decisive to opt for the keyword approach. Storing the keywords opens up the possibility of conducting other sorts of analysis by storing what the speeches were focusing on among those variables such as looking at the appearance of the topics through time and the topics most correlated to hawkish or dovish stances.

From the prompt, the model was instructed a desired output which comprised each label mentioned separated by a comma. While this was mostly successful there were many

instances of new categories or more verbose answers. In order to tackle this issue I followed a similar technique used in the paper (Kojima, Gu e Reid), with a first prompt and query to the model being reasoning extraction and a second prompt and query dedicated to answer extraction. In this case, this cleaning pipeline was very effective, delivering exactly the desired output. Finally, for the sake of the analysis, the labels were structured as dummy variables. New columns were made for each label and values of 1 and 0 represent the topics mentioned within the text.

### 2.4 Classification Methodology – Classifying the sentiment of the speeches and determining the score

The second part of the analysis is aimed at assessing the sentiment of each communication. The approach used for the classification is similar as (Hansen e Kazinnik). The model is tasked with labeling the text among these categories: Mostly Dovish, Dovish, Neutral, Mostly Hawkish, Hawkish. As previously mentioned, word categories were used instead of numbers due to the difficulty of making the model consistently output numbers. To then translate the classification into a quantifiable measure, a hawk/dove score is developed similarly to (Lucca e Trebbi):

| Category | Score |
|---|---|
| Dovish | -1 |
| Mostly Dovish | -0.5 |
| Neutral | 0 |
| Mostly Hawkish | 0.5 |
| Hawkish | 1 |

*Table 2: Shows how the scoring is translated into numbers from the words classification*

The reason for the positive number representing negative sentiment is to be aligned with the sign of an increase in interest rates. In other words, an increase in the rates is symbol of hawkish policy. The half measures of 0.5 also creates more granularity in the score. Due to the fact that the speeches were divided into smaller parts, the mean of the scores was taken to represent the sentiment score of the speech.

## 3. Results

This section will be divided into 3 main parts, firstly the results of the two classification tasks will be discussed. Lastly, the comparison among the classification results and market prices will be thoroughly analyzed.

### 3.1 Classification of Topics

The ability of having displayed different topics opens up possibilities of getting some insights on how the model has classified the data. Let's look at the categories of the classification in Figure 8.

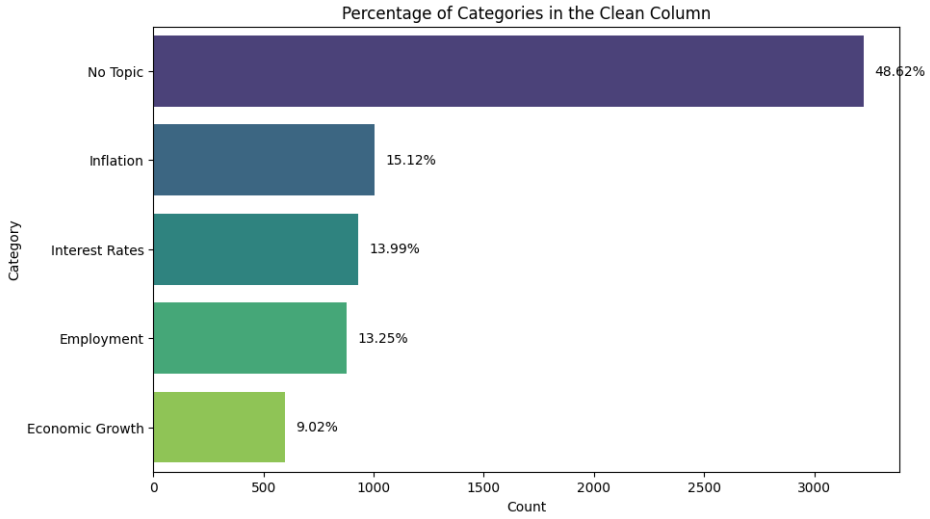


*Figure 8: This is the distribution of topics the model has made after having "cleaned" the raw output through answer extraction.*

Slightly less than half of the total speeches were deemed irrelevant according to our criteria. This might seem disappointing, but the interesting result lies delving deeper into what kind of speech was classified. As stated previously, included in the data are generic speeches made by FED officials, FOMC press conference transcripts and FOMC statements. Below the division of topics among categories in Tables 3 to 5.

| **FOMC Statements** | |
|---|---|
| Inflation | 25% |
| Interest Rates | 25% |
| Economic Growth | 25% |
| Employment | 22% |
| No Topic | 3% |

*Table 3: Distribution of topics of the statements of the federal open market committee (FOMC)*

| **Press Conferences** | |
|---|---|
| Inflation | 21% |
| Interest Rates | 19% |
| Economic Growth | 11% |
| Employment | 19% |
| No Topic | 30% |

*Table 4: Distribution of topics of the press conferences held after each FOMC.*

| **Generic Speeches** | |
|---|---|
| Inflation | 11% |
| Interest Rates | 11% |
| Economic Growth | 6% |
| Employment | 10% |
| No Topic | 62% |

*Table 5: Distribution of topics of any communication made by FED officials not including FOMC statements and related Press Conferences*

From these tables we get a much better picture of whether the GPT3.5 model output is sensible or not. From intuition we can deduce that the FOMC statements will always be relevant for the future stances on monetary policy and in fact, the "No topic" classification is

the lowest with only 3% which is reasonable. We get additional confirmation of the results by looking at the press conference categories which also seem more relevant. It is important to note that these are also divided into multiple parts and much of the text contains answers of the speaker to journalist questions, thus it is reasonable to assume chunks of press conferences can be considered irrelevant.

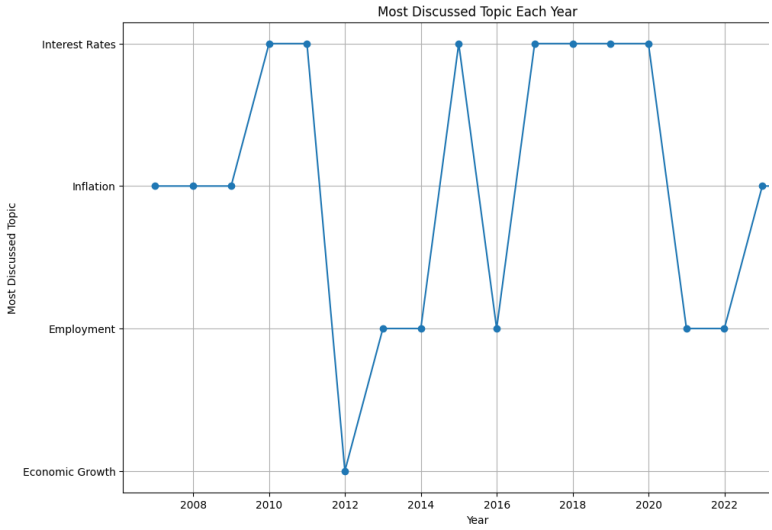


*Figure 9: The graph indicates the most discussed topics each year. These are represented by the blue dots. The y axis contains the topics.*

Figure 9 explores the variation of topics in time. This can be useful to evaluate as it is a confirmation of an accurate classification by the model if it coincides with a storyline of the main interests of the FED through time. Details to notice are economic growth and eployment being relevant in 2012 to 2014, time where monetary expansion was relevant within the world economy. Furthermore, the topic of an overheating economy in 2018 with starting discussions on coming out of the low interest rate regime. Topic of employment becoming the most relevant during the period of covid 2021 to 2022 when unemployment had skyrocketed only to come to a reversion shortly. Finally, the topic of inflation being the most discussed by FED officials in 2022/23.

This topic classification not only provides us with a first step towards our final result on assessing the impact communication has on asset prices, but also depicts a picture of the most relevant topics of FED officials related to monetary policy.

### 3.2 Classification of Speech Tone

As mentioned in the methodology section, each speech has been classified with the tone on a scale from dovish to hawkish. In Figure 10 the total results of the classification:

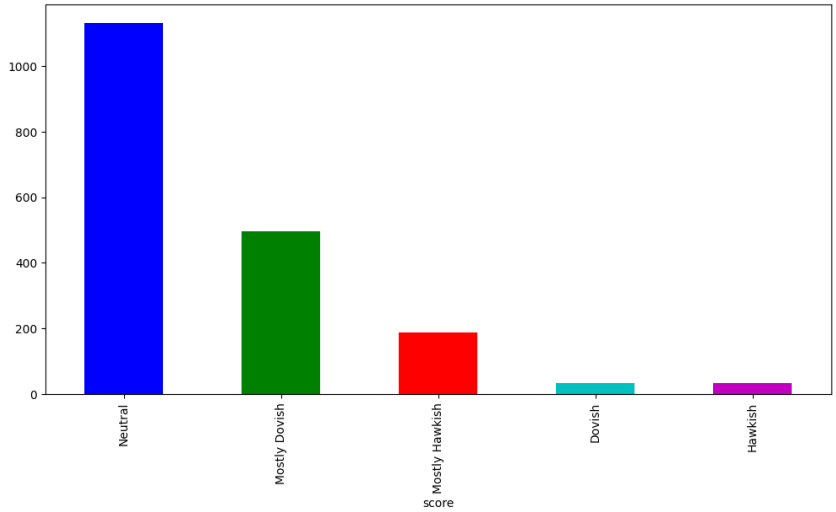


*Figure 10: Final distribution of classifications. Each bar represents a category of classification. The y-axis are the value counts.*

Expectedly most of the text analyzed was categorized as “Neutral”, additionally it is also reported within (Hansen e Kazinnik, Can ChatGPT Decipher Fedspeak?) that the GPT3.5 model tends to classify sentences as neutral and also the second most prominent classification being “Mostly Dovish”, aligns with their paper. We will get a better picture by looking at the classification among the categories of speeches looked at previously. Figure 11 is the division of classifications among the FOMC statements which were the only category where “Neutral” was not the most occurring sentiment. A key insight we gain from this result is the importance given to how the statements are formulated by FED officials. The statements are carefully weighted to produce an effect (combined with market intervention) on the markets. From this result we could infer a mostly dovish sentiment is preferred, which aligns with the stability officials want to bring to the markets. Additionally, it is noticeable that the model does not tend to classify sentiment at the extremities with low dove and hawk classifications.

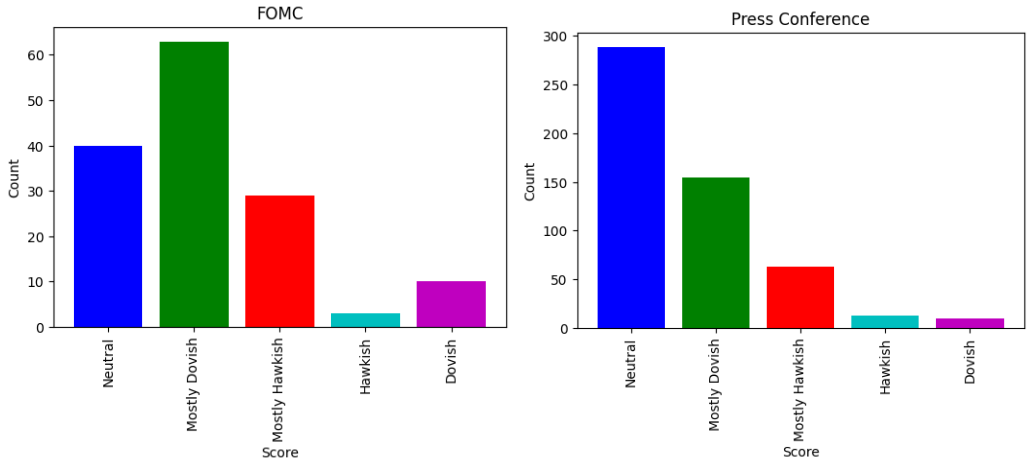


*Figure 11: Final distribution of classifications of FOMC and Press Conference statements. Each bar represents a category of classification. The y-axis are the value counts.*

After having classified the speeches and looked at the categories it is time to turn towards the hawk dove score built which will give us insights on the sentiment of the FED officials through time.

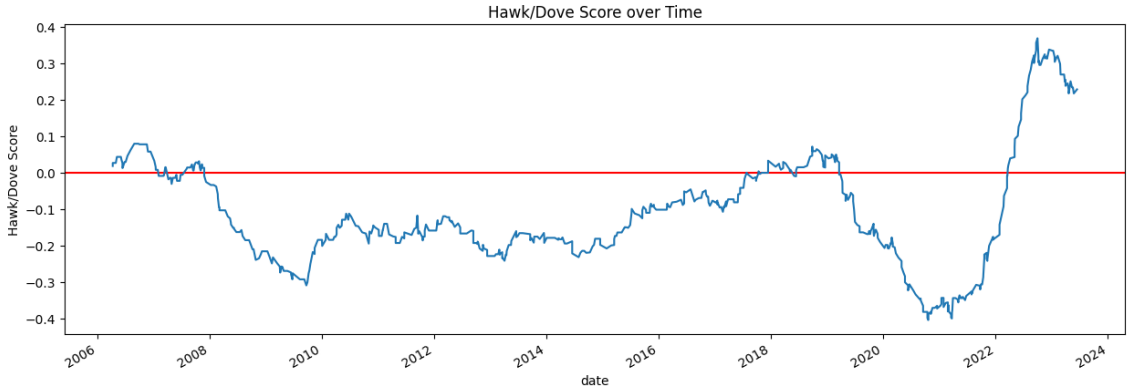


*Figure 12: The rolling average of the hawk dove score is the blue line. The red horizontal line represent the 0.*

Looking at Figure 12 we see the change in sentiment over time according to our hawk dove score. We immediately notice important historical details. Around the 2008 period, sentiment changed from hawkish to dovish quickly. This coincides with the 08 crisis. The overheated economy that forced the FED to raise rates in 2006/7 is followed by the economic crisis which matches a period of dovish and expansionary monetary policy carried through by the FED. This dovish period continued for most part of the decade and was a time in which the FED (and other central banks) kept rates low. The trend then reversed once more after 2016. The fiscal policy of the time, which included radical tax cuts and advantages for stock buybacks during the Trump presidency, led to a peak in asset prices which was followed by hawkish remarks of interest rate hikes. In the last years instead, we see the drastic and rapid shift that COVID brought. This result should be further analyzed. First it shows the rapidity at which the sentiment went from the most dovish score in 2021 to the most hawkish in 2023. It also shows a potential change in the way statements by FED officials are drafted as Chairman Powell came into his office in 2018. The graph shows a clear change in pattern since that year. Changes are much more drastic compared to the previous decade. This can be a sign of a shift in communication policy for the FED.

Our analysis also allows us to get deeper insights into the different officials of the FED and their personal hawk/dove score as shown in Figure 13.

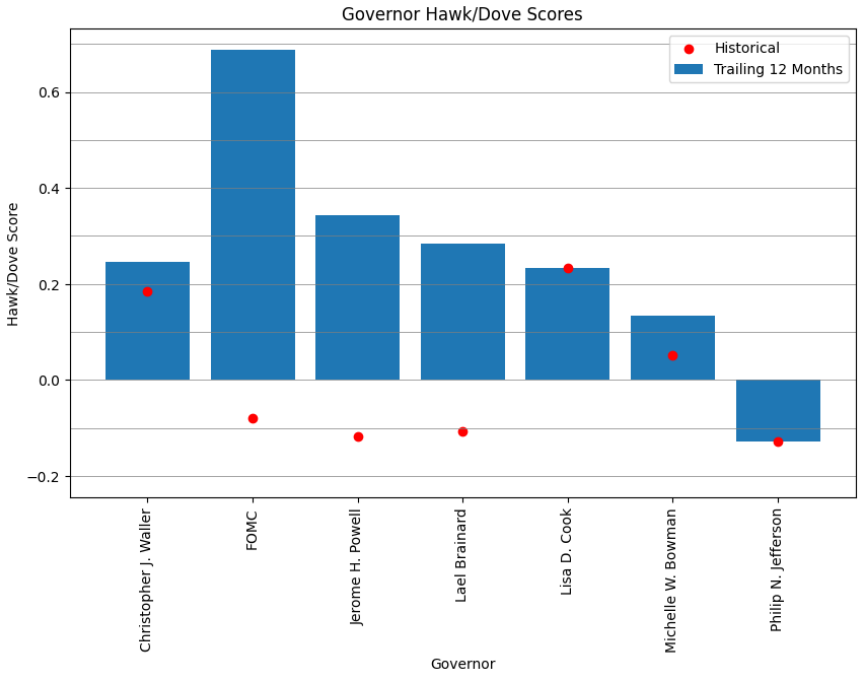


*Figure 13: Each blue bar represents the hawk dove score average of each governor whom has held a speech in the past 12 months. The red dots represent the hawk dove score average of each governor since they took office. The FOMC voice are the released statements of the committee.*

Figure 13 displays all the members that have given relevant speeches in the past year and compares the average of their hawk dove score in the past 12 months with their historical hawk dove score since in office. Additionally the FOMC entry reflects the statements issued after each meeting. This graph provides tremendously interesting insights.

Powell has a hawk dove score which historically is dovish. This might be caused by the tendency of the model to estimate more mostly dovish entries. On the other hand, critics have said that Powell has been too slow to intervene with monetary policy causing the current rapid hikes that have been going on in the past 2 years. Accurately, his score in the past year is instead the most hawkish of them all along with the committee statements. Apart from the economic insight into Powell's rhetoric which, rightfully so, has been hawkish in the past months, the analysis also shows an accuracy of the model to deliver an understanding of his stances through time without much effort of manual classification or training on particular economic data.

Another committee member which has shifted ideals in the last period is Brainard which historically has had a dovish sentiment, he is also regarded as a dove within newspapers and public opinion, has switched to a hawkish sentiment in the past year. Other hawkish committee members include Waller and Cook.

### 3.3 The Hawk/Dove score and Asset Prices

The first result that has to be highlighted is how well the hawk/dove score, autonomously created through few-shot learning classification, follows interest rates. Figure 14 has graphs of the entire interest rate structure.

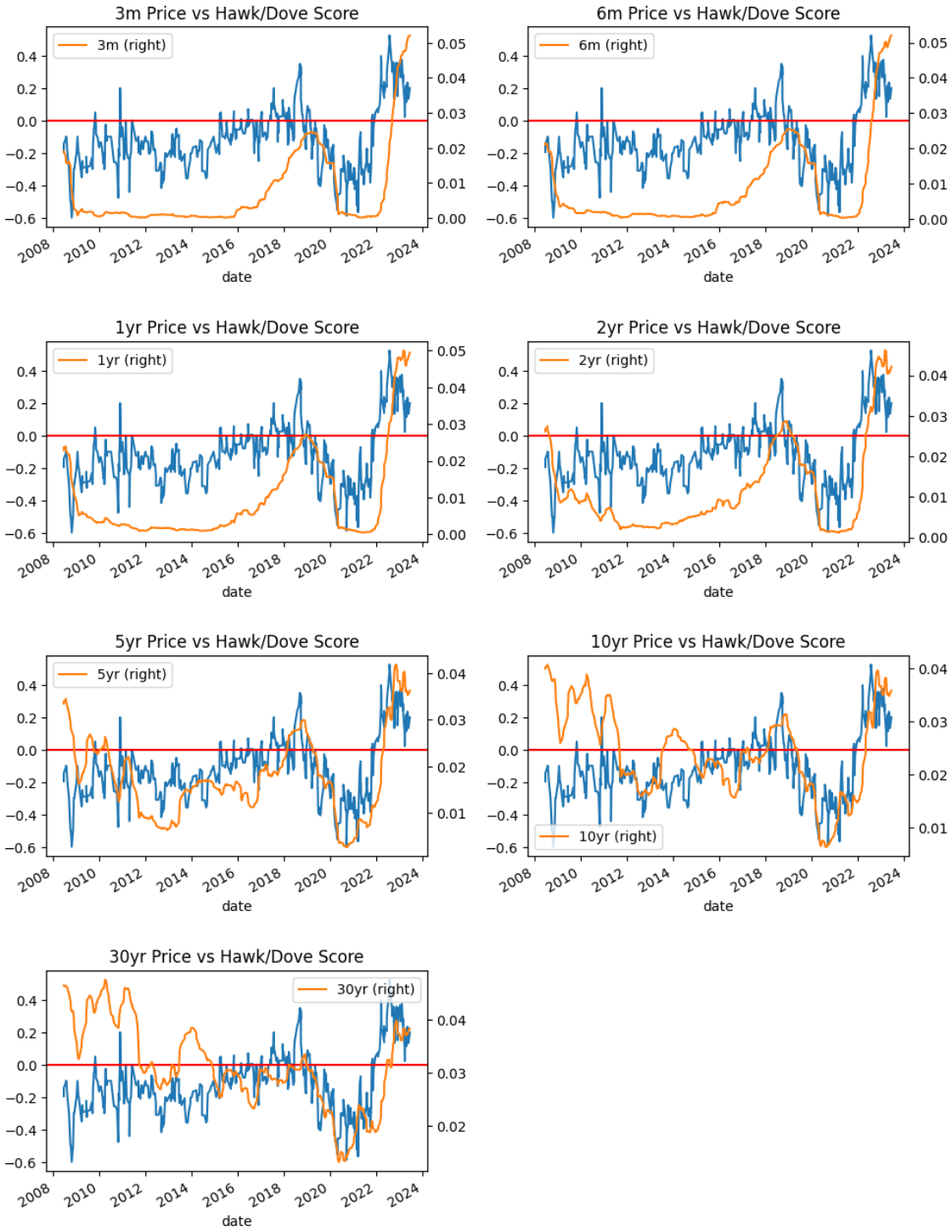


*Figure 14: Each graph is a comparison of the hawk dove score through time (blue line, left y-axis) and the Treasury Market Yield of the respective security (Orange line, right y-axis)*

At a first glance results seem impressive considering how the score was built. This section will delve deeper into an analysis of these results but one point has to be mentioned before. The score is entirely built based on an interpretation of speech. Machine or not to have done the interpretation the first important result is the importance of speech and communication for monetary policy. This is to the point where the direction of change in market yields can be (approximately) estimated just by reading the communications. It highlights the importance of communication for the financial world and to a greater extent how relevant interpretation of communication is to asset prices.

Graphically we have two results: firstly, the 5, 10 and 30 year maturities seem to follow more closely the HD score, Secondly the score seems to follow the rates more closely since 2018. It is surprising that the longer end of the curve follows more closely the score than the shorter end. The short end of the curve is the one that is more sensitive to changes in interest rates as the FED controls the short term lending rate. This observation can be explained by what investors consider when buying longer term maturities. These longer date maturities are a stronger reflection of future states of the economy and a reflection of where expected interest rates are going in the future. The prices created by the models are in fact based on future expected rates built on the shorter end of the curve. With this in mind, it is reasonable to point to the possibility that the informational content of FED officials communications has more relevance in the longer end of the curve, in which investors invest on their beliefs on future monetary policy actions.

The second observation can also reveal some interesting insights. It is evident that from 2018 onwards the score follows more closely all the rates in the term structure. Coincidentally, it is the same year in which Powell become chair of the FED. One first explanation of what we observe is a voluntary shift in the informational content FED officials release. If voluntary this change has created a more clear and transparent communication which reflects on the asset prices. Another factor which could explain the difference between the periods 2010-2016 and 2017-2023 regards the data and it is bifold. Firstly, observations in the first period which are not neutral are only 97 whilst in the latter 168. More datapoints in the first period could have brought the score at lower levels to reflect the rates markets. Secondly, the score may not be able to reflect well the extreme low-rate environment within that period.

### 3.3.1. Regression on Market Interest Rates

I now turn to the attempt of understanding the numerical relationship between the hawk dove score (HD score) and the market yields at the different maturities. To do this, an OLS

regression to find the significant relationship between the variables is effective as there is only one variable we are looking at. I run a regression with the independent variables being the changes in market yields for the securities at different maturities. The market yields are taken at different time periods 1, 7, 14, 50 days ahead of the hawk dove score. The score is taken as the mean of the latest 30 days to keep the sentiment as relevant for the present time as possible. The objective of running the regression is to see whether the score can give a rough estimate of the rates movement. Results of the regression were mostly statistically significant.

| **Regression Results for 1 day look ahead period** | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD_score coefficient** | **p-value** |
| **3m** | 0.0008 | -0.4554 | 0.099 |
| **6m** | 0.0001 | 0.3386 | 0.581 |
| **1yr** | 0.0001 | 0.8165 | 0.501 |
| **2yr** | 0.0011 | 0.0544 | 0.056 |
| **5yr** | 0.0012 | 0.0426 | 0.047 |
| **10yr** | 0.0015 | 0.0337 | 0.026 |
| **30yr** | 0.0018 | 0.0257 | 0.015 |

| **Regression Results for 7 days look ahead period** | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD_score coefficient** | **p-value** |
| **3m** | 0.0004 | 0.6442 | 0.254 |
| **6m** | 0.0007 | 0.6324 | 0.13 |
| **1yr** | 0.0004 | 0.9561 | 0.252 |
| **2yr** | 0.0072 | 0.3166 | 0.0 |
| **5yr** | 0.0022 | 0.1351 | 0.008 |
| **10yr** | 0.0012 | 0.0723 | 0.053 |
| **30yr** | 0.0019 | 0.0649 | 0.014 |

| **Regression Results for 14 days look ahead period** | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| **3m** | 0.0002 | 0.6238 | 0.409 |
| **6m** | 0.0013 | 1.2799 | 0.038 |
| **1yr** | 0.0009 | 1.4728 | 0.096 |
| **2yr** | 0.0114 | 0.5487 | 0.0 |
| **5yr** | 0.0019 | 0.1713 | 0.013 |
| **10yr** | 0.0018 | 0.1174 | 0.015 |
| **30yr** | 0.0036 | 0.1187 | 0.001 |

| **Regression Results for 50 days look ahead period** | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| **3m** | 0.0024 | 6.0454 | 0.005 |
| **6m** | 0.0090 | 5.5102 | 0.0 |
| **1yr** | 0.0041 | 4.4032 | 0.0 |
| **2yr** | 0.0728 | 2.7902 | 0.0 |
| **5yr** | 0.0036 | 0.4971 | 0.001 |

| | | | |
|---|---|---|---|
| **10yr** | 0.0034 | 0.3335 | 0.001 |
| **30yr** | 0.0067 | 0.3218 | 0.0 |

*Table 6: The results of the regressions of the 30 day average HD score on the different treasury market rates differences with different look ahead periods into the future.*

We see the results in table 6. For the 1 day look ahead period, only the longer date maturities have p values below the 0.05 threshold. The longer maturities produce the most significant results (higher p-values) for all look ahead periods. Increasing the look ahead period results in more significant values, with the 50 day look ahead period having the highest p values across all different maturities.

Both p values and r squared (R2) values are valuable knowledge to assess our model. The coefficient of determination (R2) determines how much of the variation in the market yields can be explained by the hawk dove score. A result of 1 would mean the model perfectly explains the observed variable. In our case results are all less than 0.1, which shows a weak ability of the model to explain the variance within our data. The R2 values change depending on look ahead period and maturities. For all look ahead periods the 2yr maturity has the best R2 score in all look ahead periods. The p-value is also statistically significant except for the 1 day look ahead.

The coefficients of the HD score in our regression are particularly important for this section of our analysis. They determine how much the yields changes when the score changes. In our case the changes are all important. All statistically significant coefficients are positive. The shorter dated maturities in the 50 day look ahead period exhibit the highest coefficients with values above 1. This result seems contrasting to what the graphs were showing us previously where we notice the hawk dove score mean follow more closely the longer dated maturities. This leads to think that the period of the low interest rates could not be easy to visualize graphically overlapping the HD score. Additionally, the result that monetary policy sentiment affects shorter dated maturities is very established in the literature. The FED does control the short term borrowing rates making the short end of the curve more susceptible to monetary policy discussions and decisions.

For completeness let’s take a look at how changing the rolling window parameter within our regression affects the results. This is interesting to look at as the differences can also provide us a glimpse into the time span of market participants and how long their context for interpreting and reacting to monetary policy forward guidance is. Similarly for large language models, investors have a context for their decisions. Changing the average value window can

give us a glimpse into how long this context is. As done previously we use the 1,7,14,50 days look ahead periods, this time with a 14 day rolling mean for the HD score.

| Regression Results for 1 day look ahead period | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| **3m** | 0.0006 | -0.2716 | 0.181 |
| **6m** | 0.0000 | 0.1753 | 0.697 |
| **1yr** | 0.0001 | 0.3911 | 0.66 |
| **2yr** | 0.0017 | 0.0495 | 0.018 |
| **5yr** | 0.0023 | 0.0428 | 0.007 |
| **10yr** | 0.0022 | 0.0300 | 0.007 |
| **30yr** | 0.0014 | 0.0167 | 0.033 |

| Regression Results for 7 days look ahead period | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| **3m** | 0.0002 | 0.3421 | 0.41 |
| **6m** | 0.0003 | 0.3124 | 0.308 |
| **1yr** | 0.0002 | 0.5132 | 0.402 |
| **2yr** | 0.0080 | 0.2450 | 0 |
| **5yr** | 0.0047 | 0.1454 | 0 |
| **10yr** | 0.0033 | 0.0896 | 0.001 |
| **30yr** | 0.0028 | 0.0579 | 0.003 |

| Regression Results for 14 days look ahead period | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| **3m** | 7.37E-05 | 0.271 | 0.625 |
| **6m** | 0.000 | 0.405 | 0.37 |
| **1yr** | 0.000 | 0.575 | 0.376 |
| **2yr** | 0.008 | 0.333 | 0 |
| **5yr** | 0.002 | 0.118 | 0.02 |
| **10yr** | 0.001 | 0.072 | 0.044 |
| **30yr** | 0.002 | 0.066 | 0.009 |

| Regression Results for 50 days look ahead period | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| **3m** | 0.002 | 3.562 | 0.024 |
| **6m** | 0.004 | 2.692 | 0 |
| **1yr** | 0.002 | 2.132 | 0.016 |
| **2yr** | 0.047 | 1.639 | 0 |
| **5yr** | 0.006 | 0.477 | 0 |
| **10yr** | 0.008 | 0.376 | 0 |
| **30yr** | 0.011 | 0.298 | 0 |

*Table 7: The results of the regressions of the 14 day average HD score on the different treasury market rates differences with different look ahead periods into the future.*

Interestingly, the p values do change slightly between the two cases. The smaller context window shows it is more significant for the shorter look ahead periods.

This appears to be the pattern also for the R2 values. These improve for the shorter look ahead periods. Particularly noticeable for the 1 day ahead case. Overall neither p-values nor R2 values significantly improve or worsen when shortening the time span.
The coefficients show us a different picture than the larger context window. These are all much lower than the previous results. This is noticeable for all the look ahead periods. Although the relationships between the variables is weak, the lower coefficients couple with higher R2 values for the shorter look ahead periods enable some conclusions from this difference in the context window. The differences highlight that the shorter context might explain better the shorter look ahead periods, indicating investors look at more immediate sentiment from the FED in the very short term. The lower coefficients instead indicate a weaker impact of the latest communications on the interest rate markets.
Looking at the results, monthly monetary policy communication has a relationship with market prices more than 1 month ahead. The communication also displays a much stronger impact and relationship further ahead in the future than in the short term. In the longer look ahead periods, the shorter term maturities present a higher coefficient of the HD score showing a larger impact on the short end of the curve by monetary policy.

### 3.3.2. Regression on Interest Rate Differentials

One of the important economic indicators FED officials and investors look at is the term structure of interest rates. The difference between the long and short end of the yield curve. In theory, investors should be rewarded with a higher rate for lending money for longer periods. The normal yield curve, constructed by the interest rates of the treasury notes at different maturities, is in normal times upward sloping, with higher interest rates towards the long end of the curve. There are times in which an “inversion” of the yield curve occurs, the short-term rates become higher than the long end and the curve flattens and ultimately becomes downward sloping. The FED and Investors see this inversion as sign of recession, as lenders become bearish and flock to longer term fixed income securities. There are multiple theories on the reasons behind the shape of the yield curve, but an inversion is seen as a sign of a recession for the FED and investors alike.
I turn the regression of our HD score to two metrics: the spread between the 10 year treasury note yield and the 2 year notes. Additionally, I also consider the spread between the 10 year and the 3 months yields. The mean of the HD score is of 30 days.
The p-values for all the look ahead periods are statistically significant.

| Regression Results for 1 day look ahead period | | | |
|---|---|---|---|
| **Maturity Differential** | **R-squared** | **HD score coefficient** | **p-value** |
| 10 yrs – 2yrs | 0.123 | -9.762 | 0 |
| 10yrs – 3 mths | 0.048 | -7.599 | 0 |

| Regression Results for 7 days look ahead period | | | |
|---|---|---|---|
| **Maturity Differential** | **R-squared** | **HD score coefficient** | **p-value** |
| 10 yrs – 2yrs | 0.114 | -9.374 | 0 |
| 10yrs – 3 mths | 0.042 | -7.086 | 0 |

| Regression Results for 14 days look ahead period | | | |
|---|---|---|---|
| **Maturity Differential** | **R-squared** | **HD score coefficient** | **p-value** |
| 10 yrs – 2yrs | 0.103 | -8.884 | 0 |
| 10yrs – 3 mths | 0.036 | -6.464 | 0 |

| Regression Results for 50 days look ahead period | | | |
|---|---|---|---|
| **Maturity Differential** | **R-squared** | **HD score coefficient** | **p-value** |
| 10 yrs – 2yrs | 0.063 | -6.812 | 0 |
| 10yrs – 3 mths | 0.010 | -3.232 | 0 |

*Table 8: The results of the regressions of the 30 day average HD score on the treasury market rate differentials with different look ahead periods into the future.*

The R2 values for all the look ahead periods all seem to explain the model better than the regression on the market interest rates. The shorter look ahead periods display higher R2 values above 0.1 which was the maximum achieved with the previous variables the regression was ran on.

All coefficients indicate a negative relationship of the interest rate spread between the long and short end of the curve and the HD score. This result is very positive for our analysis as it shows the sentiment portrayed by the HD score reflects an important economic indicator in the proper manner. The increase in the hawkishness of the FED leads to an increase in the short-term interest rates which decreases the spread between the maturities (turns more negative). The opposite holds, dovish statements are aimed at increasing the spread positively. All coefficients in this regression are negative and large indicating the score has a strong negative effect on the spreads. The result is very promising and a good proof of the effects of forward guidance and the ability of ChatGPT to identify accurately sentiment.

### 3.4 Daily Price Changes and the Hawk Dove Score

It is also interesting to take the price graph of different securities and see if there are any movements and in which direction are these movements during the days the communication takes place.

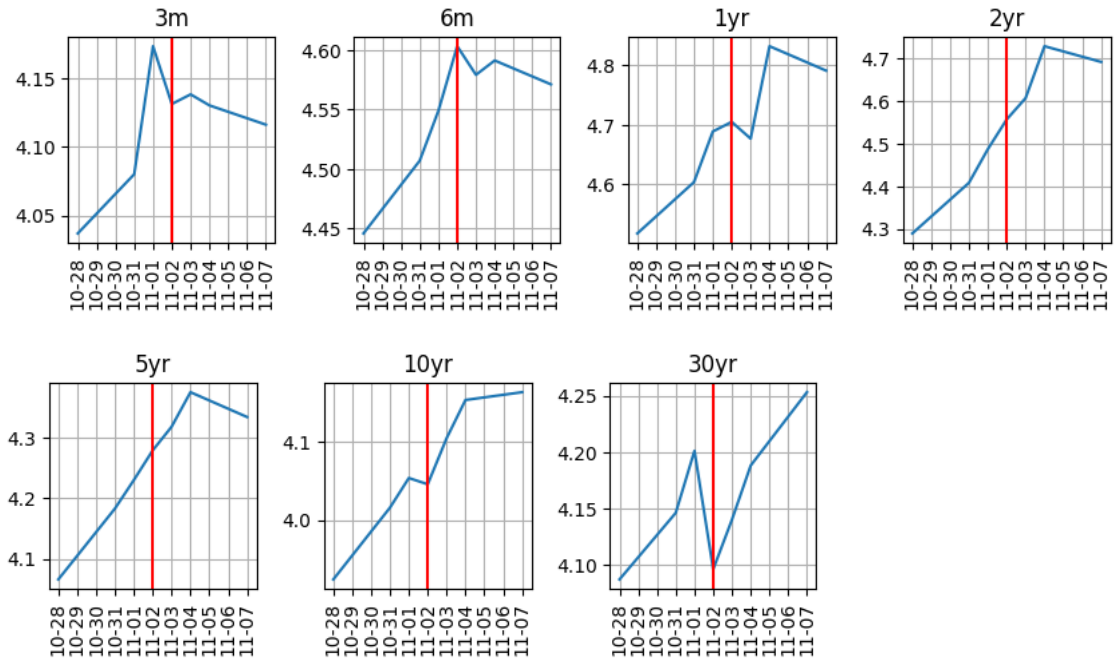


*Figure 15: Graphs illustrate treasury market rates changes on a day with hawkish remarks. The blue line represents the interest rate movement for each treasury. The vertical red line marks the day a communication by the FED has taken place and classified as hawkish by the model.*

These graphs in Figure 15 are taken from the day of a FOMC statement in 2022 which was classified as hawkish by the model. Due to the lack of intraday data, it is difficult to assess the real impact of the speech. This could be observed by looking at the minute data on the day. Here we take the opening price 5 days earlier and 5 days after the statement. With a hawkish remark we should see a rise in the rates. This pattern seems evident for the longer ended 5, 10 and 30 year maturities. This could be caused by the maturities being more sensitive to interest rate changes having a higher duration. Duration is the sensitivity of a bond to interest rate changes. Longer dated bonds have higher duration. What is true for all the maturities is that the days before the FOMC statement there seems to be a rise in the yield.

Looking at the graphs is not enough. As a measure of the LLM in classifying sentiment we can take a look at how many of the times a hawkish statement led to an increase in the markets yields and vice versa for dovish remarks. For this, the methodology is straightforward: count the number of times a hawkish remark led to an increase of the price in 1 day, 2 days and 7 days later in percentage terms of the opening price the day of the statement. This was performed for all the maturities and can be observed in Figures 16 and 17.

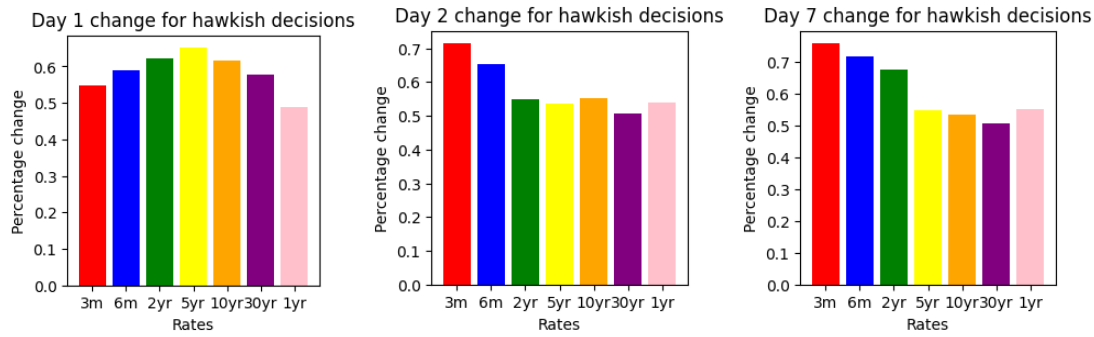


*Figure 16: Each bar in the graphs represent the number of times the 1, 2, and 7 day treasury rate change rose when a hawkish classification was picked up by the model.*

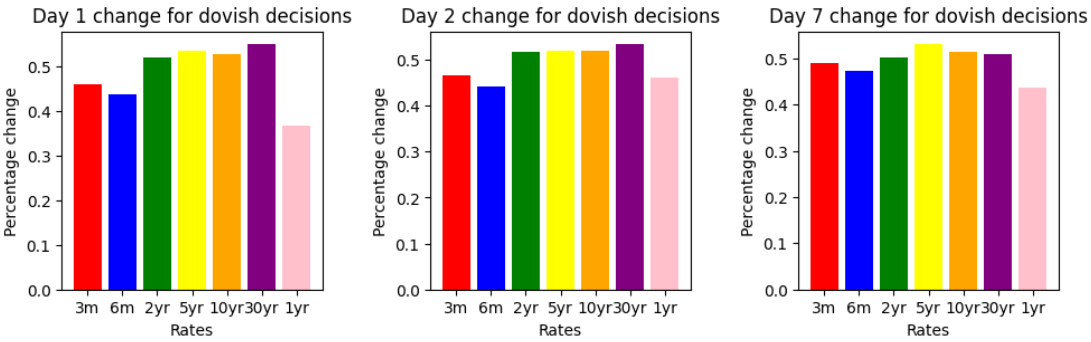


*Figure 17: Each bar in the graphs represent the number of times the 1, 2, and 7 day treasury rate change rose when a dovish classification was picked up by the model.*

For the hawkish remarks, the 3 months and 6 months displayed a price change in the expected direction more than 60% in the 2 day difference and 7 day difference.

The dovish remarks had less promising results. This can be due to the larger amount of datapoints with this classification which lead to a better depiction of reality. None of the daily changes across all maturities led to a price change in the right direction more than 50% of the time. The results appear slightly positive in the case of the hawkish remarks and not conclusive for the dovish case. This seems reasonable as hawkish (seen as negative) remarks could have a wider impact on the markets as opposed to a dovish statement which is aimed at keeping volatility lower.

Other than looking at the amount of price changes in the expected direction, I run regressions on each single hawk dove score. The dependent variable of the regressions is the 1, 2, and 7 day opening price difference in percentage.

| Regression Results for 1 day price change after FED communication | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| 3m | 0.001 | -0.039 | 0.485 |
| 6m | 0.002 | 0.016 | 0.302 |
| 1yr | 0.000 | 0.003 | 0.823 |
| 2yr | 0.020 | 0.030 | 0 |
| 5yr | 0.021 | 0.024 | 0 |
| 10yr | 0.018 | 0.016 | 0.001 |
| 30yr | 0.011 | 0.009 | 0.008 |

| Regression Results for 2 day price change after FED communication | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| 3m | 0.000 | 0.024 | 0.637 |
| 6m | 0.000 | -0.043 | 0.624 |
| 1yr | 0.009 | 0.041 | 0.023 |
| 2yr | 0.006 | 0.027 | 0.047 |
| 5yr | 0.002 | 0.012 | 0.279 |
| 10yr | 0.001 | 0.005 | 0.48 |
| 30yr | 0.001 | 0.005 | 0.405 |

| Regression Results for 7 day price change after FED communication | | | |
|---|---|---|---|
| **Maturity** | **R-squared** | **HD score coefficient** | **p-value** |
| 3m | 0.001 | 0.101 | 0.373 |
| 6m | 0.001 | 0.062 | 0.414 |
| 1yr | 0.025 | 0.095 | 0 |
| 2yr | 0.009 | 0.048 | 0.017 |
| 5yr | 0.005 | 0.031 | 0.064 |
| 10yr | 0.005 | 0.020 | 0.087 |
| 30yr | 0.007 | 0.017 | 0.036 |

*Table 9: The tables show the results of the regression run for each day a hd score is present on the 1, 2, and 7 day interest rate difference for the treasuries listed.*

Not all p-values reach the 0.05 threshold of statistical significance. The results for the 2 year maturities and up looking at the 1 day price change have low p values yet their R2 is low, similarly to the previous regressions which were ran. The 2 day and 7 day price difference yields the least significant results in terms of p values.

Among the coefficients of the regression considering the 1 day price difference, the significant results show a positive price relationship between the score and the yields. The coefficients are lower than the previous regression. Smaller coefficients indicate a smaller impact of the HD score on these price changes. As stated previously, looking at intraday data could yield more positive results as we would be looking at the portion of the price graphs directly concerning the moment the speech was made.

# 4. Literature Review

As the objectives and findings of the paper are bifold the section will be divided into two segments. Firstly, I will display the use of ChatGPT for economic analysis that can be found in the literature. Secondly, I will provide an overview the field of measuring central bank communication broadly and then showcase some literature which uses ChatGPT for this purpose.

## 4.1. ChatGPT for Economic Analysis & Academia

There are 3 articles which cover interest use cases of ChatGPT relevant within the context of this paper. These span also into other contexts other than solely financial research, these findings and applications are relevant for all social sciences (and not only). Three main topics are of interest: extracting categorical information from open ended survey answers, methodically summarizing and analyzing information from large corporate disclosures and lastly generating synthetic labelled data for improving model performance.

### 4.1.1. Open End Survey Coding

In the study: "Does GPT-3 know what the Most Important Issue is? Using Large Language Models to Code Open-Text Social Survey Responses At Scale" (Mellon, Bailey e Scott), the researchers perform an investigation into the potential of GPT-3.5, in coding open-text survey responses. This is useful in the field of academia as well as other business applications and political studies as closed survey questions may prime respondents towards an answer or prejudge answers that the respondents might give. The issue with this method is it requires an extensive amount of time consuming work to label or code the answers to the survey. Taking data from the British Election Study Internet Panel (BESIP), the researchers compared the efficacy of GPT-3 against human coders and traditional supervised learning algorithms (Support Vector Machines) with human labelled datasets of responses. The researchers use 3 shot learning for ChatGPT in their research paper. Notably, while the human coder exhibited superior performance, the margin of this superiority depended on various conditions examined in the paper. When considering the most relevant variable, according to the researchers, for a wide range of applications, the human coder displayed a 97% agreement with the original coder and GPT-3.5 showcased a performance close to this human coder's accuracy of 95%. The SVM fitted to 576,000 previous cases had a 94% agreement. It is important to note how the SVM with only 1000 cases substantially underperformed all other methods.

While there are certain niche research domains demanding the precision and well rounded understanding of a human coder, a vast majority of use cases benefit immensely from GPT-3.5's performance. This assertion is particularly salient when the discourse pertains to topics that have wide discussion in the English language, given that such topics would have been extensively represented in GPT-3's training data. The researchers argue this could be the reason why the model performed well within their study as it is based on a dataset encompassing public opinion. They also argue most open ended surveys will be in assessing public opinion, making the use case very relevant. The study highlights the potential LLMs have to analyze vast datasets without the need for human coders which not only renders the process economically viable but also ensures a level of consistency and standardization that might be difficult to achieve with multiple human coders.

#### 4.1.2. Extracting Information from Company Financials

The study: "Bloated Disclosures: Can ChatGPT Help Investors Process Financial Information?" (Kim, Muhn e Nikolaev) , explores the efficacy of ChatGPT in summarizing intricate corporate disclosures. The authors use the management discussions & analysis section of annual reports and earning calls transcripts to perform their analysis. They use the model to make summaries of these documents and later attempt to focus these on ESG topics and financial performance only.

The summaries on average are 30% in length of the initial documents. One of the primary findings is the notable contrast between the sentiment of the summarized versus the original document in explaining market reactions. The ChatGPT summaries tend to be aligned in sentiment direction with the original texts. The interesting point is the strength of the sentiment, positive sentiment in summaries is more positive than in the original text (and vice versa). The sentiment derived from the summarized content exhibited a stronger association with abnormal returns, especially in the context of conference calls.

The relevance of these findings lies in the potential of LLMs to assist investors in processing financial information. Despite ChatGPT not being specifically trained to summarize corporate disclosures or predict stock performance, the study shows its capability to navigate through the complexities of corporate disclosures and produce summaries that are not only concise but also more informative. This has significant implications for investors with information processing constraints, as it offers a more efficient way to extract and understand the essence of corporate disclosures, ultimately aiding in informed decision-making.

### 4.1.3. Creating Synthetic Data to improve model performance with ChatGPT

The paper, “Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks ” (Moller) confronts human labelled data against synthetic training data for low-resource classification tasks. The author’s methodology involves testing real data vs. synthetic data on 3 different classification tasks. The model used is a smaller LLM and it is confronted with zero shot GPT4 on all the tasks. The tasks differ in nature and in the data available, reflecting the issue of imbalanced datasets when training machine learning models. The synthetic data is obtained by prompting the GPT4 model with few examples. The approach to the data augmentation is bifold in order to assess differences. Proportional data is made which respects the distribution of classes found in the original dataset. Balanced data is also created which changes the amount of labels in each class distorting the original dataset.

Results are mixed depending on the 3 tasks at hand. In the sentiment analysis task, the all the synthetic data performed worse than the original, whilst zero shot GPT4 performed better than the smaller fine-tuned LLMs. On the task involving the identification of hate speech balanced synthetic data appears to perform slightly better than the original data which was highly unbalanced. Zero shot GPT4 in this case performed worse, with a tendency to label inputs as hate speech more often. Lastly, researchers performed a social dimensions task involving labelling what a text conveyed among some categories. Zero shot GPT4 outperforms the fine-tuned models. The synthetic data yields similar results to the original data, also when generating balanced synthetic data.

The results show the strong potential of GPT4 on a wide variety of classification tasks without the need of data which can be hard to obtain. The study also highlights the power of GPT4 in performing tasks also in other languages (Danish for the hate speech classification task). Additionally, the study highlights the increasing power of smaller open-source LLM models which are inexpensive and easy to set up while being more privacy focused.

## 4.2. Central Bank Communication and Asset Prices

Trying to understand the information within the communications of central banks is a widely studied topic. In particular the relationship between communication and asset prices. For this segment I propose two interesting papers delving into measuring central bank communication and how it affects asset prices.

### 4.2.1. Creating a Measure of Central Bank Communication

Many researchers build a measure of central bank communication, particularly relevant in the literature are (Nardelli, Mertens e Tobback) for the ECB or (Jegadeesh e Wu) for the FED. Here I will focus on the paper by (Lucca e Trebbi): “Measuring central bank communication: an automated approach with application to fomc statements ”.

The researchers set to create a hawk dove measure in a similar fashion as the methodology described in this paper. The data used for the FED communication are the FOMC statements since 1999. Additionally the researchers create a market expectations measure as well, which can provide interesting insights when compared to the FED sentiment. They do so by collecting any information available (news, commentaries) regarding the FOMC and policy rate decisions in two different ways: using a google search database and using the dow jones factiva database of economic news.

The research displays several findings. Firstly, short-term Treasuries predominantly react to policy decisions close to policy announcements, while longer-term yields are more influenced by changes in policy communication. Communications also prove to be influential in predicting longer-term yields 1 year ahead. Additionally, the communication effect on yields when compared to the change in the policy rate effect, holds a greater effect. This new perspective suggests that traditional monetary models might have previously missed the vital dimension of communication in their analyses.

Lastly, the more restricted Dow Jones database provided more precise scores than the broader Google data. These measures, which are rooted in expert commentaries, present themselves as alternative markers of market expectations and sentiments, with potential applications extending beyond monetary policies to areas such as political campaigns and corporate communications.

### 4.2.2. Central Bank Tone and Asset Prices

The authors Maik Schmeling and Christian Wagner write the paper: "Does Central Bank Tone Move Asset Prices?" (Schmeling e Wagner), in which they measure the tone of European Central Bank policymakers. They focus on the verbal communication of these during press conferences following policy meetings. The central premise rests on using a systematic approach to define the tone of the ECB president. Their measure is focused on the change in tone between press conferences. They use a well-known financial dictionary (Loughran e Macdonald) for finding negative words within the speeches. The metric they use for tone in a speech is then:

$$Tone = 1 - \frac{Total\ Number\ of\ Words}{Number\ of\ Negative\ Words}$$

The change of tone is the difference between this metric among two consecutive speeches. Their findings suggest that a positive tone surprise correlates with a significant uptick in stock prices, increased interest rates, a reduction in volatility risk premiums, and narrower credit spreads. Notably, these tone effects persist even when accounting for policy actions, monetary policy shocks, and central bank information effects, leading to the conclusion that ECB tone changes transmit price-relevant news to the global markets.

### 4.3. Measuring Central Bank Communication

In this last section of the Literature Review I focus on two papers which have had a significant effect on my research as they pioneer the methodology of classifying central bank communication with ChatGPT.

#### 4.3.1. Federal Reserve Sentence Classification

The researchers Hansen and Kazinnik in their paper: "Can ChatGPT decipher Fedspeak" delve into assessing the ability of GPT3.5 to classify FED sentiment. They set up their research to enable a comparison between GPT3.5 with zero shot learning, GPT3.5 with fine tuning, BERT and also dictionary based methods such as (Loughran e Macdonald). The classification is performed sentence by sentence and involves a word classification on a scale of dovish, slightly dovish, neutral, slightly hawkish, hawkish. Towards the end of their paper they also use GPT4 for the classification and prompt the machine to provide explanations to the classifications. They are able to assess the quality of the classifications through a dataset of labelled sentences on the same scale, performed by their staff.

Their results indicate that GPT3.5 with zero shot learning outperforms BERT and dictionary-based methods by a good margin on all metrics. This margin widens dramatically with the fine-tuned version of GPT3. It is important to point out that the better results with fine-tuned GPT3.5 compared to zero shot learning could change with better prompting techniques and more trial and error in this sphere. The section in which they compare the GPT explanations and the human explanations for why the classification was assigned is also interesting, showing how GPT4 mostly follows the human explanations and GPT3.5 slightly lagging behind.

#### 4.3.2. Measuring Federal Reserve Sentiment with ChatGPT

The paper: "Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis " provides a new comparison of ChatGPT vs other NLP models. The authors create a dataset of

all communications created by the FED (from which this paper takes the press conference meetings) and classifies them with different LLMs. GPT3 is used in a zero-shot manner. The model is compared to other fine tuned pre trained models, including FinBert, pre-trained BERT fine tuned to a financial words dataset.
They manually classify the dataset to obtain a results comparison. The results show that zero shot GPT3.5 outperforms all but the fine-tuned BERT models. Only these are then used for the underlying economic analysis. It is important to mention that also this paper does not put emphasis on the importance of prompting and does not take advantage of few shot learning with GPT3.5. Additionally, the paper classifies each sentence individually.
The underlying economic analysis runs regression on treasury yields and CPI data, showing statistically significant relationships among the variables and the classifications performed by the fine-tuned BERT models.

## 5. Limitations of the Study & Further Studies

The sentiment classification is a part of the study which could have been approached in different ways. In this section I focus on two main issues where I see room for improvement.

### 5.1.1. Splitting the text to classify

Firstly, when deciding how to classify each piece of communication I decided to conduct the study splitting each speech or statement into smaller pieces as described in the methodology section to stay within the API limits yet preserve the meaning of the speech in its entirety as much as possible. The method also reduces the number of total API calls to make to classify the entire dataset which leads to smaller timeframe for the study. This is the same approach as the study conducted by (Kim, Muhn e Nikolaev). Another approach could have been to split each communication into individual sentences and perform the classification by analyzing each sentence. This could have provided more granularity as smaller sections of a speech could be relevant for monetary policy yet they are not picked up by the model in the entirety of the speech. This approach is also used in the paper: (Hansen e Kazinnik) and in the JP Morgan note as explained by (Elder). While this approach would grant more speeches to be classified as relevant it would also be more prone to error as it would fail to capture the combination of sentences within each speech. The model could classify a hawkish sentence such as: “the FED will raise rates when inflation is high” which could be irrelevant within a very long conference speech whose aim is purely to explain how the FED operates. Inputting

larger pieces of text removes this problem yet results in less pieces of communication being deemed as relevant by the model.

### 5.1.2. Finding the appropriate prompt

Secondly, the methodology for finding the best prompt can be made more empirical. To select the best prompt to perform a section of the analysis I used 100 classifications randomly sampled through time as a test to determine whether the model output was sufficient. In addition, a lot of the experimentation with building prompts and setting the correct temperature was done with the "playground" function of OpenAI. Although testing prompts and their effectiveness without having labelled data is hard to make more empirical, the tests performed and their results could be recorded in a more methodical manner. This can also be a significant informational result for the paper as there is a lack in the literature of this process being well recorded and shared with readers. Of the papers cited within the literature review which made use of ChatGPT, none of them present a detailed section on prompting and how the researchers came up for the final prompt used for their classification.

## 5.2. Further Studies

There are 2 main improvements which can be interesting to further study and provide more insights on the capabilities of ChatGPT in classifying economic text data.

### 5.2.1. Use of Intraday Data

The use of intraday price data opens the possibility to two main further studies. Firstly, the major price fluctuations created by the FOMC press conferences or statements releases occur during the time of the conference and at its end when market participants are closely weighing each word the speakers. Trying to find the relationship between these intraday fluctuations and the classified speeches would shed a further light on both the relationship of monetary policy communication on market prices and the ability of ChatGPT to capture this relationship.

Additionally, having the intraday data also gives more importance to the decision on how to split the text to classify. A sentence-by-sentence approach with minute data can reveal more informative as it would capture the impact of sentences which shock investors.

### 5.2.3. Use of synthetic data for fine-tuning

With a different approach than the paper (Moller), using synthetic data created by ChatGPT to fine-tune the GPT3 model is an interesting area for further studies. This would add the step of understanding how to build the synthetic dataset through prompts. Additionally, the comparison between the few-shot learning model and the fine-tuned model with synthetic

data of different kinds is also an interesting area to further explore. Fine tuning models with synthetic data can be a field which will develop in the coming period as it can improve model performance especially when there is a lack of high-quality data.

### 5.2.4. Fine tuning the speech to the price changes

Model fine-tuning also introduces the possibility of directly associating each speech with the relative price change as label. The price change would replace in the dataset the classification on the hawkish to dovish scale. With this method a model which predicts the daily price change after a FED press conference can be obtained. This kind of study can be very important not only for academic purposes but also within the industry as a reliable model would be a good indicator for traders to follow.

## 6. Conclusion

This paper attempts to make one main contribution to the literature. It shows the ability of GPT3.5 in successfully performing a classification task which can be used for economic analysis without the need of a labelled dataset.

The paper also illustrates in detail a method of using the ChatGPT API to classify a dataset without the use of human labelled data. The study reveals that forming the right combination of prompts to input into the API is key to successfully using the tool. Providing examples of each classification inside the prompt, a technique known as few-shot learning, greatly improves the reliability of responses which the AI is going to provide back. Additionally, in order to create a clean dataset out of these classifications, another prompt is needed for the so called “answer extraction”. This step cleans many of the classifications which might have misrepresented the original category desired or have provided a more verbose answer. The combination has successfully created a classified dataset of FED communications.

Additionally, the task at hand, classifying FED communication, involves complex language which can be hard to grasp also for humans. Thus, the results may be even more relevant for researchers and industry professionals.

This paper does not test the classifications against a labelled dataset but delves into measuring FED communication and assessing its level of information on future market treasury yields. The measure created “hawk/dove” score or HD score captures relationships with market yields which are statistically significant.

Firstly, the 30-day average value of the HD score has a positive coefficient on market yield returns particularly 50 days ahead of the score. These coefficients are quite large with 5-6

points increase with the shorter dated maturities, with these values becoming smaller going along the curve.

Secondly, the most significant and interesting result is the impact of the mean value of the HD score on the interest rate differentials of the longer term 10 year maturity with the 2 year and 3 months short term maturities. This is one of the lead indicators policy makers try to control and investors look at in the markets. The relationships between the score and these market yields are all significant no matter the look ahead period. R2 values are also higher among all results. The coefficients are also negative and present higher values which is the relationship which should be expected as higher index values should decrease the value of long minus short maturities.

Thirdly, the model also shows a relationship with the daily yield changes for each communication. The relationship with the 1 and 7 day yield differences have statistical significance, even if with lower coefficients and R2 values than the previous relationships.

The ability of the model to capture these relationships gives confidence to using LLMs, ChatGPT in particular, for classifying datasets without the need for human classification. The model is able to capture complex relationships well studied in the literature, at statically significant levels comparable to other more complex and time consuming methodologies. The paper also shows how the informational content of policy makers communications is highly relevant for market prices and effectively serves as an additional tool for monetary policy.

## Bibliography


Pierce, John R. *Symbols, Signals and Noise*. 1979.

Roberts, Eric. *Entropy and Redundancy in English*. n.d. <https://cs.stanford.edu/people/eroberts/courses/soco/projects/1999-00/information-theory/entropy_of_english_9.html>.

OpenAI. *tiktoken*. n.d. <https://github.com/openai/tiktoken>.

FED. n.d. <https://www.federalreserve.gov/newsevents/pressreleases/monetary20110324a.htm>.

Pooter, Michiel De. "The Information Content of the Post-FOMC Meeting Press Conference." (n.d.).

OpenAI. n.d. <https://platform.openai.com/docs/introduction>.

Gnan, Philipp, et al. "Decpihering Monetary Policy Shocks." (2022).

Hansen, Anne Lundgaard and Sophia Kazinnik. "Can ChatGPT Decipher Fedspeak?" (2023).

Kojima, Takeshi, et al. "Large Language Models are Zero-Shot Reasoners." (2023).

Shah, Agam, Suvan Paturi and Sudheer Chava. "Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis." (2023).

Elder, Bryce. *AI wants to know what dove is*. 2023. <https://www.ft.com/content/4a6d60bd-62a2-4c98-a0c9-5d58814ce0bf>.

Hansen, Anne Lundgaard and Sophia Kazinnik. "Can ChatGPT Decipher Fedspeak?" (2023).

Lucca, David O. and Francesco Trebbi. "Meauring Central Bank Communication: an automated approach with application to FOMC statements." *National Bureau of Economic Research* (2009).

Bernanke, Ben. *21st Century Monetary Policy*. n.d.

Campbell, et al. "Macroeconomic Effects of Federal Reserve Forward Guidance." (n.d.).

Mellon, et al. "Does GPT-3 know what the Most Important Issue is? Using Large Language Models to Code Open-Text Social Survey Responses At Scale." (2023).

Kim, Muhn and Nikolaev. "Bloated Disclosures: Can ChatGPT Help Investors Process Financial Information?" (2023).

Moller, Dalsgaard, Pera, Aiello. "Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks." (2023).

Jegadeesh and Wu. "Deciphering Fedspeak: The Information Content of FOMC Meetings." (2015).

Schmeling and Wagner. "Does Central Bank Tone Move Asset Prices." (2022).

Nardelli, Mertens and Tobback. "Between Hawks and Doves: Measuring Central Bank Communication." (n.d.).

Wolfram. *What does chatgpt do and how does it work?* 2023.

Brown, et al. "Language Models are Few Shot Learners." (2020).

Loughran and Macdonald. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." (2011).