ISSN 1936-5330

kc FED Research Working Papers

Evaluating Local Language Models: An Application to Financial Earnings Calls

they basically used 3 metrics to guage the trade signals:
- what language nuances it uses (positive/negative/neutral)
- degree of clarity it engages (vague/precise)
- its temporal emphasis (backward/contemporaneous/forward oriented).

Thomas R. Cook, Sophia Kazinnik, Anne Lundgaard Hansen, and Peter McAdam

November 2023
RWP 23-12
http://doi.org/10.18651/RWP2023-12

Federal Reserve Bank of Kansas City

Evaluating Local Language Models:

An Application to Financial Earnings Calls *

Thomas R. Cook
Federal Reserve Bank of Kansas City

Anne Lundgaard Hansen
Federal Reserve Bank of Richmond

Sophia Kazinnik
Federal Reserve Bank of Richmond

Peter McAdam
Federal Reserve Bank of Kansas City

Abstract

This study evaluates the performance of local large language models (LLMs) in interpreting financial texts, in comparison to closed-source, cloud-based models. Our study is comprised of two main exercises. The first exercise benchmarks local LLM performance in analyzing financial and economic texts. Through this exercise, we introduce new benchmarking tasks for assessing LLM performance and explore the refinements needed to improve local LLM performance. Benchmarking results suggest that local LLMs are viable as a tool for general NLP analysis of financial and economic texts. In the second exercise, we leverage local LLMs to analyze the tone and substance of bank earnings calls in the post-pandemic era, including calls conducted during the banking stress of early 2023. Using local LLMs, we analyze remarks in bank earnings calls in terms of topics discussed, overall sentiment, temporal orientation, and vagueness. In response to the banking stress of early 2023, bank calls tended to converge to a similar set of topics and conveyed a distinctly less positive sentiment.

Keywords: Large Language Models (LLMs), Banking, Natural Language Processing (NLP)

JEL Codes: C45, G21

* We thank Stefan Jacewitz, Blake Marsh, and seminar audiences for valuable feedback. We also thank Bryson Alexander, Nadia Audzeichuk, Ethan Butler, Nicole Lindsay, and Kevin Muhic for excellent research assistance. The views expressed are those of the authors and not necessarily those of the Federal Reserve Bank of Kansas City, the Federal Reserve Bank of Richmond, or the Federal Reserve system.

Corresponding Author. Email: peter.mcadam@kc.frb.org

1 Introduction

In this paper, we evaluate whether locally deployed large language models (henceforth, local LLMs) are able to understand financial texts. While large language models in general are not explicitly trained for this purpose, models such as OpenAI's ChatGPT, have demonstrated a strong performance on this task (Hansen and Kazinnik, 2023; Lopez-Lira and Tang, 2023; Jha et al., 2023).[1] However impressive, there are several important drawbacks when it comes to using closed-source, cloud-based LLMs (henceforth, closed LLMs), such as ChatGPT. We provide empirical evidence that local LLMs can effectively replace closed LLMs in financial language tasks.

There are several considerable drawbacks when it comes to using closed LLMs. First, closed LLMs are not appropriate for scientific research[2] because they are not transparent and, crucially, their outputs are not reproducible. Second, use of closed LLMs carry privacy concerns making them ill-suited for confidential materials. Third, in most cases, closed LLMs do not allow for model customization as the underlying code is not accessible for modification. Finally, depending on the task at hand, closed LLMs can be prohibitively costly. While local LLMs might not rival closed LLMs in every task, they are designed to be fine-tuned and executed on consumer hardware and offer an appealing balance of convenience and privacy for specific tasks. Our focus in this paper is therefore on local LLMs and their capabilities.

We conduct two empirical exercises. The first exercise answers the questions of whether local LLMs could be used for performing natural language processing (NLP) tasks on the nuanced, highly specialized language present in economic and financial texts. And if so, what are the refinements necessary to sufficiently boost the performance of a local LLM?

[1] The impressive "zero-shot" performance attracted much attention because of the challenges and costs typically associated with data labeling, a time-consuming and often expensive process.
[2] See, for example Rogers (2023).

To answer these questions, we first benchmark several local LLMs against the task from Hansen and Kazinnik (2023) which prompts closed LLMs (namely, GPT-3 and GPT-4) to label portions of Federal Open Market Committee (FOMC) statements as "hawkish" or "dovish." We also benchmark local LLMs against a sentiment analysis exercise on financial texts. Specifically, we task local LLMs with scoring passages from the Financial Phrase Bank dataset (Malo et al., 2014) on overall sentiment[3] and on two new dimensions for which we have manually created labels: temporality and vagueness.

The second exercise explores the tone and substance of bank earnings calls in the post-pandemic era, including calls conducted during the banking stress of early 2023. Using refined prompts from the first exercise, local LLMs are prompted to score earnings calls on overall sentiment, temporality (i.e., the extent to which the content is forward- or backward-looking), and vagueness. To more precisely understand the substance of earnings calls, we also use LLMs to label earnings calls with the topics discussed (i.e., to produce a topic model).

Overall, our paper contributes to the intersection of several lines of research. First, we outline the steps needed to get local LLMs to produce high quality, automated natural language analysis on economic and financial texts. Second, we produce two new benchmark tasks for evaluating local LLM performance in that regard. Third, the applied exercise gives us a better understanding of the role of earnings calls in the relationship between banks and their investors.

Our research spans several strands of literature. In the field of economics and finance, there are several papers that explore the performance of LLMs for domain-specific tasks. Hansen and Kazinnik (2023) evaluate the ability of GPT models to classify the policy stance of FOMC announcements relative to human assessment. They find that GPT models

[3]The publicly released version of this dataset already has annotations for this task.

significantly outperform other NLP methods based on the distribution of labels across the classification methods. Lopez-Lira and Tang (2023) investigate the capabilities of ChatGPT in predicting stock market returns using sentiment analysis of news headlines and compare its performance to a suite of models, including BERT and GPT-2. Jha et al. (2023) extract firms' outlooks on corporate policies from earnings call transcripts using ChatGPT prompts. They highlight the capabilities of ChatGPT in processing and interpreting large volumes of textual data, specifically earnings call transcripts, to provide insights into firms' future corporate policies.

We also contribute to the research on local, open-source LLMs. By benchmarking against datasets that have not been released publicly, we can provide an assessment of model accuracy that is free from data-leakage.

Organization We discuss LLMs in Section 2, and in particular we describe the 5 models used in our exercises, and their comparative properties. Section 3 performs a benchmarking evaluation of these models. A key challenge being whether we can refine the models, or the "prompting", to reproduce a natural thought process and a satisfactory, informative textual analysis. This leads to a classification of texts along several dimensions: the *sentiment* of the text (is it, positive, negative, neutral?), the *clarity* of the text (is it vague or clear), its *temporal* dimension (does it refer to the past, present, or future?), and finally an analysis of topics.

key indicator =>

Section 4 then discusses our empirical application. This involves a textual analysis of banks' financial earning calls against the backdrop of the 2023 banking crisis. We also draw upon game-theoretic concepts of pooling and separating equilibria to interpret how different banks with different vulnerabilities and exposures might tailor the language of their earnings calls. Section 5, our results section, uses our LLMs to perform these four separate thematic analyses of the earnings call data. Our results resonate quite well with

earning calls are:

- more heterogenous and forward looking (expansion) during low stress periods. ==> +ve outlook

- more homogenous in outlooks during stressful ==> -ve outlook

the game theoretic backdrop. In the absence of a crisis, banks pool on promote as a strategy: discussing their own individual agendas in an effort to shape investors' expectations. This implies that earnings calls during a period of low stress are more heterogeneous, forward-looking, and positive in sentiment. Conversely, as the probability of bank stress increases, banks pool on reassurance. This implies earnings calls become more homogeneous in topics discussed, less forward-looking, and less positive in sentiment. Finally, Section 6 concludes. Additional material is in the appendices.

2 Large Language Models

We examine performance across the following five local LLMs, most of which derive from the LLaMA base model (Touvron et al., 2023) released by Meta in April 2023:

Table 1: Overview of Examined Models

Model	Parameters	Authors	Implementation
Wizard-Vicuna Uncensored	30B	Hartford (2023)	4-bit Quantized
Guanaco	33B & 65B	Dettmers et al. (2023)	4-bit Quantized
Fin-LLaMA	33B	Todt et al. (2023)	4-bit Quantized
Vicuna	7B & 13B	Chiang et al. (2023)	Mixed Precision

Due to hardware constraints, we quantize most models to 4-bits using Generative Post-Training Quantization (GPTQ) as suggested by Frantar et al. (2022).[4] This technique reduces the size of the model while limiting the impact on the quality of the model output and allows for significant improvements in the speed of inference. Recent research suggests that quantization to 4-bits provides an ideal reduction in memory without a substantial sacrifice in accuracy (Dettmers and Zettlemoyer, 2023). Other models in our sample were sufficiently small to be effectively run at a mixture of half (16-bit) and full (32-bit) precision.

[4]Very broadly, quantization is a process of reducing the amount of bits that represent a number.

All models explored in this paper are in some way derivatives of the LLaMA foundation models (Touvron et al., 2023).[5] Released in February 2023, LLaMA is a suite of models ranging in sizes from 7 billion to 65 billion parameters. The models are trained on over a trillion data points using publicly available data.[6] The model is open in the sense that the parameter values have been disclosed and in the sense that the model can be downloaded, modified, and deployed locally.[7] At the beginning of this project, LLaMA was the largest and most widely used foundation model that was free and available for local deployment.[8]

Vicuna by Chiang et al. (2023) is a fine-tune of LLaMA that aligns the model for interactive chat and instruction following. It is a full fine-tune (in contrast to the LoRA models discussed below) of LLaMA, meaning that parameters of the base LLaMA model were directly adjusted in response to new data. Vicuna was among the first and most popular instruction-tuned iterations of LLaMA. It is only available in 7 and 13 billion parameter versions. We the 7 billion parameter version at full (32-bit) precision, and the 13 billion parameter version at half (16-bit) precision.

Wizard-Vicuna Uncensored, initially created by Lee (2023) and later uncensored by Hartford (2023), is also a full fine-tune of LLaMA.[9] It combines different approaches to dataset construction and model training in an effort to produce a well-rounded model. Specifically, the model combines the evolutionary prompt generation strategy used in syn-

[5] A foundation model is a model that is not preconditioned to a particular domain.

[6] The training data used for GPT-3 (Brown et al., 2020) is disclosed and public, but many subsequent models (e.g. ChatGPT, GPT-4, Bard) have not disclosed the precise nature of their training data and/or have used data that is not otherwise publicly available.

[7] It is notable, however, that the license for LLaMA is more restrictive than typical open-source licenses (restrictions on commercial use). The subsequently released LLaMA 2 carries a less restrictive license but is still arguably not compliant with the principals of free and open-source software (FOSS).

[8] In the time since the start of this project, other foundation models have been released. At the time of this writing, however, LLaMA and LLaMA 2 still remain in widespread use.

[9] Many models are trained to avoid topics or reject requests that the researchers have determined to be potentially harmful, misleading, or otherwise unethical. This results in models sometimes rejecting benign requests. For example, ChatGPT will routinely reject requests for forecasts of stock prices or the weather. Training models to execute this form of censorship (prompt refusal) may result in a degradation in overall performance.

thesizing prompts for WizardLM (Xu et al., 2023) with the multi-turn conversation training approach used by the Vicuna model (Chiang et al., 2023). Authors argue that the resulting model performance exceeds the performance of the Vicuna models. We use the uncensored version of this model to avoid any degradation in model quality that may arise from the censoring process.

The Guanaco Dettmers et al. (2023) model places a quantized low rank adapter (QLoRA) on top of LLaMA. The low rank adapter method essentially creates a relatively small adapter model that interprets and adjusts LLaMA model outputs. The quantized version of this method (QLoRA) allows the adapter model to be trained using a quantized model. The QLoRA trained for Guanaco fine tunes the LLaMA model for instruction following. Because the QLoRA itself is relatively small, it can be trained more quickly and on less data than a full fine-tuning of the base LLaMA model. As a result, the Guanaco models can be built on the largest class of LLaMA model (65 billion parameters). For this paper, we use the 65 and 33 billion parameter versions of Guanaco, quantized to 4-bit. They are the largest parameter models we explore in this paper.

Fin-LLaMA by Todt et al. (2023) is a QLoRA for LLaMA, released recently and tailored for financial applications. We chose this model because the training data used was specifically oriented towards the subject of this paper: finance and economics. We expect that the use of financial texts as training data may yield superior performance in deciphering the baroque language patterns typically used in FOMC statements and earnings calls.

3 Performance Evaluation

3.1 Initial Performance and Refinements

Our initial testing examines whether and to what extent local models can perform complex economic NLP tasks. To assess performance, we take the exercise outlined in Hansen and Kazinnik (2023) as a benchmark.[10] Local LLMs are given the same dataset used in that paper and asked to perform the same task: label sentences from FOMC statements as "hawkish" or "dovish".[11]

The prompt used was the same across each model and tasked the model with identifying the sentiment of a sentence from an FOMC statement as "hawkish," "mostly hawkish," "neutral," "mostly dovish," or "dovish". Using the prompt in Hansen and Kazinnik (2023), we obtain results similar to those presented in Table 3 of their paper (see the "Initial" columns in Table 2). These results are underwhelming.[12] A key challenge (and a key contribution) of our paper is whether we can refine the models, or refine the prompting, to produce performance that is good enough to be useful.

Our initial experiments focus on adjusting the prompt in two ways. First, we refine the prompt by inducing chain-of-thought (COT; Wei et al., 2022) reasoning by the model. In this approach, the model is urged to outline its intermediate reasoning steps before delivering the final solution to a problem requiring multiple steps. The intent is for the model to reproduce a natural thought process as it navigates through multi-step reasoning. The accuracy scores for the COT prompt are shown in Table 2. Compared to the initial prompt, 3-category accuracy scores improve across all models, and 5-category accuracy

[10] This dataset has not been released publicly and therefore cannot have been incorporated into the training data used for any of the models we study here.

[11] We use both 3-category and 5-category assessments.

[12] Three of the six models tested produce accuracy scores that are worse than what we'd expect from a model that chose labels at random The exception to this is Vicuna 13B, which manages to perform surprisingly well using the initial prompt.

scores improve across all models except Vicuna 13B (which degrades modestly). In some cases, as with Vicuna 7B, the COT prompt produces as much as a 20-fold improvement in accuracy.

Second, we consider few-shot prompting, where we present the model with examples of input-output pairs (phrased as questions and answers) and then ask it to predict an answer based on a new example. In practice, we discovered that performance gains from this approach varied considerably depending on the precise structure of the prompt and the examples used. Some sets of examples would produce superior performance for one LLM and cause a degradation in performance for another. Our working explanation for this behavior is that the conceptually nuanced nature of hawk-dove classification makes some LLMs more sensitive to prompt design and possibly more likely to misinterpret examples.[13] To correct for this, we created a battery of eight different few-shot prompts and generated hawk-dove scores for each prompt using each model.[14] Table 2 provides the accuracy scores for the best performing few-shot prompt for a given model. Compared to the initial prompt, and using the few-shot prompt that best suits each model, the few-shot adjustment improves 3-category and 5-category accuracy in all cases except for the 5-category accuracy of Vicuna 13B. Compared to COT prompting, few-shot prompting produces comparable or improved performance for every model except Fin-LLaMA, which degrades considerably.

3.2 Benchmarking on Financial Data

With our refined prompting techniques, we expand upon the findings in Hansen and Kazinnik (2023) to examine whether and to what extent local LLMs can interpret finance-related statements. Specifically, we examine the performance of local LLM models in predicting

[13] We did not have the same difficulty with few-shot prompts in the Financial Phrase Bank exercise discussed below.

[14] See Table B.1 in the appendix for accuracy results for each prompt.

Table 2: Model performance with 3- and 5-categories using initial prompt, chain of thought (COT) prompting, and few-shot prompting

	3-category accuracy			5-category accuracy		
Model	Initial	COT	Few-shot	Initial	COT	Few-shot
Wizard-Vicuna	0.47	0.69	0.83	0.30	0.39	0.53
Guanaco (33B)	0.16	0.70	0.60	0.14	0.41	0.40
Guanaco (65B)	0.62	0.84	0.80	0.30	0.38	0.61
Fin-LLaMA	0.16	0.80	0.40	0.02	0.46	0.29
Vicuna (7B)	0.04	0.80	0.78	0.04	0.34	0.46
Vicuna (13B)	0.80	0.77	0.81	0.48	0.32	0.40

Notes: Scores reported under the Few-Shot column reflect the best performance for the model across a battery of few-shot prompts. The complete list of accuracy scores for the few-shot prompts used in this exercise can be found in appendix B.1.

the sentiment of financial phrases. We use data from the Financial Phrase Bank (Malo et al., 2014), which provides sentiment labels. To highlight the versatility of the LLM models, we extend the labels in the dataset to include two additional dimensions – vagueness and temporal perspective – which we use to test the ability of models to capture further nuances of text.

3.2.1 The Financial Phrase Bank Dataset

The Financial Phrase Bank dataset (Malo et al., 2014) provides high quality training data to evaluate the performance of language models. This data has been widely used in the literature, see, e.g., Huang et al. (2023), Shang et al. (2023), and Peng et al. (2021).

The dataset consists of 4,840 sentences selected randomly from English financial news articles and company press releases on companies listed in OMX Helsinki. These sentences are labeled manually as *positive*, *neutral*, or *negative* from an investor's standpoint. This task was distributed among 16 annotators with adequate background knowledge in finance; each sentence is annotated by 5-8 people.

The dataset is published in four different versions: (i) sentences for which all annotators agree; (ii) sentences for which more than 75% of the annotators agree; (iii) sentences with more than 66% agreement; and (iv) sentences with more than 50% agreement. The choice among these versions is not innocuous for our exercise. The dataset in which all annotators agree likely represents sentences that are easier to classify. Hence, using these sentences presents the models with an easier task. To challenge our models, we therefore aim for some disagreement among the annotators. Using the 50% agreement dataset, however, raises concerns about the reliability of the benchmark. It is not clear whether a model that agrees with just half[15] of the annotators can be considered accurate. We consider the 75% agreement dataset as a balanced compromise.

3.2.2 New dimensions of sentiment

While the classification of text into a positive, neutral, or negative sentiment is a popular way to quantify text data, we argue that other dimensions of sentiment are equally important to consider. Besides providing interesting insights into the wider types of exercises for which LLMs can be applied, considering different aspects of text also poses an additional challenge for the models and allows us to test how well they can capture nuance.[16]

We consider two additional aspects of sentiment beyond the positive– neutral–negative dimension. First, we classify the text into clear and vague language. When making statements, companies may resort to vague or ambiguous language to make circumstances sound more favorable (or less unfavorable) than they actually are. This practice is so widespread that the Securities and Exchange Commission (SEC) publishes *A Plain English Handbook*, which contains guidelines for companies' disclosures. The handbook calls for avoiding long

[15] We similarly disregard the 66% split which is usually separated from the 50% split by the agreement of only one or two additional annotators.

[16] As with the FOMC dataset, labels for these two new dimensions are not publicly available and thus cannot have been used for the training of any of the local LLMs we investigate in this paper.

sentences, superfluous words, jargon, the passive voice, and abstract words. Statements may also be made vague by creative use of synonyms. Suslava (2021) studies the usage of corporate euphemisms in earnings calls. She defines corporate euphemisms as negative sentiment coupled with obfuscation. Determining whether language is vague is relatively straightforward for humans, but historically it is a substantial challenge for computer models.

The other dimension that we examine is a temporal perspective. Generally, most financial statements are about past performance, future expectations, or present conditions. The temporal perspective has a relationship to the meaning and sentiment of a statement. In ordinary language, it is often easy for an NLP model to discern whether a remark is about the past, present, or future. But phrases such as "We have revised our earnings forecast upward" can easily create confusion for computer models.

We assess the local LLM performance in rating the vagueness and temporal perspective of financial statements. To provide a useful benchmark, we tasked five human annotators (research analysts, RAs[17]) with labeling a subset of 1,000 sentences drawn randomly from the Financial Phrase Bank dataset according to the vagueness and temporal perspective of each sentence. The RAs are asked to treat each sentence as independent statements without further context. We provided the RAs with the following definitions of the vagueness and temporal dimensions:

Definition 1 (Vagueness). A sentence is either clear or vague. With this label we are trying to identify language that is intentionally convoluting the message or talking around the subject.

- A vague sentence can be subject to misinterpretation and more than one interpretation. A clear sentence is NOT subject to either.
- A vague sentence can convey contradictory messages. A clear sentence does NOT.

[17]The RAs all have educational backgrounds in economics, finance, or political science and are employed by the Federal Reserve Bank of Richmond.

– A vague sentence uses language that is shadowy, dim, obscure, indistinct, hazy, or uncertain and may use weasel words (words that are used to make a dubious claim appear to be a strong claim and avoid outright lying, i.e., intentional vagueness) or words that make statements sound less factual. A clear sentence uses unambiguous and transparent language.

– A vague sentence can leave the reader confused, urged to do a double take, and the reader may wonder what the author really meant or wanted to say. The reader may have difficulties explaining to a third party what the sentence was about. A clear sentence generally makes the reader feel confident about having understood the message in the first take. This does not mean that sentences written in technical language or that long sentences are vague. Sentences can be difficult to read because they are written in complex language but still be clear.

Definition 2 (Temporality). A sentences is focused on the past, the present, or the future.

– "Backward-looking" text talks about history, interpreting past events, learning, or reflecting on the past.

– Text that focuses on the "present" interprets or comments upon ongoing events.

– "Forward-looking" text talks about the future, plans, expectations, or strategies.

Accordingly, we consolidate the results from the five RAs by converting labels into numeric values, where

$$vague = 1, clear = 0$$

and

$$future\ focus = +1, present\ focus = 0, past\ focus = -1$$

and then taking the average value. Table 3 reports the number of sentences annotated with each label. First, we note that the sample is representative of the full Phrase Bank dataset in which 32% of the sentences are labeled "positive", 54% have the "neutral" label, and 15% are labeled "negative"; in our subset, there are 33% positive, 52% neutral, and

14% negative sentences. Next, we describe the distribution of labels along the two new dimensions.

As expected, there is an imbalance in the scoring on vagueness, with most sentences being scored as "clear" and only about 10% being scored as "vague". Half of the "vague" sentences have positive investor sentiment, and a majority of them focus on the present. The scores for temporality are almost evenly distributed. They exhibit a small skew toward present- and past-focused sentences.

The table also summarizes the degree of disagreement among the reviewers. All five annotators agree on both labels in about a third of the sample, and more than four out of the five reviewers (>75% agreement) agree in about 60% of the sample. Finally, there is more than 50% agreement for nearly all sentences.

3.2.3 Results

We evaluate the ability of the considered LLMs to classify the sentiment of Phrase Bank sentences along the three dimensions: (i) positive–neutral–negative, (ii) clarity–vagueness, and (iii) past–present–future. The performance evaluation focuses on sentences for which at least 75% of the human reviewers agree for the reasons discussed above. We focus on results obtained using the few-shot prompt, which tends to generate higher accuracy in this context[18] than chain-of-thought prompting.[19]

Table 4 reports results for the positive–neutral–negative dimension. The accuracy exceeds 75% for all models, but the Guanaco model with 65B parameters. Accuracy is highest

[18] This improvement was much more consistent than for the hawk-dove exercise – nearly all models responded similarly to the few-shot prompt. In the few instances where the few-shot prompt did improve model performance (compared to COT prompting) the degradation in accuracy was quite small (less than 5%). Though we are currently still researching this, we suspect that the improved consistency in response to the few-shot prompt (compared to the hawk-dove exercise) is attributable to the fact that the Financial Phrase Bank tasks are less nuanced and require classification into fewer categories.

[19] For comparison, Table B.6 in Appendix B shows results for positive–neutral–negative classification using the chain-of-thought prompt.

Table 3: Number of sentences classified with each label.

	Sentiment			Language clarity		Temporality		
	Positive	Neutral	Negative	Vague	Clear	Future	Present	Past
Total	333	523	144	107	893	238	438	324
Conditional:								
Positive	–	–	–	54	279	64	160	109
Neutral	–	–	–	24	499	160	212	151
Negative	–	–	–	29	115	14	66	64
Clear	279	499	115	–	–	216	360	317
Vague	54	24	29	–	–	22	78	7
Future	64	160	14	22	216	–	–	–
Present	160	212	66	78	360	–	–	–
Past	109	151	64	7	317	–	–	–
All agree	–	–	–	4	443	175	240	199
>75% agree	–	–	–	35	725	213	306	250
>50% agree	–	–	–	265	735	262	349	360

Notes: The table reports both total count and count conditional on other labels. The positive–neutral–negative labels are from the Financial Phrase Bank dataset. The vague–clear and future–present–past labels are results from our annotation exercise involving five human reviewers. For these, the table also reports number of sentences for which all reviewers agree on both the vague–clear and future–present–past labels and for which more than 75% and 50% agree.

for the Vicuna model with 13B parameters. Since the categories are distributed almost evenly across the sentences (see Table 3), the Vicuna 13B model also achieves the highest F_1 scores and balanced accuracy.[20] This is consistent with the initial performance results reported in Section 3.1. At first, it seems surprising that the Vicuna model can outperform the other models, which are much larger with at least 30B parameters. However, the differences between the models are not just their size, but also their training datasets. Along the same lines, comparing the performance of the Guanaco models suggests that more parameters worsen classification accuracy. Since these models are trained on the same dataset, we conjecture that this result can be attributed to an overfitting problem of the 65B parameter model.

Comparing the results with those obtained on the same data set using different methods by Malo et al. (2014) (see Table 4, panel B), we note that all LLMs outperform the dictionary-based methods. This is unsurprising as we expect LLMs to better capture nuance and context than word-count-based methods. We also note that the performance of the Vicuna models is comparable to – and in some aspects, such as the F_1 score, exceeds – the performance of the Linearized Phrase-Structure (LPS) model of Malo et al. (2014).

Next, we discuss the ability of the models to quantify texts along other dimensions than positive–neutral–negative sentiment. Table 5 reports the performance measures for classifying the sentences as clear or vague. When considering these results, recall from Table 3 that this is a highly unbalanced classification exercise with just less than 5% of the sentences labeled as vague.

Thus, while the Vicuna model with 13B parameters achieves high precision in detecting clear sentences, it fails to classify sentences as "vague," which lowers its overall accuracy and F_1 scores. The Vicuna model with 7B parameters balances this trade-off better and

[20]The F_1 score is the harmonic mean of the precision and recall.

thus achieves higher overall accuracy. The results do, however, reflect the difficulty of detecting vagueness: the Vicuna 7B model detects just slightly more than half of the "vague" sentences. This is likely due to the fact that annotators disagree more on the classification of "vague" sentences. For the subset of sentences for which all annotators agree, the models are much better able to capture vagueness (see Table B.3 in Appendix B).[21]

Finally, we consider the classification of temporality in Table 6. The models all perform well in this exercise. The Guanaco 65B model achieves the highest overall accuracy, and it has high F_1 scores and balanced accuracy for all categories. It is notable that, unlike the classification of clarity and sentiment, more parameters is beneficial for the classification of temporality, i.e., the Guanaco model with 65B parameters outperforms that with 33B parameters, and the Vicuna 13B model outperforms the Vicuna 7B model.

[21] These results are, however, based on a highly unbalanced data set with just 4 out of 447 sentences classified as "vague".

Table 4: Model performance with few-shot prompting for classifying positive–neutral–negative sentiment of Phrase Bank sentences

	Wizard-Vicuna	Guanaco (33B)	Guanaco (65B)	Fin-LLaMA	Vicuna (7B)	Vicuna (13B)
Accuracy	0.766	0.784	0.694	0.784	0.798	**0.826**
Precision:						
Positive	0.587	**0.891**	0.468	0.833	0.815	0.717
Neutral	0.839	0.756	**0.962**	0.763	0.779	0.854
Negative	**1.000**	0.968	0.904	**1.000**	**1.000**	0.932
Recall:						
Positive	0.766	0.383	**0.969**	0.508	0.516	0.711
Neutral	0.784	**0.981**	0.552	0.959	0.953	0.881
Negative	0.660	0.566	**0.887**	0.396	0.547	0.774
Specificity:						
Positive	0.766	0.383	**0.969**	0.508	0.516	0.711
Neutral	0.784	**0.981**	0.552	0.959	0.953	0.881
Negative	0.660	0.566	**0.887**	0.396	0.547	0.774
F_1 score:						
Positive	0.664	0.536	0.631	0.631	0.632	**0.714**
Neutral	0.810	0.854	0.701	0.850	0.858	**0.867**
Negative	0.795	0.714	**0.895**	0.568	0.707	0.845
Balanced accuracy:						
Positive	0.785	0.683	0.791	0.735	0.736	**0.805**
Neutral	0.759	0.710	0.756	0.717	0.739	**0.807**
Negative	0.830	0.782	**0.935**	0.698	0.774	0.883

Notes: The table reports classification performance for the subset of Phrase Bank sentences for which at least 75% annotators agree (a total of 3,448 sentences). The considered LLMs are implemented with few-shot prompting. **Bold-faced** values indicate the winning model per performance metric.

Table 5: Model performance for classifying clarity of Phrase Bank sentences

	Wizard-Vicuna	Guanaco (33B)	Guanaco (65B)	Fin-LLaMA	Vicuna (7B)	Vicuna (13B)
Accuracy	0.861	0.725	0.696	0.839	**0.963**	0.846
Precision:						
Clear	0.990	0.992	0.994	0.971	0.974	**0.995**
Vague	0.228	0.133	0.124	0.139	**0.640**	0.221
Recall:						
Clear	0.863	0.718	0.686	0.857	**0.987**	0.843
Vague	0.824	0.882	**0.912**	0.471	0.471	**0.912**
F_1 score:						
Clear	0.922	0.833	0.812	0.910	**0.981**	0.913
Vague	0.357	0.231	0.219	0.215	**0.542**	0.356

Notes: The table reports classification performance for the subset of the 1000 sentences for which at least 75% of the annotators agree (a total of 760 sentences). **Bold-faced** values indicate the winning model per performance metric.

Table 6: Model performance for classifying temporality of Financial Phrase Bank sentences

	Wizard-Vicuna	Guanaco (33B)	Guanaco (65B)	Fin-LLaMA	Vicuna (7B)	Vicuna (13B)
Accuracy	0.746	0.706	**0.767**	0.745	0.605	0.671
Precision:						
Positive	0.767	0.632	0.851	0.824	0.853	**0.905**
Neutral	0.674	0.717	0.732	**0.825**	0.527	0.709
Negative	**0.828**	0.750	0.771	0.673	0.698	0.618
Recall:						
Positive	0.573	**0.596**	0.484	0.507	0.136	0.268
Neutral	**0.939**	0.883	**0.939**	0.781	0.951	0.709
Negative	0.711	0.639	0.826	0.882	0.652	**0.921**
Specificity:						
Positive	0.924	0.848	0.964	0.953	**0.989**	0.987
Neutral	0.752	0.789	0.807	**0.902**	0.519	0.824
Negative	**0.887**	0.841	0.817	0.697	0.754	0.571
F_1 score:						
Positive	**0.656**	0.614	0.617	0.628	0.235	0.413
Neutral	0.785	0.791	**0.823**	0.802	0.678	0.709
Negative	0.765	0.690	**0.797**	0.763	0.675	0.739
Balanced accuracy:						
Positive	**0.748**	0.722	0.724	0.730	0.562	0.627
Neutral	0.845	0.836	**0.873**	0.842	0.735	0.766
Negative	0.799	0.740	**0.822**	0.789	0.703	0.746

Notes: The table reports classification performance for the subset of the 1000 sentences for which at least 75% of the annotators agree (a total of 769 sentences). **Bold-faced** values indicate the winning model per performance metric.

4 Earnings calls and the banking stress of 2023

Having established the performance of local LLMs, we turn to an empirical application. We use LLMs to analyze earnings call transcripts from banks to establish patterns in language and communication around the 2023 banking crisis (for an excellent discussion see Acharya et al., 2023).

4.1 Background

We now provide some background explanation of financial earning calls, against the backdrop of the 2023 banking crisis.

4.1.1 Earnings calls

A financial earnings call is a conference call during which the management of a public company announces and discusses their financial results for the quarter. External participants are investors, equity analysts, and journalists. Investors frequently plan trades around the date of a call and use the information to update their earnings estimates and portfolios. Accordingly, they are considered to have some predictive and narrative power (Correa et al., 2021; De Amicis et al., 2021; Roozen and Lelli, 2021).

A typical earnings call has three elements: Safe Harbor Statement; Welcome and Financial Results; and Q&A. The first is a disclaimer relating to the uncertainty around future projections. The second contains both a welcome and a detailed discussion of the enterprise's financial situation. The welcome portion is a general sales pitch for the company (often given by the CEO/President) and presents an overall narrative explaining the company's strategic position in its market and how the audience might contextualize the subsequent financial results. This is followed by a detailed description of the firms' results: how its current performance relates to past performance and future developments. The

final and longest part of the call allows analysts to request additional information. The enterprise is not required to answer all questions and can call upon analysts in their preferred order.

Earnings calls thus provide a rich testing ground in which to analyze language patterns and information signals. Although the firm (in our case, banks) must report its activities accurately, it has considerable discretion over how it does so: for instance, what language nuances it uses (positive/negative/neutral), which degree of clarity it engages (vague/precise), and its temporal emphasis (backward/contemporaneous/forward oriented). Moreover, the bank must decide how much information to reveal about its own situation relative to its peers. To illustrate, if there is an industry-wide shock, a particular bank may use the call to differentiate itself from the rest of the industry and convince clients of its own soundness. However, if a shock is specific to the bank, it may instead tailor its communication to insulate itself against such interpretations.

4.1.2 Specific features of the 2023 banking crisis

These considerations are interesting in the light of recent bank turmoil. Silicon Valley Bank (SVB) and First Republic Bank (FRCB) both failed in March 2023, constituting two of the largest failures in US banking history. Moreover, both banks had highly idiosyncratic business models. SVB invested in the tech industry (especially startups), offered mortgage loans to high-net-worth individuals, and had its capital base mainly in long-term treasury and mortgage-backed securities. A downturn in the tech industry and the Fed's rate hiking cycle starting in March 2022 (causing it to sell its bond portfolio at a loss to maintain liquidity) prompted a bank run (a run that was intensified given that most of its deposits were uninsured). FRCB failed for many of the same reasons: an unusually high proportion of uninsured deposits from informed wealthy clients and tech startups,

considerable interest-rate risk, and illiquid assets. Silvergate Bank and Signature Bank also failed in this period (with the twist of having large cryptocurrency exposure). Although the ensuing banking stress negatively impacted the wider economy, its effect was surprisingly muted compared to comparable financial crises such as the Savings and Loan turmoil (from the mid-1980s onward). One driver of that fortuitous outcome might be that investors and media commentators were able to separate those local risks from risks in the broader banking industry.

4.2 Banking crises as a signaling game

To develop some basic expectations about this question, we can think of the bank earnings call in the context of a simple signaling game.

4.2.1 Game Setup

Figure 1 provides an illustration of the game. Assume there are two sets of players: companies (banks), denoted by c, and investors, denoted by i. Further, let banks be divided randomly into types $t \in \{0, 1\}$ indicating whether they are vulnerable to some external stress or risk; denote stressed banks as $c_{t=1}$, and banks that are not stressed as $c_{t=0}$. At the beginning of the game, the state of nature (N) determines whether a bank is subject to stress $c_{t=1}$ with probability γ, or escapes stress, $c_{t=0}$ with probability $1 - \gamma$.

At this point, each bank must choose what to discuss during their earnings call. In an effort to satisfy investor expectations they can *reassure* investors, discussing potential risks and explaining how they have been mitigated (denoted as option A). Alternatively, in an effort to shape future expectations, they can *promote* their own agenda, discussing topics of their choosing (option B). In response, investors, i, can decide upon an investment position. They can take a short position, selling their stake in the bank and receiving an

immediate return, $y \geq 0$. Alternatively, they can choose to take a long position, continuing to hold shares and selling them at some future date. If the bank is vulnerable to the stress ($c_{t=1}$), the returns to a long position are $-L \leq 0$. If, however, the bank is not vulnerable ($c_{t=0}$), the returns to a long position are $\alpha y \geq y$.

Payouts to banks depend on the choice of the investor as well as whether or not they chose to *promote* their own agenda (B) or *reassure* their investors (A). If investors chose to take a short position, banks receive a payout of 0. If investors chose to take a long position, banks receive a payout of 1 if the bank chose to promote its own agenda (B), and some lesser amount, $1 \geq 1 - x_t \geq 0$ for $t \in \{0, 1\}$, if they could not promote their chosen agenda (B) and instead chose to reassure investors (A). We can think of x_t as the cost of reassuring investors.

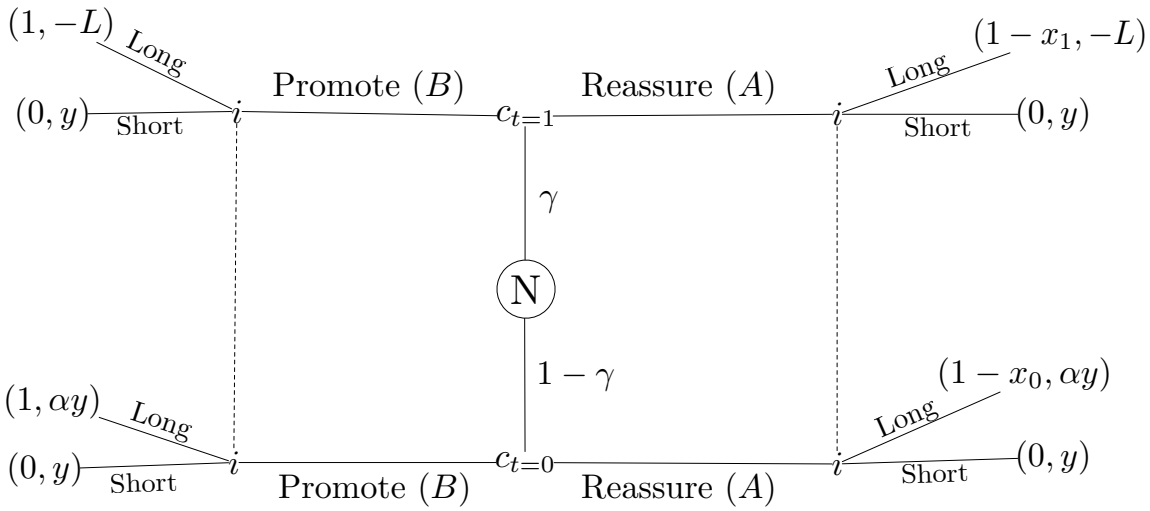
4.2.2 Equilibria

As is common with signaling games of this sort, various equilibria are possible. We focus on equilibria under the two pooling strategies available to the bank:[22] (1) all banks reassure and (2) all banks promote. Under either strategy, perfect Bayesian equilibrium (PBE) revolves around the threshold quantity $\Phi^* = \frac{y-\alpha y}{-L-\alpha y}$, which lies between 0 and 1 when $L > y$. Let $\phi = \Pr(c_{t=1}|A)$, the posterior belief that a bank is under stress given reassurance, and denote the corresponding posterior if the bank promotes its own agenda as $\theta = \Pr(c_{t=1}|B)$. In equilibrium, banks can pool on promoting their own agenda[23] if $\theta < \Phi^*$, which under pooling reduces to the condition that $\gamma < \Phi^*$. Intuitively, this suggests that, as long as the probability of a bank being stressed (γ) and the consequent losses from holding investment

[22] There is a mixing equilibrium wherein investors and banks behave probabilistically with regard to their actions. Strategies that sustain the mixed equilibrium rely on knife-edge conditions that $\gamma = \Phi^*$ and $x_0 = x_1$. While plausible, these circumstances seem to require considerable coincidence, and we will not consider them further here.

[23] This is valid under any set of off-path beliefs, ϕ, and meets the robust criteria of belief refinement.

Figure 1: Bank-Investor Signaling Game. For sufficiently high losses, L, and a sufficiently low probability of realized bank stress γ, banks will pool on reassurance.



Assumptions: $x_1, x_0 \leq 1$

$\alpha \geq 1$

$L, y \geq 0$

$0 \leq \gamma \leq 1$

in a stressed bank, are sufficiently low, then investors will take long positions in response to a bank promoting its own agenda.

The other pooling strategy is one in which all banks choose to reassure. This is a PBE strategy if $\phi < \Phi^*$ and the investor maintains the off-path belief[24] $\theta > \Phi^*$. Intuitively, this equilibrium would occur if investors grew suspicious that a bank that avoids reassuring is likely trying to hide something. These suspicions would arise in response to stress during times of heightened economic uncertainty, exogenous shocks to the banking industry, or other changes that generally shift the likelihood of γ towards Φ^*.

4.2.3 Hypotheses

The framework as discussed above carries empirical implications for the content of earnings calls. In the absence of a crisis (broader, industry-wide shock), we expect banks to pool on promote, discussing their own individual agendas in an effort to shape the expectations of investors. This implies that earnings calls during a period of low stress should be more heterogeneous, forward-looking, and positive in sentiment. Conversely, as the probability of bank stress increases (or the consequences of stress grow more severe), we would expect banks to shift to pool on reassure. This would imply that earnings calls would be more homogeneous in topics discussed, less forward-looking, and less positive in sentiment. The next section will use LLMs to examine bank earnings call data and consider whether these baseline expectations are met.

[24] This set of off-path beliefs survives refinement by the intuitive criterion (Cho and Kreps, 1987) since reassurance (A) does not strictly dominate promotion (B).

5 An LLM study of bank earning calls

5.1 Data

The data we use is a collection of LISCC (Large Institution Supervision Coordinating Committee) banks along with several publicly traded large, regional, and community banking organizations. In total, our sample represents about 100 banks per quarter. The current exercise examines quarterly earnings statements from 2021 to the second quarter of 2023.[25] During this time, there were a number of potential sources of bank stress, including rapid increases in inflation, corresponding rises in the Federal Funds Rate, tightness in the housing market, and various other issues associated with the COVID-19 pandemic. One specific event that we will focus on is the banking stress brought about by the collapse of SVB, FRCB, and Signature Bank in the first quarter of 2023. Because it was unexpected, this event presents a unique scenario where we might expect to see banks rush to reassure investors, i.e., *satisfy* investor expectations.

On average, each earnings call is about 12 thousand words long (around 50 pages). The length of the average earnings call transcript exceeds the capacity of the local LLM models we use in this paper, and it is otherwise useful to separate transcripts into smaller discrete units for analysis. Accordingly, we separate each transcript into smaller parts corresponding to a passage of unbroken text spoken by an individual speaker. Breaking up the transcripts in this way allows us to individually analyze portions of the transcript that are associated to different parts of the meeting, i.e., the investor presentation and the Q&A section. In total, the broken-up transcripts yield a corpus of about 120 thousand individual passages, roughly 12 thousand passages for each quarter.

[25] Transcripts of meetings are available from many different sources including wire services, financial news websites, and individual banks themselves. We manually assembled a corpus of transcripts that covers the sample and added further annotations to delineate the roles of the participants and the portions of the meeting.

5.2 LLM analysis

We use our LLMs to perform four separate analyses of the earnings call data: a *sentiment* analysis, a scoring of passages on *clarity*, a scoring of passages on *temporal orientation*, and *topic* summarization/modeling. Examples of the prompts used here are presented in Appendix A.

To score the passages on sentiment, clarity, and temporal orientation, we prompted the model using the prompts developed for classifying FOMC statements and the Financial Phrase Bank datasets discussed above. Because the corpus of text that we want to analyze is large, we only use some of the models from the benchmark exercises for this scoring. Specifically, we report here the results for the model with the best benchmark performance for each scoring task. For sentiment, we use Vicuna 13B; for clarity, we use Vicuna 7B; and for temporal orientation, we use Guanaco 65B. For each scoring task, we use the few-shot version of the prompt.

multiple models

The way we use LLMs to construct a topic model is more involved. We begin by prompting the LLM to describe a passage in terms of five broader categories. This produces a list of terms ("categories") which are usually between one and three words long each. Because we do not restrict the possible responses of the LLM, the categories identified tend to be similar but still vary between passages. For example, the LLM might label one passage with the category "Real Estate Market" and another as "Housing Market". Semantically, these categories refer to the same thing and should be treated as the same topic. To achieve this, we generate sentence embeddings[26] then extract clusters of embeddings using K-means clustering. We take each cluster of category embedding as identifying a topic, and we name the topic for the category embedding closest to its center. The topic modeling

[26] Embeddings are generated using the model "all-mpnet-base-v2" from the sentence-transformers library, see Reimers and Gurevych (2019).

results discussed below use Vicuna 7B as the LLM for generating categories.[27] At the end of the process, each passage has up to five topics associated with it.

5.3 Results

The result of the LLM analysis is a substantial corpus of labeled passages, which we use to evaluate the implications of the signaling game.

As a point of illustration, Figure 2 lists the top 15 topics by frequency. As we might expect, for bank earnings calls, the most frequently occurring topic is "deposits," which includes passages labeled by the LLM as "deposit balance," "deposit base," "deposit flows," and so on. The topic of deposits is associated with roughly 5% of all passages. The broad topic of "growth" is associated with about 2.5% of all passages. Further disaggregation of the data reveals information about the distribution of topics over time. The color of each bar reflects the positive-neutral-negative sentiment attached to the overall discussion of each topic. Lighter colors indicate more positive sentiment.

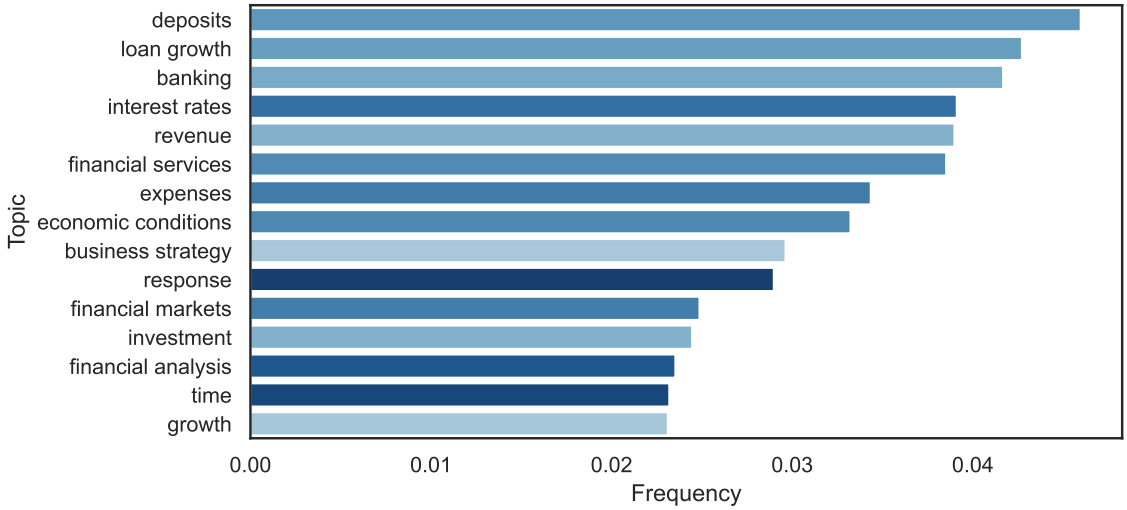
Figure 3 shows the mean similarity of topics discussed over time.[28] To calculate this measure, we begin by aggregating topics discussed by a company at a given quarter's earnings call into a 150-element vector where each position in the vector reflects the relative frequency of that topic during that earnings call. We then use this vector to calculate cosine similarity with all other companies for that same quarter and take the mean value of this

how did they generate topics?

[27] We generate four sets of topics using four different LLM models: Vicuna 7B, Vicuna 13B, Fin-LLaMA, and Wizard-Vicuna. This is a computationally costly operation. After the analysis was complete, we discovered that the four models generated topic distributions that were fairly similar to one another. To maintain focus in the discussion of the model results, we will focus on the results from the Vicuna 7B model.

[28] Shaded regions represent 95% Confidence intervals. In this figure and all remaining figures, confidence intervals reflect only the sampling error at the *bank level*, taking the LLM response as given. That is, they reflect *data* uncertainty and do not reflect *model* uncertainty (see Chatfield, 1995) in the underlying LLM. These two sources of uncertainty are separable because the LLM was pre-trained on separate data. Future development of this project will incorporate (approximate) model uncertainty into confidence interval estimates.

Figure 2: An example of the top topics identified by the model



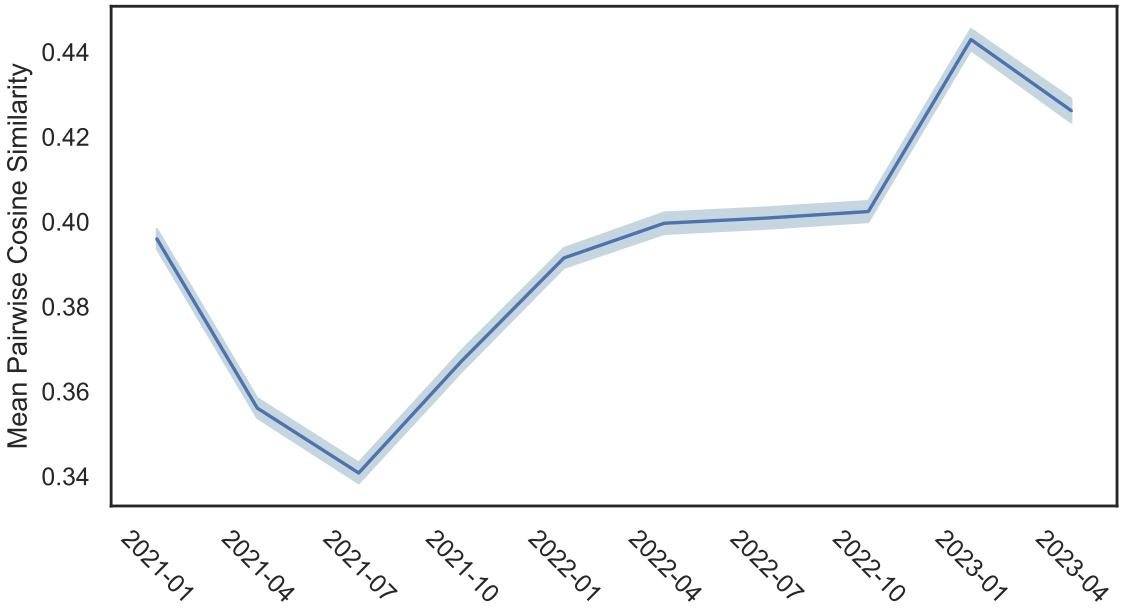
similarity score to indicate the extent to which companies tended to discuss similar topics at a given time.

Figure 3 portrays trends that are broadly consistent with the model expectations outlined above. To see this, first consider the SVB collapse in 2023q1 and the subsequent period of banking stress. Under these circumstances, we might expect from the signaling game that banks reassure investors about deposit risk and duration risk (which were drivers of the collapse of SVB), and we could reasonably expect that investors would be suspicious of any bank that did not provide reassurance. Accordingly we might expect the topics discussed in the earnings calls to become more similar and tend towards topics of deposits and associated risks. Indeed, this is precisely what we see in Figure 3, where the similarity in topics discussed by banks jumps to its peak in the first quarter of 2023.

We also note from Figure 3 that the similarity in topics discussed descends to a nadir in q2 and q3 of 2021. This period is notable as it corresponds to a brief lull between the

uncertainty of the pandemic[29] and stress associated with the acceleration in inflation.[30] Viewing this as a period of low stress, our expectations from the signaling game setup are that banks should pool on promoting their own preferred agenda and, consequently, should be less similar to one another in terms of the topics discussed.

Figure 3: Similarity in topics discussed



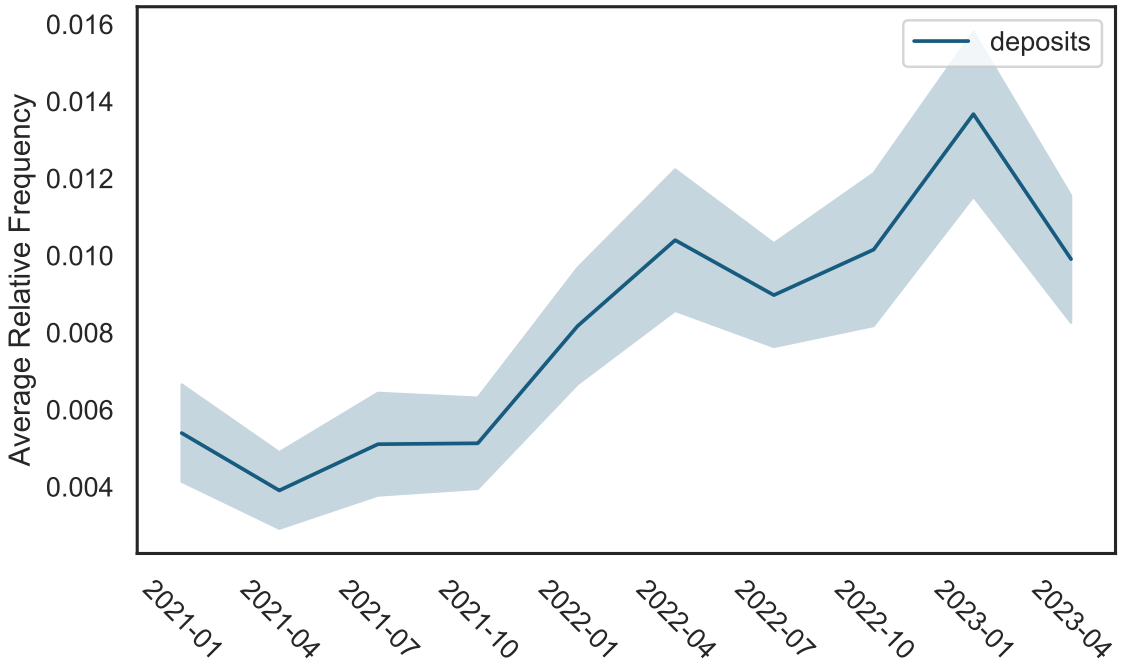
Other trends in the data produced in this analysis further conform to our expectations about pooling during these two episodes (the low-stress episode in q2-q3 2021 and the high-stress episode in 2023q1). Figure 4 shows the average relative frequency with which a bank discusses deposits during its earnings calls. We calculate the relative frequency of a topic as the proportion of remarks in a bank's earnings call that are labelled with that topic.

[29] Perhaps most notably, the COVID-19 vaccine became widely available in the first quarter, and the government made a strong effort to encourage vaccination, administering 200 million doses of the vaccine by the end of April (Biden, Joseph, "Remarks by President Biden on the COVID-19 Response and the State of Vaccinations." 21, April 2021).

[30] While inflation during this period was elevated, it had not yet reached the historic highs seen at the end of 2021. Furthermore, there was a fairly widespread expectation during this period that inflation would be transitory. During this period, the Federal Funds Rate remained low as did mortgage rates and loan delinquencies.

The average relative frequency of a topic is the mean value of this quantity across all banks. Deposits are discussed less frequently during the low-stress episode; on average, they are the topic of less than one percent of a bank's remarks. More strikingly, the frequency of the discussion of deposits jumps by over 200%, to a peak, in response to the high-stress episode. Other topics, included in Appendix C, produce similarly informative patterns.

Figure 4: Average relative frequency with which a bank discusses deposits.



Turning to the LLM scoring of the earnings calls on sentiment, clarity, and temporal orientation, we see trends that are similar to those that emerge from the topic analysis. These trends fit our expectations from the signaling game and also point towards distinct high-stress and low-stress episodes.

Figure 5 shows the mean meeting-level[31] scores for clarity, sentiment, and temporal orientation. Clarity is mostly unchanged. Temporal orientation declines slightly from

[31]The meeting-level score is calculated as the mean score across all remarks made in a bank's meeting.

forward-looking to present-looking. Overall, sentiment declines most substantially from its high during the low-stress episode to its lowest during the high-stress episode.

The presentation portion of the earnings call occurs before questions, and we can interpret it as revealing what the bank would prefer to discuss before responding to additional investor concerns. In terms of sentiment, we would expect banks promoting their own agenda to convey a more positive sentiment, while reassurance, in part because of its focus on topics like risks, headwinds, etc., would likely be associated with a more neutral or negative sentiment. The red line in Figure 6 shows high sentiment during the low stress episode and a more neutral sentiment during the high stress episode. The difference between the sentiment of the presentation and the question portion is charted in blue. The difference is highest during the low-stress episode, suggesting a focus on promoting preferred agendas during the investor presentation.

Lastly, we can associate sentiment with individual topics to enhance our understanding of how topics were discussed. In Figure 7, we focus on two topics that were central to the high-stress episode: "deposits" and "interest rates". The figure portrays the mean meeting-level sentiment associated with the bank's discussion of a topic,[32] with higher values indicating more positive sentiment. Banks espoused a more positive sentiment towards these topics (reflecting optimism) during the low-stress episode. For both topics, sentiment declined after the low-stress episode, reaching its minimum during the high-stress episode.

[32]For this figure, we exclude passages attributed to external participants.

Figure 5: Mean meeting-level scores for clarity, sentiment, and temporal orientation. Meeting-level scores are the aggregated (mean) score across all passages attributed to bank managers (and excluding passages attributed to external participants).

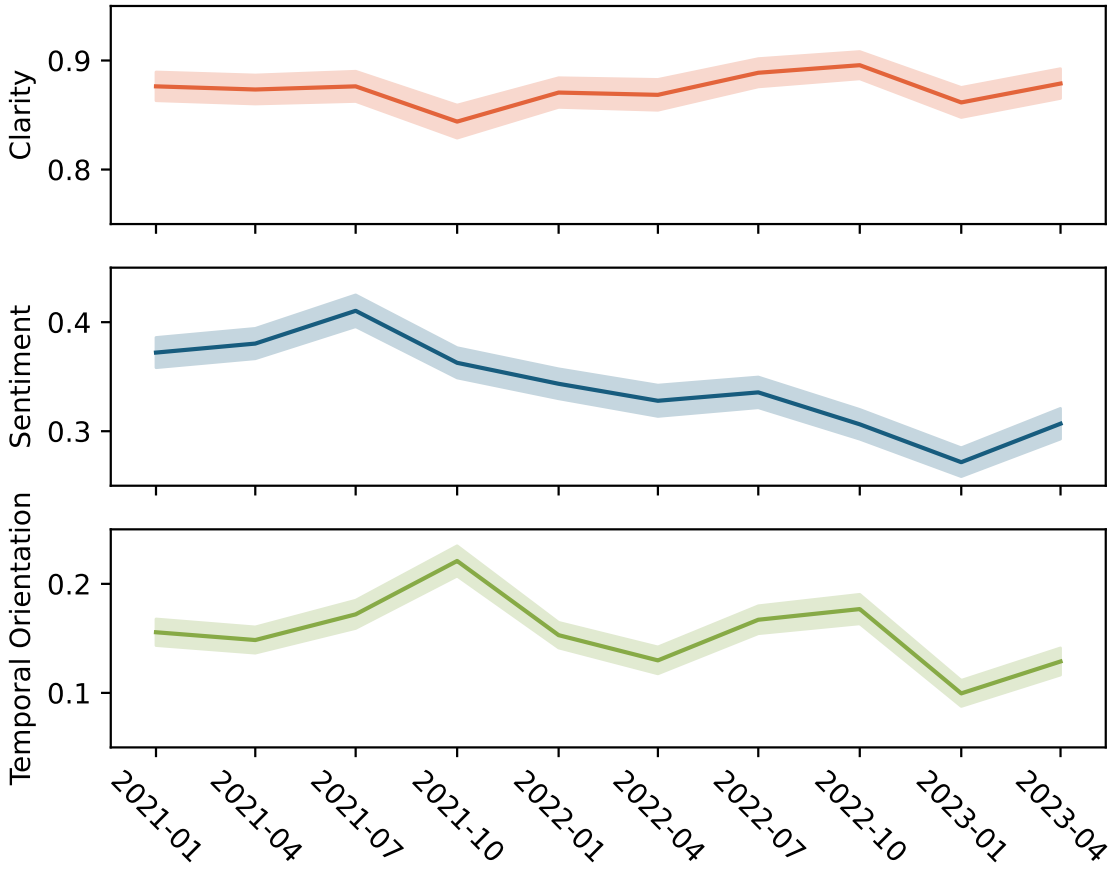


Figure 6: Sentiment of presentation only, and difference with Q&A sentiment.

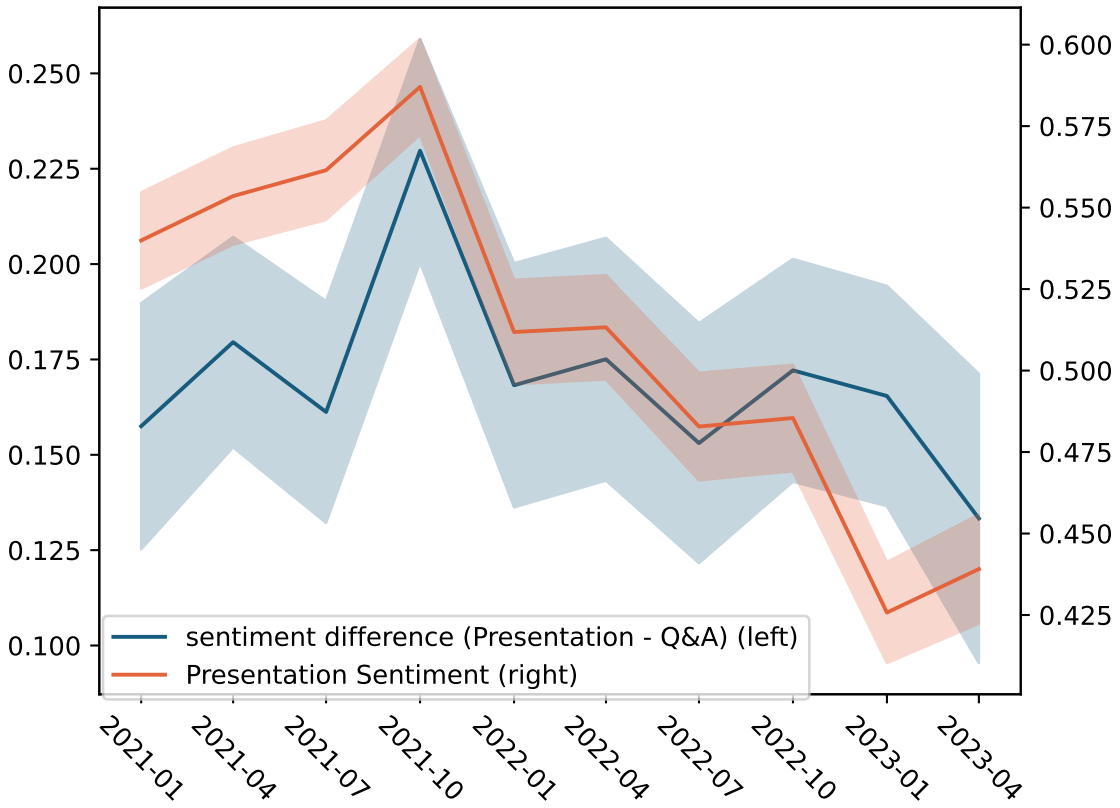
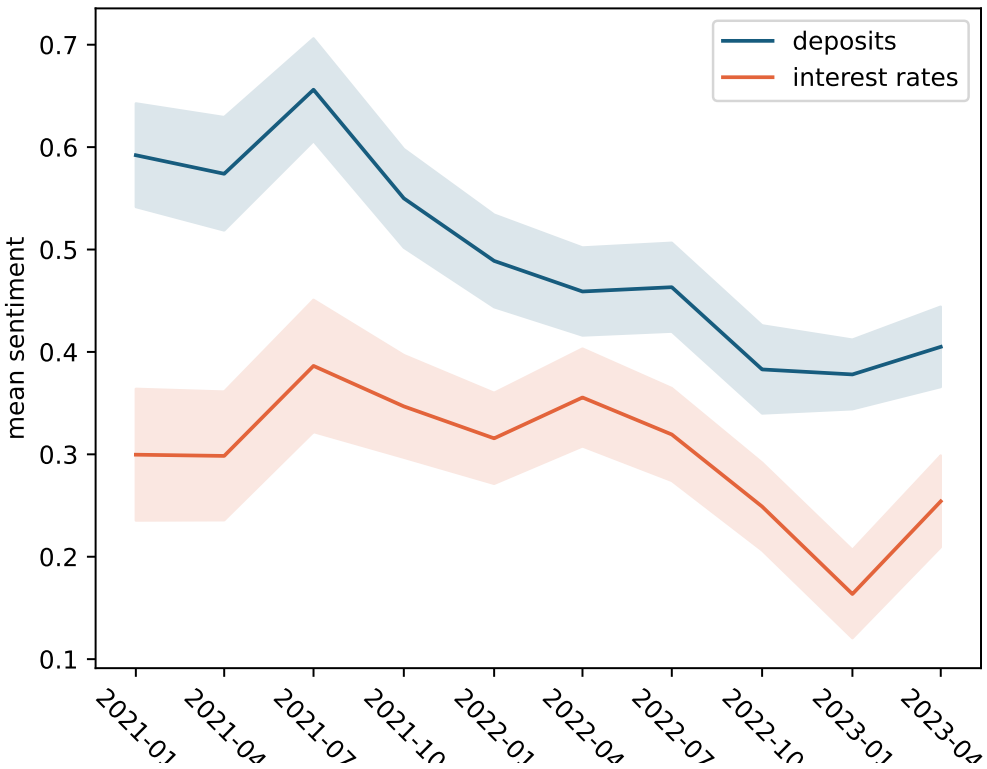


Figure 7: Mean meeting-level bank sentiment associated with deposits and interest rates.



6 Concluding remarks

Why should we care about local LLMs? For organizations and developers who prioritize data autonomy and security, local, open-source LLMs present a compelling option. Though cloud-based or closed systems might offer considerable convenience, they bring forth data privacy concerns, especially when handling confidential data. Utilizing local, open-source models mitigates such worries by ensuring exclusive control over data: keeping it in-house and free from external threats. The transparency of open models ensures trust, aligning with many organizations' goal to bypass the risks of unauthorized data access and breaches.

As data privacy and proprietary concerns intensify, locally deployed LLMs offer a compelling alternative to traditional cloud-based models. By ensuring that data remains in-house, local LLMs align closely with modern organizational needs, laying the groundwork for a more secure digital future. Organizations can process confidential or sensitive information without the need to transmit data externally, substantially diminishing exposure risks. This is especially pertinent in domains where user privacy and data protection are non-negotiable, for example in the financial and medical fields.

At the same time, while local deployment guarantees data privacy, reducing exposure risks, it demands substantial computational resources and expertise. While the merits are evident, it is essential to broach the challenges of local LLM deployment.

As an application of the LLM framework, we examined how the models interpreted financial texts, namely banks' quarterly earnings calls against the backdrop of the banking crisis in 2023. We used our models to classify the data across a number of labels: sentiment, clarity, temporality, and topic structure. To motivate our analysis, we drew upon game-theoretic models of information signaling. Marrying the two, we found that in calm times, earnings calls across different banks are more heterogeneous, forward-looking, and positive in sentiment. As the probability of bank stress increases, however, banks pool

on reassurance, meaning that calls become more homogeneous in topics discussed, less forward-looking, and less positive in sentiment.

Regarding future work, the models and framework developed here could be used in a wide variety of applications in economics and finance to analyze texts, speeches, and news media in both historical and real time. They could allow analysts and researchers faced with heterogeneous enterprises exposed to different economic events to rapidly gain disciplined insights into how information is structured and presented.

References

Acharya, V. V., S. G. Cecchetti, and K. L. Schoenholtz (2023). Overview of recent banking stress. In *SVB and Beyond: The Banking Stress of 2023*, pp. 14–32. London, UK: CEPR Press.

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

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society 158*(3), 419–444.

Chiang, W.-L., Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing (2023, March). Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cho, I.-K. and D. M. Kreps (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics 102*(2), 179–221.

Correa, R., K. Garud, J. M. Londono, and N. Mislang (2021). Sentiment in Central Banks' Financial Stability Reports. *Review of Finance 25*(1), 85–120.

De Amicis, C., S. Falconieri, and M. Tastan (2021). Sentiment analysis and gender differences in earnings conference calls. *Journal of Corporate Finance 71*(C).

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Dettmers, T. and L. Zettlemoyer (2023). The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pp. 7750–7774. PMLR.

Frantar, E., S. Ashkboos, T. Hoefler, and D. Alistarh (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Hansen, A. L. and S. Kazinnik (2023). Can ChatGPT Decipher Fedspeak? *SSRN Working Paper*.

Hartford, E. (2023, May). Uncensored models. Available at erichartford.com/uncensored-models.

Huang, A. H., H. Wang, and Y. Yang (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research 40*(2), 806–841.

Jha, M., J. Qian, M. Weber, and B. Yang (2023). ChatGPT and Corporate Policies. *Chicago Booth Research Paper* (23-15).

Lee, J. (2023, May). Wizardvicunalm.

Lopez-Lira, A. and Y. Tang (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *SSRN Working Paper*.

Malo, P., A. Sinha, P. Korhonen, J. Wallenius, and P. Takala (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology 65*(4), 782–796.

Peng, B., E. Chersoni, Y.-Y. Hsu, and C.-R. Huang (2021). Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks. *Proceedings of the Third Workshop on Economics and Natural Language Processing*, 37–44.

Reimers, N. and I. Gurevych (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rogers, A. (2023). Closed AI models make bad baselines. *Towards Data Science*. Available at hackingsemantics.xyz/2023/closed-baselines.

Roozen, D. K. and F. Lelli (2021). Stock values and earnings call transcripts: a dataset suitable for sentiment analysis. Papers, arXiv.org.

Shang, L., H. Xi, J. Hua, H. Tang, and J. Zhou (2023). A lexicon enhanced collaborative network for targeted financial sentiment analysis. *Information Processing & Management 60*(2), 103187.

Suslava, K. (2021). Stiff Business Headwinds and Uncharted Economic Waters: The Use of Euphemisms in Earnings Conference Calls. *Management Science 67*(11), 7184–7213.

Todt, W., R. Babaei, and P. Babaei (2023). Fin-llama: Efficient finetuning of quantized llms for finance. https://github.com/Bavest/fin-llama.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample (2023). Llama: Open and efficient foundation language models.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, F. Xia, Q. Le, and D. Zhou (2022). Chain of thought prompting elicits reasoning in large language models. *ArXiv abs/2201.11903*.

Xu, C., Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang (2023). Wizardlm: Empowering large language models to follow complex instructions.

A Prompt Examples

A.1 Initial Prompt

An example of the initial prompt used for the FOMC exercise and the Financial Phrase Bank Exercise is provided below. Programatically inserted portions and model responses are enclosed in brackets.

Algorithm 1 Initial Prompt

SYSTEM: You are an economist. Read the following passage by a member of the FOMC {PASSAGE}
USER: Classify the passage into one of the 5 categories (hawkish, mostly hawkish, neutral, mostly dovish, dovish):
ASSISTANT: Rating: {CONSTRAINED LLM RESPONSE 1}
ASSISTANT: Explanation: {LLM RESPONSE 2}

The LLM response after "Rating:" is constrained such that only tokens corresponding to the given scale are considered. Of those tokens, we take the tokens with the highest likelihood as the model's rating. The model response after "Explanation:" is not constrained in this way, though it is constrained to a total length of 500 tokens.

A.2 COT Prompt

An example of the (V2) chain of thought prompt used for the FOMC exercise is provided below. Programatically inserted portions and model responses are enclosed in brackets.

As in the initial prompt, the response to "Rating:" is constrained so that the LLM response falls within the desired scale.

A.3 Few-shot Prompt

An example of the few-shot prompt used for the Financial Phrase Bank exercise is provided below. Programatically inserted portions and model responses are enclosed in brackets.

Algorithm 2 COT Prompt

SYSTEM: You are an economist. A hawkish stance on monetary policy it favors higher interest rates to keep inflation in check. A dovish stance on monetary policy supports low interest rates to stimulate investment and spending and to maintain low unemployment.
USER: Read this passage from a speech by an FOMC member:
{PASSAGE}
Discuss the monetary policy stance that seems to be endorsed in the passage. Does it seem to be hawkish or dovish?
ASSISTANT: {LLM RESPONSE 1}
USER: Based on your response, rate it as hawkish, slightly hawkish, neutral, slightly dovish, or dovish
ASSISTANT: Rating: {CONSTRAINED LLM RESPONSE 2}

Algorithm 3 Few-Shot Prompt

USER: Rate this passage from a financial newspaper and score it's sentiment as positive, negative or neutral.
{EXAMPLE PASSAGE 1}
ASSISTANT: {EXAMPLE RATING 1}
USER: Rate this passage from a financial newspaper and score it's sentiment as positive, negative or neutral.
{EXAMPLE PASSAGE 2}
ASSISTANT: {EXAMPLE RATING 2}
USER: Rate this passage from a financial newspaper and score it's sentiment as positive, negative or neutral.
{PASSAGE}
ASSISTANT: {CONSTRAINED LLM RESPONSE 1}

In practice, a total of six examples are shown to the model: two with positive ratings, two with negative ratings, and two with neutral ratings. As in the other prompts, the response to "Rating:" is constrained so that the LLM response falls within the desired scale.

We found that when using the few-shot prompt, including a system message did not substantially improve performance. Accordingly, we omit it from the prompt used to produce the results discussed in this paper.

A.4 Topic Model Prompts

An example of the topic model prompt used is shown below

Algorithm 4 Topic Model Prompt

USER: Read the following passage and describe it in terms of a broader set of categories:
{PASSAGE}
ASSISTANT:
Topics:
1. {LLM RESPONSE 1}
2. {LLM RESPONSE 2}
3. {LLM RESPONSE 3}
4. {LLM RESPONSE 4}
5. {LLM RESPONSE 5}

As with the few-shot example, we did not find that including a system message substantially improved performance, and we choose instead to embed the entirety of the model instruction in the user message. To encourage the model to provide concisely named categories, we restrict all LLM responses to fifteen tokens. LLM responses in our exercise are generally between one and four words long.

B Additional Results

B.1 Few-shot accuracy of FOMC classification

Few-shot prompting of LLMs for the FOMC classification exercise provided varying results depending on the model and prompt used. Several prompts were tried with varying examples and slight variations on the prompt template in Appendix A.3. Table B.1 reports the three and five class accuracy for each model and prompt. Overall, the best three-class accuracy was achieved by Wizard-Vicuna (0.831) while the best five-class accuracy was achieved by Guanaco 65B. Vicuna 13 provided the most consistent three class performance.

Table B.1: Multi-class accuracy scores for few-shot prompts (FOMC exercise)

	Fin-llama		Guanaco 65		Guanaco 33		Vicuna 13		Vicuna 7		Wizard Vicuna	
prompt	5-class	3-class	5-class	3-class	5-class	3-class	5-class	3-class	5-class	3-class	5-class	3-class
v03	0.194	0.244	**0.612**	**0.804***	0.184	0.242	0.401	0.706	0.309	0.559	0.497	0.814
v04	0.284	0.376	0.562	0.733	0.334	0.505	0.395	0.729	**0.468**	**0.785***	**0.526**	0.729
v05	0.284	0.359	0.537	0.699	0.363	0.572	0.395	0.691	0.438	0.745*	0.516	0.743
v06	**0.290**	0.365	0.610	0.770	**0.399**	**0.603**	**0.403**	0.716	0.447	0.749	0.524	**0.831***
v10	0.169	0.217	0.255	0.401	0.157	0.253	0.324	0.708*	0.213	0.273	0.190	0.305
v11	0.219	0.355	0.217	0.422	0.273	0.472	0.296	0.754*	0.332	0.672	0.278	0.528
v12	0.192	0.313	0.188	0.355	0.286	0.537	0.290	0.758*	0.328	0.681	0.280	0.537
v13	0.255	**0.409**	0.209	0.403	0.261	0.409	0.290	**0.810***	0.340	0.714	0.265	0.572

Notes: **Bold-faced** numerals indicate the best performing prompt for a model (best score in a column).

* Indicates model with the best 3-category accuracy for a given prompt (best score in a row)

B.2 Classifying Phrase Bank Sentences with 100% Agreement

Table B.2: Model performance with few-shot prompting for classifying positive–neutral–negative sentiment of Phrase Bank "all-agree" sentences

	Wizard-Vicuna	Guanaco (33B)	Guanaco (65B)	Fin-LLaMA	Vicuna (7B)	Vicuna (13B)
Accuracy	0.844	0.840	0.780	0.840	0.844	**0.902**
Precision:						
Positive	0.705	**0.984**	0.529	**0.859**	0.848	0.829
Neutral	0.895	0.807	**0.995**	0.825	0.833	0.923
Negative	**0.951**	0.946	0.919	**0.968**	**0.944**	0.941
Recall:						
Positive	0.880	0.538	**1.000**	0.675	0.667	0.872
Neutral	0.862	0.991	0.665	**0.957**	0.954	0.926
Negative	0.672	0.603	**0.983**	0.517	0.586	0.828
Specificity:						
Positive	0.880	0.538	**1.000**	0.675	0.667	0.872
Neutral	0.862	0.991	0.665	**0.957**	0.954	0.926
Negative	0.672	0.603	**0.983**	0.517	0.586	0.828
F_1 score:						
Positive	0.783	0.696	0.692	0.756	0.746	**0.850**
Neutral	0.878	0.890	0.797	0.886	0.890	**0.925**
Negative	0.788	0.737	**0.950**	0.674	0.723	0.881
Balanced accuracy:						
Positive	0.881	0.768	0.862	0.819	0.814	**0.908**
Neutral	0.836	0.775	0.829	0.790	0.799	**0.892**
Negative	0.834	0.799	**0.984**	0.757	0.791	0.910

Notes: The table reports classification performance for the subset of Phrase Bank sentences for which all annotators agree (a total of 2,259 sentences). The considered LLMs are implemented with few-shot prompting.

Table B.3: Model performance for classifying clarity of Phrase Bank "all agree" sentences

	Wizard-Vicuna	Guanaco (33B)	Guanaco (65B)	Fin-LLaMA	Vicuna (7B)	Vicuna (13B)
Accuracy	0.941	0.816	0.784	0.906	**0.984**	0.934
Precision:						
Clear	**0.997**	**0.997**	**0.997**	0.995	0.995	**0.997**
Vague	0.143	0.049	0.042	0.073	**0.375**	0.129
Recall:						
Clear	0.943	0.817	0.783	0.910	**0.988**	0.936
Vague	**0.800**	**0.800**	**0.800**	0.600	0.600	**0.800**
F_1 score:						
Clear	**0.969**	0.898	0.877	0.950	0.992	0.966
Vague	0.242	0.093	0.080	0.130	**0.462**	0.222

Notes: The table reports classification performance for the subset of the 1000 sentences for which at all annotators agree (a total of 447 sentences).

Table B.4: Model performance for classifying temporality of Phrase Bank "all agree" sentences

	Wizard-Vicuna	Guanaco (33B)	Guanaco (65B)	Fin-LLaMA	Vicuna (7B)	Vicuna (13B)
Accuracy	0.771	0.743	**0.801**	0.776	**0.618**	0.690
Precision:						
Positive	**0.797**	**0.692**	**0.890**	0.861	0.852	**0.925**
Neutral	0.702	0.745	0.772	**0.851**	**0.538**	0.734
Negative	**0.845**	0.779	0.794	0.700	0.715	0.631
Recall:						
Positive	**0.606**	**0.617**	**0.509**	0.531	0.131	**0.280**
Neutral		0.919	0.980	0.838	**0.975**	0.741
Negative	**0.729**	**0.692**	**0.867**	0.904	0.679	**0.946**
Specificity:						
Positive	0.931	0.878	0.973	0.962	**0.989**	**0.989**
Neutral	**0.776**	0.815	0.839	**0.914**	0.530	0.839
Negative	**0.903**	0.860	0.839	0.735	**0.768**	0.595
F_1 score:						
Positive	**0.688**	0.653	0.647	0.657	0.228	0.430
Neutral	**0.814**	0.823	**0.864**	0.844	0.693	0.737
Negative	0.783	0.733	**0.829**	0.789	0.697	0.757
Balanced accuracy:						
Positive	**0.769**	0.748	0.741	0.747	0.560	0.635
Neutral	0.873	0.867	**0.909**	0.876	0.752	0.790
Negative	0.816	0.776	**0.853**	0.820	0.724	0.770

Notes: The table reports classification performance for the subset of the 1000 sentences for which all annotators agree (a total of 615 sentences).

B.3 Classifying Phrase Bank Sentences: Alternative Prompts

Table B.5: Model performance with zero-shot prompts for classifying positive–neutral–negative sentiment of Phrase Bank "at least 75% agree" sentences

	Wizard-Vicuna	Guanaco (33B)	Guanaco (65B)	Fin-LLaMA	Vicuna (7B)	Vicuna (13B)
Accuracy	0.670	0.712	0.708	**0.796**	0.708	0.744
Precision:						
Accuracy	0.672	0.712	0.710	**0.798**	0.708	**0.744**
Precision:						
Positive	0.450	**0.822**	0.488	**0.674**	0.498	0.535
Neutral	0.861	0.696	**0.931**	0.843	**0.945**	0.864
Negative	**0.855**	**1.000**	0.920	**0.830**	**0.731**	0.882
Recall:						
Positive	0.805	0.289	**0.945**	0.680	0.906	0.773
Neutral	0.583	**0.975**	0.589	**0.840**	0.592	0.715
Negative	0.887	0.151	**0.868**	0.830	**0.925**	0.849
Specificity:						
Positive	0.805	0.289	**0.945**	0.680	**0.906**	0.773
Neutral	0.583	**0.975**	0.589	**0.840**	0.592	0.715
Negative	0.887	0.151	**0.868**	0.830	**0.925**	0.849
F_1 score:						
Positive	0.577	0.428	0.644	**0.677**	0.643	**0.633**
Neutral	0.695	0.812	0.722	**0.841**	0.728	**0.782**
Negative	0.870	0.262	**0.893**	0.830	0.817	0.865
Balanced accuracy:						
Positive	0.727	0.632	**0.797**	0.781	0.788	**0.767**
Neutral	0.708	0.612	0.756	**0.782**	0.765	**0.757**
Negative	**0.930**	0.575	**0.928**	0.903	0.934	0.916

Notes: The table reports classification performance for the subset of Phrase Bank sentences for which at least 75% of the annotators agree (a total of 3,448 sentences). The considered LLMs are implemented with zero-shot prompts.

Table B.6: Model performance with COT prompting for classifying positive–neutral–negative sentiment of Phrase Bank "at least 75% agree" sentences

Accuracy	0.644	0.678	0.650	**0.638**	0.620	**0.720**
Precision:						
Positive	**1.000**	**0.750**	0.500	**0.000**	0.448	0.738
Neutral	0.642	0.670	**0.649**	0.638	**0.663**	**0.711**
Negative	**1.000**	**1.000**	0.688	**0.000**	**0.000**	0.929
Recall:						
Positive	0.016	0.164	**0.008**	0.000	0.336	**0.352**
Neutral	**1.000**	**0.978**	0.981	**1.000**	0.837	0.947
Negative	0.019	0.113	**0.208**	0.000	**0.000**	**0.245**
Specificity:						
Positive	0.016	0.164	**0.008**	0.000	**0.336**	**0.352**
Neutral	**1.000**	**0.978**	0.981	**1.000**	0.837	0.947
Negative	0.019	0.113	**0.208**	0.000	**0.000**	**0.245**
F_1 score:						
Positive	0.031	0.269	0.015	**0.000**	0.384	**0.476**
Neutral	0.782	0.795	0.782	**0.779**	0.740	**0.812**
Negative	0.037	0.203	**0.319**	0.000	0.000	**0.388**
Balanced accuracy:						
Positive	0.508	0.571	**0.502**	0.500	0.585	**0.652**
Neutral	0.508	0.564	0.524	**0.500**	0.539	**0.634**
Negative	**0.509**	0.557	**0.596**	0.500	0.498	**0.621**

Notes: The table reports classification performance for the subset of Phrase Bank sentences for which at least 75% of the annotators agree (a total of 3,448 sentences). The considered LLMs are implemented with COT prompting.

C Earnings Calls: Additional Materials

Additional tables and figures related to the earnings call analysis are presented below.

Figure C.1 shows the mean meeting-level frequency of the discussion of interest rates. The figure shows The discussion of interest rates jumped immediately after the end of the low-stress episode, in Q4 2021 and remained near that level as the FOMC raised the federal funds rate through 2023. Figure C.2 portrays the frequency of discussion of risk management, a topic that suggests reassurance. Discussion of this topic increased after the low-stress episode (Q2-Q3 2021) and jumped even higher with the onset of the high-stress period (Q1 2023).

Figure C.1: Mean meeting-level frequency of discussion of interest rates

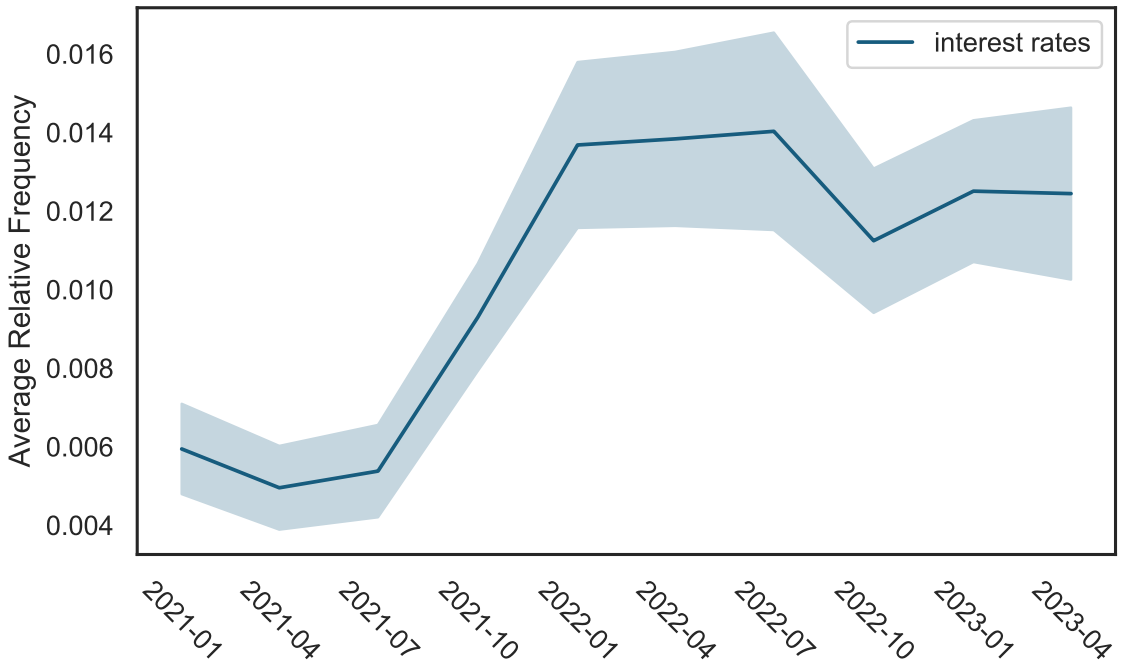


Figure C.2: Mean meeting-level discussion of risk management.