Analysis and Mitigation of ChatGPT's Dovish Bias on Classifying Fedspeak

Amaury Mantion mantion.amaury@gmail.com
University College London
Melvin Kianmanesh Rad melvin.kian@effixis.ch
Effixis SA
Christophe Morel cmorel@groupama-am.fr
Groupama Asset Management
Romain Faure rfaure@groupama-am.fr
Groupama Asset Management
Zachary Schillaci zachary.schillaci@effixis.ch
Effixis SA

Abstract

The work "Can ChatGPT Decipher Fedspeak?" (Hansen and Kazinnik, 2023) marked a significant milestone in the application of Generative Pre-trained Transformer (GPT) (Alec and Karthik, 2018) language models in deciphering and classifying communications from the United States Federal Reserve, commonly referred to as *Fedspeak*. The paper showed an exceptional improvement in classifying policy stances into categories dovish, neutral, or hawkish sentiments compared to other methods. However, despite the impressive performance gains, an underlying issue in the approach was left unaddressed: a pronounced bias towards dovish classification. Our study builds on this foundation by proposing a potential explanation for the observed bias, hypothesizing that the predominance of economically dovish conditions over the past two decades has led to an imbalance in the corpus used for training large language models such as OpenAI's ChatGPT (OpenAI, 2023). This era's economic trends may therefore have a disproportionate influence on the text generations of such models. This observation presents a rare and tangible evidence of data drift in machine learning models, with implications that extend beyond the context of classifying Federal Reserve communications. The objectives of this paper are twofold: firstly, to analyze and quantify the aforementioned bias and, secondly, to explore corrective techniques to build a more robust Fedspeak classifier.

Keywords: Natural Language Processing (NLP), Generative Pre-trained Transformer (GPT), ChatGPT, GPT-4, Artificial Intelligence (AI), Prompt Engineering, Supervised Fine-tuning (SFT), Bias, Classification, Central Bank Communication, Federal Reserve, Monetary Policy

Electronic copy available at: https://ssrn.com/abstract=4769112

Contents

1 Introduction **3**
1.1 Large Language Models . . . 3
1.2 Objective . . . 3

2 Literature Review **5**
2.1 GPT . . . 5
2.2 Bias in Machine Learning . . . 5
2.3 Evolution of LLMs . . . 6
2.4 Prompting Techniques for LLMs . . . 6
2.5 Classification of Fedspeak . . . 7

3 Methods **8**
3.1 Data . . . 8
3.2 Classification . . . 9
3.3 Model Generation . . . 10

4 Results **11**
4.1 Classification Bias . . . 11
4.2 Bias Mitigation . . . 12
4.2.1 Threshold shifting . . . 12
4.2.2 Prompt engineering . . . 13
4.2.3 Combined approach . . . 16
4.2.4 Supervised fine-tuning . . . 18
4.3 Summary . . . 20

5 Interpretation **20**
5.1 Potential Origin of Bias . . . 20
5.2 Limitations . . . 22

6 Conclusion **23**

A Appendix **24**
A.1 Bias in Hansen and Kazinnik (2023) . . . 24
A.2 Dependence on Sampling Temperature . . . 24
A.3 Quantifying the Impact of different Prompting Techniques . . . 24

Electronic copy available at: https://ssrn.com/abstract=4769112

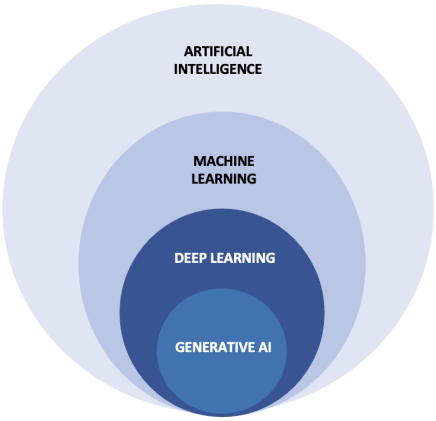


Figure 1: Illustration of the field of Artificial Intelligence (AI) and its various sub-disciplines. Machine learning is the branch of AI which uses data to train models, effectively teaching systems how to learn independent of human oversight. Deep learning is a further subset of machine learning, synonymous with the specific application of neural networks and gradient descent, often using massive quantities of data. Generative AI (GenAI) encompasses a variety of techniques for training models to generate novel content, built primarily on a deep learning paradigm.

1. Introduction

1.1 Large Language Models

Large Language Models (LLMs) are a type deep learning model suited for text generation and chat functionalities. As an LLM-powered chatbot, OpenAI's ChatGPT can write essays, compose emails, generate creative content, engage in coherent dialogue, and even produce functional programming code. Of relevance here, ChatGPT and other similarly trained models are particularly adept at solving natural language tasks, such as translation, named entity recognition, and text classification, when given clearly specified instructions. Prior to the current generation of instruction-tuned LLMs, addressing any one of these tasks typically required the use of specialized approaches in natural language processing (NLP) to collect task-specific datasets and train custom models. Now, for some applications, what would previously require a team of experienced machine learning engineers, can now handled by an individual using ChatGPT. In this study, the capabilities of ChatGPT as a zero and few-shot text classifier are explored.

1.2 Objective

The language used by the United States Federal Reserve in their speeches and announcements, commonly referred to as *Fedspeak*, is infamously challenging to interpret due to its complex and indirect nature. Nonetheless, deciphering the true meaning of these policy communications is vital for those involved in financial markets, government policy-making, and the broader public. In theory, AI systems may help in the automatic classification and analysis of Fedspeak, but, in practice, building well-performing models remains a challenge

Electronic copy available at: https://ssrn.com/abstract=4769112

due to the demanding NLU requirements. Traditional NLP approaches relying on syntactic analysis often fail to capture the nuance in such complex text. Machine and deep learning approaches may be more suitable, but typically require large, high-quality datasets to learn from. Given their unmatched NLU and few-shot learning capabilities, modern large language models, such as ChatGPT, are therefore an excellent choice for the task.

In terms of general methodology, it is important to understand the fundamental difference in approach between Hansen and Kazinnik (2023) and this paper. In the former, the authors perform single sampling classification, that is, they prompt ChatGPT to classify each quote exactly once. In this paper, we perform 10-sample classification to better quantify and correct the bias. Further analysis into the bias is performed through a study of the confusion matrices and probability density functions (PDFs) of the ChatGPT-based classifier model. To correct for the bias, we introduce a multi-pronged approach that includes shifting classification thresholds, prompt engineering, and, lastly, supervised fine-tuning (SFT). These methods successfully correct the bias, leading to a system that not only preserves the strengths of the original ChatGPT classifier, but simultaneously improves accuracy to levels previously unattained in Hansen and Kazinnik (2023).

Beyond the methodology, there is also a major difference between the Hansen and Kazinnik (2023) paper and this one in terms of dataset. The Hansen and Kazinnik (2023) paper uses sentences taken from FOMC press releases, which are extremely coded and standardized. This paper, on the other hand, uses quotes from central banker speeches, sentences that are far more complex than sentences from FOMC statements.

The implications of this research are broad and multifaceted, with relevance to both the economics and AI domains. On a technical level, the methodology demonstrates how seemingly entrenched biases can be diagnosed and corrected using relatively straightforward techniques. This paves the way for more robust and trustworthy machine learning models and a better understanding of modern LLM systems. On a practical level, the enhanced classification process provides a tool that is more accurate and accessible to a wider audience, including non-experts who often find difficulty interpreting Fed communication. This has far-reaching implications for the financial markets, policymakers, and the general public, offering a more transparent lens through which to view the often-cryptic language of the Federal Reserve. The reproducible nature of our method adds further value, inviting other researchers to explore and apply it in other domains.

The importance of finding more automated ways of interpreting Fedspeak is accentuated by the volume of text-based information (publications, speeches) produced by central bankers that has considerably increased. Over the years, for example, they have raised the frequency of their speeches. Before the financial crisis of 2007-2009, the FOMC statements consisted of just a few sentences, and central bankers' speeches were measured out. Today, central bankers speak several times a day, to the point where it is almost impossible to read everything. Economic and financial professionals therefore need systematic tools that can read for them.

Our paper is organised to guide readers through our discovery and resolution of the bias. Section 2 provides a comprehensive literature review, situating our work within the broader field of natural language processing and central bank communication. Section 3.1 details the data used in the process, elaborating on the specific challenges of analysing Fedspeak and the steps taken to ensure a representative dataset. Section 3 outlines the methods

Electronic copy available at: https://ssrn.com/abstract=4769112

used to both identify and correct the bias, supported by statistical analyses. Section 4 presents the results based on our various approaches. Finally, Section 6 concludes the paper, offering reflections on the research's broader societal and technological impact, and laying a foundation for future exploration and innovation in unbiased machine learning.

2. Literature Review

2.1 GPT

Generative pre-trained transformers (GPT) (Alec and Karthik, 2018) are a family of artificial intelligence (AI) models and a well-known example of generative AI (see Figure 1) developed for natural language understanding (NLU) and natural language generation (NLG). Specifically, they are a type of large language model (LLM) based on a particular deep learning model architecture, known as the decoder-only transformer (Ashish et al., 2023). The original transformer (Ashish et al., 2023), created by researchers at Google, was developed specifically for the task of machine translation, and later adapted for general purpose language generation in the GPT-series of models created by researchers at OpenAI (Alec and Karthik, 2018). Research into the third generation GPT model, GPT-3, (B. et al., 2020) demonstrated that sufficiently large language models are capable of solving a wide variety of natural language tasks, provided they are *prompted* with the right instructions and examples. This discovery motivated OpenAI to continue the development of larger, more capable models which were further fine-tuned to follow user instructions, eventually leading to the commercial success of ChatGPT. The expanding capabilities and increasing sophistication of LLMs has become a focal point in both AI research and practical application.

2.2 Bias in Machine Learning

In the context of machine learning, there are many forms of bias which may affect the reliability and fairness of a model's predictions. In Ninareh et al. (2022), the authors detail a comprehensive survey on bias and its many manifestations in machine learning. As the authors note, one major class of bias is that of *user-to-data* bias, wherein biases present in a user population are reflected in the data in which they generate (Ninareh et al., 2022). Models which are trained on such data may learn the same biases.

State-of-the-art LLMs are by no means immune to bias and suffer from many of its prominent manifestations, as outlined in Ninareh et al. (2022). Indeed, given the massive datasets of unstructured text used to train LLMs, documenting and understanding the multitude of emergent biases remains a challenge. O. et al. (2023) specifically focus on the social biases affecting LLMs and further outlines detailed taxonomies of bias evaluation and mitigation techniques. A specific example of user-to-data bias is *temporal bias*, which occurs when the data used to train a model reflects the specific attitudes, beliefs, or trends of a particular time period (Ninareh et al., 2022). In the context of this project, if a language model is trained on internet data from a period largely characterized by dovish economic perspectives, it may have disproportionately learned these viewpoints, leading to a bias in its classification of economic statements.

Electronic copy available at: https://ssrn.com/abstract=4769112

2.3 Evolution of LLMs

OpenAI's seminal paper "Language Models are Few-Shot Learners" (B. et al., 2020) demonstrated the *few-shot* learning capabilities of GPT-3, wherein the model could effectively learn how to solve novel natural language tasks with only an instruction and anywhere from zero (*zero-shot*), one (*one-shot*), or a handful (*few-shot*) of examples provided at inference time. Crucially, this did not require training the model on the specific tasks, a process which is normally compute and data intensive. Research into supervised fine-tuning (SFT) and reinforcement learning from human feedback (Paul et al., 2023), further expanded the instruction-following capabilities of models, leading to the development of InstructGPT (Long et al., 2022). OpenAI's ChatGPT is a direct descendent of this research and follows the now standard recipe for self-supervised pre-training followed by SFT and RLHF, as outlined in Long et al. (2022). During self-supervised pre-training (B. et al., 2020), models are trained with the language-modeling objective to predict missing words in text. This is the most expensive step of training and requires vast data resources to effectively train models. While such pre-trained models have a highly evolved understanding of language and are very capable of sophisticated language generation, they are poorly equipped to follow user instructions. This significantly limits the direct usability of these so-called *foundation models* for many downstream tasks. SFT (Long et al., 2022) contains the training of language models over much smaller supervised datasets consisting of high-quality, often manually curated prompt-response pairs. Lastly, RLHF (Paul et al., 2023; Long et al., 2022) further fine-tunes the model to align with human preferences, with a specific focus on maximizing helpfulness.

2.4 Prompting Techniques for LLMs

Prompting techniques are strategies used to effectively elicit a desired response - in terms of content, format, and style - from a generative AI model, such as ChatGPT. The outputs of generative models are directly conditioned on the prompts provided as input, and therefore, carefully crafted prompts can significantly impact the quality and relevance of the generated responses. Effective prompting is crucial for steering the model's creativity towards desired outcomes, minimizing the risk of irrelevant or inappropriate output, and ensuring that the AI-generated content aligns with the intended purpose and context.

On a technical level, language model prompt engineering is an emergent property of the in-context learning capabilities of such models, as demonstrated by B. et al. (2020). During the compute-intensive pre-training stage, LLMs develop a rich understanding of the statistical and semantic patterns of language purely by optimizing for next token prediction. Subsequently, during supervised fine-tuning, LLMs are trained along the direction of instruction-following to better answer user queries. This fine-tuning generally reinforces desired behaviors such as truthfulness, helpfulness, and harmlessness. Crucially, however, the next token prediction objective remains unchanged. As a result, LLMs are fundamentally probabilistic sequence generators and not question-answering oracles. This means that, technically speaking, the primary *objective* of a generative language model is not to respond correctly, but instead to generate a likely continuation of the user's prompt, based on the entirety of its training data. For this reason, additions such as "respond in the manner of an expert", "provide a detailed reasoning", and "let's think step by step" (Takeshi et al.,

Electronic copy available at: https://ssrn.com/abstract=4769112

2023) can significantly improve the quality of output by conditioning the model's generation along more desired directions. Furthermore, due to their sequential nature, instructing a model to expand on its reasoning or provide step-by-step thinking allows it to process and generate longer, more coherent responses. This approach leverages the auto-regressive nature of LLMs, where each new token is generated based on the preceding context, thereby enriching the quality of the final response.

2.5 Classification of Fedspeak

The Federal Reserve, commonly known as the "Fed", is the central banking system of the United States. One of its primary roles includes managing the country's monetary policy which it achieves through the manipulation of interest rates and of money supply, aiming to control inflation, maximize employment, and stabilise long-term interest rates. The Fed's decisions on interest rates and other policies have far-reaching implications for the overall economy, influencing everything from inflation rates to employment, bank lending, and exchange rates. "Fed watchers" therefore pay close attention to what is said by the Fed to get a better idea of what actions are likely to be taken.

Asset managers and hedge funds in particular will use this information to formulate a view on what will happen in the economy and take positions that would be profitable if correct. These institutions therefore try to listen for "hawkish" and "dovish" sentences to describe the tone or stance of language used by central banks. A "hawkish" sentence implies a stance that favors tighter monetary policies, often associated with concerns about inflation. This might include language suggesting an increase in interest rates or a reduction in the money supply. On the other hand, a "dovish" sentence indicates a preference for looser monetary policies, typically to encourage economic growth, and might include indications of lowering interest rates or increasing the money supply.

Naturally, a position will be worth more if the investors is the first to have this view and other investors follow suit. Therefore quickly and correctly interpreting Fedspeak is of great monetary value and significant investment has gone into developing automated solutions in this endeavour.

Before the rise of deep learning approaches, the main method of classification was dictionary based, relying upon a predefined list of words with an associated sentiment score. Some well-known dictionaries, as mentioned in Hansen and Kazinnik (2023), are Loughran and McDonald (LM) (Tim and Bill, 2010), Henry (Elaine, 2006), and the NRC Word-Emotion Association Lexicon developed by Mohammad and Turney (M. and D., 2013).

These are simple and sometimes effective methods, but struggle to understand the nuanced meaning of sentences in their broader economic context. Hence AI, and LLMs in particular, provide a potentially new avenue of improvement in the search for a quicker and more accurate interpretation method of Fedspeak. Our study builds upon the work of Hansen and Kazinnik (2023), wherein the authors demonstrate the efficacy of transformer-based (Ashish et al., 2023) language models in deciphering the intricate and technical language of the Federal Reserve. The authors use a combination of BERT-style (Jacob et al., 2019) encoder-only models and GPT-style Alec and Karthik (2018) decoder-only models. The BERT-based approaches of Hansen and Kazinnik (2023) use a few-shot, supervised learning pipeline, known as SetFit (Lewis et al., 2022), to adapt general purpose sentence

Electronic copy available at: https://ssrn.com/abstract=4769112

transformer models (Nils and Iryna, 2019) to the specific classification objective. SetFit provides an efficient mechanism for few-shot classification without the need for advanced prompting and, thanks to its reliance on supervised learning, often outperforms in-context learning based approaches with much larger models (Lewis et al., 2022). This model serves as a reference in Hansen and Kazinnik (2023), offering a comparative perspective against the few-shot, in-context learning capabilities of ChatGPT.

Label	**Count**	**Percentage**
Dovish	261	46%
Hawkish	310	54%
Total	571	100%

Table 1: Breakdown of dovish and hawkish quotes in Morel's binarised dataset.

Binarising the data in this manner helps reduce the subjectivity of labels, providing a clearer distinction between the two policy stances. However, it is important understand that the labels remain the product of the subjective interpretation of Morel. Hence the objective of this paper should be understood as a study of classification bias against the chief economist's expert reasoning.

Metadata were added to each quote, including the speaker, date, keywords, and a specific indicator reflecting the state of the U.S. economy at the time of the quote. This indicator, known as a composite of business climate in the US economy, is constructed with a weighting of 80% on the ISM in industry, 15% on the ISM in services, and 5% on the NAHB in construction. These weights are determined to minimise the gap between the composite indicator and the difference between actual and potential GDP. Unlike other potential conditioning variables, such as inflation indicators or interest rate series, which might introduce biased opinions, the composite indicator offers a more objective perspective on the state of the economy. Therefore, it has been chosen to provide context to the quotes without conditioning the opinion extracted by ChatGPT to another biased financial market opinion.

This dataset, with its unique combination of quotes, contextual information, and straightforward classification method, provides the foundation for our study. It enables an exploration into ChatGPT's ability to decipher and classify Federal Reserve communications, taking into account not only the language of the quotes but also the broader economic context. The inclusion of the composite economic indicator adds an additional layer of complexity and relevance, grounding the analysis in an objective measure of the U.S. economy's state.

3. Methods

3.1 Data

To facilitate our study, we use an expert-labelled dataset of Fedspeak curated by Christophe Morel, Fed watcher and chief economist of a European asset management company. The dataset consists of approximately 600 quotes taken from speeches delivered by central bankers at the Federal Reserve System. The quotes come from Goldman Sachs Fed Chat-

Electronic copy available at: https://ssrn.com/abstract=4769112

terbox reports published before each FOMC meeting. Those reports highlight the most notable comments on key topics from Fed officials. Our dataset only includes quotes on the theme of growth. It is important to note that these quotes span the time period from 2018 to 2023, and represent some of the most cryptic and idiosyncratic examples of Fedspeak. By comparison, the quotes from Hansen and Kazinnik (2023) were taken from the Federal Open Market Committee (FOMC) press releases, which are generally more coded and standardised. The quotes of Morel's dataset are manually annotated with a ordinal score ranging from –3 to 3 in increments of 0.5, indicating the sentiment of the quote from dovish to hawkish, respectively. Hence, the more negative the value, the more dovish the quote; and the more positive the value, the more hawkish the quote is interpreted to be. For the purposes of our analysis, all quotes with a neutral label of 0 are removed and the remaining quotes are assigned a binary label of dovish or hawkish. Quotes with a negative value are relabelled as dovish, while quote ones with a positive value were relabelled as hawkish. The resulting dataset consists of 571 quotes split into two fairly well-balanced classes: 261 (46%) dovish quotes and 310 (54%) hawkish quotes (see Table 1). This well-balanced dataset reflects the fact that there are as many as periods with tightening Fed Funds (2018; 2022-2023) as there are with monetary easing (2019-2021).

3.2 Classification

To quantify the presence of a bias in the classification of quotes as either dovish or hawkish, we analyze the confusion tables, which are an efficient way to summarise how the quotes are correctly or incorrectly classified into the two classes. This is instrumental in giving an overview of how well the classification process performed. To better visualise the potential bias, we plot the probability density functions (PDFs) of $P(D|D)$ and $1 - P(H|H)$, where D and H represent dovish and hawkish quotes, respectively. For an unbiased classifier, the expectation is to see two distinct distributions at 0 and 1 with minimal overlap; however, a considerable overlap between these distributions would be a clear indication of bias. Analysing this overlap helps in understanding the degree and nature of the bias present in the model.

Given the complexities of using token log probabilities from large language models in binary classification, we adopt a more straightforward approach based on sampling statistics. For each quote, we sample the model's classification response 10 times, while maintaining a constant sampling temperature. Our comparative analysis extended to 20 samples revealed no notable differences, reinforcing our choice of 10 runs as a balance of efficiency and effectiveness. This approach allows us to calculate empirical probability estimates for each quote by evaluating the frequency of each classification outcome across the 10 generated samples.

After determining the probabilities of a quote being classified as either dovish or hawkish, we introduced the concept of threshold shifting. This technique allows for adjustments to the classification thresholds to align with observed probabilities. By default, the usual threshold is implicitly set as 50%, meaning quotes are assigned the category with the highest probability. However, using this default option does not help to reduce the observed bias on our task. To ensure accurate classification, we first tune the threshold for the dataset. For instance, if a quote that is genuinely hawkish is mistakenly classified as dovish 70% of

Electronic copy available at: https://ssrn.com/abstract=4769112

the time and as hawkish 30% of the time, we modify the threshold; originally set at 50%, we lower it to 30% for hawkish classifications. This change aims to correctly identify such quotes as hawkish. To determine the threshold to be used for the entire dataset, we use a numerical technique based on finding the optimal level between the true positive rate (TPR) and the false positive rate (FPR). To measure the improvement of the same model before and after threshold shifting, we compute the change in accuracy and the F_1 score.

Receiver operating characteristic (ROC) curves are graphical representations used to illustrate the diagnostic ability of binary classifiers. An ROC curve is plotted with the true positive rate against the false positive rate at various threshold settings. The area under the ROC curve (AUC) provides a scalar value that quantifies the overall ability of the model to discriminate between positive and negative classes. The AUC value is particularly valuable for model comparison because it captures the overall performance across all threshold settings, unlike other metrics which might only be valid for a specific threshold. A model with an AUC value closer to 1 indicates higher discriminative power, making it an excellent metric for unbiased model comparison. It ensures a holistic view of model performance, taking into account both sensitivity (TPR) and specificity (1 - FPR) over a range of decision thresholds. In our study, we leverage the AUC as a standard metric to compare different classifiers. By evaluating and comparing the AUC values of the classifiers under consideration, we are able to identify the model that performs best in distinguishing between the hawkish and dovish quotes. This comparison is not only statistically rigorous but also provides an intuitive visual insight into the relative performances of the classifiers.

3.3 Model Generation

The non-deterministic, open-ended generation of ChatGPT presents some additional hurdles in building robust classifiers. In instances where the output of the ChatGPT model is not aligned with our prompted instructions (e.g. outputting "neutral" when asked to classify the quote as either "hawkish" or "dovish"), we opt to re-run the generation until an acceptable answer is obtained. This iterative process helps maintain the integrity and reliability of our analysis.

We perform all generations at fixed sampling temperature. The temperature largely determines the *randomness* of the model generations, with higher values making the output more random and lower values more deterministic. However, even at a temperature of 0, model generations from OpenAI's ChatGPT API are not guaranteed to be deterministic. We conduct an initial ablation study scanning the temperature across the range of 0 to 2 that is allowed by the API. We conclude that, rather interestingly, the temperature has no significant influence on the classification accuracy on our task. However, it was observed that temperatures above 1.25 produce longer and often nonsensical outputs. For the consistency and clarity of our results, we fix the temperature at 0.75 for all subsequent model generations.

For reproducibility purposes, we use two version-locked instances of ChatGPT models available from OpenAI's developer API: `gpt-3.5-turbo-0613`, based on GPT-3.5, and `gpt-4-0613`, based on GPT-4. For brevity, the former model is referred to as GPT-3.5 and the latter as GPT-4. Fine-tuning refers to the process of refining a pre-trained model on specific datasets, allowing the model to specialise and improve performance on particular

Electronic copy available at: https://ssrn.com/abstract=4769112

tasks. OpenAI allows for the fine-tuning of GPT-3.5 series models via its user-friendly API, which is explored for use in this study. However, during the course of our study, the fine-tuning of GPT-4 models was not widely supported. This limitation influenced certain aspects of our methodology and the scope of analyses we could undertake.

Generative language models such as ChatGPT are directly conditioned on the user input which is provided in the form a prompt. Ensuring robust and reliable performance on specific tasks typically requires the tailoring of a specialized prompt, complete with detailed instructions and, sometimes, relevant examples. Hence, *prompt engineering* forms a vital component of our research methodology. Prompt engineering refers to the practice of carefully designing and modifying the input prompts to the model to elicit specific responses or behaviour. In the context of this project, this process required iterative experimentation while taking into account the biases observed in the initial runs, the underlying semantics of the quotes, and the empirical probabilities derived from repeated classifications. By balancing these factors, we are able to craft prompts that guide the model to a more accurate and nuanced understanding of the quotes, thereby reducing the observed biases and aligning the predictions more closely with the classification objective.

4. Results

4.1 Classification Bias

In Hansen and Kazinnik (2023), the authors highlight the data imbalance, but it also seems solely through the F1 scores obtained that they demonstrate that it is harder for GPT to classify "Hawkish" quotes. We invite the reader to examine their results which can be found in Table 1.

Throughout this publication, when talking about the "Base Prompt" we refer to the following:

```
You are a financial analyst, specialized in interpreting communications
from the Federal Reserve (Fed). Your task is to assess the Fed's stance
on the economy using their statements as a source. Based on the given
quote, categorize their sentiment as 'Hawkish' or 'Dovish'. Only print
out the sentiment in one word.
```

Note that the arrangement of the terms 'Hawkish' or 'Dovish' is deliberately structured to mitigate any potential positional bias within GPT, as empirical research indicates that LLMs tend to favor the initial option presented (Peiyi et al., 2023).

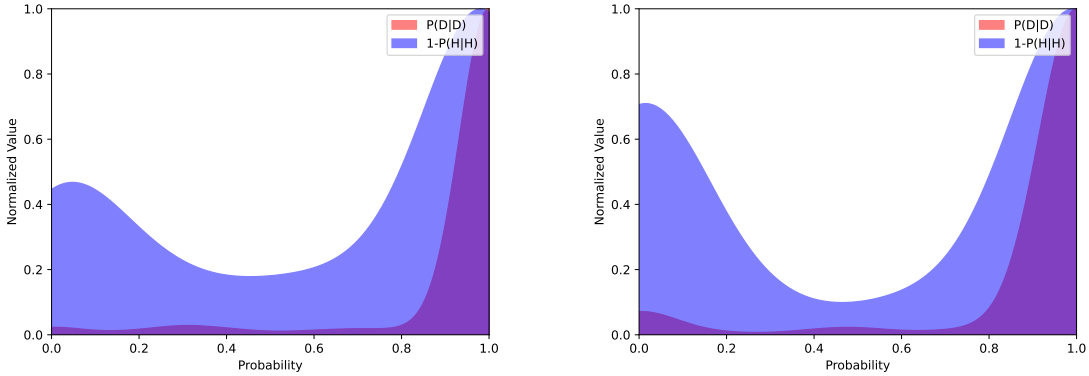
In Table 2, GPT-3.5 is observed to accurately identify most of the dovish quotes. However, a noticeable bias emerges in its classification of hawkish quotes. It not only misclassifies a significant portion of these quotes, but, alarmingly, it gets them wrong 65% of the time.

Figure 2a provides another visualisation of the bias. Ideally, for an unbiased model, we would expect two distinct concentrations at 0 and 1. However, in this case, both concentrations are skewed towards 1. This suggests that, for both dovish and hawkish quotes, GPT often assigns a 100% probability that the quote is hawkish. Yet, the figure also illustrates that GPT is seldom fully confident when classifying hawkish quotes.

Electronic copy available at: https://ssrn.com/abstract=4769112

Table 2: Combined confusion table for GPT-3.5 and GPT-4 in the binary classification task of identifying quotes as either dovish or hawkish using the Base Prompt. Bold values indicate the highest counts of correct classifications for the 'Dovish' and 'Hawkish' categories across both models. GPT-3.5 classifies more dovish quotes correctly, while GPT-4 classifies more hawkish quotes correctly. The results of GPT-4 are observed to be less biased.

	Classified as Dovish	Classified as Hawkish
GPT-3.5		
Dovish	**242**	19
Hawkish	201	109
GPT-4		
Dovish	238	23
Hawkish	182	**128**



(a) GPT-3.5, Base Prompt

(b) GPT-4, Base Prompt

Figure 2: Here we plot the condition probabilities of a quote being dovish and being classified as such, $P(D|D)$, in red and the complement of a quote being hawkish and being classified as such, $1 - P(H|H)$, in blue. Plots are shown separately for GPT-3.5 (a) and GPT-4 (b).

Table 2 and Figure 2b illustrate the improvements of GPT-4 over GPT-3.5, with the updated model accurately classifying an additional 19 hawkish quotes compared to its predecessor. Furthermore, a pronounced concentration of blue mass near 0 indicates that GPT-4 is often more confident in identifying a quote as truly hawkish. Given these findings, it is clear that GPT-4 stands out as a more capable model for quote classification than GPT-3.5. As such, with the exception of model fine-tuning, we exclusively utilise GPT-4 moving forward.

4.2 Bias Mitigation

4.2.1 Threshold shifting

To be able to assess the impact of threshold shifting, we measure the accuracy and F_1 score over several generations, as in Table 3. Specifically, the chosen model and prompt are

Electronic copy available at: https://ssrn.com/abstract=4769112

sampled at a fixed temperature 10 times for each quote. This results in each quote being independently classified 10 times, allowing for an estimate of the classification probability. The Base Prompt has an accuracy of 65% and an F_1 score of 0.70. Importantly, both the

Table 3: Accuracy and F_1 score for GPT-4 using the Base Prompt on 10 separate runs.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	**Average**
Accuracy	0.642	0.643	0.647	0.65	0.65	0.65	0.631	0.649	0.657	0.656	0.6475
F_1 Score	0.698	0.701	0.704	0.704	0.705	0.692	0.703	0.711	0.708	0.708	0.7034

accuracy and F_1 score consistently maintain these levels across different tests.

We determine the optimal threshold for dovish quotes is 100%. Here, the optimal threshold is defined to be the threshold that maximises Youden's J statistic (D et al., 2008), defined as $J = \text{TPR} - \text{FPR}$, which has values ranging from 0.0 to 1.0. For an ideal classifier, the TPR tends to its maximal value of 1.0, while the FPR tends to its minimum of 0.0; but in practice, there is a trade-off between the two. This implies that for a quote to be classified as dovish, it must be predicted as dovish on each of the 10 sampled responses. Any quotes which are predicted hawkish on one or more generations are labelled hawkish.

Table 4: Confusion table for GPT-4 using the Base Prompt with optimal threshold shifting.

	Classified as Dovish	Classified as Hawkish
Dovish	231	30
Hawkish	157	153

Table 4 presents the outcomes after adjusting the threshold. This adjustment leads to an additional 25 hawkish quotes being correctly classified. As a result, the bias is noticeably reduced. The updated accuracy is 67%, and the F_1 score has improved to 0.71. Our findings clearly demonstrate that altering the threshold is an effective strategy to mitigate bias. In the subsequent section, we will explore how prompt engineering can not only enhance the accuracy but also introduce variability, particularly in instances where the model might otherwise be consistently erroneous.

4.2.2 Prompt engineering

We use prompt engineering to further debias the classification. To compare different models, we measure the AUC of each binary classifier. Figure 3 displays the ROC curve and associated AUC value for the GPT-4 classifier using the Base Prompt.

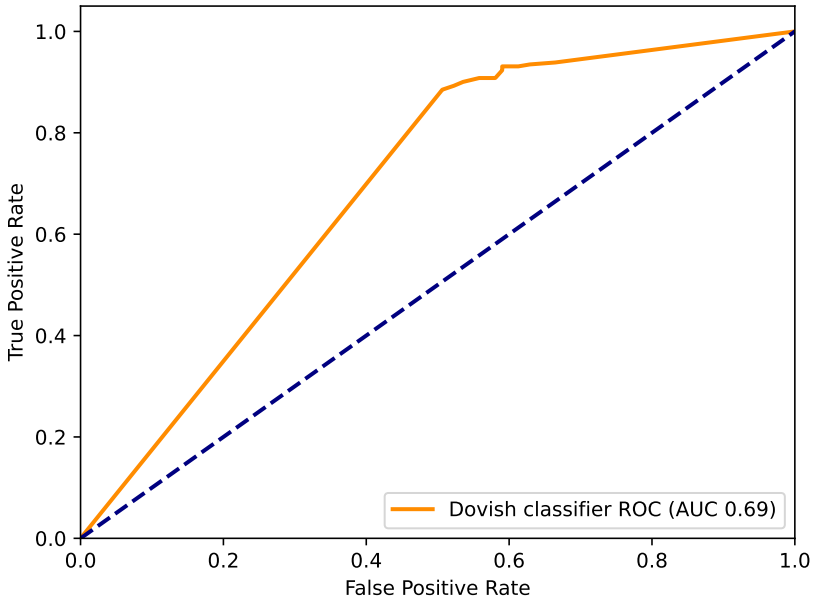
Following the simplistic Base Prompt, we use prompt engineering techniques inspired by Takeshi et al. (2023), arriving at a more detailed instruction, referred to as Prompt 1:

```
You are a financial analyst, specialized in interpreting communications
from the Federal Reserve (Fed). Your task is to assess the Fed's stance
on the economy using their statements as a source. Based on the given
quote, categorize their sentiment as 1 for 'Hawkish' or -1 for 'Dovish'.
Give a reasoning in no more than 50 words.

Speaker's CITATION:
```


Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 3: ROC curve and AUC value for GPT-4 using the Base Prompt. The horizontal reflects the performance of a baseline dummy classifier that is no better than chance.




```
{text}

OUTPUT FORMAT:

REASONING: <Your reasoning>
CLASSIFICATION: 1 for 'Hawkish' or -1 for 'Dovish' <Only one number>

Let's think step by step.
```

Figures 4 and 5 illustrate the significant impact of minor modifications to the prompt, such as requesting a numeric label and an explanation for the classification. The AUC is increased by 0.18, and the PDF distribution now exhibits nearly perfect separation between the hawkish and dovish masses, indicating that the bias has been substantially reduced.

The reader may rightly wonder how we arrived upon such a prompt. The answer is a mix of motivated experimentation and creativity. Indeed, it is well known that asking LLMs to explain their *thinking* helps in improving the quality of the output Takeshi et al. (2023). As for asking ChatGPT to output a number instead of a word, that was the fruit of multiple failed attempts. The results of these tests show that, individually, the changes do not yield the same level of improvement. A detailed description of the tests and their results can be found in Figures 5, 6, 7 and 8.

Instructing the model to return a numerical value, instead of a class label, is, to our knowledge, not a widely used prompt engineering technique. However, we believe that it is effective because it conditions ChatGPT to respond in a more classification-oriented manner, which encourages more objective behavior. In a similar way, we know that a

Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 4: ROC curve and AUC value for GPT-4 using Prompt 1.

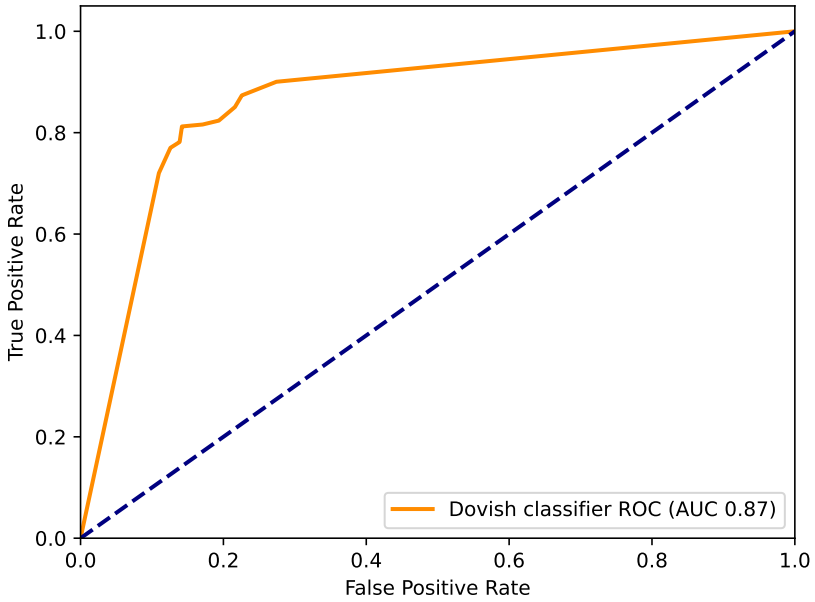
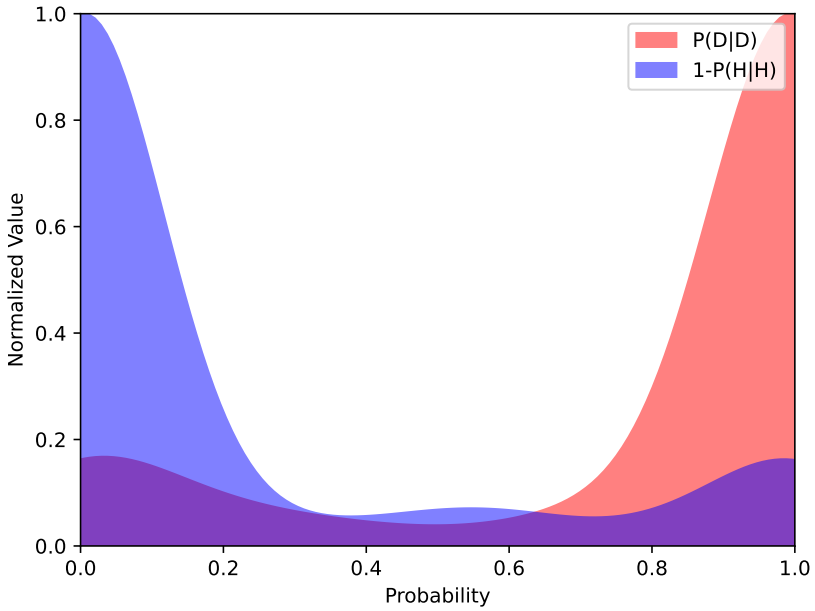


Figure 5: PDF for GPT-4 using Prompt 1.



Electronic copy available at: https://ssrn.com/abstract=4769112

relatively simple addition to the prompt like "let's think step by step" Takeshi et al. (2023), leads to improvements in accuracy. This can be understand as conditioning the model to respond in a more textbook or pedagogical style, which tends to elicit more accurate responses.

To further guide the model in the classification objective, we add additional context into the instructions. Specifically, we use the composite indicator provided by Morel to re-contextualise the model towards a more hawkish economic environment in an attempt to suppress the inherent dovish bias. The resulting new prompt, will be referred to as Prompt 2 [1]:

```
You are a financial analyst, specialized in interpreting communications
from the Federal Reserve (Fed). Your task is to assess the Fed's stance
on the economy using their statements as a source. Based on the given
quote, categorize their sentiment as 1 for 'Hawkish' or -1 for 'Dovish'.
Give a reasoning in no more than 50 words. The economic sentiment
at the time was [...] where values above 1 mean that the economy is
doing very well and below -1 mean there is a recession.

Speaker's CITATION:

{text}

OUTPUT FORMAT:

REASONING: <Your reasoning>
CLASSIFICATION: 1 for 'Hawkish' or -1 for 'Dovish' <Only one number>

Let's think step by step.
```

Shown in Figures 6 and 7, the results for Prompt 2 are very similar to those from Prompt 1. In this section we have shown that the bias can be significantly corrected simply through prompt engineering.

4.2.3 Combined approach

We now apply the threshold shifting technique outlined in Section 4.2.1 with the improved prompt (Prompt 2) and analyze the change in classification performance. As a reminder, applying the same technique with the Base Prompt yields an accuracy of 65% and an F_1 score of 0.70. Interestingly, threshold shifting has no improvement on accuracy nor F_1 score with Prompt 1. This technique relies on an underlying variability in the model generations, and we notice that various prompts lead to different levels of variability. In Table 5, we can see that the combination of Prompt 2 and threshold shifting results in an accuracy of 82% and an F_1 score of 0.80. Note that the analogous results for the Base Prompt in Table 3 are consistent. The new threshold for dovish quotes is now determined to be 40%.

Using this threshold we obtain the results summarized in Table 6. The changes yield a massive reduction in classification, resulting in a new accuracy of 83.5% and an F_1 score of 0.81. This presents a relative improvement of 28% and 16% in accuracy and F_1 score, respectively, over the Base Prompt.

1. An example of the additional data given: Williams - 14/02/2023 - "I expect real GDP growth to come in around 1 percent for 2023. And I anticipate the unemployment rate to edge up over the next year to between 4 and 4.5 percent." had an associated indicator value of -0.926.

Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 6: ROC curve and AUC value for GPT-4 using Prompt 2.

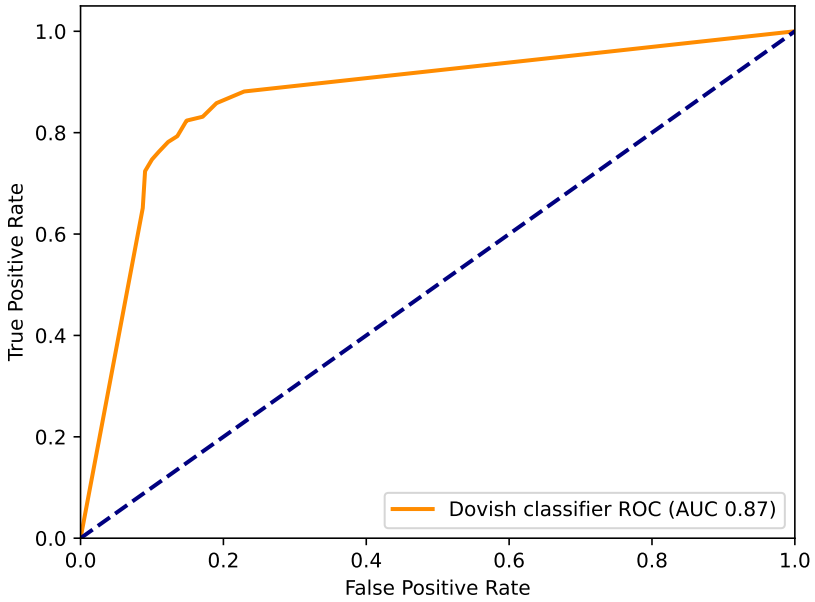
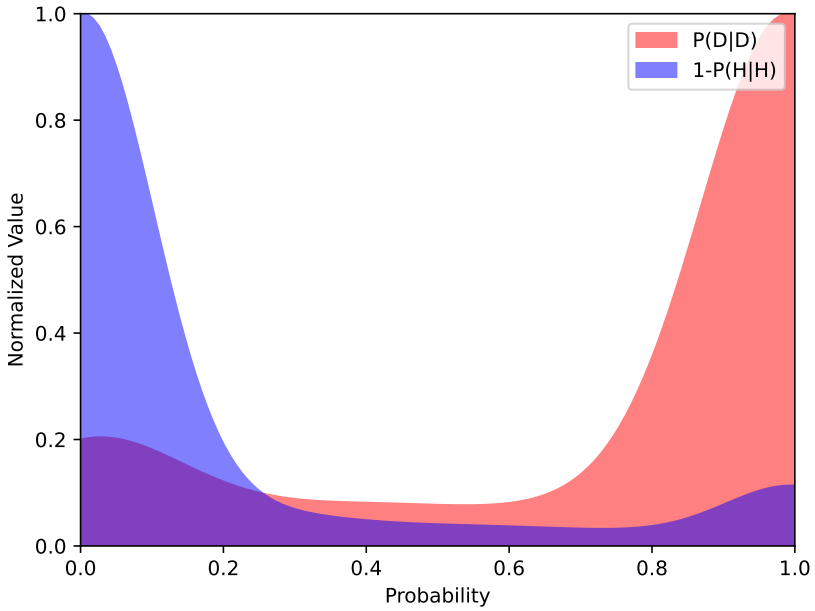


Figure 7: PDF for GPT-4 using Prompt 2.



Electronic copy available at: https://ssrn.com/abstract=4769112

Table 5: Accuracy and F_1 score for GPT-4 using Prompt 2 on 10 separate runs.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	**Average**
Accuracy	0.824	0.828	0.815	0.817	0.819	0.824	0.819	0.819	0.824	0.814	0.82
F_1 Score	0.811	0.813	0.801	0.801	0.803	0.809	0.802	0.801	0.809	0.799	0.805

Table 6: Confusion table for GPT-4 using Prompt 2 with optimal threshold shifting.

	Classified as Dovish	Classified as Hawkish
Dovish	204	57
Hawkish	38	272

4.2.4 Supervised fine-tuning

In this section we explore the use of supervised fine-tuning of GPT, as described in Section 2.3, to further improve the classification performance. OpenAI provides a fine-tuning API[2] which makes the instruction training of GPT-3.5 models straightforward, requiring only a small dataset of conversation examples encoding the desired model behaviour. Following OpenAI's suggestion for training set size, we select 100 random quotes with the same proportion as the dataset to train the model on.

We fine-tune GPT-3.5 with the Base Prompt to be able to clearly disentangle its impact from more advanced prompt engineering. Moreover, because supervised fine-tuning requires curating a dataset of model responses, we choose a format which is conducive to automatic generation. Fine-tuning using Prompt 1 or 2 would require the more labor intensive task of writing explanations for each classification example. An example from the training dataset is shown in Figure 8. Note that the quote and classification label are presented in the user-assistant conversation dialogue, as expected by the ChatGPT fine-tuning API. We fine-tune GPT-3.5 for a total of 3 epochs, totalling 39,975 tokens. The results from fine-tuning GPT-3.5 are shown in Figures 9 and 10.

Although we see from Figure 9 that the AUC is lower by 0.01 compared to Prompt 1 (4) and 2 (6), the PDFs in Figure 10 are better separated than in Figures 5 and 7. Importantly, there is no overlap in the middle and just a few quotes that GPT classifies incorrectly every time.

Table 7: Accuracy and F_1 score for fine-tuned GPT-3.5 using the Base Prompt on 10 separate runs.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	**Average**
Accuracy	0.857	0.859	0.859	0.857	0.857	0.859	0.857	0.853	0.857	0.855	0.857
F_1 Score	0.841	0.844	0.844	0.842	0.842	0.844	0.842	0.837	0.841	0.839	0.842

We further apply threshold shifting (see Section 4.2.1) to the fine-tuned model. Shown in Table 7, the accuracy without threshold shifting is 85.7% and the F_1 score is 0.84. As there are little to no quotes for which GPT is uncertain about (see Figure 10), threshold shifting does not improve the results. Furthermore, it is clear from Table 8 that fine-tuning has essentially eliminates the dovish bias.

2. https://platform.openai.com/docs/api-reference/fine-tuning

Electronic copy available at: https://ssrn.com/abstract=4769112

```
{
  "messages": [
    {
      "role": "user",
      "content": "\nYou are a financial analyst,
         specialized in interpreting communications from
         the Federal Reserve (Fed).Your task is to assess
         the Fed's stance on the economy using their
         statements as a source.Based on the given quote,
         categorize their sentiment as 'Hawkish' or '
         Dovish'.Only print out the sentiment in one word
         .\n```We are seeing some indications of a slowing
          in the pace of recovery in states that are
         currently experiencing large-scale outbreaks.
         This is a valuable reminder that the economy's
         fate is inextricably linked to the path of the
         virus.```\n"
    },
    {
      "role": "assistant",
      "content": "Dovish"
    }
  ]
}
```

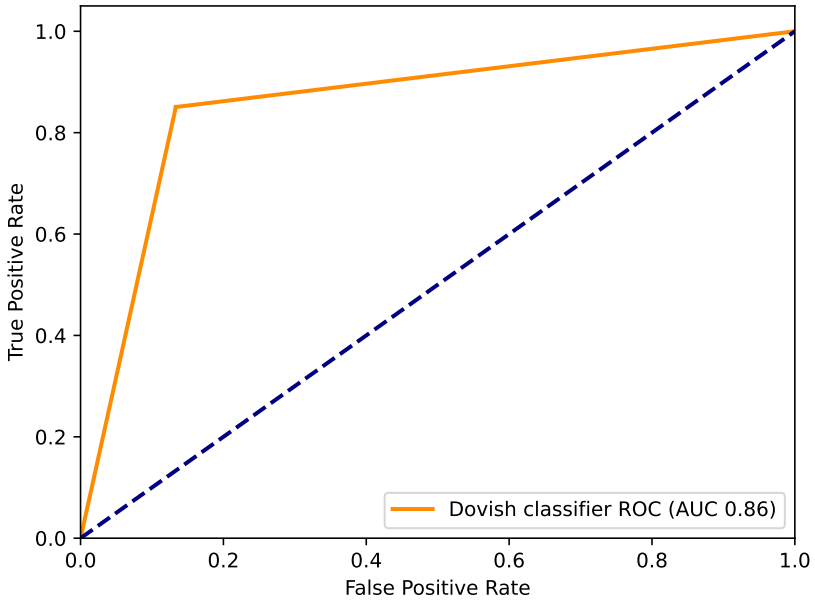
Figure 8: Example 'Dovish' quote from the fine-tuning dataset. The task instruction is provided with the quote as the user message, while the classification label is provided as the assistant response. No system message is supplied.

Table 8: Confusion table for fine-tuned GPT-3.5 using the Base Prompt with optimal threshold shifting.

	Classified as Dovish	Classified as Hawkish
Dovish	181	34
Hawkish	34	222

Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 9: ROC curve and AUC value for fine-tuned GPT-3.5 using the Base Prompt.



4.3 Summary

Table 9 presents an overview of the classification results from the various bias mitigation techniques outlined in Section 4.2.

Table 9: Overview of the classification results from the different bias mitigation techniques described in Section 4.2. The model GPT-3.5* denotes the custom, fine-tuned version of GPT-3.5 describe in Section 4.2.4. The initial accuracy F_1 score are obtained by averaging over and Bold values denote the maximal value of the respective metric obtained across the different strategies.

Model	Prompt	AUC	Accuracy	F_1 Score	Shifted Accuracy	Shifted F_1 Score
GPT-3.5	Base Prompt	0.72	0.62	0.69	-	-
GPT-4	Base Prompt	0.69	0.65	0.7	0.67	0.71
GPT-4	Prompt 1	**0.87**	0.82	0.8	0.82	0.8
GPT-4	Prompt 2	**0.87**	0.82	0.8	0.84	0.81
GPT-3.5*	Base prompt	0.86	**0.86**	**0.84**	**0.86**	**0.84**

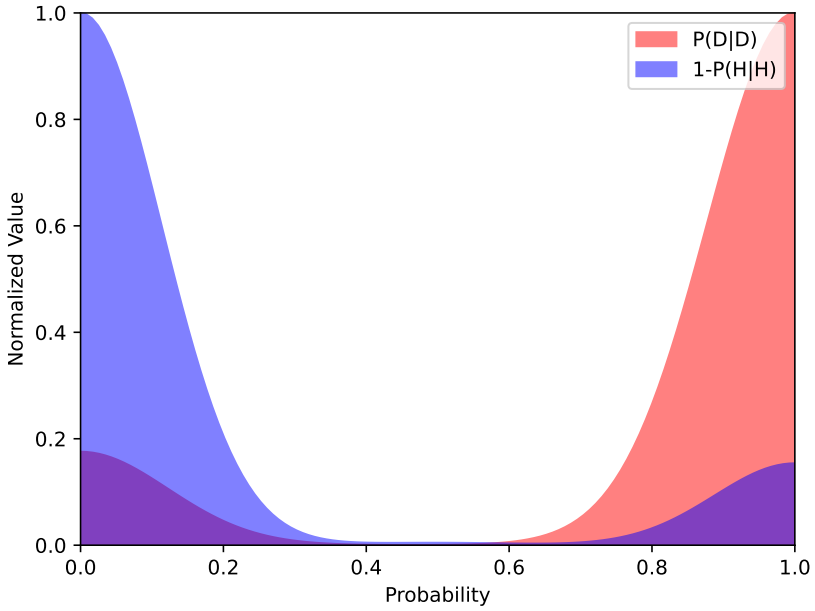
5. Interpretation

5.1 Potential Origin of Bias

As machine learning models are reflections on the data on which they are trained, the dovish bias uncovered in ChatGPT almost certainly originates from its training dataset.

Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 10: PDF for fine-tuned GPT-3.5 using the Base Prompt.

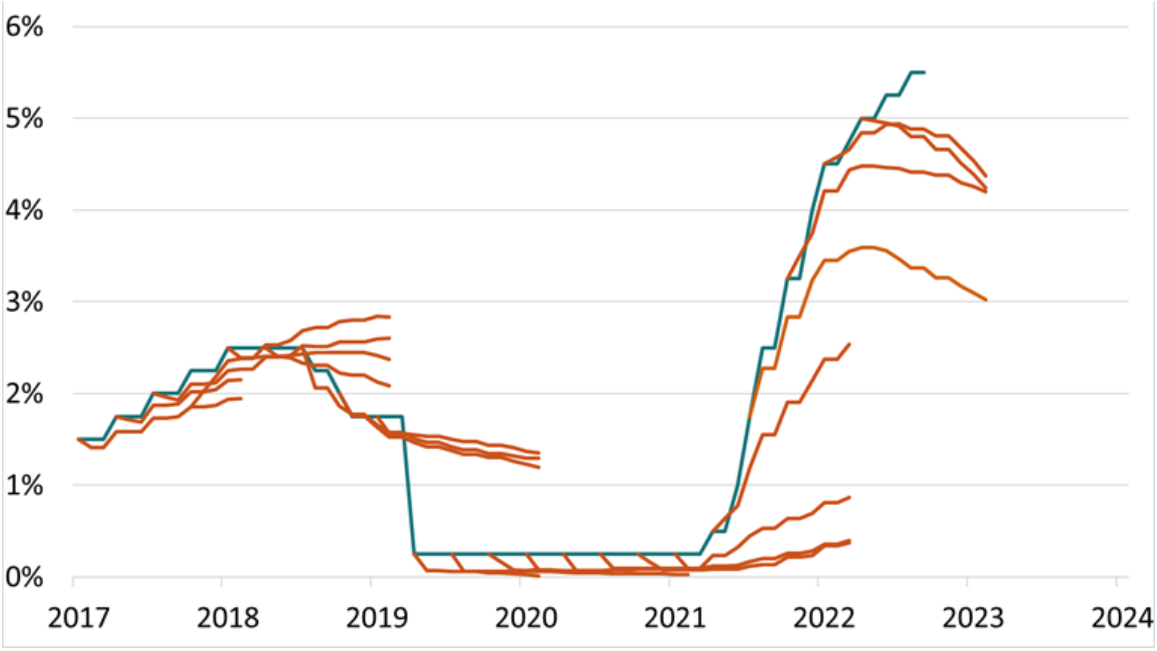


Though the specifics of this dataset are not publicly known, following the trend of GPT-3 (B. et al., 2020) and scaling trends forward, it likely consists of a snapshot of the entire open Internet up to September 2021[3]. Given recent economic trends and the exponential growth of data over the same years, this dataset is bound to consist of overly dovish sentiments of the recent economic era. For example, Figure 11 shows that expectations about the future path of Fed Funds have generally fallen short of realised Fed Funds over the 2018 to 2023 period. Such an imbalance in opinion in its training set, may lead ChatGPT to preferentially classify Fed quotes as dovish. Arguably, this is a reflection of recenecy bias towards the current economic view. This is discussed in Filippo (2024), a paper that aims to explore the market's bias in expectations of monetary policies: "Investor's perception of the central bank reaction function changes over the business cycle. Typically, monetary policy expectations are asymmetric with a dovish bias. During a tightening monetary cycle, investors tend to expect a less aggressive monetary policy than what eventually occurred while at the start of an easing cycle, expectations for the policy rate adjust downwards quite fast".

3. The cutoff date is specific to the model generation and version. In this study, we exclusively use GPT-3.5 (`gpt-3.5-turbo-0613`) and GPT-4 (`gpt-4-0613`), both of which share a cutoff date of September 2021. For more information, please refer to OpenAI's API documentation, https://platform.openai.com/docs/models.

Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 11: Fed Funds and forward expectations, based on Fed Funds future.
Data Source: Bloomberg



5.2 Limitations

This study focuses on the detection and correction of bias in ChatGPT on a particular task of Fedspeak classification and has several limitations that must be considered. Firstly, the methodologies of threshold shifting (Section 4.2.1) and prompt engineering (Section 4.2.2), while effective for bias mitigation, were applied without a distinct train-test split in our dataset. This approach, diverging from traditional machine learning practices, raises concerns about potential over-fitting, especially in the context of prompt engineering where iterative refinement could lead to test set leakage. Our findings here might thus be over-specialized to our dataset, impacting their general lisibility. In contrast, for the fine-tuning aspect of our study, outlined in Section 4.2.4, we implemented a strict train-test split, aligning with standard machine learning protocols. This provides a more robust evaluation for this part of our research. However, the absence of external validation for the prompt engineering and threshold shifting techniques limits our ability to assert the broader applicability of our findings to other financial datasets.

While our study contributes to understanding inherent biases in LLMs used in financial analysis, particularly ChatGPT, the scope is primarily within the confines of our dataset. Our insights into bias reduction through threshold adjustments and prompt engineering should be interpreted within this context. Future research would benefit from employing these techniques across diverse datasets, enhancing the understanding of biases in LLMs and their implications in the financial sector and beyond.

Electronic copy available at: https://ssrn.com/abstract=4769112

6. Conclusion

We demonstrate an inherent dovish bias in the classification of quotes from the Federal Reserve for OpenAI's ChatGPT series of models, namely GPT-3.5 and GPT-4. and of various bias mitigation techniques and their impact. Though this bias is present under certain prompts, it is a well established fact that small changes in prompts can have significant changes in the output, with certain phrases (eg "let's think step by step") eliciting more accurate and insightful responses. This implies that the only source of bias isn't just the model in itself but also the prompt used. Carefully selecting the right prompt for a given task is known as "prompt engineering". It is worth mentioning that while the optimal results are derived from the inclusion of supplemental information, such as fine-tuning on a specific dataset and integrating a composite indicator, substantial improvements can be achieved solely through prompt engineering, as demonstrated by the efficacy of Prompt 1. For the specific dataset of quotes curated by Morel, we achieve an accuracy rate of 85.7% and an F_1 score of 0.84, faithfully reproducing the interpretation of Fedspeak by the Chief Economist. More importantly, we successfully eliminate the bias of classifying hawkish quotes as dovish. We are optimistic that the methodologies and tools we have introduced will serve as invaluable resources for those attempting to decode the strategic manoeuvres of the Federal Reserve and analogous institutions, particularly in the intricate economic climate we currently navigate.

Acknowledgments

This article has benefited from the contributions of E. Paty from Groupama Asset Management Financial Engineering department, who provided useful discussion on theoretical frameworks. The views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of their institutions, employers, or funders.

Electronic copy available at: https://ssrn.com/abstract=4769112

Appendix A. Appendix

A.1 Bias in Hansen and Kazinnik (2023)

The F1 score of "Hawkish" quotes is markedly lower than for the "Dovish" quotes in Hansen and Kazinnik (2023), which suggests it is harder for GPT to classify them correctly. This is what we refer to as a bias, as detailed in Table 1.

Table 1: Summary of GPT-3 classification results from Hansen and Kazinnik (2023).

GPT-3	F_1 score
Dovish	0.49
Mostly Dovish	0.43
Neutral	0.15
Mostly Hawkish	0.36
Hawkish	0.1

These results were obtained with the following prompt, taken from Hansen and Kazinnik (2023):

```
"Imagine you are a research assistant working for the Fed. You have
a degree in Economics. Your task is to classify the following [x number
of] sentences into one of the five categories (dovish, mostly dovish,
neutral, mostly hawkish, hawkish). These sentences are taken at random
from the texts of FOMC announcements: [...]"
```

A.2 Dependence on Sampling Temperature

To improve our results, we tested the prompts with different temperatures in the hopes of getting more variability in the output. Figures 1, 2, 3 and 4 demonstrate that changing the temperature of model sampling has (surprisingly) no effect on AUC and variability.

1. These results are obtained using a sample of 50 random quotes with 25 quotes from each category.

2. As there are only two categories in the binary classification task, there is no need to analyze the hawkish results.

3. We were unable to extract results for temperatures above 1.5 as they start to become complete gibberish.

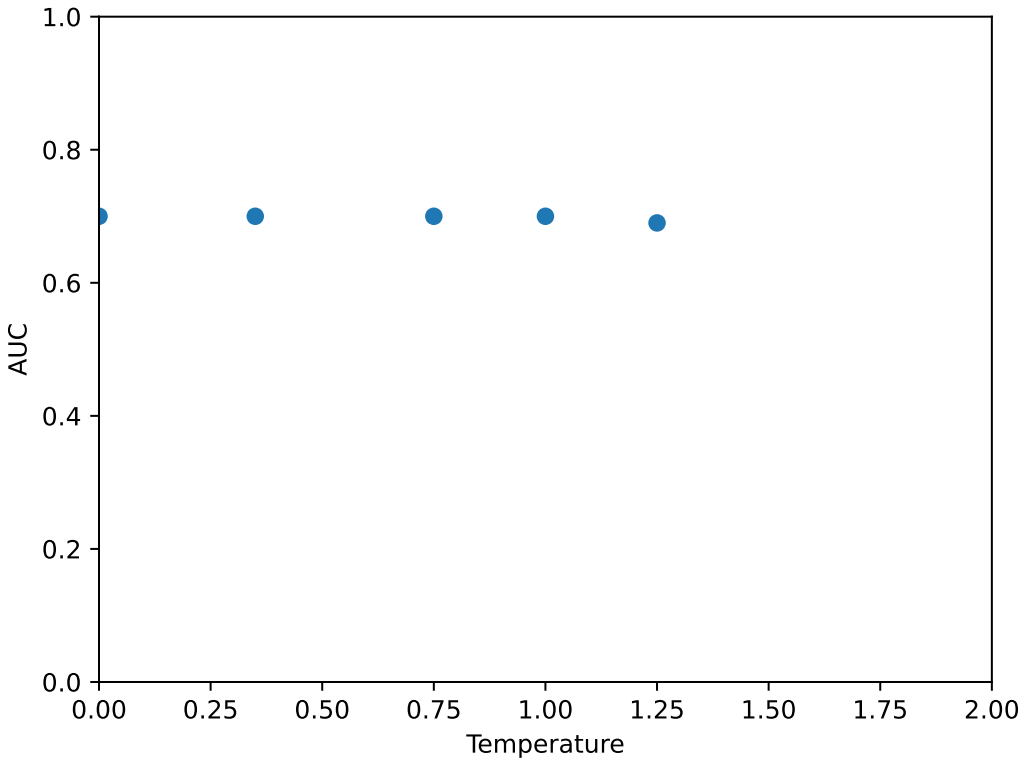
A.3 Quantifying the Impact of different Prompting Techniques

The results in Figures 5 and 6 are from running the following prompt which tests the effect of asking for numeric conversion only:

```
You are a financial analyst, specialized in interpreting communications
from the Federal Reserve (Fed). Your task is to assess the Fed's stance
```


Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 1: AUC values for GPT-4 with different temperatures using Base Prompt.




```
on the economy using their statements as a source.  Based on the given
quote, categorize their sentiment as 1 for 'Hawkish' or -1 for 'Dovish'.

Speaker's CITATION:

{text}

OUTPUT FORMAT:

CLASSIFICATION: 1 for 'Hawkish' or -1 for 'Dovish' <Only one number>
```

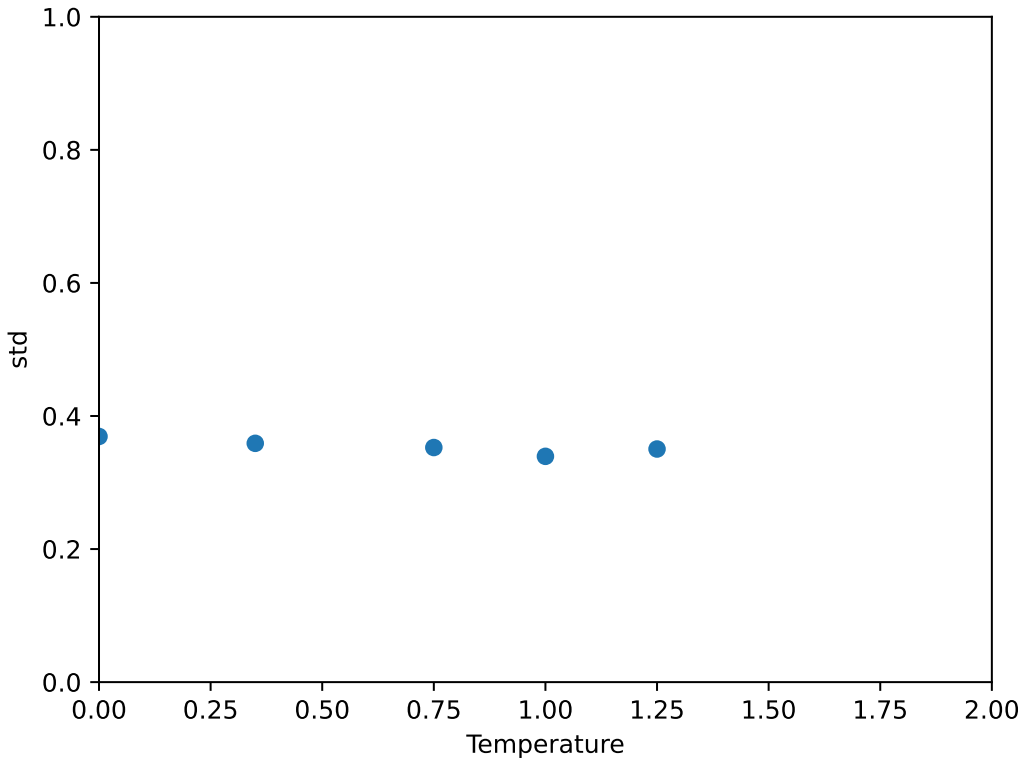
The results in Figures 7 and 8 are from running the following prompt which tests the effect of asking for an explanation only:

```
You are a financial analyst, specialized in interpreting communications
from the Federal Reserve (Fed).  Your task is to assess the Fed's stance
on the economy using their statements as a source.  Based on the given
quote, categorize their sentiment as 'Hawkish' or 'Dovish'.  Give a
reasoning in no more than 50 words.

Speaker's CITATION:
```


Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 2: Standard deviation values for GPT-4 with different temperatures using Base Prompt.




```
{text}
OUTPUT FORMAT:
REASONING: <Your reasoning>
CLASSIFICATION: 'Hawkish' or 'Dovish' <Only one word>
Let's think step by step.
```

Results from Figures 5, 6, 7 and 8 indicate that asking only for an explanation compared to only numeric conversion, is more effective, corresponding to an AUC difference of 0.04. These figures also show that explanation is the main factor in removing the problem of having hawkish quotes with a very high probability of being classified as dovish. In Table 2 we have tested the statistical significance of this result by doing a DeLong's test. The alternative hypothesis is that the true difference in AUC is not equal to 0. Since the p-value of the test is 0.00144, we deduce that the difference is significant.

Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 3: AUC values for GPT-4 with different temperatures using Prompt 1.

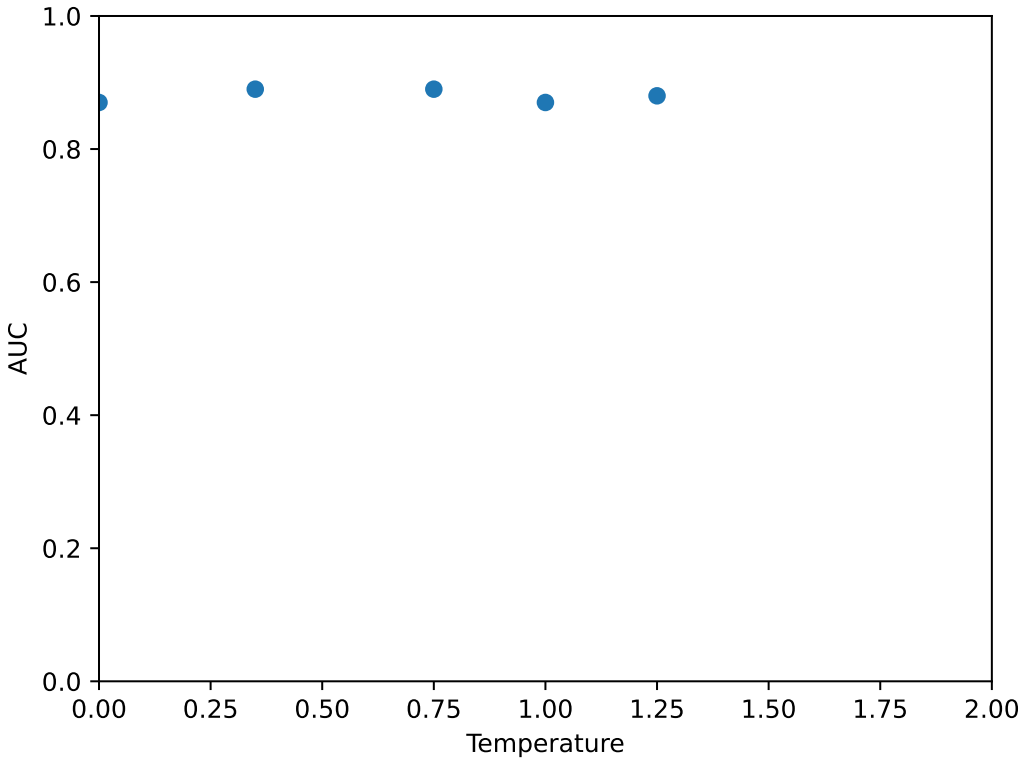
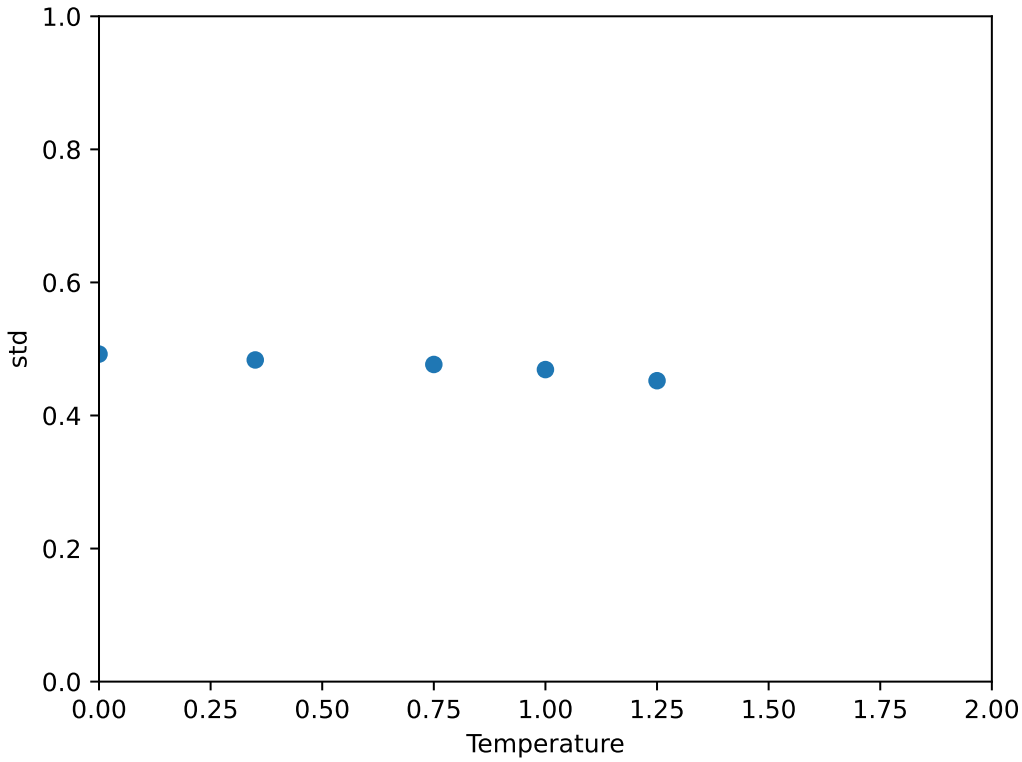


Table 2: We use DeLong's test for two correlated ROC curves to determine the statistical significance of the AUC improvement observed in the aforementioned prompting techniques.

Measure	**Value**
Z value	3.1848
p-value	0.001449
95% Confidence Interval	
Lower bound	0.01755362
Upper bound	0.07373299

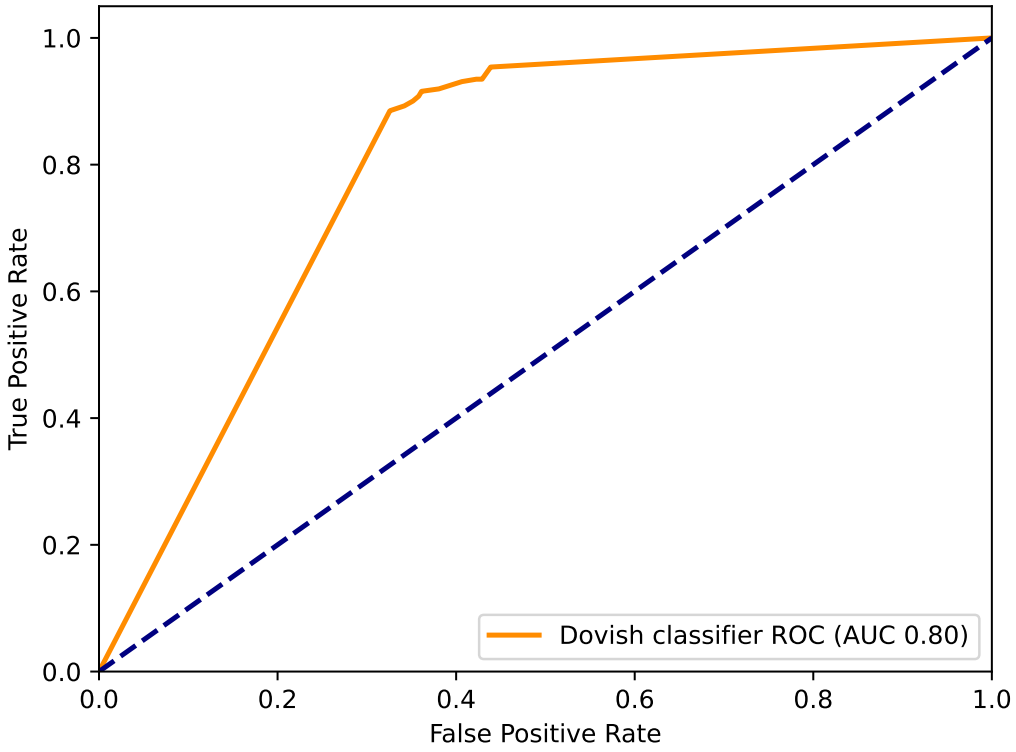
Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 4: Standard deviation values for GPT-4 with different temperatures using Prompt 1.



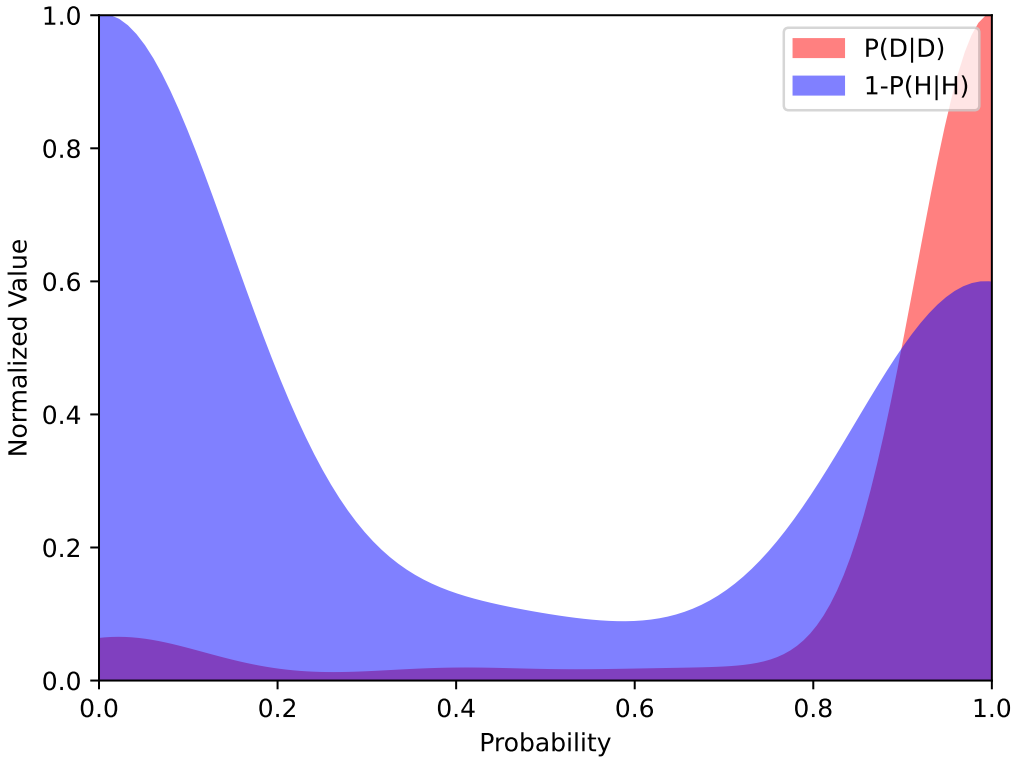
Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 5: ROC curve and AUC value for GPT-4 using numeric conversion only.



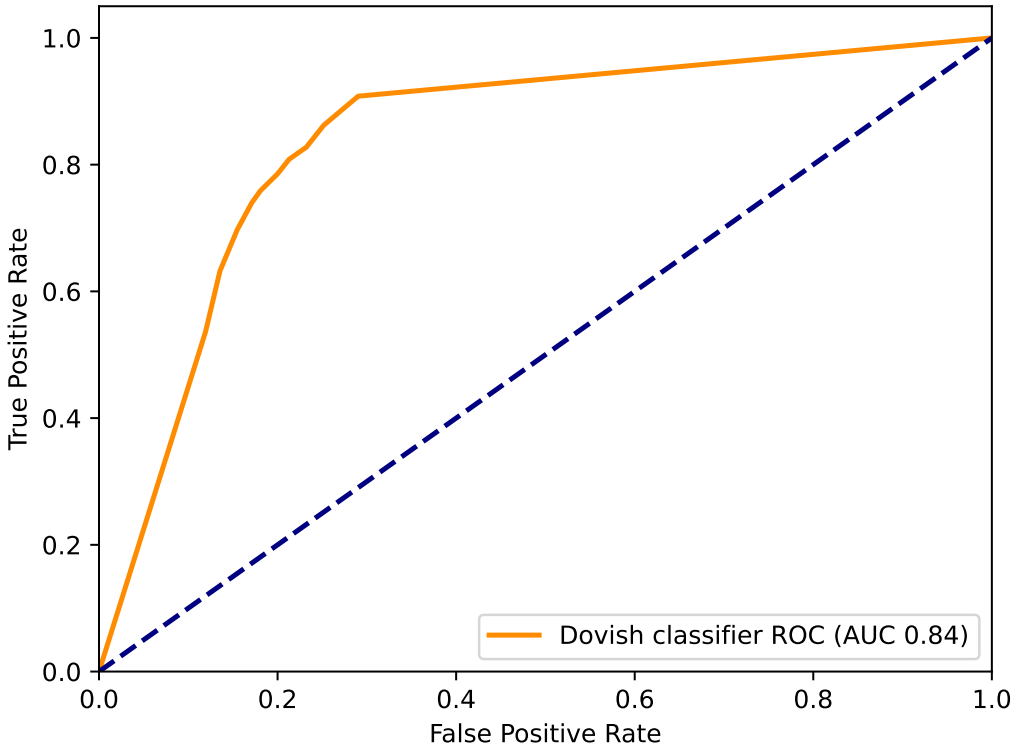
Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 6: PDF for GPT-4 using numeric conversion only.



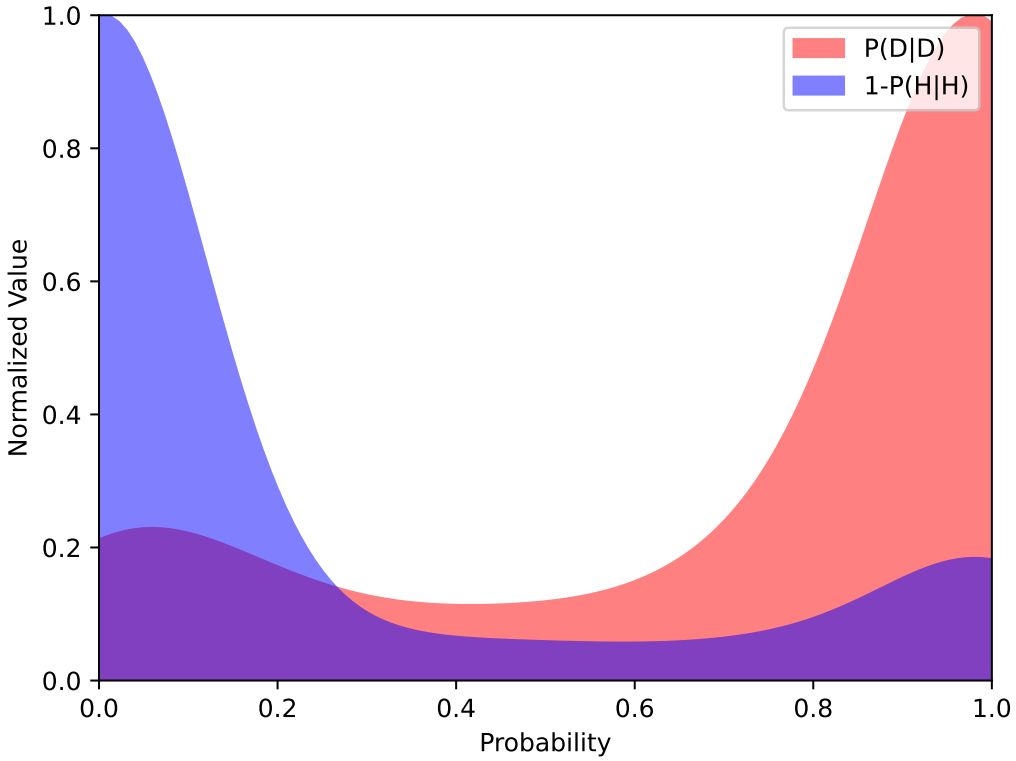
Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 7: ROC curve and AUC value for GPT-4 using explanation only.



Electronic copy available at: https://ssrn.com/abstract=4769112

Figure 8: PDF for GPT-4 using explanation only.



Electronic copy available at: https://ssrn.com/abstract=4769112

References

Radford Alec and Narasimhan Karthik. Improving Language Understanding by Generative Pre-Training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, and Polosukhin Illia. Attention Is All You Need, 2023.

Brown Tom B., Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Ziegler Daniel M., Wu Jeffrey, Winter Clemens, Hesse Christopher, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, McCandlish Sam, Radford Alec, Sutskever Ilya, and Amodei Dario. Language Models are Few-Shot Learners. 2020.

Ruopp Mark D, , Perkins Neil J, Whitcomb Brian W, and Schisterman Enrique F. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*, 50(3):419–430, 2008. doi: 10.1002/bimj.200710415.

Henry Elaine. Are Investors Influenced by How Earnings Press Releases are Written? *SSRN Electronic Journal*, November 2006. doi: 10.2139/ssrn.933100. URL https://ssrn.com/abstract=933100. Available at SSRN.

Busetto Filippo. Asymmetric expectations of monetary policy. *Bank of England*, 2 2024. doi: 10.2139/ssrn.4399406. URL https://www.bankofengland.co.uk/working-paper/2024/asymmetric-expectations-of-monetary-policy.

Anne Lundgaard Hansen and Sophia Kazinnik. Can ChatGPT Decipher Fedspeak? *SSRN Electronic Journal*, 10 2023. doi: 10.2139/ssrn.4399406. URL https://ssrn.com/abstract=4399406.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

Tunstall Lewis, Reimers Nils, Jo Unso Eun Seo, Bates Luke, Korat Daniel, Wasserblat Moshe, and Pereg Oren. Efficient Few-Shot Learning Without Prompts. 2022.

Ouyang Long, Wu Jeff, Jiang Xu, Almeida Diogo, Wainwright Carroll L., Mishkin Pamela, Zhang Chong, Agarwal Sandhini, Slama Katarina, Ray Alex, Schulman John, Hilton Jacob, Kelton Fraser, Miller Luke, Simens Maddie, Askell Amanda, Welinder Peter, Christiano Paul, Leike Jan, and Lowe Ryan. Training language models to follow instructions with human feedback, 2022.

Mohammad Saif M. and Turney Peter D. NRC Emotion Lexicon. Technical Report 21270984, National Research Council of Canada, 2013. URL https://doi.org/10.4224/21270984.

Reimers Nils and Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 2019.

Electronic copy available at: https://ssrn.com/abstract=4769112

Mehrabi Ninareh, Morstatter Fred, Saxena Nripsuta, Lerman Kristina, and Galstyan Aram. A Survey on Bias and Fairness in Machine Learning. 2022.

Gallegos Isabel O., Rossi Ryan A., Barrow Joe, Tanjim Md Mehrab, Kim Sungchul, Dernoncourt Franck, Yu Tong, Zhang Ruiyi, and Ahmed Nesreen K. Bias and Fairness in Large Language Models: A Survey, 2023.

OpenAI. GPT-4 Technical Report, 2023.

Christiano Paul, Leike Jan, Brown Tom B., Martic Miljan, Legg Shane, and Amodei Dario. Deep reinforcement learning from human preferences, 2023.

Wang Peiyi, Li Lei, Chen Liang, Cai Zefan, Zhu Dawei, Lin Binghuai, Cao Yunbo, Liu Qi, Liu Tianyu, and Sui Zhifang. Large Language Models are not Fair Evaluators, 2023.

Kojima Takeshi, Gu Shixiang Shane, Reid Machel, Matsuo Yutaka, and Iwasawa Yusuke. Large Language Models are Zero-Shot Reasoners. 2023.

Loughran Tim and McDonald Bill. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, March 2010. URL https://ssrn.com/abstract=1331573. Available at SSRN: https://ssrn.com/abstract=1331573.

Electronic copy available at: https://ssrn.com/abstract=4769112