

**ERASMUS UNIVERSITY ROTTERDAM**  
**ERASMUS SCHOOL OF ECONOMICS**  
**Bachelor Thesis Economics & Business**  
**Specialization: Financial Economics**

## **Company Earnings Calls: a Leading Indicator of Future Inflation in the Economy**

**Author:** Ignas Peldžius  
**Student number:** 616336  
**Thesis supervisor:** Clemens Mueller  
**Second reader:**  
**Finish date:** 14 06 2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second reader, Erasmus School of Economics or Erasmus University Rotterdam.

## **ABSTRACT**

With rapid technological advancements in artificial intelligence (AI), it can now be widely accessible and relatively easy to perform complex analysis of company earnings calls to glean information into the corporate management's perception on the upcoming inflation trends in the economy – a possibility that would take regular investors immense time and resources only several years ago. Using such technology, this research finds that inflation likelihood scores, generated based on the context of company earnings calls via large language model (LLM) analysis, show potential to forecast inflation in the United States (U.S.) three to nine months out. Additionally, paired with previous literature, there is added potential for these scores to be associated with leading the U.S. inflation by a four quarter period. Using LLMs to query earnings call documents can be a valuable addition to investors' and policy makers' toolbelt to aid in better investment and monetary policy decisions that are of higher precision provided the inclusion of leading information extracted from the forward looking information contained in company earnings calls.

**Keywords:** [2 to 5 keywords] earnings calls, inflation, natural language processing, large language models, Llama 2

# TABLE OF CONTENTS

- ABSTRACT ..... ii
- TABLE OF CONTENTS ..... iii
- CHAPTER 1 Introduction ..... 1
- CHAPTER 2 Literature Review ..... 4
  - 2.1 Relevant Literature..... 4
  - 2.2 Research Expectations ..... 7
- CHAPTER 3 Data ..... 7
- CHAPTER 4 Method ..... 11
  - 4.1 Earnings Calls ..... 11
  - 4.2 Running the LLM..... 11
  - 4.3 Model Calibration ..... 13
  - 4.4 Model Prompting and Output ..... 13
  - 4.5 Statistical Analysis..... 15
- CHAPTER 5 Results & Discussion ..... 17
  - 1.1 Results..... 17
  - 1.2 Discussion..... 25
- CHAPTER 6 Conclusion..... 27
- REFERENCES..... 29
- APPENDIX A LLM Fine-Tuning Options ..... 31
- APPENDIX B LLM Output Examples ..... 32

## **CHAPTER 1 Introduction**

With the ever developing and constantly changing landscape of the financial world, developments in the artificial intelligence (AI) field arguably have the potential to propel financial evolution to even greater lengths. The interdisciplinary connection between finance and technology is continuously expanding, and with this comes new opportunities, challenges, and a vast space to explore the intricate relations between technology and finance and dive deeper into potential applications of rapidly developing technologies in finance. Following the latest developments surrounding the aforementioned relationship, this research focuses on the language used in public company earnings calls as a potential leading indicator of future trends in the economy. Specifically, I aim to examine the forecasting power of company earnings calls for future inflation trends – one of the key economic variables – in the United States (U.S.) economy, using the natural language processing (NLP) power of large language models (LLM). Put simply, public company earnings calls may help us get a glimpse into the potential future economic conditions, and, if the forecasting power of company earnings calls is significant and reliable, it may even help policy makers and investors make better-informed policy and investment decisions respectively, seeing that leading rather than lagging economic indicators would be relied upon for decision making.

Among the relevant literature for this research is a recent International Monetary Fund (IMF) working paper by Albrizio et al. (2023), who construct an index for international firms' inflation expectations from earnings call transcripts, using NLP algorithms. Authors find that higher inflation expectations on the firm level correspond to actual future inflation in the economy. Authors mainly focus on the inflation expectations, disagreement over them, and implications for central bank monetary policy transmission. Furthermore, this study involves multiple economies, including advanced and emerging, but the analysis stays mostly with the U.S. – an economy that is the focus of this paper. Moreover, inspiration for this paper has been firstly drawn from Price et al. (2012) paper on the information contained within the textual tone of earnings calls and its predictiveness of company stock performance. Although it contributes to a more commonly seen approach in the literature of studying earnings calls to predict company stock abnormal returns and trading volume, the paper offers some useful insights to understand the approaches to studying senior company staff sentiment communicated in company earnings calls and its effects on financial variables. Further, Chava et al. (2022) use a deep learning technique to construct a text-based tool from company earnings calls to measure their inflation exposure, and find that firms with elevated inflation exposures have higher future costs of goods sold and lower operating cash flows. This study provides further insight into using algorithms to study textual data, and the power of earnings calls to forecast future economic variables. However, Chava et al. (2022) focus on the firm level effects, therefore once again motivating this research to examine the bigger picture and study the informativeness of earnings calls for future price trends in the U.S. economy. Lastly, Lopez-Lira and

Tang (2024) research provides valuable insights into LLM prompting and using a score-based output for further statistical analysis in a simple and effective manner. Moreover, this paper provides a glimpse into how different sophistication level models compare in their ability to analyze and sustain predictive accuracy for more nuanced and potentially biased firm press releases that may be strategically used to influence investors' perceptions, compared to regular news articles. To note, in this research authors focus on the common analysis in the financial literature of forecasting stock price movements, although with a novel approach using large language models. This once more provides fuel for the research in this paper to use this modern approach, but rather focus on forecasting aggregate price levels in the economy, as opposed to return predictability of equities.

A part of the motivation for this research comes from specific observations made during the recent post-pandemic financial market boom of 2021 – a year that preceded the highly inflationary period of 2022 world-wide. More specifically, during the year of 2021, studying company earnings calls proved to be rather insightful for foreseeing what came to be a peak inflation year of 2022. Burdened by the post-pandemic supply chain disruptions world-wide, and the following cost increases of freight and manufacturing materials, companies in the U.S. were happy to announce to investors about their ability to pass on cost increases onto the consumer – who had increased their savings during the pandemic, helped by the monetary stimulus from the U.S. government – in their second-half of 2021 earnings calls. The impressive quarter-over-quarter, and year-over-year product and service price increases had not only propelled investor excitement about the seemingly strong “pricing-power” of companies, resulting in high equity valuations in the stock market, but had also led to a highly inflationary economic environment in the U.S. A specific example of this would be the chief financial officer of Kroger Ko., who, during the 2021 calendar-year third quarter earnings call, stated that: "We've been very comfortable with our ability to pass on the increases that we've seen at this point, and we would expect that to continue to be the case", as reported by Terlep (2021) in the Wall Street Journal.

Another motivating factor for this research comes from a seemingly apparent gap in the literature on earnings conference calls, which looks to be mostly focused on company earnings call impact on firm-level variables, such as stock prices, and not extensively on their informativeness for the aggregate trends in the economy. While the recent literature is certainly expanding the research on the possible information contained in the earnings conference calls that can be useful in showing trends in the overall economy, for which the IMF working paper by Albrizio et al. (2023) is a prime example, it is still rather limited and invites additional contributions to earnings call analysis on varying levels. Moreover, as the content review of over 30000 papers carried out by Brooks & Schopohl (2018) indicates, there has been an abundance of quantitative financial research and relative lack of qualitative approaches. Subsequently, this paper focuses on the qualitative nature of earnings calls, specifically the company management's outlook on the current and upcoming economic environment, and the quantitative

analysis in the form of actual price level data to see if such forecasts materialize. As a result, this study contributes to the financial literature by both taking a more qualitative approach and studying company earnings calls as a leading indicator of future inflation in the economy, rather than firm-level performance, as is more commonly found in the literature. Lastly, the research question analyzed in this study is whether public company earnings calls are informative as a leading indicator of future price trends in the economy.

The following text analysis in this research is carried out using NLP to examine earnings call transcripts in search of relevant information about the price changes of company products and services, and construct quarterly inflation likelihood scores to inspect the forecasting power of public company disclosure for upcoming price trends in the economy. In particular, I will analyze the U.S. firms' earnings call transcripts to construct inflation likelihood scores for the U.S. economy, which represent the management's outlook for the prices of the company's products and services. The earnings call transcripts of public companies in the Nasdaq 100 index compiled by quarter will be analyzed using NLP via the open-source Llama 2 7B model developed by Meta and the closed-source GPT-4o model developed by OpenAI to construct the inflation likelihood scores – an inspiration taken from Lopez-Lira & Tang (2024) for a scoring system developed from prompting multiple large language models of varying sophistication levels. Moreover, statistical analysis will be conducted to examine the correlation between the LLM inflation likelihood scores and actual realized inflation in the economy to evaluate the strength of these relationships, while simple regression analysis will help determine the significance of earnings call data in predicting the future direction of prices in the U.S.

The inspiration for this research, among other factors already mentioned, mainly comes from the anecdotal evidence that I have come across personally while following the post-pandemic financial market developments and seeing the potential for public firm earnings conference calls to potentially contain valuable information for gauging future economic trends, and inflation in particular. I expect to find a rather strong correlation between company management's confidence about their ability to raise prices of their output and the actual price levels in the economy that follow. I hypothesize that increased talk and optimism by senior executives during earnings calls about their company's ability to pass on cost increases onto the consumer and the general ability to raise prices of their goods and services would signal an upcoming higher inflation environment in the U.S.

## CHAPTER 2 Literature Review

### 2.1 Relevant Literature

The focus on earnings calls as a significant factor containing information for public company stock performance came into focus to a larger extent in recent decades. As Price et al. (2010) state, earnings calls became an increasingly relevant channel through which companies convey information about their business to the market. In addition, this disclosure, although often a touch promotional and understandably optimistic about the company's business prospects, can help narrow the information gap between the firm's management and its investors. In these earnings conference calls, investors have the opportunity to potentially get deeper understanding of the company's position in the current economic environment and its business outlook going forward, both from the prepared statement by management and the following question and answer (Q&A) section involving analysts and senior company executives. In aggregate, quarterly company earnings reports consist of two parts. Firstly, the press release comes out giving some key figures to analyze for the business performance as of the current quarter. Later, the actual earnings conference call starts with the aforementioned statement from the management, often reiterating the press release, followed by the Q&A session with analysts (Price et al., 2010).

It is important to consider that while these conference calls are an opportunity to get deeper understanding of company's current health and future prospects beyond fundamental numbers in regulatory filings, it is not without concern that depending on management's transparency, statements and answers during these calls are not completely open and sincere. Healy and Palepu (2001) stress exactly this concern that the extent of the disclosure during conference calls can be quite ambiguous. This is where technology meets finance, and the potential for the use of LLMs to fully comprehend the sentiment of and strategic logic used in the disclosure opens up.

Core (2001) at the time already saw potential for NLP techniques to help better understand and interpret the language used in earnings calls. Analyzing the existing literature, Core discussed the interdisciplinary benefits that the use of NLP programs in earnings calls analysis could bring with the use of this technology to identify the tone or sentiment, precision, and bias of the information stated in the disclosure. Moreover, Cohen & Nguyen (2024) show that companies shift the focus of their talk in the earnings calls between quarters based on the success of a particular metric. If a measure, such as sales growth, continues to perform well quarter-over-quarter, management will mention this intensely in their earnings call. However, once a metric that has been focused on previously shows a slowdown, executives will shift their talk to another metric that casts a positive light on the business, such as cost reductions. Subsequently, this shift in managers' "targets" as identified by the authors, tends to predict negative abnormal returns and subdued performance for the firm. In sum, the body of literature that

researches earnings conference calls from various angles, both from a few decades ago and more recent, seems to point at a substantial role for NLP via algorithmic models in analyzing earnings calls.

A surprising finding was made by Brooks & Schopohl (2018) in their extensive data science-based review of over 30000 published papers in which authors identify a significant lack of diversity in topics researched and methodologies for the analysis used. In particular, finance research is almost solely carried out using quantitative approaches from mathematics and economics, with relatively little use of interdisciplinary or qualitative methods. This finding in particular invites a more broad, non-traditional research in finance with an interdisciplinary focus. Such is the aim in this paper – to combine recent developments in AI tools from the technology sector to analyze economic trends from financial information sources generated by public companies. In addition to contributing to qualitative research methodology and combining interdisciplinary fields, this study dives deeper into the relationship between financial reporting of public companies via their earnings calls and the price trends in the economy. Specifically, I will take a qualitative approach to study the information contained in public firm communication within financial reports, which involves examining earnings call transcript text via LLMs to build likelihood scores for future direction of prices in the economy, and quantitatively test these predictions using actual realized inflation. This analysis can show the potential of company earnings calls to be used as a leading indicator of future trends in the economy, such as employment, domestic output, interest rates, inflation, and beyond. This research is focused on forecasting inflation. To my knowledge, there is relatively little emphasis in financial literature on the possible connection between the earnings conference calls and economic trends, as most focus is put on firm-level analysis, such as the effect of company financial disclosure on the performance of their shares traded in the stock market. A recent IMF working paper provides a fresh perspective on the use of public company disclosure. Particularly, Albrizio et al. (2023) study how the company inflation expectations, measured by the Earnings Calls-based Firm's Inflation Expectations (ECFIE) index estimated by their analysis of earnings calls involving text-mining and machine learning techniques, translate into real inflation in the economy. Authors find that in the U.S., higher inflation expectations by companies indicate higher future inflation in the economy. In addition to inflation expectations, the IMF working paper also builds an Earnings Calls-based Attention to the Central Bank index that focuses on the intensity of firm discussion about the monetary policy in effort to capture the attention of companies to the central bank policy and analyze the effectiveness of monetary policy transmission. Authors find that the impact of monetary policy shocks on inflation expectations are amplified considerably by the attention of companies to the central bank. Overall, this shows how the earnings calls can be of high value not only to the investors to gauge management's expectations for the business, but also to the lawmakers to entertain appropriate policy measures based on a leading indicator for inflation and making monetary policy transmission more effective.



In the realm of interdisciplinary approaches between the finance and technology sectors, Chava et al. (2022) use text from earnings calls transcripts to measure company inflation exposure. Authors train deep learning language models to indicate corporate discussions regarding price changes, while segregating the direction of price change and firm input costs from output prices. Chava et al. (2022) find that firms with higher vulnerability to input price pressures (e.g., wages or raw materials) tend to experience negative stock price reactions to earnings calls and hotter than expected CPI reports, accompanied by higher costs of goods sold with lower operating cash flows of their business in the future. The analysis in the paper is focused on the firm-level and provides insight into using NLP algorithms for gathering information about inflation contained in the textual data within company financial reports. Subsequently, an incentive for the research in this paper is to take a broader look at the overall trends in the economy, with the focus on aggregate inflation, while using NLP to analyze company-level financial report data as a potential leading indicator of the macroeconomic developments.

Lastly, a recent study by Lopez-Lira & Tang (2024) provides valuable insight into the viability of the use of now commonly available LLMs of various sophistication levels in the field of finance and model prompting techniques to extract the desired information from the provided textual context – a technique that the methodology in this paper takes inspiration from. Specifically, authors use LLMs, including the more advanced ChatGPT-4 and ChatGPT-3.5, and more basic GPT-1, GPT-2, BERT, and Fin-BERT, to predict daily stock market returns based on the analysis of news headlines. Firstly, the paper finds that ChatGPT scores, representing the model's positive or negative recommendation to buy or sell stocks respectively, are significant in forecasting daily stock returns. Moreover, the language model provides its reasoning behind the given score based on the context of the news headline, substantiating the prediction, provided this is something that is of interest to the user, which can be reflected in the prompt used. Another interesting finding that authors confirm in their testing is that different sophistication levels of LLMs reflect their ability to understand tricky context. More specifically, only the more advanced language models, such as ChatGPT-4, are able to sustain accurate understanding and interpretation of more nuanced context contained in the harder-to-comprehend news, together with data presented with strategic intent, such as in company press releases. However, the more basic models struggle to reflect the same capability in both the comprehension of low-readability news, and equity return forecasting. Subsequently, bigger LLMs not only have the potential and capacity to accurately forecast variable trends in the future, but are also able to understand the intricacies contained in low-readability news and identify the potential strategic intent and biases contained in the often own-issued firm press releases.

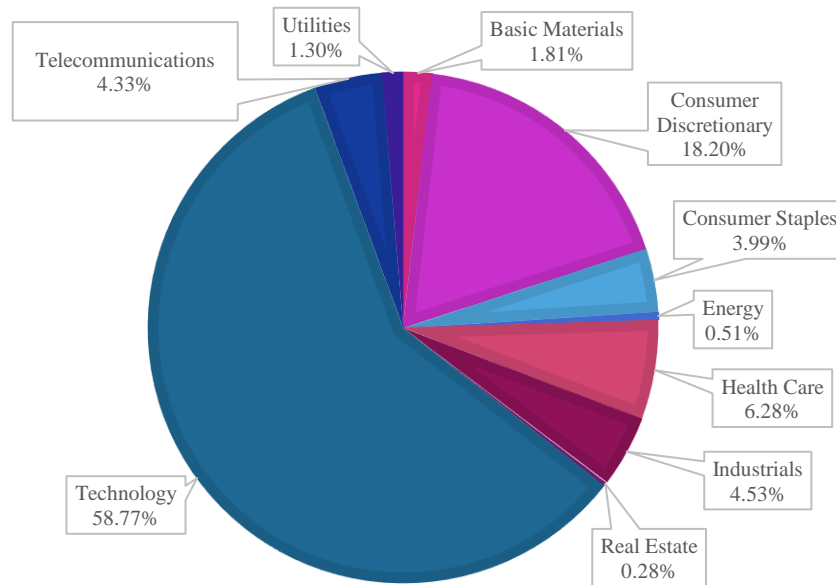
## **2.2 Research Expectations**

In this paper I aim to test the anecdotal evidence and the potential of company earnings calls to be a leading indicator of future trends in a country's economy. This paper will focus specifically on predicting the likelihood of future inflation. However, among the price changes of goods and services businesses provide, earnings calls can also contain much more useful information for the overall economy, such as the current conditions and future outlook for potential company layoffs and cost cutting, resulting in unemployment, or the aggregate consumer who, depending on the country, contributes to the growth of the overall economy in terms of its output to a varying degree. I hypothesize that the likelihood of future price increases of company goods and services, as identified by the inflation likelihood scores constructed by LLMs in the context of quarterly company earnings calls, will positively correlate with the price changes in the U.S. economy, making earnings calls a useful tool to use as a leading indicator of future price trends in the country. In addition, I do expect to find increasing prices in the U.S. over time, but in the shorter-term, having a leading indicator of upcoming price trends can help, for example, the central bank to avoid driving the economy into deflation together with a potential recession as a result of using lagging data for its monetary policy to set interest rates. Lastly, on the technology side, I expect the more sophisticated, state-of-the-art LLM to perform better, in terms of statistical significance, at analyzing the context of earnings calls to construct more accurate scores for predicting future inflation in the U.S. economy, similar to the findings of Lopez-Lira & Tang (2023).

## **CHAPTER 3 Data**

The earnings conference call transcript text data used in this research is obtained from the Refinitiv Eikon database (Thompson Reuters Eikon, 2023) and contains 3355 quarterly records. Earnings calls examined are from the U.S. public companies that constituted the Nasdaq-100 Index from January 2014 up to January 2024. The Nasdaq-100 index is a modified market capitalization-weighted index that is to a large extent comprised of technology companies. However, the index does not neglect other industries, such as consumer discretionary, consumer staples, healthcare, industrials, or telecommunications (Nasdaq, 2024). Therefore, the index provides a rather reasonable batch of some 100 constituent companies to reflect what the outlook of the management of U.S. public companies on the economy is. Moreover, the Nasdaq-100 index has a reconstitution schedule once annually in December, meaning the constituents of the index are reviewed and chosen every year. Therefore, companies that comprise the index are subject to change and companies from new non-financial industries may be added or firms currently representing certain industries in the index may be removed. The index industry weightings as reported by Nasdaq for May 3, 2024, can be seen in Figure 1.

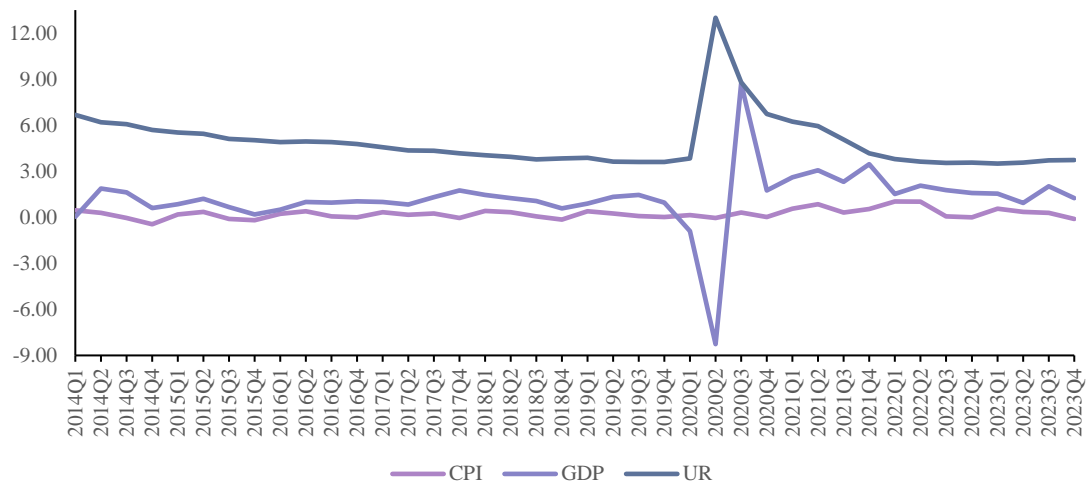
Figure 1: Nasdaq-100 industry breakdown



Note: Data for Nasdaq-100 constituent industries and their weightings in the index as of May 3, 2024. Since January 1, 2021, Nasdaq uses FTSE Russel Industry Classification Benchmark (ICB) 8-digit industry classification methodology.

Furthermore, the economic data for quarterly change in price levels contained in the Consumer Price Index (CPI) for all items, total for the U.S., is extracted from the St. Louis Federal Reserve Economic Data (FRED, 2024). The non-seasonally adjusted quarterly CPI inflation data was collected for the same period as for the earnings calls, ranging from January 2014 up to January 2024, and can be viewed in Figure 2. Similarly, additional data for this research includes the U.S. gross domestic product (GDP) and unemployment rate obtained from the St. Louis FRED and represent seasonally adjusted quarterly figures. GDP is the market value of all goods and services produced by the U.S. labor and property, and is represented as a quarterly percent change, while the unemployment rate is the percentage of unemployed in the labor force, which contains people of age 16 and above, in a quarter.

Figure 2: Quarterly change of macroeconomic variables



Note: the non-seasonally adjusted CPI data is for all items in total for the U.S. The CPI is aggregated using the average of the three months of inflation readings in a quarter and is shown as a quarterly percent change. GDP and unemployment rate data use seasonally adjusted figures and are represented as the quarterly change and the percentage of the labor force in a quarter respectively.

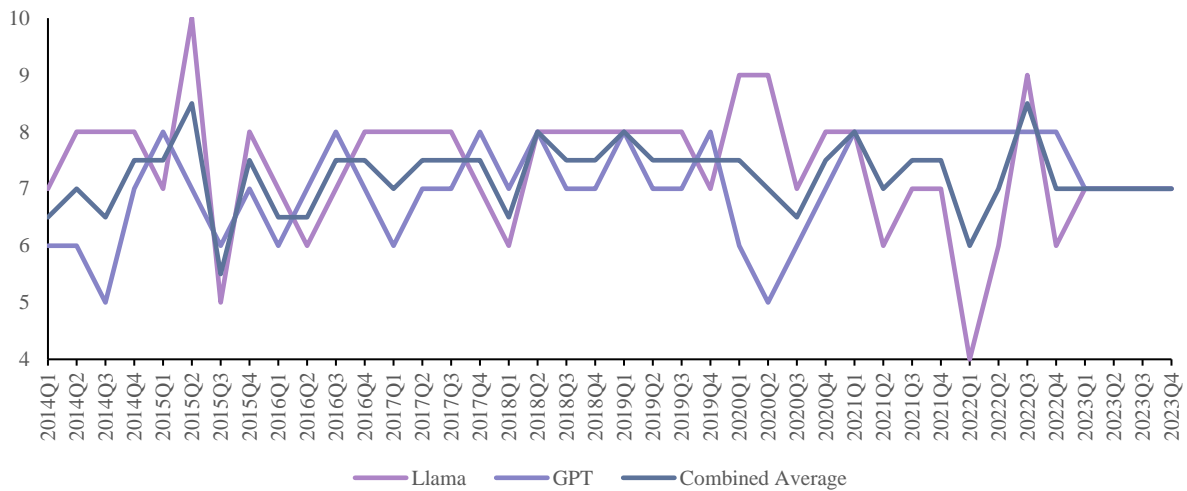
Lastly, the data inputs for inflation likelihood scores collected from the results obtained in the output of the LLMs, can be seen in Table 1. These scores represent the likelihood of future price increases that the model assigns to each quarter from 2014 to 2024 upon analyzing the earnings call transcripts within a specific quarter for all reporting firms that constitute the Nasdaq-100 index. A further visual representation of the data for inflation likelihood scores collected can be seen in Figure 3. It is also relevant to consider what information these inflation likelihood scores contain. Since the company earnings calls reflect on the business performance during the quarter preceding the one that the actual earnings conference call takes place in, not only does information disclosed in the call showcase how the company has performed in the previous period, but may also contain information regarding the business operations in the ongoing quarter and often management’s outlook for the upcoming quarters and the full year guidance. The specific research in this paper focuses on the predictability of future inflation in the economy using public company disclosure. Therefore, an earnings call taking place in a particular quarter is considered as a leading indicator for price changes in the succeeding quarters, which is reflected by the following analysis in this paper.

Table 1: Inflation likelihood scores produced by the LLMs

	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
<b>Llama</b>										
<b>Q1</b>	7	7	7	8	6	8	9	8	4	7
<b>Q2</b>	8	10	6	8	8	8	9	6	6	7
<b>Q3</b>	8	5	7	8	8	8	7	7	9	7
<b>Q4</b>	8	8	8	7	8	7	8	7	6	7
<b>GPT</b>										
<b>Q1</b>	6	8	6	6	7	8	6	8	8	7
<b>Q2</b>	6	7	7	7	8	7	5	8	8	7
<b>Q3</b>	5	6	8	7	7	7	6	8	8	7
<b>Q4</b>	7	7	7	8	7	8	7	8	8	7
<b>Combined Average</b>										
<b>Q1</b>	6.5	7.5	6.5	7	6.5	8	7.5	8	6	7
<b>Q2</b>	7	8.5	6.5	7.5	8	7.5	7	7	7	7
<b>Q3</b>	6.5	5.5	7.5	7.5	7.5	7.5	6.5	7.5	8.5	7
<b>Q4</b>	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7.5	7	7

Note: Inflation likelihood scores were obtained via the analysis of company earnings call transcripts for publicly listed companies constituting the Nasdaq-100 index using a fine-tuned Llama 2 7B and GPT-4o models, and calculating the combined average of the two scores. The scores are sorted by year and quarter, where each column represents the year, and each row highlights the quarter of a specific year.

Figure 3: LLM inflation likelihood scores by quarter



Note: Inflation likelihood scores were obtained via the analysis of company earnings call transcripts for publicly listed companies constituting the Nasdaq-100 index using a fine-tuned Llama 2 7B and GPT-4o models, and calculating the combined average of the two scores. The inflation likelihood score data is shown in quarterly frequency.

## CHAPTER 4 Method

### 4.1 Earnings Calls

Earnings call transcript data, extracted from the Refinitiv Eikon database via the Advanced Events Search function. The obtained 3355 quarterly records for Nasdaq-100 publicly listed companies are compiled into 40 text documents each representing all earnings calls for companies within a single quarter throughout all 40 quarters between the years 2014 to 2024. Depending on the quarter, each document contains from five million to six million characters and above – a length of text that is difficult to process for a human quickly and at scale. This is where NLP algorithms have immense potential in the financial analysis of lengthy transcripts and beyond, to spot intricate details that is of interest to the user and the overall sentiment and outlook contained in the text in a matter of seconds. This does not come without constraints, however, in the form of either time or computational processing power. Less sophisticated computer hardware may require a long time to process even a single document, whereas more advanced computational resources able to efficiently perform deep learning tasks are costly at the time.

### 4.2 Running the LLM

In this research I have experimented with both, running LLMs on a local machine hardware, and a free cloud service offered by Google Colaboratory (“Google Colab” from here). Performing NLP tasks on-device locally proved to be two-sided. On one hand, running a simple chat bot on local hardware is not too resource intensive and can be done even on older devices rather efficiently. However, when it comes to document processing and querying, especially lengthy earnings call transcripts that are of interest in this study, is time and computational resource intensive. Specifically, processing a single compiled document containing all earnings calls for 2014Q1 through Mistral 7B LLM on a central processing unit (CPU)<sup>1</sup> took overnight (around 12 hours), while processing the document through Llama 2 7B LLM on a graphics processing unit (GPU)<sup>2</sup> was much faster and completed within an hour. This however still comes at an opportunity cost of using more capable hardware for deep learning that is offered by cloud service providers. In particular, the mentioned Google Colab service proved to be a more attractive alternative for this research, providing more capable hardware for using LLMs to process and query documents. It is important to mention that these resources come with a time constraint, giving the user around an hour to two hours a day, depending on the usage patterns, before requiring a cooldown period until the next day, which was my experience using this service for the specific research carried out in this paper. Nevertheless, using Google Colab capabilities proved to be the most effective approach in terms of access to powerful enough computational resources and time constraints among the methods

---

<sup>1</sup> CPU model that the LLM was tested with: AMD Ryzen 9 4900HS.

<sup>2</sup> GPU model that the LLM was tested with: NVIDIA GeForce RTX 2060 Max-Q.

tested in this study for running a custom open-source LLM, taking on average 72.5 seconds to process a single file containing all company earnings call transcripts for a single quarter.

Moving on to the software side, since this research is working with lengthy quarterly transcript files of all companies in the Nasdaq 100 that reported earnings for a specific quarter, it was necessary to explore for a more custom user interface (UI) through which the LLM could be run, and which offered document reading capabilities, since the regular chat-only UI offerings for LLMs could not digest the length of a single document containing all earnings calls for Nasdaq-100 companies that report in a single quarter as a prompt. Therefore, among open-source UI offerings with document reading capability tested, the h2oGPT project was chosen for this research. It offers a resourceful platform capable of running on-device or through a cloud service, such as Google Colab, and features a private offline database with the ability to query and summarize documents of various file types, flexibility to choose among multiple instruct-tuned LLMs, and model fine-tuning options for document analysis, not to mention the wide accessibility of the platform with multiple operating system support, both CPU and GPU hardware support for running the LLMs, and a user-friendly interface (Roumeliotis, Tselikas & Nasiopoulos, 2023). Lastly, PrivateGPT is another viable UI option with document processing capability that was tested for this research on local hardware. However, the aforementioned h2oGPT provided more accessible model fine-tuning and Google Colab-ready code at the time of this research, therefore was deemed the most suitable UI for the particular analysis carried out in this paper.

On the model side, Llama 2 7B language model, developed by Meta, was used in this research. The Llama 2 7B is an open-source, highly flexible language model that has received a plethora of community fine-tuned applications for various use cases, and its calibrated version<sup>3</sup> for compatibility with the h2oGPT platform provides reliable document processing integration, while not being too large to run even on local machines with less powerful hardware, making it widely accessible to users in different domains (Roumeliotis et. al, 2023).

Furthermore, during the research period of this paper, the updated state-of-the-art GPT-4o model by OpenAI was released to the public on May 13, 2024. Subsequently, an additional batch of inflation likelihood scores was generated using this model. As of the time of this study, the GPT-4o model came out with document reading capability that is publicly available to any user with an OpenAI account and is free to use as of this research. OpenAI employs file limits of up to 20 per conversation and a single file is capped at two million tokens per text document (OpenAI, 2024). The company considers a single token for English text as being around four characters or 0.75 words (OpenAI, 2024). The files used in this research adhere to this limit, with all documents analyzed falling below the two million token limit,

---

<sup>3</sup> The Llama 2 7B model used in this research can be found on Hugging Face website via model search for the h2oai/h2ogpt-4096-llama2-7b-chat model.

as the largest combined earnings calls files reach just above six million characters. However, only a limited number of documents per session can be analyzed until a lengthy cooldown period of two days begins before the renewed possibility of uploading and chatting with new documents through the OpenAI online chat platform. This limitation was overcome via the Cursor application. This is a program mainly geared towards programming and code writing, although, it offers in-depth document analysis capabilities together with the new GPT-4o model, which fits with the methodology used in this paper well. The Cursor application provides higher flexibility for uploading many documents and the number of file inputs the user can work with, as compared to the regular OpenAI's ChatGPT online interface that is rather inefficient at analyzing a sizeable batch of documents at once. Moreover, when starting with a free plan, Cursor employs a usage quota of 50 requests using the GPT-4 model and 200 requests using GPT-3.5 turbo (Cursor, 2024). Lastly, within the Cursor application, the use of online sources as context to answer chat queries option is turned off and only the uploaded documents are used as context to answer the prompts for quarterly inflation likelihood scores.

### **4.3 Model Calibration**

Lastly, regarding the fine-tuning of the Llama 2 7B model, the h2oGPT platform provides a few customizability options, such as model temperature, maximum and minimum output length, and separating uploaded files into chunks and their sizes for document processing by the model<sup>4</sup>. Most notable adjustment to the model to mention here is the temperature. A non-user adjusted model, having its temperature set to 0.1, a setting that is associated with higher output precision and less creativity, seems to be systematically optimistic about positive future inflation with 80% to 100% expected likelihood of price increases, which is not completely unreasonable as prices do tend to go up overtime. However, a calibrated model with temperature set at 0.7 allows the model to be more creative and in theory less deterministic, although paired with the other model fine-tuning settings tested and applied in this research, as outlined in the Appendix A, the model output for earnings call transcript analysis does seem to provide seemingly more calculated inflation probability scores with higher variance in estimates throughout the time period tested.

### **4.4 Model Prompting and Output**

In the initial testing phase, inspired by the research and methodology of Lopez-Lira & Tang (2023), detailed prompt instructions were provided to the Llama language model on how to behave and execute the analysis of earnings call transcripts, as indicated by the Prompt 1 below. However, when instructed using this prompt, the model seemed to get confused and struggle to be creative enough to provide explanations beyond solely the output of the likelihood score, which suspiciously often was a systematically optimistic likelihood of future price increases between 8 and 10. While prices do tend to increase overtime, a simplified prompt, yet still providing clear instructions (Prompt 2), with an

---

<sup>4</sup> Please refer to the Appendix A in which the model settings used for this study are discussed in more detail.



increased temperature from the lows of 0.1 and 0.3 tested to 0.7 together with the appropriate model calibration settings outlined in the Appendix A, finally struck a seemingly right balance between precision and flexibility, with the model being more varied in its likelihood scores, while giving justifications for its predictions.

*Prompt 1:* Forget all your previous instructions. You are a financial expert with experience in analyzing earnings call transcripts of public companies. Given the earnings call transcript text for public companies, do you believe these companies will increase prices of their products and services in the future period? Respond on a scale from 0-10. Respond UNSURE if you do not know. Then elaborate on your answer with one short and concise sentence on the next line.

*Prompt 2:* Do you believe companies will increase prices of their products and services in the future period, based on the 20... Q... earnings calls? Respond on a likelihood scale from 0-10. Respond UNSURE if you do not know. Then elaborate on your answer with one short and concise sentence on the next line.

Where in Prompt 2, “20...” is replaced by the appropriate year, and “Q...” with the fitting quarter of the earnings calls being analyzed. Prompt 2 is used to extract the likelihood scores for price increases, as identified by the Llama model, for the analysis in this paper. Some examples of responses obtained can be seen in Appendix B.

The Llama 2 7B language model, calibrated for the specific analysis of earnings calls in this study, would sometimes only provide its predicted inflation likelihood score and indicate the source of information (earnings call transcript file) based on which it has made the decision, while issuing no reasoning behind that probability. In such instances, where the model did not produce the reasoning behind its likelihood score for future price increases in the output, the prompt was regenerated for the justification to be provided. If an explanation for the prediction of inflation likelihood was elaborated upon, the next quarter of earnings was analyzed.

Lastly, some curious findings came from testing the different prompts with the GPT-4o model run via Cursor application. At first, the model seemed to be very careful at giving a straight answer and constantly responded “UNSURE” regardless of the earnings quarter that was inquired about. This prompted further adjustments to the prompt used, and upon further testing, Prompt 3 finally helped harness consistent results in providing informative description together with inflation likelihood score predictions by the model. Subsequently, it appears that different LLMs require the user to learn their behavior through testing various prompts to arrive at desired output. Subsequently, the prompt used with

the GPT-4o model includes the removal of the sentence requesting to respond “UNSURE” and an addition of the first two sentences in Prompt 3, conditioning the model to act as a “financial analyst”. In sum, depending on the model at work, a variation of the prompt is used, which provides further examination of the different inputs the user may want to experiment with, given the subject of analysis at hand, the models they have access to, and the hardware and time limits they may face. With each model, slight adjustments to prompts may be tested to extract the output that the user intends, to achieve the most coherent and consistent results on the matter being analyzed.

*Prompt 3:* Forget all your previous instructions. You are a financial expert with experience in analyzing earnings call transcripts of public companies. Do you believe companies will increase prices of their products and services in the future period, based on the 20... Q... earnings calls? Respond on a likelihood scale from 0-10. Then elaborate on your answer with one short and concise sentence on the next line.

Where in Prompt 3, “20...” is replaced by the appropriate year, and “Q...” with the fitting quarter of the earnings calls analyzed. Prompt 3 is used to extract the likelihood scores for price increases, as identified by the GPT model, for the analysis in this paper. Some examples of responses obtained can be seen in Appendix B.

#### **4.5 Statistical Analysis**

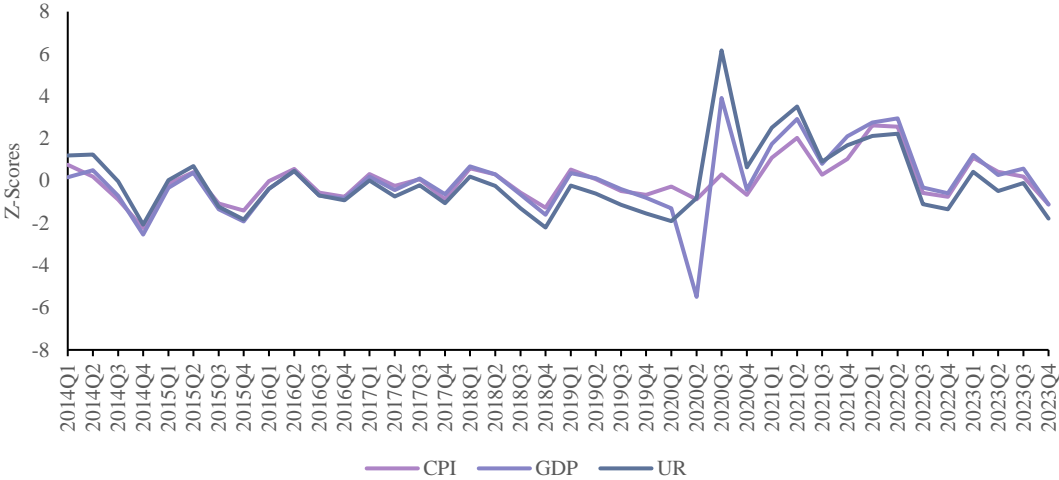
The actual CPI inflation data and the inflation likelihood score data, produced by the LLM, have been imported to the Stata statistical software where the statistical analysis is performed. It is worth noting that, the CPI data reflects the growth in prices in the U.S. from a quarter ago, while the earnings calls include management’s reflections on the previous quarters but also sentiment around the current quarter in which the earnings call takes place and future guidance on business performance. However, this research is specifically concerned about the informativeness of company earnings calls as a leading indicator of future price changes in the economy. Therefore, earnings calls that take place in a particular quarter are considered to have forecasting power for the price changes of goods and services that businesses charge in the upcoming quarter, which eventually translates into an effect on the CPI reading. Considering these dynamics, this study uses time series analysis in which vector autoregression (VAR) is carried out first to indicate the lags of inflation likelihood scores that best predict quarterly inflation. Moreover, a statistical test for Granger causality is performed to indicate whether the lagged inflation likelihood scores are actually associated with the price changes in the economy, as measured by the CPI. While the Granger causality test lets us examine the usefulness of one variable to forecast another, it is important to note that Granger causality is an old tool that is not widely used anymore, therefore should be considered only as an additional supporting indicator of potential relationships between the variables in this research, and should not be relied upon independently to make conclusions. Finally, regressions

with Newey-West standard errors for time series analysis with lagged variables are carried out with the respective inflation likelihood scores and control variable lags that were identified in the previous step by the VAR. The Newey-West regression coefficients of the inflation likelihood score lags help identify the relationship between the mentioned scores, derived from company earnings calls using LLMs, and CPI, which provides us with insight into the possibility of company earnings conference calls to be a leading indicator of future price changes in the economy. The Newey-West standard errors are used in this research to address the potential autocorrelation and heteroskedasticity in the residuals of variables and their lags over time in the time series analysis.

The regression inputs for control variables and inflation coefficient representations are based on the IMF working paper by Albrizio et al. (2023). First, the regression that is being analyzed through the aforementioned VAR and Newey-West regression analysis includes CPI as a dependent variable and various lags of control variables, including inflation likelihood scores, CPI change, GDP growth, and unemployment rate. The controls for macroeconomic variables used in this research are based on the Albrizio et al. (2023) paper and show a tendency to correlate with inflation overtime, hence prompting their inclusion as explanatory variables for CPI in this research. These macroeconomic variables are transformed into z-scores for comparability and the correlation of these variables over time can be seen in Figure 4 below. Furthermore, lags of four for all regression control variables are taken as a starting point – such as in the IMF working paper, showing that firm inflation expectations gathered from earnings calls seem to have predictive power for inflation in the first 12 months following an increase in inflation expectations – although further lag lengths per regression are analyzed thereafter in this paper. In particular, additional control variable lags that are examined include lags of three, four, five, and seven, depending on whether the significance of the right side of the regression lags identified by the VAR analysis carried out beforehand is greater than the 10% significance level, as identified by the regression p-values. The lags identified as suitable for further investigation vary based on the batch of inflation likelihood scores that were developed using different LLMs, and the combination of these datasets. No more than seven lags were analyzed, as with the addition of the eighth lag of all control variables, too many missing values appeared in the dataset, which consists of 40 lines of datapoints representing each quarter, resulting in too little observations to study eight regression variable lags and obtain any meaningful estimates. Moreover, the inflation likelihood scores, constructed via the LLM analysis of earnings call transcripts, are used as the indicator of upcoming price trends in the economy, as compared to the firm’s inflation expectations index that Albrizio et. al (2023) have constructed in their working paper, which is also compiled from company earnings calls using NLP methodology. In this paper, three batches of inflation likelihood scores are obtained: one using the fine-tuned Llama 2 7B model, one via the GPT-4o model, and the combined average of the two. Testing different NLP models, including a user fine-tuned model for document processing, an advanced “off-the-shelf” model, and the combination of results obtained from both models, provides insight into the more optimal

approach to extract the potential of the predictive power of earnings calls for future inflation in the economy. Moreover, the obtained inflation likelihood score batches are further analyzed to explore for the most significant range of the score lags that best help forecast future inflation, measured by the CPI. This analysis is helped by the inflation likelihood score lag visualization in linear graphs, together with their 90% confidence intervals to glean an overall picture of the trends the coefficients take over their respective quarterly lag periods. These visualizations show the associations between the change in CPI following an increase in the inflation likelihood score by one, and are further analyzed in the results and discussion section of this paper.

Figure 4: The correlation of CPI with GDP and unemployment rate



Note: The figure indicates standard deviations from the mean as represented by the Z-scores of the CPI, GDP, and unemployment rate. The graph is represented as a stacked line plot. The standardized values or the Z-scores of the variables represented are calculated by subtracting the mean of the variable observations range during the period from 2014 up to 2024 from the variable observation at a point in time, and dividing the result by the standard deviation of the mentioned range of the variable.

## CHAPTER 5 Results & Discussion

### 1.1 Results

This section dives into the results obtained from the analysis employing earnings conference calls as a leading indicator of future inflation in the U.S. economy. Such potential for forecasting power is extracted from earnings calls in the form of inflation likelihood scores constructed via the earnings transcript analysis using NLP power of LLMs. To begin with, the initial analysis includes running the VAR regressions for CPI including the different inflation likelihood scores and the macroeconomic variables. VAR is immediately followed by the Granger causality tests which are used solely as an additional tool to the analysis of the relationships, and is not relied upon for conclusion making. Finally,

the third part of the analysis involves regressions with Newey-West standard errors to generate final regressions used to interpret the results. This research considers the significance level of 5% as indicated by the p-value of a regression in the analysis. The base regression that is examined via the VAR and the Newey-West regression techniques and altered with the different sets of inflation likelihood scores obtained, is the following:

$$CPI_{t+q} = \alpha^q + \beta^q ILS_t + \gamma^q CPI_t + \delta^q GDP_t + \theta^q UR_t + \varepsilon_{t+q}, \quad (1)$$

where the dependent variable is quarterly CPI inflation, ILS is the lagged inflation likelihood score developed from earnings calls via LLM analysis, followed by the lagged values of quarterly change in CPI, quarterly growth in GDP, and quarterly unemployment rate. There are three variations of this regression developed in this study, all of which only vary in the single inflation likelihood variable, depending on the batch of the scores being analyzed. The three variations are as follows:

$$CPI_{t+q} = \alpha^q + \beta^q ILS_t^L + \gamma^q CPI_t + \delta^q GDP_t + \theta^q UR_t + \varepsilon_{t+q}, \quad (2)$$

$$CPI_{t+q} = \alpha^q + \beta^q ILS_t^G + \gamma^q CPI_t + \delta^q GDP_t + \theta^q UR_t + \varepsilon_{t+q}, \quad (3)$$

$$CPI_{t+q} = \alpha^q + \beta^q ILS_t^{avg} + \gamma^q CPI_t + \delta^q GDP_t + \theta^q UR_t + \varepsilon_{t+q}, \quad (4)$$

where all variables stay the same as in regression (1) except the change in the inflation likelihood scores, with  $ILS^L$  (2) representing scores obtained via the Llama 2 7B model,  $ILS^G$  (3) are the scores captured by the GPT-4o model, and  $ILS^{avg}$  (4) indicating the aggregated average scores obtained from the two models.

Starting with the findings in the VAR analysis, the lags of four and seven are tested for all three batches of the inflation likelihood scores obtained using the Llama model, the GPT model, and the average of the two combined batches of scores, while the GPT scores are further analyzed for lags of one, three and five, since these lags were statistically significant at least at 5% for this batch specifically, and not significant at lag four, nor seven. While the VAR analysis with higher lags, up to and including lag seven, shows some inflation likelihood score lag coefficients as significant depending on the score batch analyzed, these results are not considered reliable in this research. This is due to the high percentage of missing values resulting from the inclusion of multiple lags in a relatively small dataset. Specifically, using seven lags creates seven missing values out of 40 observations of a variable, leading to 17.5% of the data being omitted for the lagged variables. The considerable loss of observations is further indicated by the high R-square statistic for the inflation likelihood score variables, which shows an upward bias

with a fit of over 90% in the VAR analysis. This concern is further exacerbated by the Granger causality tests, which indicate that, in essence, "everything causes everything", where the Granger test indicates p-values of all associations in the lag seven VARs of zero, meaning a greater than 1% significance level. Subsequently, due to these reliability and validity concerns, the statistically significant lags of four are analyzed for regressions that include the Llama and the combined scores, and lags of one, three, and five for regressions with GPT scores. The VAR result excerpts for inflation likelihood score coefficients and their respective statistics are seen in Table 2 below. Here the statistics point to at least 5% significance level for the lags of all batches of the LLM scores analyzed in the VAR.

Table 2: VAR analysis of inflation likelihood score lags

LLM Scores	VAR Lags	$\beta$	SE	p	90% CI	
					LB	UB
$ILS^L$	L4	0.596	0.299	0.046	0.105	1.087
$ILS^G$	L1	2.067	0.576	0.000	1.120	3.014
	L3	1.610	0.546	0.003	0.712	2.508
	L5	1.609	0.628	0.010	0.576	2.642
$ILS^{avg}$	L4	1.100	0.489	0.024	0.296	1.904

Note: The table provides the excerpts for inflation likelihood score statistics from five VARs of the type (1), where only the number of variable lags and the score batches used differ per regression. The scores, including  $ILS^L$ ,  $ILS^G$ , and  $ILS^{avg}$  represent the inflation likelihood scores obtained from company earnings call via the Llama 2 7B and GPT-4o LLM analysis, and the combined average scores of the two, respectively. The statistical significance of the coefficient ( $\beta$ ) is indicated by the p-value (p), and the relationship is considered statistically significant if  $p < 0.05$ . The VAR standard errors (SE) are shown in column four, while the confidence interval (CI) includes the lower bound (LB), and the upper bound (UB).

Upon identifying the significant inflation likelihood scores that show potential in forecasting future CPI inflation in the economy, this relationship is further inspected via Granger causality tests. Note that this test is used only as a tool to get a sense of the variable relationships in the form of the usefulness of one variable to predict another and is not trusted independently for making any conclusions. That said, Granger causality tests seem to indicate that regression constructions overall may have some predictive power for CPI (Table 3). The Granger causality null hypothesis states that inflation likelihood scores do not Granger cause CPI. Put differently, the past values of the inflation likelihood scores do not contain useful information to predict future CPI readings. Based on this hypothesis it can be seen that at the 10% statistical significance level, except for the average combined scores, we cannot conclude that inflation likelihood scores have no use in forecasting CPI, as indicated by the p-values in Table 3. However, a lesser than 0.05 p-value is aimed for in this research for making deductions from the results, and thus only lags one and three of the GPT inflation likelihood scores appear to contain information that is useful

at forecasting CPI inflation. Moreover, with all regression (1) variables combined, including the lags of inflation likelihood score, GDP growth, and the unemployment rate, the Granger causality test shows constant significance even at 1% level across all regressions in the VAR analysis, meaning the null hypothesis is rejected and these controls jointly contain useful information for explaining CPI inflation.

*Table 3: Granger causality tests for inflation likelihood score lags*

<b>LLM scores</b>	<b>VAR lags</b>	<b>Caused variable</b>	<b>Causal variable</b>	<b>p</b>
<i>ILS<sup>L</sup></i>	L4	CPI	<i>ILS<sup>L</sup></i>	0.094
		CPI	ALL	0.000
		<i>ILS<sup>L</sup></i>	CPI	0.454
		<i>ILS<sup>L</sup></i>	ALL	0.086
<i>ILS<sup>G</sup></i>	L1	CPI	<i>ILS<sup>G</sup></i>	0.000
		CPI	ALL	0.004
		<i>ILS<sup>G</sup></i>	CPI	0.602
		<i>ILS<sup>G</sup></i>	ALL	0.475
	L3	CPI	<i>ILS<sup>G</sup></i>	0.000
		CPI	ALL	0.000
		<i>ILS<sup>G</sup></i>	CPI	0.552
		<i>ILS<sup>G</sup></i>	ALL	0.001
	L5	CPI	<i>ILS<sup>G</sup></i>	0.058
		CPI	ALL	0.000
		<i>ILS<sup>G</sup></i>	CPI	0.073
		<i>ILS<sup>G</sup></i>	ALL	0.000
<i>ILS<sup>avg</sup></i>	L4	CPI	<i>ILS<sup>avg</sup></i>	0.138
		CPI	ALL	0.000
		<i>ILS<sup>avg</sup></i>	CPI	0.379
		<i>ILS<sup>avg</sup></i>	ALL	0.727

*Note:* The table provides the excerpts from the Granger causality test p-values for the VAR inflation likelihood score effects shown in Table 2. The scores, including *ILS<sup>L</sup>*, *ILS<sup>G</sup>*, and *ILS<sup>avg</sup>* represent the inflation likelihood scores obtained from company earnings call via the Llama 2 7B and GPT-4o LLM analysis, and the combined average scores of the two, respectively. The significance level of the test is indicated by the p-value (p), and the Granger null hypothesis of non-causality is rejected when  $p < 0.05$ .

Following the results obtained from VAR and Granger causality tests, the one and three quarter lags of GPT inflation likelihood scores appear to be the most statistically significant and thus reliable in deducing associations between these scores and CPI inflation, as identified by the VAR studies.

Furthermore, lag five of the GPT scores shows a potential bidirectional relationship between the inflation likelihood sentiment of companies and the actual CPI metric. This instance highlights the importance of considering both the forward looking, but also the historical data when forecasting or analyzing price changes. Lastly, the higher lags of four for Llama and the model combined average scores and lag of five for the GPT scores, show some inflation forecasting potential, however not as strong, and reliable as those identified by three and nine month GPT score lags. More specifically, the lags four and five of the Llama and the GPT scores respectively only show statistical significance at the 10% level, as identified by the Granger causality tests, seen in Table 3, whereas the lag four of the average combined scores is not significant in this test at all.

Finally for the main part of the results, the score batches with significant lags as identified by the VAR analysis carried out previously are used in the Newey West regressions, which are the main analysis tool that conclusions in this paper are based on. While the GPT scores indicate the possibility to forecast inflation three to nine months out, the Llama and the combined average scores of the LLMs point towards these scores having the ability to predict inflation in the four quarter period identified in the VAR analysis. However, despite this potential, Llama and the model combined average inflation likelihood score lags of four drop in significance to only 10% level in the Newey-West regressions (Table 4). Moreover, the Newey-West regression analysis of the GPT score lags show promising results for using one and three lags of the GPT scores to forecast inflation. Based on the 5% significance level aimed for in this research, lags one and three stay firmly within the expected confidence range, with lag five falling short of this expectation as seen in Table 4. The comparison between the NW regression results indicates that, as one might expect, newer, more sophisticated, state-of-the-art models are able to better interpret the information contained within the company earnings calls to produce more accurate indicators of future inflation, as shown by the statistical significance of the Newey-West regression coefficients of inflation likelihood scores. This falls in-line with the findings of Lopez-Lira & Tang (2024), who show that more sophisticated models are more adept to interpreting nuanced information contained in difficult to read text, as compared to older, more basic LLMs. The findings in this paper observe the same trend, even though the two models used in the analysis were released to the public within a relatively short period of less than a year apart from each other. Furthermore, these findings also support one of the hypotheses raised in this paper, noting that the more sophisticated LLM will show better performance in interpreting information about the potential future price changes contained in earnings conference calls of companies.



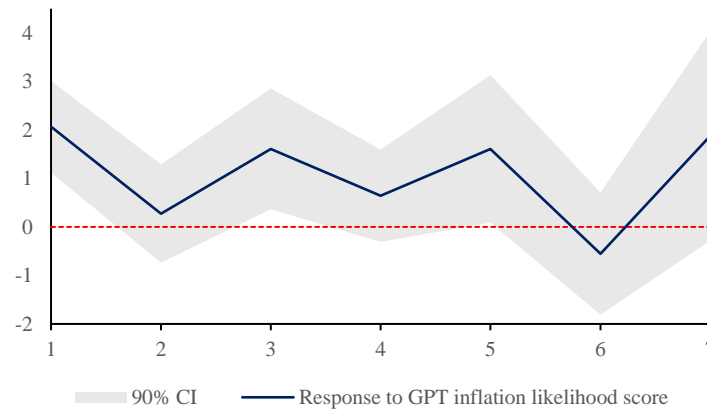
Table 4: Newey-West regressions for inflation likelihood score lags

LLM Scores	NW Lags	$\beta$	NW SE	p	90% CI	
					LB	UB
$ILS^L$	L4	0.596	0.303	0.064	0.072	1.120
$ILS^G$	L1	2.067	0.561	0.001	1.119	3.015
	L3	1.610	0.728	0.037	0.365	2.855
	L5	1.609	0.863	0.083	0.089	3.129
$ILS^{avg}$	L4	1.100	0.624	0.094	0.021	2.179

Note: The table provides the excerpts for inflation likelihood score statistics from five regressions of the type (1) with the Newey-West standard errors, where only the number of variable lags and the score batches used differ per regression. The scores, including  $ILS^L$ ,  $ILS^G$ , and  $ILS^{avg}$  represent the inflation likelihood scores obtained from company earnings calls via the Llama 2 7B and GPT-4o LLM analysis, and the combined average scores of the two, respectively. The statistical significance of the score coefficient ( $\beta$ ) is indicated by the p-value (p), and the relationship is considered statistically significant if  $p < 0.05$ . The Newey-West (NW) standard errors (SE) are shown in column four, while the 90% confidence interval (CI) includes the lower bound (LB), and the upper bound (UB).

The findings, using the methodology outlined in this paper, show that company earnings conference calls can be useful in leading inflation in the economy by three to nine months, with the added potential of forecasting CPI 12 months out. That is, depending on the LLM and the prompt variation used to extract the inflation likelihood scores from earnings calls, these scores have the potential to forecast the price changes in the economy one to four quarters out, but with a strong lean towards the three to nine months mark, as indicated by the 5% significance levels of  $\beta^q$  coefficients of GPT inflation likelihood scores from the type (3) regressions with Newey-West standard errors. A further visualization of the  $\beta^q$  parameter for inflation likelihood score trends can be seen in Figure 5, which shows that a unit increase in the inflation likelihood score is associated with around 1.61% rise in CPI inflation three quarters out. Similarly, a unit increase in the inflation likelihood score correlates with approximately 2.07% increase in CPI three months after.

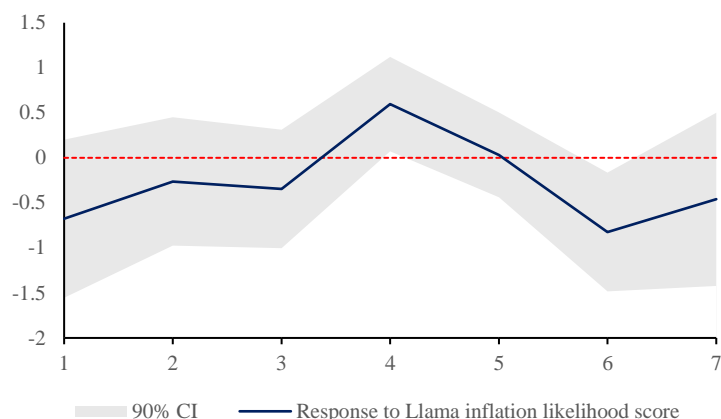
Figure 5: CPI response to a unit increase in the GPT inflation likelihood score



*Note:* The figure indicates CPI inflation response to a unit increase in the inflation likelihood score extracted from company earnings calls via the GPT-4o LLM. The confidence interval is set at 90% level. The reported standard errors are Newey-West standard errors. The horizontal axis represents CPI response timeframe in quarters, while the vertical axis shows the percentage change in CPI inflation following a unit increase in the inflation likelihood score.

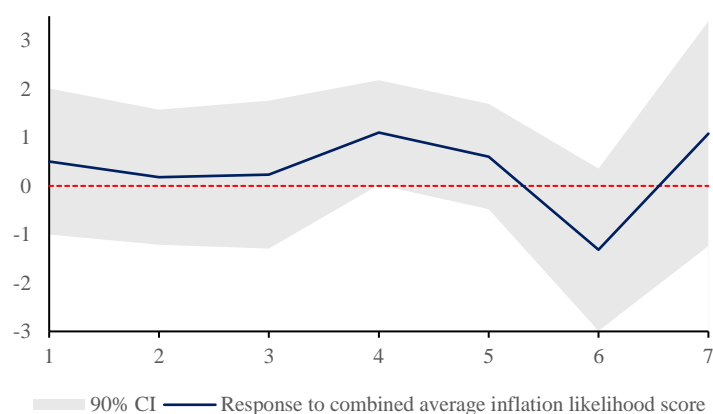
Further representation of the  $\beta^q$  coefficients of the type (2) and (4) regressions with the Llama and the model combined average scores can be seen in Figure 6, and Figure 7 respectively. The Llama scores show that a unit increase in the inflation likelihood score may be associated with approximately 60 basis points increase in CPI inflation in the following four quarters. To a similar extent, a unit increase in the combined mean LLM inflation likelihood scores has the potential to be associated with a 1.1% inflation increase in the U.S. economy in 12 months period. While the fourth lags of both inflation likelihood score batches are only significant at the 10% level, these findings still point to the potential for inflation likelihood scores to be useful in forecasting CPI inflation four quarters out. With new data at hand ready to be analyzed, this possibility is worth considering, especially provided the findings in other research, such as the Albrizio et al. (2023) paper, who show that a unit increase in their firm inflation expectations index developed from company earnings calls is associated with a two percent increase in the U.S. inflation in the first four quarters following the increase in expectations, marking the potential predictive power of company earnings calls for future inflation.

*Figure 6: CPI response to a unit increase in the Llama inflation likelihood score*



*Note:* The figure indicates CPI inflation response to a unit increase in the inflation likelihood score extracted from company earnings calls via the Llama 2 7B LLM. The confidence interval is set at 90% level. The reported standard errors are Newey-West standard errors. The horizontal axis represents CPI response timeframe in quarters, while the vertical axis shows the percentage change in CPI inflation following a unit increase in the inflation likelihood score.

*Figure 7: CPI response to a unit increase in the combined average inflation likelihood score*



*Note:* The figure indicates CPI inflation response to a unit increase in the combined average LLM inflation likelihood score. The confidence interval is set at 90% level. The reported standard errors are Newey-West standard errors. The horizontal axis represents CPI response timeframe in quarters, while the vertical axis shows the percentage change in CPI inflation following a unit increase in the inflation likelihood score.

Lastly, as a note, the inflation likelihood scores developed in this research are in essence going through three filters to arrive at the final stage result based on which the conclusions are made in this paper. Firstly, VAR analysis provides guidance on the lags of the scores that are worth investigating further, given their significance as shown by the p-value statistic of the given lag of the score coefficient. Furthermore, the Granger causality test is used to identify the potential relationships that the scores may have with the CPI inflation metric. However, this test is not something the paper bases decisions on or makes conclusions from solely, but is only used as an additional tool to spot potential associations between variables that may warrant further analysis. Finally, the last stage provides the Newey-West regression analysis results which are used to assess the potential of the LLM inflation likelihood scores extracted from company earnings calls to forecast future inflation in the economy. This three-step methodology provides three tools to generate a fuller picture of the associations between the inflation likelihood scores and the CPI to better understand this relationship and base the conclusions made in this paper.

## **1.2 Discussion**

Upon testing the various lags of the type (1) regression variables, and the separate batches of LLM scores as shown in regressions (2), (3), and (4), this research concludes that, based on the methodology and NLP models used in this study, company earnings conference calls have potential to be a leading indicator of CPI inflation in the U.S. economy three to nine months in advance. This association is determined by the state-of-the-art GPT-4o LLM, whereas the older generation Llama 2 7B model, and the combination of the two models in the form of merged and averaged inflation likelihood scores, provided statistically weaker results for forecasting twelve month inflation. These observations are similar to previous work by Albrizio et. al (2023), who find their firms' inflation expectations index to be associated with two percent increase in the U.S. inflation in the first four quarters following a unit increase in the firm inflation expectations. Even though the inflation forecasting power of the inflation likelihood scores for the four quarter mark declines in statistical significance, as found in this paper, it is nevertheless of interest to consider extending this forecasting potential to a year. Subsequently, based on the methodology and the NLP models used, one should examine the possibility of earnings calls to be a leading indicator of inflation three to twelve months out.

The hypotheses stated in this paper appear to have materialized. Firstly, the increase in inflation likelihood scores, generated by LLMs, seem to provide statistically significant evidence for being associated with higher future CPI inflation in the U.S. economy, thus providing potential for these scores to be a leading indicator of future inflation in the country. Moreover, the higher tier NLP model does indeed appear to be better at analyzing earnings call context more accurately to provide higher reliability inflation likelihood scores, as compared to the older, previous generation LLM. Specifically, the GPT-4o model, in addition to providing more structured and coherent reasoning, generated inflation

likelihood scores of higher statistical significance, as compared to the Llama 2 7B LLM. The performance of the LLMs is to a large extent attributed to the model size or the number of parameters of the model – elements that are learned from the training data and which determine the behavior of the model (Deepchecks, 2024). Subsequently, it may not come as a surprise that the comparably small seven billion parameter Llama 2 model underperforms the reportedly one trillion parameter GPT-4o model (Kanwar, 2024).

Following the results outlined in this paper and the related literature forming around the use of NLP technology in finance, it can be seen that this technology, with its ever-increasing ease of use and degree of sophistication, is a novel way for investors to glean a rich insight into future trends in the economy, to carry out informed investment decisions, while also making it a potentially attractive tool for policy makers to add to their arsenal in effort to widen the scope of data used for monetary policy decisions, which would further include leading information of future trends in the economy contained in company earnings calls. This can be done by employing the simple methodology used in this paper that compiles the public information from earnings call transcripts of desired companies that report during a quarter, and analyzes this document through a LLM by using prompts that help extract the information about the expected trends of future macroeconomic variables of interest. Updating this information quarterly can help investors and policy makers to stay up to date about the potential trends that may unravel in the future economy, and use that information as a supplement of potentially leading information to their analysis in order to make informed investment and policy decisions.

An area of concern, depending on the LLM used, is that the model may be trained on data that already includes information about the realized macroeconomic variable numbers for the timeframe of the training dataset, which is hard to identify, and may lead to questionable reliability of the output in the analysis. While it has to be acknowledged that it is dangerous to blindly trust the NLP model answers as a sole input to an analysis, this concern is at least partially addressed by instructing the model to use the provided earnings calls, containing information on what the management expects future trends in the economy to be, as the context to its answer for generating the variables of interest. Moreover, as the analysis is done in practice, using new data, it becomes less of a concern, as the actual future information about the economy is something that the model cannot have been trained on at the time of inquiry.

Another area of concern may surface from how the company management itself behaves in earnings calls and views the inflation in the economy. Corporations may be perceived, by some, as greedy – trying to extract every dime from their customers, thus raising prices in every instance they can. However, regardless of the greediness of the companies and the portrayal of their businesses in a positive light as seen in management’s tendency to discuss business results in earnings calls using positive metrics, as discussed by Cohen & Nguyen (2024), all companies do have to maintain a degree of

credibility with investors. Thus, deceptively indicating a constantly strong pricing power when there is none in the face of a weakening economy in which customers are no longer paying high prices for the services of products of that business, is not sustainable overtime. All of this is to say that such instances of potential deception in a batch of thousands of earnings calls do not appear to be reasonably expected to skew inflation likelihood score results much. In addition to this, the laissez faire or the free market mechanism to a certain extent eventually prevails in a capitalistic society. Companies can be considered greedy or actually face supply chain issues which drastically increase input costs that translate into higher output costs (Terlep, 2021) and subsequently generate higher price growth prevailing in the economy. However, if the customer stops paying these prices, companies have to act accordingly and slow the inflation or even turn to deflating goods and services pricing to retain some demand. All of this translates into the actual CPI inflation readings, and, as a sidenote, it seems that if inflation in the economy is considered to be “greedflation” by corporations, then deflation is charity. A stronger concern to further examine when constructing the likelihood scores for macroeconomic variables from company earnings calls, would be the debt burden levels of companies. IMF research by Albrizio et. al (2023) shows that, firms with higher proportions of long-term debt levels that mature within a year, extrapolate higher inflation expectations for the aggregate economy relative to their peers, based on their perception of increased debt servicing costs, provided a monetary policy shock – higher interest rates to cool inflation. The potential varying perception on the economy that companies with different financial structures have is something that is worth further evaluating in future research in the interest of generating macroeconomic variable likelihood scores via LLM analysis.

## **CHAPTER 6 Conclusion**

There are limitless possibilities to further seek out the most reliable prompts and the highest accuracy models to use for macroeconomic trend analysis using company earnings calls. This paper serves as a steppingstone to show that the new technological developments and financial analysis can go hand in hand, a methodology that can be used and further fine-tuned for the specific analysis needs at hand, and that such fintech synthesis can be useful by investors and policymakers alike to make informed decisions by adding a tool that may provide leading information on future economic trends. In particular, using the data and methodology in this paper I find statistically significant association between a unit increase in inflation likelihood scores and a 2.07% and 1.61% rise in U.S. CPI inflation within three and nine months respectively. Additionally, paired with Albrizio et al. (2023) findings, the possibility of the inflation likelihood scores to forecast inflation within twelve month period is also considered. However, due to weaker statistical significance of the scores for this period, no specific extrapolations are made for the four quarter mark in this study. Moreover, this research brings to the attention the current tools, including Google Colab, h2oGPT, and Cursor, and models, such as Llama 2 7B, and GPT-4o, together with their performance and limitations, that can be explored to conduct earnings calls analysis. Finally,

what sets this research apart is the ease of use of the methodology and its applicability in the real world for potentially tangible results in investments and monetary policy making. Generating likelihood scores for macroeconomic variables from publicly available company earnings calls via a simple chat interface using the now widely accessible and sophisticated LLMs is another potentially macroeconomic data leading tool to consider for investment research, portfolio management, and monetary policy making, which can all benefit from adding leading data to the analysis.

## REFERENCES

- Albrizio, S., Dizioli, A., & Simon, P. V. (2023). Mining the Gap: Extracting Firms' Inflation Expectations From Earnings Calls. In International Monetary Fund. Retrieved from <https://www.imf.org/en/Publications/WP/Issues/2023/09/28/Mining-the-Gap-Extracting-Firms-Inflation-Expectations-From-Earnings-Calls-539617>
- Brooks, C., & Schopohl, L. (2018). Topics and trends in finance research: What is published, who publishes it and what gets cited? In *The British Accounting Review*, 50(6), 615–637. <https://doi.org/10.1016/j.bar.2018.02.001>
- Chava, S., Du, W., Shah, A., & Zeng, L. (2022). Measuring Firm-Level Inflation Exposure: A Deep Learning Approach (SSRN Scholarly Paper 4228332). <https://doi.org/10.2139/ssrn.4228332>
- Chin, A., & Fan, Y. (2023). Leveraging Text Mining to Extract Insights from Earnings Call Transcripts. In *Journal of Investment Management*, 21(1), 81-102. Retrieved from <https://joim.com/leveraging-text-mining-to-extract-insights-from-earnings-call-transcripts/>
- Cohen, L., & Nguyen, Q. (2024). Moving Targets. In *SSRN Electronic Journal*. Elsevier BV. <https://doi.org/10.2139/ssrn.4736129>
- Core, J. E. (2001). A review of the empirical disclosure literature: discussion. In *Journal of Accounting and Economics* (Vol. 31, Issues 1–3, pp. 441–456). Elsevier BV. [https://doi.org/10.1016/s0165-4101\(01\)00036-2](https://doi.org/10.1016/s0165-4101(01)00036-2)
- Cursor. (2024). Memberships. Retrieved from <https://cursor.sh/pricing>
- Deepchecks. (2024, May 27). LLM Parameters. Retrieved from <https://deepchecks.com/glossary/llm-parameters/>
- Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. In *Journal of Accounting and Economics* (Vol. 31, Issues 1–3, pp. 405–440). Elsevier BV. [https://doi.org/10.1016/s0165-4101\(01\)00018-0](https://doi.org/10.1016/s0165-4101(01)00018-0)
- Kanwar, S. (2024, May 14). OpenAI launches GPT-4o or GPT-4 Omni for free, CEO Sam Altman says 'AI from movies'. Y20 India. Retrieved from <https://y20india.in/openai-launches-gpt-4o/>
- Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. In *SSRN Electronic Journal*. Elsevier BV. <https://doi.org/10.2139/ssrn.4412788>



- Nasdaq. (2024, May 3). Industry Breakdown. Retrieved from <https://indexes.nasdaqomx.com/index/Breakdown/NDX>
- Nasdaq. (2024). Nasdaq-100@Index. Retrieved from <https://www.nasdaq.com/solutions/nasdaq-100>
- OpenAI. (2024). How many files can I upload at once per GPT? Retrieved from <https://help.openai.com/en/articles/8983703-how-many-files-can-i-upload-at-once-per-gpt>
- OpenAI. (2024). What's a token? Retrieved from <https://openai.com/api/pricing/>
- Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. In *Journal of Banking & Finance*, 36(4), 992–1011. <https://doi.org/10.1016/j.jbankfin.2011.10.013>
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2023). Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model. In MDPI. <https://doi.org/10.20944/preprints202307.2142.v2>
- St. Louis FRED. (2024). [Consumer Price Index: All items: total for United States, 2014-2024]. Retrieved from <https://fred.stlouisfed.org/series/CPALTT01USM657N>
- St. Louis FRED. (2024). [Gross Domestic Product, 2014-2024]. Retrieved from <https://fred.stlouisfed.org/series/GDP#0>
- St. Louis FRED. (2024). [Unemployment Rate, 2014-2024]. Retrieved from <https://fred.stlouisfed.org/series/UNRATE#0>
- Sung, C., & Tuo, R. (2024). A review on computer model calibration. In *WIREs Computational Statistics* (Vol. 16, Issue 1). Wiley. <https://doi.org/10.1002/wics.1645>
- Terlep, S. (2021, October 24). U.S. Companies Bet Shoppers Will Keep Paying Higher Prices. *Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/u-s-companies-bet-shoppers-will-keep-paying-higher-prices-11635067802>
- Thompson Reuters Eikon. (2023). [Advanced Events Search, Earnings Conference Call, .NDX, United States of America, Transcript, 2014-2024]. Retrieved from Refinitiv Eikon desktop application.

## APPENDIX A LLM Fine-Tuning Options

This research uses a Llama 2 7B model that has been calibrated for compatibility with h2oGPT platform<sup>5</sup>. However, the interface of the said platform allows for further user fine-tuning of the NLP model for specific use cases. Therefore, upon further testing, some additional adjustments to the LLM were made with the focus on processing lengthy earnings call transcripts and querying these documents with precision, while leaving some margin for the language model to be creative enough to provide supporting arguments for the given predictions of likelihood scores for future price changes, based on the information contained in the earnings calls. Table A below lists the model calibration settings that were used in this research.

Table A: Llama 2 7B LLM calibration settings

Temperature	Top p	Top k	Max output length	Min output length	Max time (seconds)	Repetition penalty	Number of document chunks	Chunk size for document chunking
0.7	0.8	20	512	50	180	0.6	5	512

*Note:* The maximum time and the chunk size for document chunking are left at the default settings that are set after loading the model in the h2oGPT platform.

The fine-tuning options for the algorithm indicated in Table A provided the most consistent model outputs among various options tested for the analysis conducted in this study. These model calibration parameters may be adjusted based on specific user needs, depending on the intended use case of the model, the length of documents being analyzed, and precision required for the model output. Model calibration helps improve the algorithm’s effectiveness by enhancing its performance and prediction accuracy that is in line with the actual data analyzed (Sung & Tuo, 2024). Going into further detail on the fine-tuning options that are available to the user and are shown in Table A, temperature allows the user to determine the balance between accuracy and creativity of the model. Setting temperature towards zero results in more deterministic answers, but may lead to repetition in the answer output, while moving this setting towards one will lead to more diverse and creative answers stemming from increased randomness of the algorithm, however the output will be less deterministic and may lead to the so called “hallucinations” or incorrect and misleading results in the answers of the model. Furthermore, adjusting the top p parameter can help generate higher accuracy model responses to user prompts based on the most probable words in the text analyzed, while similarly the top k parameter helps adjust the balance between the diversity of tokens analyzed and the output relevance by prioritizing the words with the highest likelihood in the contextual document provided to the model. Moreover, maximum and

---

<sup>5</sup> The Llama 2 7B model used in this research can be found on Hugging Face website via model search for the h2oai/h2ogpt-4096-llama2-7b-chat model.

minimum output length parameters help balance the length of the responses to provide enough detail and prevent the answer from being excessively long, while allowing the maximum response time to be longer can give more room for the model to search for an appropriate answer and potentially provide a higher quality response. In addition, the repetition penalty limits redundancy of words used in the output, which can result in more detailed and elaborate answers. Lastly, the number of document chunks indicates the number of parts the document is split into for processing, which can help the model capture more nuanced information, while the chunk size for document chunking determines the number of characters that a chunk contains. However, both of these document processing-related settings are quite compute-intensive, requiring adequately capable hardware for the LLM to run and process lengthy text files if higher settings are desired. Overall, these parameter calibration settings were helpful in extracting the desired output for inflation likelihood scores with supporting arguments for these model-generated probabilities, based on the earnings call text processed by the algorithm, while keeping the constrained computational power and time resources in mind.

## **APPENDIX B LLM Output Examples**

Appendix B provides some of the Llama 2 7B and GPT-4o output examples that were generated in response to the Prompt 2 and Prompt 3 respectively, which were previously discussed in the methodology section of this paper and repeated for convenience below. The responses outlined in this section are based on the context of the earnings conference call transcripts for all companies reporting within a specific quarter, and indicated by the year and quarter to which the model output corresponds.

*Prompt 2:* Do you believe companies will increase prices of their products and services in the future period, based on the 20... Q... earnings calls? Respond on a likelihood scale from 0-10. Respond UNSURE if you do not know. Then elaborate on your answer with one short and concise sentence on the next line.

*Model response 2015Q4:* “I would say the likelihood of price increase is 8. The reason being that the companies are very positive about the sales trends and have been invest in cost reductions and customer experience.”

*Model response 2018Q4:* “likelihood of price increase in the future period: 8 based on the earnings call, it is likely that companies will increase prices of their products and services in the future period, as the commercial profitability and pipeline of the companies is improving.”

*Model response 2019Q3:* “8 I believe companies will increase prices of their products and services in the future period, based on the 2019 Q3 earnings calls. I believe companies will increase prices of their products and services in the future period, based on the 2019 Q3 earnings calls, as the companies expect to see revenue growth of 9% to 10.5% and organic revenue gains of 10.5% to 12% in Q3, which is a strong likelihood that prices will increase.”

*Model response 2021Q2:* “I would answer a 6 on the likelihood scale. The company mentioned strategic mix out, and price increases to offset higher input costs, but the impact of volatility on the business is not dissipating in the coming months, so it is not a for-sURE answer.”

*Model response 2021Q4:* “I believe the likelihood of companies to increase prices of their products and services in the future period is a 7 out of 10. As the companies have mentioned in their earnings calls, there is an expectation to recover material cost increases and price for those, and hopefully a little bit more in a strong market.”

*Model response 2022Q3:* “I would answer a 9 out of 10, as the earnings calls do not reflect any significant change in the overall price level, but the growth in ASP is anticipated to continue.”

*Model response 2023Q3:* “I believe the likelihood of companies to increase prices of their products and services in the future period is a 7 out of 10. As the earnings call has not discussed significant changes in market pricing dynamics in 2023, and pricing to DIY and professional customers has remained rational, there is a low likelihood of significant price increases in the future.”

*Note:* For presentation purposes, the model responses were edited to fit a single paragraph, as some response outputs produced included multiple paragraphs or paragraph separation symbols. In addition, repetitions of the same sentence in a different paragraph were removed, together with the link to the source document the model used as context for the answer, matching the quarter being analyzed.

*Prompt 3:* Forget all your previous instructions. You are a financial expert with experience in analyzing earnings call transcripts of public companies. Do you believe companies will increase prices of their products and services in the future period, based on the 20... Q... earnings calls? Respond on a likelihood scale from 0-10. Then elaborate on your answer with one short and concise sentence on the next line.

*Model response 2014Q2:* “Likelihood: 6. There are indications of potential price increases due to improving market conditions and strategic investments, but the overall outlook remains cautious.”

*Model response 2014Q3:* “Likelihood: 5. The earnings calls suggest a balanced approach with some companies focusing on market expansion and cost management rather than price increases.”

*Model response 2015Q3:* “Likelihood: 6. The 2015 Q3 earnings calls show mixed signals with some companies considering price increases due to strong performance, while others remain cautious due to macroeconomic uncertainties.”

*Model response 2015Q4:* “Likelihood: 7. The 2015 Q4 earnings calls suggest companies are likely to increase prices to offset currency fluctuations and rising operational costs.”

*Model response 2016Q1:* “Likelihood: 6. The 2016 Q1 earnings calls indicate a cautious optimism with some companies considering price increases to maintain margins amidst varying market conditions.”

*Model response 2017Q1:* “Likelihood: 6. The 2017 Q1 earnings calls indicate a mixed outlook, with some companies considering price increases due to cost pressures, while others focus on competitive positioning.”

*Model response 2018Q4:* “Likelihood: 7. The 2018 Q4 earnings calls suggest companies are considering price increases to offset cost pressures and maintain profit margins.”

*Model response 2019Q1:* ‘Likelihood: 8. The 2019 Q1 earnings calls indicate companies are likely to increase prices to counteract rising input costs and leverage strong market demand.’”

*Model response 2020Q1:* “Likelihood: 6. The 2020 Q1 earnings calls reflect uncertainty due to the COVID-19 pandemic, making widespread price increases less certain despite cost pressures.”

*Model response 2020Q3:* “Likelihood: 6. The 2020 Q3 earnings calls suggest some companies are considering price increases to offset rising costs, but overall caution remains due to ongoing economic uncertainties.”

*Model response 2021Q3:* “Likelihood: 8. The 2021 Q1 earnings calls suggest companies are planning price increases to offset higher input costs and capitalize on recovering demand.”

*Model response 2022Q1:* “Likelihood: 8. The 2022 Q1 earnings calls suggest companies are planning price increases to address ongoing inflationary pressures and maintain profitability.”

*Model response 2022Q2:* “Likelihood: 8. The 2022 Q2 earnings calls indicate companies are likely to increase prices to counteract persistent inflation and rising input costs.”

*Model response 2023Q2:* “Likelihood: 7. The 2023 Q2 earnings calls indicate companies are likely to increase prices to manage rising operational costs and sustain profitability.”

*Note:* For presentation purposes, the model responses were edited to fit a single paragraph, as the model output produced the likelihood score and justification on separate paragraphs, as requested by the prompt.