The Ghost in the Machine: Generating Beliefs with Large Language Models *

J. Leland Bybee

University of Chicago Booth School of Business

First Draft: February, 2023
This Draft: February, 2025

Abstract

I introduce a methodology to generate economic expectations by applying large language models to historical news. Leveraging this methodology, I make three key contributions. (1) I show generated expectations closely match existing survey measures and capture many of the same deviations from full-information rational expectations. (2) I use my method to generate 120 years of economic expectations from which I construct a measure of economic sentiment capturing systematic errors in generated expectations. (3) I then employ this measure to investigate behavioral theories of bubbles. Using a sample of industry-level run-ups over the past 100 years, I find that an industry's exposure to economic sentiment is associated with a higher probability of a crash and lower future returns. Additionally, I find a higher degree of feedback between returns and sentiment during run-ups that crash, consistent with return extrapolation as a key mechanism behind bubbles.

*leland.bybee@chicagobooth.edu. Previously titled "Surveying Generative AI's Economic Expectations." I'm thankful for my amazing dissertation committee members, Nick Barberis (co-chair), Will Goetzmann, Paul Goldsmith-Pinkham, Bryan Kelly (co-chair), and Alp Simsek. I benefited from discussions with and comments from Anna Cieslak, Anastassia Fedyk (discussant), Paul Fontanier, Stefano Giglio, Theis Jensen, Sophia Kazinnik (discussant), Pengcheng Liu, Alejandro Lopez Lira (discussant), Tianshu Lyu, Song Ma, Toby Moskowitz, Andrei Shleifer, Kelly Shue, Yinan Su, Kaushik Vasudevan, Hongyu Wu, Alex Zentefis, Daojing Zhai, Kangying Zhou, as well as participants at the AI in Finance Conference, the BlackRock Applied Research Panel, the NBER Big Data and AI Conference, Arizona State University, Bloomberg, Chicago Booth, Carnegie Mellon University, CEBRA Annual Meeting, Cornell, CQA Conference, the European Finance Association Conference, Georgetown, Harvard Business School, University of Houston, University of Illinois Urbana-Champaign, Kellogg, London Business School, London School of Economics, University of Maryland, the Insightful Minds in International Macro Seminar, the Monash-Warwick-Zurich Text-as-Data Workshop, Ohio State University, Purdue, Q-Group Conference, University of Southern California, SQA Conference, University of Texas Austin, the Advances with Field Experiments Conference, the Olin Finance Conference at Washington University (PhD Poster Session), Washington University, the Western Finance Association Conference, and the Yale SOM brown bag. I'd like to thank Jonathan Fan for his excellent research assistance. I'm grateful for the Yale SOM International Center for Finance for their financial support.

1 Introduction

Rational expectations remains the dominant model of beliefs in much of macroeconomics and finance. Its dominance is not hard to understand: rational expectations ties beliefs to realized outcomes, making it possible to build economic models without directly observing beliefs. However, for as long as there has been rational expectations, there has been evidence and alternative theories to question its dominance. Such theories place us in the "wilderness" of alternative expectations of Sims (1980).[1] In recent years, the use of surveys to tie beliefs to observable data has emerged as a prominent approach to navigate this wilderness.[2]

In this paper, I propose a new method for navigation: generating beliefs using large language models (LLMs). LLMs are a class of statistical models designed to generate human-like text. These models accomplish this goal by predicting the next term or "token" in a phrase given all previous tokens using a particular neural network architecture known as transformers. By leveraging massive corpuses of training data and a large parameter space, these models have exhibited an emergent ability to mimic human-like behavior (Brown et al. (2020), Wei et al. (2022), Bubeck et al. (2023)).[3]

I utilize this approach to form expectations with an LLM by providing a historical sample of news articles from *The Wall Street Journal* (*WSJ*) to OpenAI's GPT-3.5 instance and prompting it to forecast various financial and macroeconomic quantities based on each article. I then aggregate these article-level expectations to whatever frequency is desired to form a time-series of beliefs comparable to existing survey measures. Given a textual representation of the state of the world at a point in time, LLMs can approximate human expectations in that state of the world. These generated expectations can then be used to evaluate behavioral theories of economic behavior.

While the potential for LLMs to generate beliefs opens many new possibilities for the field of economics as a whole, nowhere is the benefit of this approach clearer than for asset pricing. Expectations are *the* central object in asset pricing. However, expectations are difficult to measure directly. Some attempts have been made to extract beliefs *from* prices but these price-based expectations have little to say about behavioral models due to the joint hypothesis problem.[4]

[1] Sims himself refers to the "wilderness of disequilibrium economics" but this wilderness has since been rephrased as Sargent (2001)'s "wilderness of bounded rationality" and Angeletos et al. (2021)'s "wilderness of alternative models for expectations formation and equilibrium".

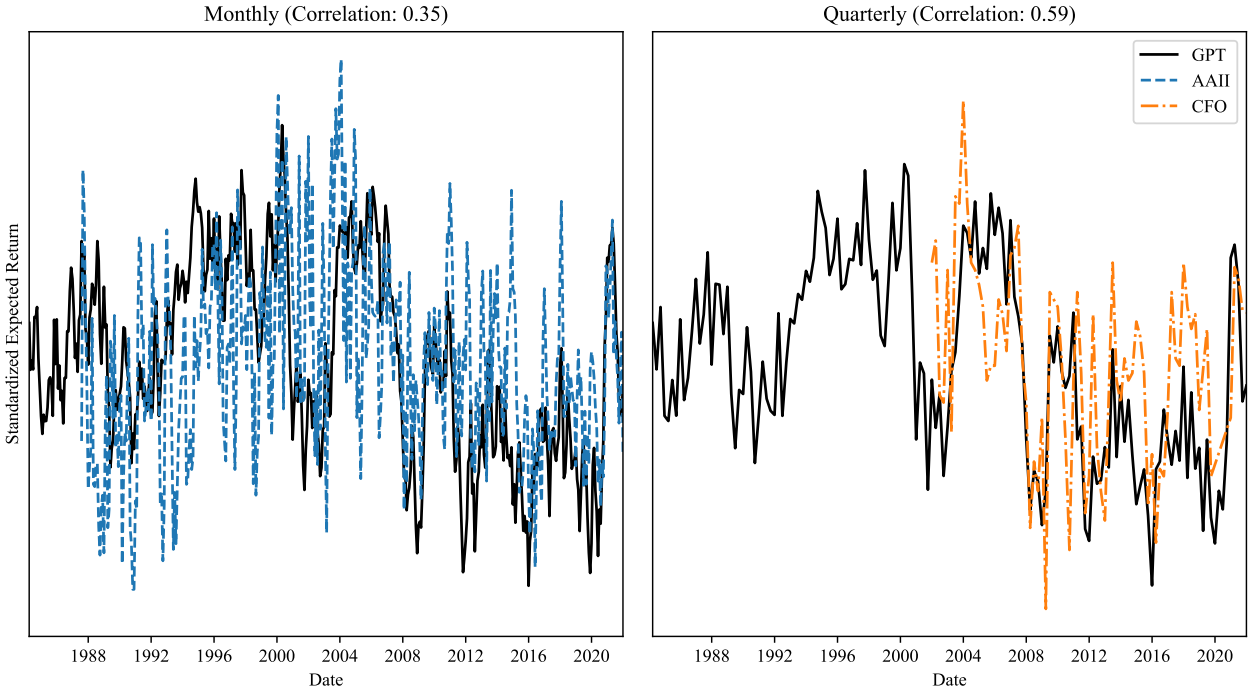
[2] Some recent examples are Greenwood and Shleifer (2014), Coibion and Gorodnichenko (2015), Bordalo et al. (2020a), Angeletos et al. (2021), Nagel and Xu (2022a), and Lochstoer and Muir (2022).

[3] Emergent refers to the scale of these models. Open AI's GPT-3 instance used roughly 45 terabytes of text for training data and has roughly 175 billion parameters.

[4] Cochrane (2017) makes exactly this point in regards to behavioral asset pricing models: "[O]ne would have thought that arguments over rational versus irrational pricing, using only price and payoff data, would have ended the minute Fama (1970) and its joint hypothesis theorem were published...The solution, of course, is to tie either probabilities or marginal utility to observable data, in some rejectable way."

Turning to my results: to evaluate the properties of my method, I start by comparing the resulting expectations against existing survey measures. First, I evaluate how well generated expectations of the stock market match the expectations of the American Association of Individual Investors (AAII) survey and Duke CFO Survey. I find generated expectations significantly correlate with these existing surveys on a level comparable to the correlation between the surveys themselves. Figure 1 reports the time series of the two benchmark return surveys overlaid with generated expectations of the S&P 500.

Figure 1: Time Series of Return Expectations



Note. Reports the time series of generated monthly/quarterly standardized expectations overlaid with the AAII and CFO surveys respectively.

I then evaluate how well generated return expectations match existing facts previously documented for survey-based return expectations. Greenwood and Shleifer (2014) document a number of facts about survey-based return expectations and I evaluate whether generated expectations match these facts. I find that generated expectations are extrapolative and significantly correlate with equity fund flows. Additionally, I find that generated expectations exhibit the same disconnect from objective measures of expected returns – matching the sign of the correlation for existing survey measures with the log dividend price ratio, Lettau and Ludvigson (2001)'s CAY measure, and several predictive proxies for expected returns. Finally, I find that generated expectations are negatively correlated with future realized returns, similar to existing survey measures while differing from objective measures of expected returns.

To help provide new insights into these systematic biases, I introduce a novel method to analyze the "mental model" underlying the generated expectations. I convert the explanations provided by the LLM along with its numerical forecasts into directed acyclic graphs (DAGs) and cluster the nodes of these DAGs to identify the core themes underlying the LLM's stated reasoning. I provide an example of this procedure in the context of generated return expectations. Based on this procedure I find evidence that positive/negative return expectations are associated with stories of positive/negative future earnings, suggesting the LLM's errors may originate from neglect of general equilibrium as documented for human survey participants in Andre et al. (2024).

Next, I compare generated expectations, for a series of macroeconomic variables studied in Coibion and Gorodnichenko (2015), to those recorded in the Survey of Professional Forecasters (SPF). I find that generated expectations are significantly correlated with revisions in all but two of the SPF series. I then evaluate whether the variation in SPF revisions associated with generated expectations is a driver of the underreaction commonly found in consensus SPF forecasts. By running Coibion-Gorodnichenko (CG) regressions using both SPF revisions and instrumenting with generated expectations, I find that GPT is able to match the observed underreaction.

An alternative story that may impact my results is the possibility of data leakage from the training corpus, resulting in look-ahead bias in generated expectations. I assume that GPT has learned some generalization of the beliefs contained in its training sample, not that it is responding with memorized text. To address this concern, I collect a sample of articles from *WSJ* after GPT's training period – after September of 2021. I then evaluate the correlation between generated expectations and the existing survey measures out-of-sample. I find the out-of-sample correlation is comparable to the in-sample correlation, indicating look-ahead bias is not the principal driver of my results.

After evaluating the properties of my generated expectations, I leverage my methodology to address one of the central constraints on existing survey data: namely the limited sample over which most surveys are available. Using an additional corpus of articles from *The New York Times* (*NYT*), extending back to 1900, I generate expectations for the same set of variables studied above.

This extended sample allows me to address a broader set of questions than those currently possible with existing survey data. In particular, I use my extended sample to construct a measure of economic sentiment designed to capture deviations from full-information rational expectations (FIRE). I examine the properties of this measure and find it is positively correlated with past returns and negatively correlated with future returns, in-line with what would be expected from a measure of economic sentiment.

After extending my sample of expectations and examining the properties of my economic sentiment measure, I apply my methodology to investigate behavioral theories of bubbles.

Many behavioral asset pricing models result in prices reflecting both rational expectations of discounted future cash flows, as well as a behavioral component disconnected from fundamental news.[5] In particular, such models have been developed to explain bubbles (Harrison and Kreps (1978), Hong and Stein (2003), Barberis et al. (2018), Bordalo et al. (2021), Bastianello and Fontanier (2023)). While many theories and much folklore exists around bubbles, rigorous empirical evidence is limited.[6]

One reason empirical evidence is so limited is that measuring mispricing is challenging but necessary for testing these theories. Historically, one approach has been to use analyst expectation errors as a proxy, but the sample of analyst expectations is limited.[7] At the same time, a number of papers have argued that these biases must be systematic to matter for asset prices (Shleifer and Summers (1990), Cochrane (2011), Kozak et al. (2018)). I combine the use of expectation errors and the potential systematic nature of these biases to propose a new proxy: *an asset's exposure to aggregate economic sentiment.*

Leveraging the extended time series and high-frequency nature of my sentiment time series, I run a series of rolling regressions of industry returns on daily sentiment to estimate time-varying sentiment betas. I test the usefulness of these sentiment betas in predicting crashes and future returns in the run-up dataset of Greenwood et al. (2019).

I find sentiment betas are significantly higher during run-ups that crash than during those that do not. This result is robust to various controls studied in Greenwood et al. (2019) as well as market beta. Additionally, I find that increased exposure to sentiment correlates with lower future returns. Again, these results are robust to various benchmarks measures as well as to my definition of sentiment.

I next explore the importance of extrapolation as a mechanism behind bubbles. To test this, I form impulse response functions (IRFs) capturing the impact of a shock to daily returns on future sentiment and vice versa. I find the response of sentiment to a return shock is significantly higher during run-ups that crash than during those that do not. Additionally, I find the response of returns to a shock to sentiment is higher (though not significantly so) during run-ups that crash than those that do not.[8]

Lastly, to evaluate the economic significance of my results, I estimate a trading strategy that holds industry portfolios that are predicted not to crash by a simple threshold rule on the sentiment beta or extrapolation IRF. I find that this strategy predicts correctly around 80% of crashes and earns significant returns relative to a naive strategy which holds all industry

[5]Some relevant examples are DeLong et al. (1990), Shleifer and Vishny (1997), Barberis et al. (1998), Bastianello and Fontanier (2022), Nagel and Xu (2022a).

[6]Recent work such as Greenwood et al. (2019), Chinco (2022) and Blank et al. (2023) attempts to make progress on this problem.

[7]See La Porta (1996), Engelberg et al. (2018) for some examples that follow this approach.

[8]This heightened feedback loop between returns and sentiment corroborates attempts to micro-found extrapolation and its relevance to bubbles with partial equilibrium thinking (PET) (Bastianello and Fontanier (2022), Bastianello and Fontanier (2023)).

run-ups. Additionally, the strategy performs comparably well to one that uses the ex-post crash realizations.

The remainder of the paper is organized as follows: In Section 2 I provide an argument for why we should expect LLMs to mimic human behavior. In Section 3 I detail my methodology for generating expectations with LLMs. In Section 4 I evaluate my generated expectations against existing survey measures. In Section 5 I extend my sample of expectations back to 1900 and construct a measure of economic sentiment from the expectations. In Section 6 I apply my methodology to investigate behavioral theories of bubbles. In Section 7 I conclude.

2 LLMs as Belief Generators

Why should we expect LLMs to mimic human beliefs, and in particular capture biases and systematic errors in expectations? The central conceptual point is that beliefs about the world can be encoded and transmitted in text. When I write the statement "prices will increase because of a supply shortage" I am encoding a particular model of the world into text.

Large language models are algorithms designed to handle and generate exactly this sort of data. More precisely, a (large) language model estimates the conditional probability of a token, s_i, given all previously observed tokens in a document:

$$p(s_i|s_1, \ldots, s_{i-1}), \tag{1}$$

where a token can be thought of as a word, punctuation, or other character. Their success in this task results from a combination of the particular neural network architecture used, namely transformers (Vaswani et al. (2017)), and the massive size of the training data and parameter space used.

Due to this objective, LLMs implicitly possess the ability to learn and apply these text-encoded beliefs to new scenarios. If a sufficiently large portion of the LLM's training sample contains discussions of supply shocks, followed by discussions of various price increases, then when prompted with a headline stating "OPEC+ Agrees to Biggest Oil Production Cut Since Start of Pandemic" and asked what it expects will happen to CPI, the LLM will likely respond that it expects CPI to increase as a result of the shock.

The training data used by modern LLMs, like OpenAI's GPT, contains a massive variety of text and reflects scenarios faced by a wide range of individuals. For the training of GPT-3, 60% of the data used comes from the Common Crawl,[9] a collection of over 240 billion webpages collected over the past 16 years.[10] With the emergence of the internet, an increasing proportion of human interactions and day-to-day scenarios are intermediated and documented via text –

[9] https://commoncrawl.org/the-data/

[10] See Brown et al. (2020) for a fuller discussion of GPT's training data.

providing a rich source of world models for LLMs to learn from.

Importantly, there is no fundamental reason that the world models encoded in LLMs need to be accurate. If a sufficient portion of the LLMs training sample reflects a particular bias this may be reasonably transmitted to the LLM as well.[11] An LLM is ultimately a statistical algorithm designed to generate new text based on what it have "seen" and as such reflects the most common patterns in its training data. In this sense, LLMs possess many similarities to work on associative memory in economics (Gilboa and Schmeidler (1995), Mullainathan (2002), Wachter and Kahana (2019), Bordalo et al. (2020b), Malmendier and Wachter (2021)).

Beyond this conceptual argument, a growing body of experimental evidence supports my claim that LLMs do indeed reflect human beliefs and biases. Aher et al. (2022) run a series of "Turing-experiments" designed to test whether LLMs can pass as humans in a variety of settings and find positive results. Horton (2023) similarly finds that LLMs behave much like humans in a series of economics experiments. Brand et al. (2023) find LLMs have very similar preferences in marketing studies. Argyle et al. (2023) finds similar results in a political science context. Overall, a growing cross-disciplinary consensus is emerging that LLMs can provide a useful proxies for human behavior.

2.1 How can Generated Beliefs be Utilized?

What are some of the potential benefits of generated beliefs? That is, given we already have a number of existing survey-based measures, what can we gain from generated beliefs? Below I consider a number of potential use cases for generated beliefs:

1. Generated beliefs can be formed wherever there is news text and as such can serve to extend the available time series of survey data. The value of this extension is explicitly acknowledged by Brunnermeier et al. (2021) – existing surveys often have a short history which limits the sorts of questions that can be answered.

2. Similarly, generated beliefs can be used to form expectations at a much higher frequency than existing surveys. Many existing surveys are conducted quarterly or monthly, whereas news is often available daily. This allows generated beliefs to be used for high-frequency identification – an approach that has seen considerable success in recent years (Bernanke (2020)).

3. Generated beliefs also open the door to surveying heterogeneous populations and forming expectations for populations that are hard to sample. By using population-specific news sources or fine-tuned LLMs, generated beliefs can address various subpopulations and expand the cross-section of beliefs available.

[11] That LLMs reflect the biases and beliefs of their training corpus is central to much of the current work on AI alignment (Bender et al. (2021), Schramowski et al. (2022)).

4. Similarly, with the increasing availability of firm-specific news in the form of news wires and earnings calls, generated beliefs can be used to form firm-specific expectations. Such expectations are of first-order importance for many economic questions so their availability is of considerable value (Coibion et al. (2018)).

5. Ideally, when participants in the Survey of Professional Forecasters release their expectations, they would also release everything they read while forming these expectations. While this isn't possible with current survey methods, with text-based beliefs, that information is directly available in the form of the news text itself. This opens up opportunities to study the narratives that drive expectations and contribute to a growing literature that centers narratives as a key driver of economic outcomes (Shiller (2019), Bybee et al. (2021), Andre et al. (2021), Bybee et al. (2022), Flynn and Sastry (2022)).

3 How to Generate Beliefs

I next document my methodology for generating beliefs with an LLM. My primary dataset consists of all articles from *The Wall Street Journal* between 1984 and 2021, purchased from the Dow Jones Historical News Archive. *WSJ* covers a wide range of topics with a strong focus on business and finance. It is the second-largest newspaper in the United States by readership and is often regarded as the newspaper of record for business and financial news. As such, it is a natural choice for a source of information on which GPT can form expectations. There is likely considerable overlap between the information contained in *WSJ* and the information used by professional forecasters to form their expectations.

A full summary of the data cleaning procedure is detailed in Appendix A.1. Beyond this initial cleaning, I sample 300 articles randomly from each month of data. This is done because each request made to OpenAI's application programming interface (API) costs a marginal fee. Queries are made to OpenAI's GPT-3.5 model instance. Figure 2 reports the prompt format used to query GPT. Each query to GPT receives a headline and a description of the desired series – for instance, for CPI I will ask about an increase or decrease in "the consumer price index in the United States". In response GPT generates a string of text corresponding to the requested format in the prompt.

Table 1 reports the surveyed series as well as summary statistics of the responses. Figure B.1 in Appendix B reports the coocurrence between the possible pairings of increase/decrease across the surveyed series.[12] Table 2 reports a series of example headlines and the corresponding responses for the S&P 500, CPI, and unemployment. Additionally, I request GPT

[12] As opposed to classical supervised machine learning, GPT is not trained on a specific task, i.e. forecasting economic quantities, but rather on a general language modeling task. As such, this sort of classification is referred to as zero-shot learning in the machine learning literature.

Figure 2: Prompt Format

```
Here is a piece of news:

"%s"

Do you think this news will increase or decrease %s?

Write your answer as:

{increase/decrease/uncertain}:
{confidence (0-1)}:
{magnitude of increase/decrease (0-1)}:
{explanation (less than 25 words)}
```

Note. Reports the prompt format for queries made to GPT. "%s" indicates where in the prompt the headline and target text are inserted.

to provide an explanation for its response which is included in Table 2. These explanations provide a useful method to analyze the reasoning behind GPT's responses. For instance, when presented with a news article about the opening of access to the Japanese phone market for U.S. firms, GPT expects this will increase the S&P by providing growth opportunities for U.S. phone companies, will decrease unemployment by creating new job opportunities in the U.S. telecommunications industry, and lower the CPI by increasing competition. Similarly, when presented with an article about tax cuts and tightening government budgets during the Bush years, GPT expects this will increase the S&P by stimulating the U.S. economy, not impact unemployment due to uncertainty about which effect may dominate, and increase the CPI by increased demand as a result of the tax cuts. These explanations highlight GPT's ability to provide a clear chain of reasoning for its answers and open up new possible routes for understanding the causal structure of beliefs.

Finally, since GPT produces article-level binary expectations, I aggregate these expectations to place them at a comparable frequency to the existing survey data. To do this aggregation I compute a "balance statistic": the proportion of articles where GPT responds with increases minus the proportion of articles where GPT responds with decreases. Let $F_t^{gpt}(X_{t+h}^k)$ correspond to the generated expectations in period t for the kth series, X_{t+h}^k, at some future horizon, h. Using this approach, generated expectations are given by:

$$F_t^{gpt}(X_{t+h}^k) = \frac{\sum_{i \in A_t} \mathbb{I}(\text{Increase})_i^k - \mathbb{I}(\text{Decrease})_i^k}{\sum_{i \in A_t} \mathbb{I}(\text{Increase})_i^k + \sum_{i \in A_t} \mathbb{I}(\text{Decrease})_i^k}, \tag{2}$$

where A_t is the set of articles published in period t, $\mathbb{I}(\text{Increase})_i^k$ indicates GPT responded

Table 1: Summary Statistics of GPT Survey

Series	Prompt	Date Range	Count	Inc. %	Dec. %	Unc. %
SNP	the S&P 500 index	1984-2021	136345	15.13	26.84	58.02
CPI	the consumer price index in the United States	1984-2021	132736	7.86	6.45	85.69
HS	housing starts in the United States	1984-2021	132212	2.50	5.68	91.82
IP	industrial production in the United States	1984-2021	132892	10.11	11.72	78.17
DEFL	the GDP price deflator in the United States	1984-2021	132760	9.63	12.50	77.88
AAA	the AAA-rated bond's rate in the United States	1984-2021	133467	11.09	14.88	74.03
C	real consumption in the United States	1984-2021	131574	11.53	17.67	70.80
GF	federal government consumption in the United States	1984-2021	132839	9.86	10.88	79.26
GY	the real GDP of the United States	1984-2021	132148	20.54	20.91	58.56
NRI	real nonresidential investment in the United States	1984-2021	132961	17.46	22.94	59.59
RI	real residential investment in the United States	1984-2021	133157	8.66	16.49	74.85
GS	state and local government consumption in the United States	1984-2021	131428	13.44	17.30	69.26
3TB	the 3-month treasury bill rate	1984-2021	134609	15.45	11.21	73.34
UE	employment in the United States	1984-2021	120102	9.81	11.24	78.95

Note. Reports summaries of the surveyed labels, the text passed to GPT, the date range, the number of non-missing responses, the proportion of responses that were "increase", the proportion of responses that were "decrease", and the proportions that were "uncertain".

with "increase" for the kth series given article i, and $\mathbb{I}(\text{Decrease})_i^k$ a comparable indicator for "decrease". This approach is used by other popular surveys such as the Gallup survey, the University of Michigan Survey of Consumers, and the American Association of Individual Investors survey.

Table 2: Example Responses

Series	Direction	Headline/Response
		(1989-06-29) U.S. Reaches Accords Widening Access To the Mobile Phone Business in Japan
S&P	1	The news signals potential growth opportunities for US mobile phone companies, which could positively impact the S&P 500 index.
UE	-1	Increased access to the Japanese mobile phone market will likely create new job opportunities in the U.S. telecommunications industry.
CPI	-1	Increased competition and access to the Japanese mobile phone market may lead to lower prices for U.S. consumers.
		(2001-02-28) Bush Offers Tax Cuts and Tight Budgets to Aid 'Faltering' Economy
S&P	1	Tax cuts stimulate the economy, which could lead to increased corporate profits and higher stock prices.
UE	0	Tax cuts may stimulate investment, but tight budgets could reduce government spending and slow job growth.
CPI	1	Tax cuts usually stimulate spending, which can increase demand and raise prices, but budget tightening may counteract that effect.

Note. Reports a sample of example article headlines and the corresponding responses from GPT. The first column reports the series queried. The second column reports the direction of the query. The third column reports the headline and the corresponding response from GPT.

4 Evaluating Generated Expectations

4.1 Return Expectations

I first evaluate generated expectations of returns by comparing them to two publicly available benchmark return expectation series used in Greenwood and Shleifer (2014). The first is the American Association of Individual Investors (AAII) Investor Sentiment Survey. The AAII survey is a weekly survey of members of the AAII running from 1987 up to the present day which measures the percentage of participants that are bullish, bearish, or neutral on the stock market for the next six months. Following Greenwood and Shleifer (2014), I aggregate the weekly responses to monthly averages for the majority of my analysis. The second survey is the Duke CFO Survey, a survey of chief financial officers (CFOs), started in 1998 by John Graham and Campbell Harvey. The survey records CFO views on a variety of macroeconomic

and firm-specific quantities including their expectations of returns for the U.S. stock market over the next twelve months. It runs from 2000 to the present day.

Figure 3 reports the correlation between each survey series and various aggregates of generated expectations. I consider aggregates over one, two, and three-month windows as well as an exponentially weighted average over daily aggregates. For the exponentially weighted average, I report correlation for the optimal smoothing parameter that maximizes the correlation with the survey series – this provides an upper bound on the correlation between generated expectations and the existing survey series.

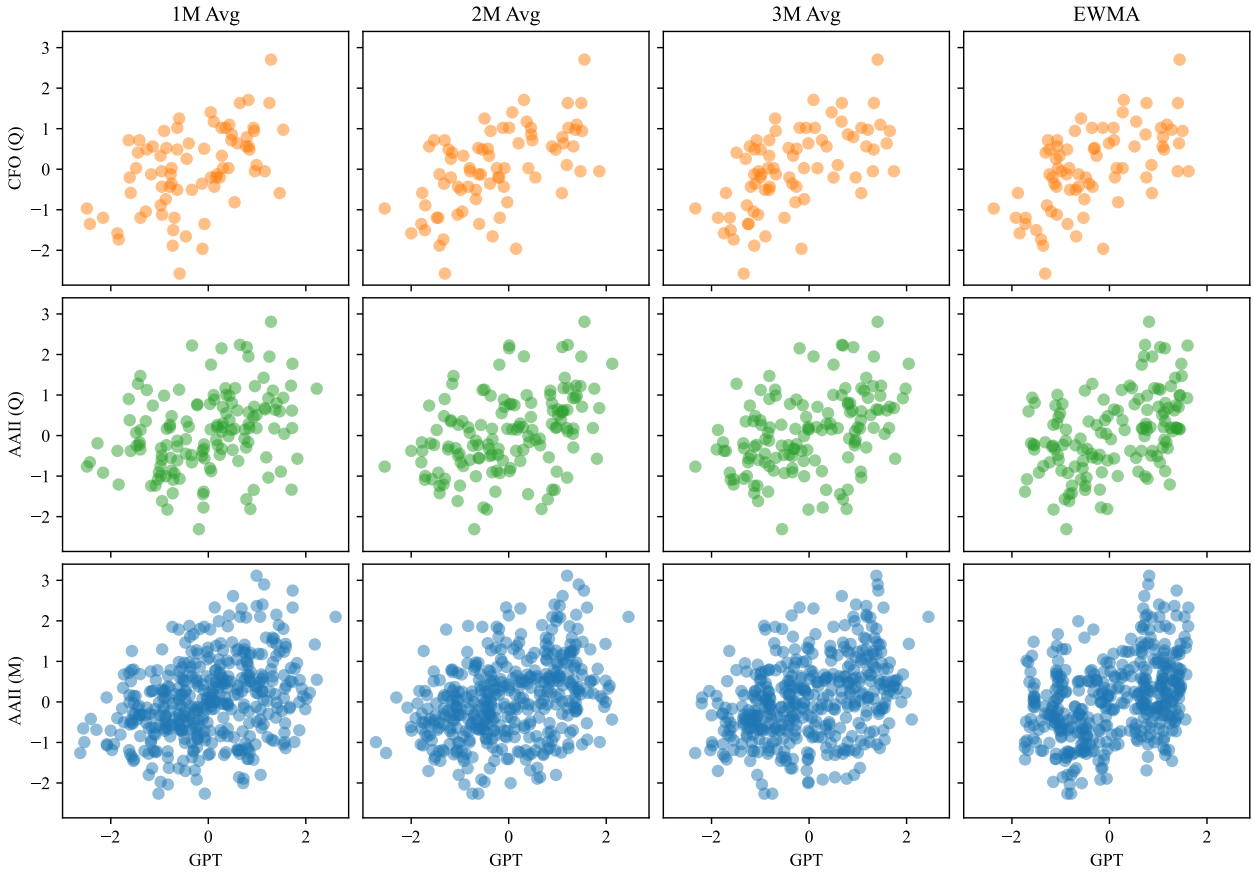
I find that for all different aggregation methods, generated expectations are significantly correlated with the existing survey measures. For the remainder of this section, I focus on the three-month aggregation window. For this window, the average correlation between generated expectations and the existing survey measures is 0.47. For comparison Greenwood and Shleifer (2014) find an average correlation of 0.43 among the full set of return expectation surveys they evaluate.

These results suggest generated expectations closely match the variation in existing survey measures. Motivated by this result, I next evaluate whether generated expectations other facts documented previously for survey-based return expectations.

An extensive literature in asset pricing has focused on the importance of extrapolative expectations (Cutler et al. (1990), Barsky and De Long (1993), Lakonishok et al. (1994), Barberis et al. (2015), Jin and Sui (2022)). As a result, I evaluate the correlation of generated expectations and different survey expectations against the past twelve-months returns (R_{t-12}) of the U.S. stock market, following Greenwood and Shleifer (2014) and Nagel and Xu (2022b). Additionally, a recent literature has leveraged the increased availability of realized trading activity as a proxy for investor beliefs (Giglio et al. (2019), Gabaix and Koijen (2021), Alekseev et al. (2022)). Following this literature, I evaluate the correlation between different return expectation measures and mutual fund flows into equities. To compute mutual fund flows I combine the Thomson Reuters Mutual Fund Holdings S12 database with the CRSP Survivor-Bias-Free US Mutual Funds and follow the data cleaning procedure detailed in Alekseev et al. (2022).

An exhaustive literature in asset pricing has also studied empirical measures of objective expected returns. Importantly, there is a well-known disconnect between these objective measures and subjective survey measures (Greenwood and Shleifer (2014), Nagel and Xu (2022b)). As a result, I next consider a number of objective expected return proxies. First, I consider the log dividend-price ratio and 12-month changes in the log dividend-price ratio following much of the empirical literature. Second, I consider the consumption wealth ratio (CAY) of Lettau and Ludvigson (2001). Finally, I consider two expected return indices formed by regressing future 12-month returns on various sets of predictors. In particular, I consider the same expected return index used in Greenwood and Shleifer (2014) (following Fama and

Figure 3: Correlation between Return Expectations



	one-month avg.	two-month avg.	three-month avg.	opt. EWMA
CFO (N=76)	0.49	0.56	0.59	0.59
	[4.84]	[5.88]	[6.21]	[6.24]
AAII (Q, N=138)	0.32	0.40	0.39	0.43
	[3.93]	[5.15]	[5.01]	[5.63]
AAII (M, N=414)	0.32	0.35	0.35	0.36
	[6.83]	[7.65]	[7.65]	[7.82]

Note. The top figure reports scatter plots of generated expectations formed over various windows (columns) against the various existing survey measures (rows). The table reports the correlation coefficients and corresponding t-stats in brackets. t-stats use Newey-West standard errors with a 12-month lag. one/two/three month average correspond to GPT expectations aggregated over one/two/three-month windows respectively. opt. EWMA corresponds to an exponentially weighted average with the optimal tuning parameter.

French (1989)) which uses the log dividend price ratio, the Treasury-bill yield, the default spread (the yield on BAA minus yield on AAA-rated bonds) and the term spread (the yield on ten-year government bonds minus the yield on three-month Treasury bill). Additionally,

I form an expected return index using the "kitchen-sink" predictors from Welch and Goyal (2008) which represent a common benchmark in the empirical asset pricing literature.

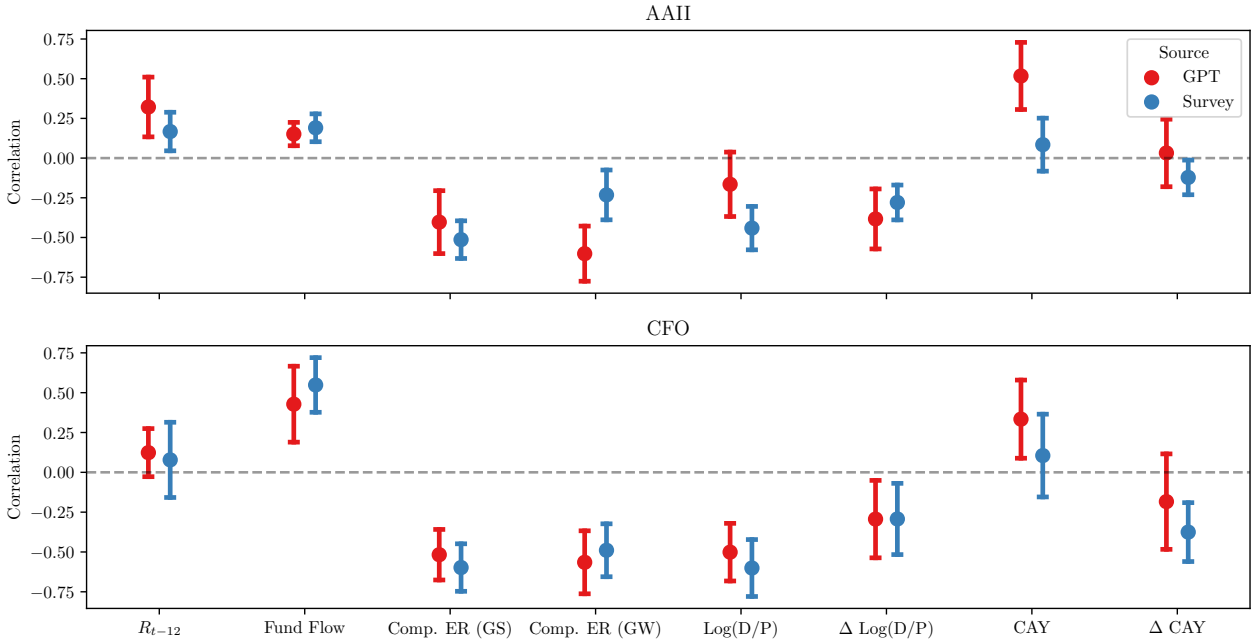
Figure 4 reports the correlations between each expectation series and the corresponding measures discussed above. For both the AAII and CFO surveys, I include correlations based on the original survey values, as well as correlations based on generated expectations at the monthly and quarterly frequencies respectively for dates where the survey is available. For instance, in panel A we can see that both the AAII survey and the corresponding GPT-based series are significantly positively correlated with the past twelve-month returns, indicating that both series exhibit return extrapolation.

For all cases except changes in CAY and the generated expectations proxy for AAII, the generated expectation correlations exhibit the same sign as the existing survey measures. These results provide evidence that GPT does not only match the variation in existing survey measures but more importantly the deviations from rational expectations noted in the literature, indicating the potential of these measures as a tool for studying nonrational expectations.

Additionally, it is the case that measures of expected returns should forecast future returns under models of rational expectations. However, if anything, existing survey measures are negatively correlated with future returns. As a result, I next compare the correlation of generated expectations with future returns to that of a number of alternative subjective and objective expected return measures. Figure 5 reports the correlation coefficients for a series of predictive regressions of future returns over various horizons on the expectation measures discussed above. GPT exhibits the same negative correlation with future returns as existing survey measures, which is in contrast to the positive correlation observed for objective measures.

Taken together these results tell a consistent story: as opposed to popular stories of LLMs as a class of super-smart AI, generated expectations are in fact quite flawed in very similar ways to human expectations. GPT exhibits extrapolative expectations that are negatively correlated with objective measures of expected returns. These results suggest a clear path forward for the use of LLMs in economics and finance: instead of being a tool for superhuman forecasting, LLMs can be used to generate the beliefs of humans and provide a new method for studying expectations.

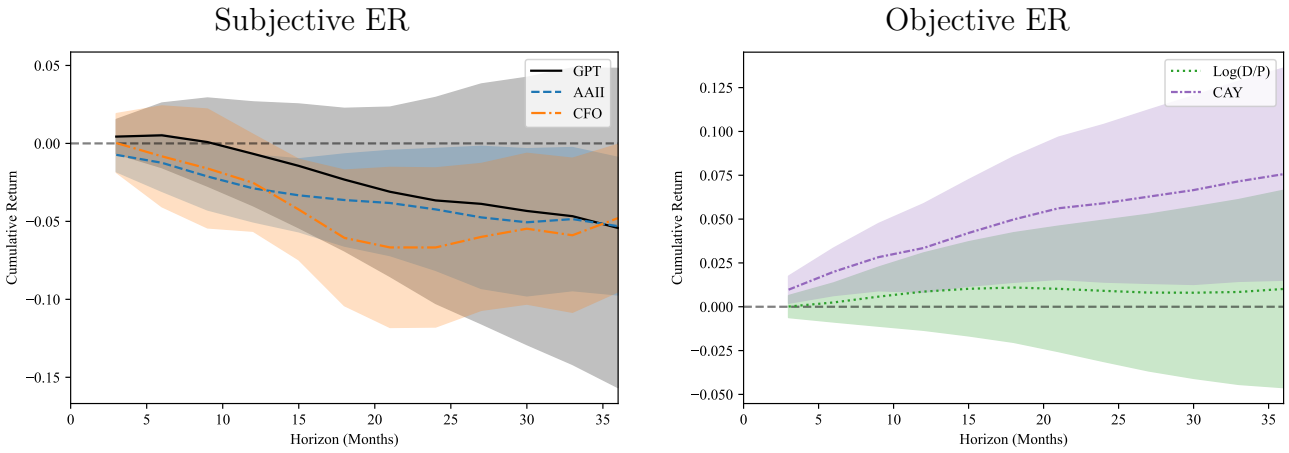
Figure 4: Survey Correlations with Existing Moments



	GPT (AAII)		AAII			GPT (CFO)		CFO		
	Corr.	t	Corr.	t	tD	Corr.	t	Corr.	t	tD
R_{t-12}	0.32	2.82	0.17	2.27	-1.14	0.12	1.35	0.08	0.54	-0.27
Fund Flow	0.15	3.39	0.19	3.58	0.57	0.43	2.95	0.55	5.26	0.68
Log(D/P)	-0.17	-1.34	-0.44	-5.31	-1.86	-0.50	-4.56	-0.60	-5.53	-0.64
Δ Log(D/P)	-0.38	-3.34	-0.28	-4.19	0.78	-0.29	-1.99	-0.29	-2.15	0.00
CAY	0.52	4.03	0.08	0.83	-2.64	0.33	2.24	0.11	0.67	-1.05
Δ CAY	0.03	0.25	-0.12	-1.84	-1.06	-0.18	-1.01	-0.38	-3.34	-0.89
Comp. ER (GS)	-0.40	-3.35	-0.51	-7.13	-0.79	-0.52	-5.36	-0.60	-6.59	-0.61
Comp. ER (GW)	-0.60	-5.70	-0.23	-2.43	2.60	-0.56	-4.70	-0.49	-4.85	0.48

Note. Reports the correlation between each expectation series and the corresponding series and 90% confidence intervals, standard errors are Newey-West with a 12-month lag. The bottom table reports the corresponding correlation coefficients and t-stats. t-stats use Newey-West standard errors with a 12-month lag. tD corresponds to the t-stat for the difference in correlation between the GPT and existing survey measures. GPT (AAII) corresponds to the correlation with generated expectations at the monthly frequency for dates where the AAII survey is available. GPT (CFO) corresponds to the correlation with generated expectations at the quarterly frequency for dates where the CFO survey is available. R_{t-12} corresponds to lagged 12-month returns. Fund Flow corresponds to the aggregate mutual fund flows into equities. Log(D/P) is the log dividend-price ratio and CAY is the aggregate log consumption-wealth ratio of Lettau and Ludvigson (2001). Δ Log(D/P) and Δ CAY correspond to the respective 12-month changes. Comp. ER (GS) corresponds to the fitted values of a regression of one-year ahead returns on log dividend price ratio, the Treasury-bill yield, the default spread and the term spread — the composite expected return measure used in Greenwood and Shleifer (2014). Comp. ER (GW) corresponds to the fitted values of a regression of one-year ahead returns on the "kitchen-sink" predictors from Welch and Goyal (2008).

Figure 5: Predictive Return Regressions



Note. Reports the coefficients for a series of predictive regressions of future cumulative returns over the given horizon on subjective and objective expected return proxies. Shaded bans report 90% confidence intervals using Newey-West standard errors with the corresponding horizon as the number of lags.

4.2 The Mental Model Underlying Generated Mistakes

Can we leverage the explanations provided along with the LLM's numerical expectations to understand the origins of the mistakes it makes? A recent and growing body of literature examines open-ended survey responses to better understand how mental models inform how individuals form beliefs.[13] The explanations provided by the LLM are an extremely rich and granular source of such data given they are formed at the article level. I next attempt to utilize these unique benefits to provide a better understanding from where the mistakes in the generate expectations may originate.

To do this I first impose some additional structure on the explanations. Recent work such as Andre et al. (2021) use directed acyclic graphs (DAGs) to represent the causal structure of narratives or a mental model. I follow a similar approach and first convert the explanations into a DAG. Here too I use the LLM to handle the processing. As an example consider the following headline:

> Machine-Tool Orders Jumped 27% in December — Results for Full Year Rose By 42% to $4.68 Billion Highest Level Since '79

The LLM expects an increase in the S&P 500 as a result of this news and provides the following explanation:

> Positive economic data suggests growth, leading to investor optimism and potential market gains.

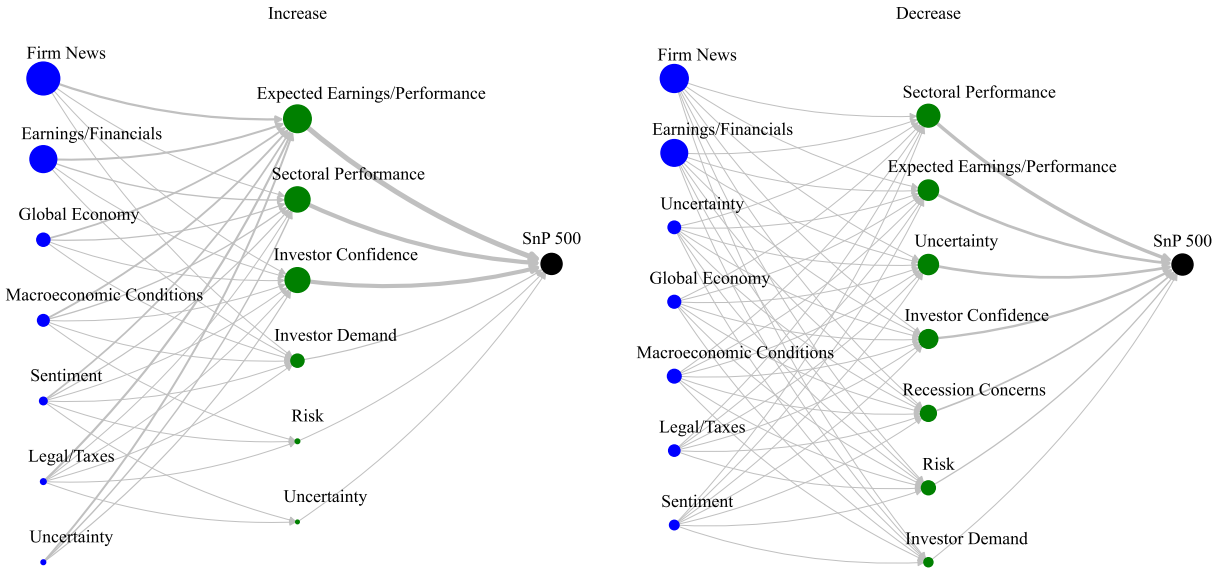
[13] See Haaland et al. (2024) for a recent survey of this literature.

I then ask the LLM to convert this explanation into a DAG yielding the following result:

> Positive economic data -> investor optimism -> potential market gains.

The full set of prompts for this procedure are detailed in Appendix C. Once the DAGs are constructed, I then aggregate the nodes into coherent clusters to summarize the result. To do this I first provide the set of nodes to the LLM and have it define labels based on the text. I then manually inspect this set of labels and select the 7 most common. I then instruct the LLM to apply these labels to each node and generate the final DAG.

Figure 6: The Mental Model Underlying Return Expectations



Note. Reports the DAGs formed from the explanations associated with the generated return expectations. The blue dots represent the initial cause, the green nodes the intermediate cause and the black nodes the outcome (increase/decrease). Edge size represents the number of connections between each node and node size represents the number of explanations associated with each node.

Figure 6 reports the results of this exercise for the generated return expectations. The LLM puts most of its weight on firm performance and earnings as the primary driver of future returns, suggesting that it expects returns will be higher in the future because earnings will be higher in the future. However, this story is incorrect, under general equilibrium, higher earnings should be reflected in higher returns today and not in the future. This mistake in reasoning provides one story for where the LLMs extrapolative beliefs may come from: it is a neglect of general equilibrium. This result is also consistent with Andre et al. (2024) who find that human survey respondents make similar mistakes in reasoning.

4.3 Macroeconomic Expectations

I next evaluate generated expectations by comparing them to a standard set of macroeconomic expectations: the Survey of Professional Forecasters (SPF). The SPF is a quarterly survey of macroeconomic forecasts conducted by the Philadelphia Federal Reserve and remains the gold standard for work studying macroeconomic expectations (Coibion and Gorodnichenko (2012), Coibion and Gorodnichenko (2015), Bordalo et al. (2020a), Angeletos et al. (2021), Farmer et al. (2021)). As professional forecasters, the respondents are some of the most informed agents in the economy and as such errors in their forecasts are notable.

The SPF covers a wide range of macroeconomic quantities during my sample, including CPI, real GDP, the unemployment rate and the federal funds rate. I compare generated expectations to that of the SPF for the 13 series examined in Coibion and Gorodnichenko (2015). For generated expectations, I use a quarterly aggregate of the article-level expectations over the quarter prior to the surveyed period. The SPF releases forecasts at a variety of horizons; however, since the horizon of generated expectations is not known, I compare generated expectations to an average over the 1-4 quarter horizon forecasts. My results are robust to this choice and results for each of the 1-4 quarter horizon forecasts are shown in Appendix E.

I first evaluate how well generated expectations correlate with SPF forecasts. I report results for both levels $F_t(X^k_{t+h})$ and revisions $F_t(X^k_{t+h}) - F_{t-1}(X^k_{t+h})$. Revisions capture the new information embedded in the forecast and given generated expectations are formed from granular pieces of recent news it is likely that they will be more closely related to revisions than levels. Figure 7 reports these correlations – a full set of time-series plots of generated expectations vs. the corresponding SPF revisions is available in Appendix D. For all but two of the sampled series, federal government consumption and state and local government consumption, generated expectations are significantly correlated with the corresponding revisions at the 90% confidence level.

Much of the literature studying deviations from full-information rational expectations focuses on the predictability of forecast errors. In particular, considerable focus is given to the rigidity of beliefs to new information measured using Coibion and Gorodnichenko (2015) (CG) regressions. CG regressions regress forecast errors on the corresponding revisions in expectations – positive coefficients capture underreaction, while negative coefficients capture overreaction.

Given the significant correlation between generated expectations and SPF revisions, I next evaluate whether the well-documented underreaction in macroeconomic expectations can be explained by the component of revisions associated with GPT. To do this, for the series that exhibit a significant correlation between GPT and SPF revisions (that is, excluding government consumption), I regress SPF forecast errors on the corresponding version of generated expectations. Figure 8 reports the correlations for these regressions and compares them to

the correlations using the original SPF revisions. For all but two of the series, the correlation between generated expectations and SPF forecast errors is positive and the overall pooled results are positive and significant at the 90% confidence level.[14]

As with generated return expectations in Section 4.1, these results suggest that generated expectations capture much of the variation in existing survey measures. Additionally, they suggest that generated expectations capture variation that is associated with deviations from full information rational expectations noted in the literature.

An open question still remains: who's beliefs are GPT's? My results suggest GPT's beliefs closely match a number of different groups, including professional forecasters, individual investors, and CFOs. GPT's training corpus is based on a wide range of sources and topics, and as such it may provide something of a "representative agent". Nevertheless, LLMs designed to more directly reflect the beliefs and interest of distinct groups are likely the future of this field. This work may be done by fine tuning existing open source LLMs like that in Touvron et al. (2023) or building from the ground up from distinct corpuses as in Wu et al. (2023).

4.4 Memorization or Generalization?

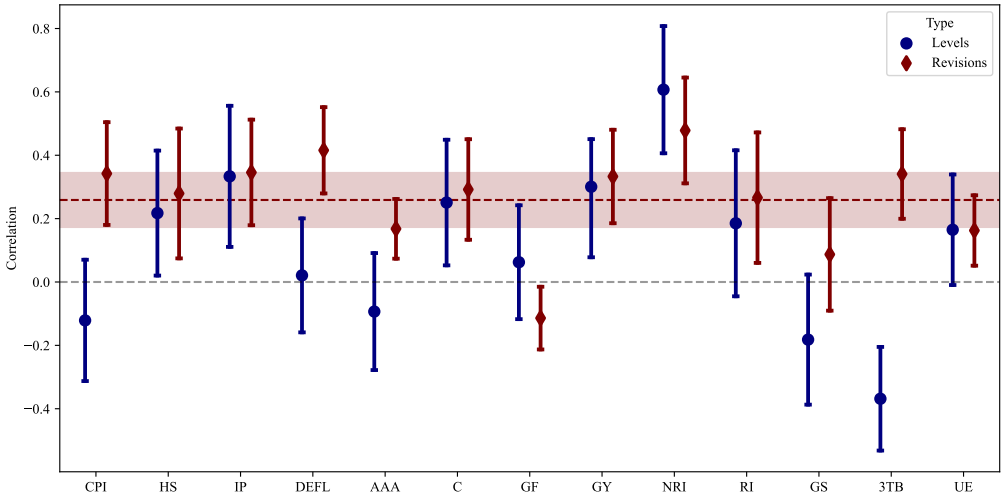
GPT's training period overlaps with my sample. As a result, one potential concern is that generated expectations may reflect memorization of the training sample. This potential for look-ahead bias is a concern as it would limit the usefulness of GPT as a tool to generate beliefs. To address this concern, I next attempt to evaluate generated expectations out-of-sample, that is, after the training period, to see if the correlations I have observed so far persist.

To evaluate generated out-of-sample expectations, I scrape all available *WSJ* articles from the *WSJ* archive between September 2021 and March 2023. I then apply the cleaning procedure detailed in Appendix A.2 and query GPT for its expectation of each article, as with the main sample. AAII return expectations are released weekly, and as a result, this gives me 79 observations outside of GPT's training period. For the SPF only five quarters of data are available – as a result I run a panel regression of revisions in SPF expectations on generated expectations for the full set of series studied above.

To measure how well GPT performs out-of-sample vs. in-sample, I take a set of 500 bootstrap samples of the same size as the out-of-sample period and compute the correlation between generated expectations and the corresponding benchmarks. I then compare the distribution of these bootstrap sampled correlations to the correlation computed using the

[14]Unemployment exhibits significant overreaction in the original SPF revisions, going against the literature. My sample includes the 2020 Covid-19 pandemic which may impact this result. To test this, I re-estimate the same procedure limiting myself to the period prior to 2019. These results are reported in Appendix F. The significant positive correlation between generated expectations and SPF forecast errors remains and the original CG results align with the literature.

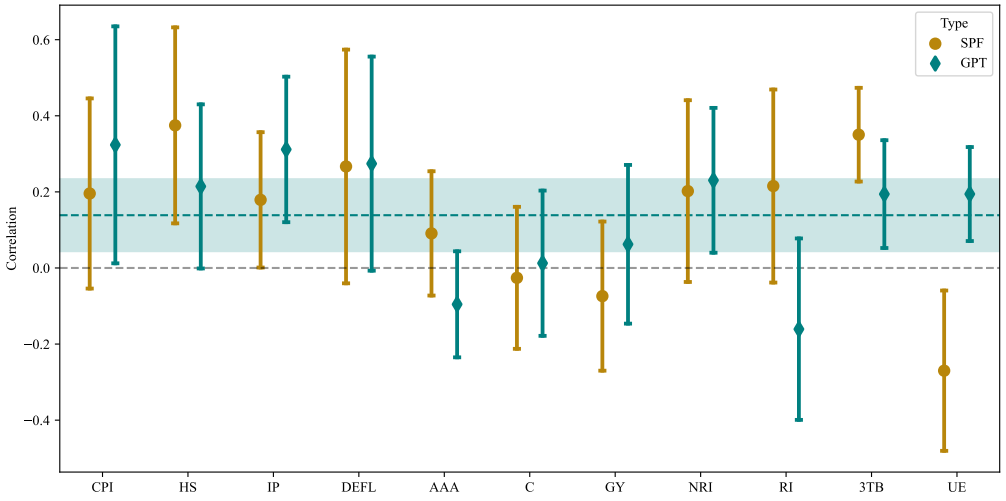
Figure 7: GPT/SPF Correlations



	Levels		Revisions	
	Corr.	t	Corr.	t
Panel	0.07	1.49	0.26	4.93
CPI	-0.12	-1.04	0.34	3.46
Housing Starts	0.22	1.81	0.28	2.24
Industrial Production	0.33	2.46	0.35	3.40
GDP Deflator	0.02	0.19	0.42	5.01
AAA Corporate Bond Yield	-0.09	-0.83	0.17	2.92
Consumption Growth	0.25	2.07	0.29	3.02
Federal Government Spending	0.06	0.57	-0.11	-1.89
GDP Growth	0.30	2.22	0.33	3.70
Nonresidential Investment	0.61	4.96	0.48	4.70
Residential Investment	0.19	1.32	0.27	2.12
State Government Spending	-0.18	-1.45	0.09	0.80
3-Month Treasury Bill	-0.37	-3.70	0.34	3.96
Unemployment Rate	0.16	1.55	0.16	2.40

Note. Reports the correlation between levels and revisions of SPF expectations and GPT expectations. Additionally reports 90% confidence intervals for the correlation coefficients. The dashed maroon line reports the panel correlation coefficient and the shaded maroon band the corresponding 90% confidence interval. Standard errors for the single variable regressions are Newey-West, and panel standard errors are Driscoll-Kraay. The bottom table reports the corresponding correlation coefficients and t-stats.

Figure 8: Coibion-Gorodnichenko Regressions



	SPF		GPT	
	Corr.	t	Corr.	t
Panel	0.12	1.31	0.14	2.39
CPI	0.20	1.28	0.32	1.70
Housing Starts	0.37	2.39	0.21	1.63
Industrial Production	0.18	1.65	0.31	2.67
GDP Deflator	0.27	1.42	0.27	1.60
AAA Corporate Bond Yield	0.09	0.91	-0.10	-1.12
Consumption Growth	-0.03	-0.23	0.01	0.11
GDP Growth	-0.07	-0.62	0.06	0.49
Nonresidential Investment	0.20	1.39	0.23	1.99
Residential Investment	0.22	1.39	-0.16	-1.11
3-Month Treasury Bill	0.35	4.67	0.19	2.25
Unemployment Rate	-0.27	-2.10	0.19	2.58

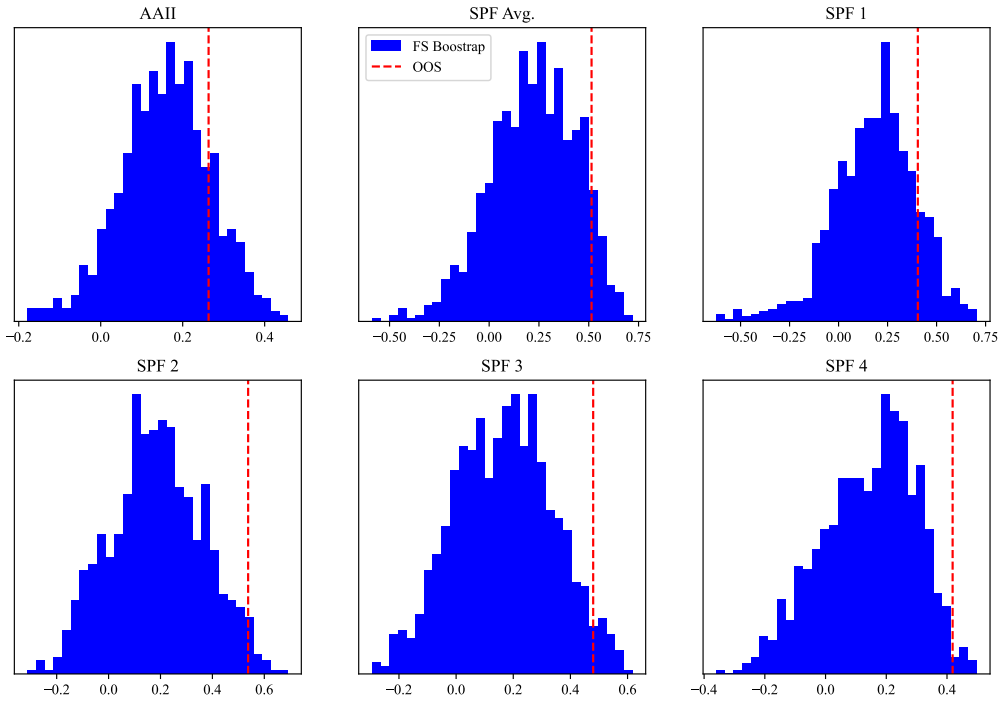
Note. Reports the CG coefficients for SPF revisions and GPT expectations. Additionally reports 90% confidence intervals for the coefficients. The dashed teal line reports the panel CG coefficient and the shaded teal band the corresponding 90% confidence interval. Standard errors for the single variable regressions are Newey-West, and panel standard errors are Driscoll-Kraay. The bottom table reports the corresponding correlation coefficients and t-stats.

out-of-sample period.

Figure 9 reports these results for the AAII return expectations as well as the pooled SPF expectations for various horizons. The results provide good evidence that generated expectations are not driven purely by look-ahead bias, the out-of-sample correlations are similar to the in-sample correlations and if anything are towards the upper tail of the in-sample correlation distribution. A simple t-test evaluating whether the out-of-sample correlation is different from zero also finds significant results for all sets of correlations considered.

A conceptual point is also worth making here: the primary focus of much of my paper is not only showing that GPT has similar expectations to humans but also that it has similar *deviations* from rational expectations. Any look-ahead bias in generated expectations would likely push generated expectations closer to rationality and work against my results. This point will become even more important in Sections 5 and 6 when I use generated expectations to extract a measure of systematic errors in expectations and use this measure to identify bubbles.

Figure 9: GPT's Correlation Out-of-Sample



Note. Reports the out-of-sample correlation between generated expectations and various benchmarks (red dashed line). Additionally, reports the distribution of 500 bootstrap sampled correlations using the same sample size as the out-of-sample period (blue histogram). Each SPF column reports a different horizon and standard errors are Driscoll-Kraay. The table reports summary statistics for the corresponding distributions.

5 Generating 120 Years of Economic Expectations

A limitation of existing survey data is that they are available only for relatively short sample periods. Generated expectations, on the other hand, can be produced wherever there is news and as a result can be used to study expectations over a much longer period. I next turn to the task of extending the sample of existing economic expectations by generating beliefs over the past 120 years.

Of particular interest are the systematic expectation errors embedded in these beliefs. The behavior of these systematic errors is of great interest for recent theoretical literature and the extended sample of expectations provides a new tool for studying them. As shown in Section 4, generated expectations exhibit many of the same deviations from FIRE documented in the literature. Bianchi et al. (2022) and van Binsbergen et al. (2023) attempt to measure these systematic errors but are limited by the sample period of their data.[15] Using my extended sample of expectations I attempt to extract a measure of these systematic errors and evaluate its dynamics over time.

I refer to this measure as *economic sentiment* and define it as the irrational component of economic expectations. Clarity on the definition of economic sentiment is important as the term has been used to mean different things in different papers. In some machine learning applications, sentiment is used to refer to the polarity of a text, that is, whether the text is positive or negative. Similarly, in macroeconomics, sentiment is sometimes used to refer to *expectations themselves*, as in the case of the University of Michigan consumer sentiment index.

5.1 Identifying Economic Sentiment with Generated Beliefs

How will I go about extracting a measure of sentiment from my generated beliefs? To motivate my approach I consider the following factor model of expectations, $e_{t,i}$, for the ith economic series. I assume that expectations can be decomposed into a component associated with the rational forecast, μ_t and a component associated with sentiment, δ_t:

$$e_{t,i} = \nu_i \mu_t + \gamma_i \delta_t. \tag{3}$$

ν_i and γ_i are series-specific loadings on the rational and sentiment components respectively.

Such single-factor models are common in macroeconomics as it is often argued that many macroeconomic series share a common source of variation (Lucas (1977), Angeletos et al. (2020)). The key innovation here is thinking of sentiment as a common component. In Section 4 I showed that generated expectations deviate from full information rational expectations.

[15] Additionally, a number of papers in finance use analyst ex-post errors in similar ways (La Porta (1996), Engelberg et al. (2018), Kozak et al. (2018)).

As such, it seems plausible that sentiment, or the irrational component of my generated expectations, is an important driver of variation in the generated expectations. The central challenge for this model is separately identifying μ_t and δ_t. If the rational expectation associated with each series was idiosyncratic, then δ_t could be directly recovered by running principal component analysis (PCA) on the expectation series themselves and taking the first principal component. Motivated by this point I consider several methods for identifying δ_t.[16]

First, given μ_t represents a rational forecast, I attempt to estimate it directly from ex-ante available information following similar work such as Bianchi et al. (2022) and van Binsbergen et al. (2023). I assume the rational forecast is formed using a dynamic factor model over a broad set of macroeconomic outcomes. Such models are common in the macroeconomics literature to model the joint dynamics of macroeconomic outcomes and as such make a reasonable choice for the rational forecast.[17] Let X_t be a vector of macroeconomic outcomes and F_t a vector of latent factors. I assume the rational forecast is formed using the following dynamic factor model:

$$\begin{aligned} X_t &= \Theta F_t + \epsilon_t \\ F_t &= \Phi F_{t-1} + \eta_t, \end{aligned} \tag{4}$$

where ϵ_t and η_t are idiosyncratic shocks. Given the rational forecaster's estimates of Φ and F_{t-1}, formed in period $t-1$, the rational forecast is then given by $F_t = \Phi F_{t-1}$. I can recover δ_t with a two-stage procedure where I first orthogonalize $e_{t,i}$ with respect to F_t and then run PCA on the orthogonalized series and take the first principal component. I refer to this estimate of δ_t as the ex-ante residual (EAR) sentiment measure.

A challenge for this first approach is that if the rational forecaster uses a different model specification or has additional information available, then the EAR estimate may be biased. As an alternative, I next consider a second approach that uses ex-post realizations of the macroeconomic outcomes to estimate μ_t, that I refer to as the ex-post residual (EPR) estimate. This approach relates more closely to work such as La Porta (1996), Engelberg et al. (2018), and Kozak et al. (2018), which use ex-post analyst errors as a measure of sentiment. This approach is comparable to assuming the rational forecaster has sufficient information available ex-ante that they can correctly recover the ex-post realization. While this assumption is likely too strong, it provides a useful benchmark for evaluating the EAR estimate and represents an upper bound on the rational forecast.

To produce the EPR estimate I once again assume that a broad set of macroeconomic

[16] It is plausible $e_{t,i}$ also contains idiosyncratic rational and sentiment components. These may be interesting to study in future work, however, I focus on the common components here as they are more directly related to the dynamics of the economy as a whole.

[17] See Stock and Watson (2011) for a fuller discussion of this literature.

outcomes, represented by X_t, are driven by a common set of latent factors G_t:

$$X_t = \Gamma G_t + \xi_t, \tag{5}$$

where ξ_t is an idiosyncratic shock. Given an estimate of G_t, recovered by running PCA directly on the macroeconomic outcomes, I can then recover δ_t by a two-stage procedure where I first orthogonalize $e_{t,i}$ with respect to G_t and then run PCA on the orthogonalized series and take the first principal component.

As a final alternative, I introduce a number of additional assumptions that allow me to directly estimate δ_t as the first principal component of the expectations themselves, labeled (EPC). In particular, if I assume that μ_t and δ_t are orthonormal and that the eigenvalue associated with δ_t is larger than that of μ_t, then the first principal component of $e_{t,i}$ will be δ_t. As a whole these three methods (EAR, EPR, and EPC) provide a range of estimates for δ_t that allow me to evaluate the robustness of my results to the identification strategy and provide evidence on the importance of the rational component for my results.

5.2 Transfer Learning with BERT

The initial sample of articles considered above covers less than 15% of the full *WSJ* corpus from 1984 to 2021. Labeling the full set of available articles would help form the most accurate estimates of daily expectations. Further, to extend my sample prior to 1984 and provide sufficient coverage to identify bubbles, a much larger set of articles would need to be labeled. However, eliciting forecasts for all these articles through GPT is prohibitively costly. As a solution: I next leverage my existing sample of labeled articles to train a simpler supervised model using the embeddings from an earlier language model, BERT, short for Bidirectional Encoder Representations from Transformers.

BERT (Devlin et al. (2018)) is an earlier transformer-based language model than GPT-3.5 – though the release of the initial GPT-1 model slightly predates it (Radford et al. (2018)). While there is overlap in the training data of both models, BERT consists of a much smaller number of parameters (340 million) as opposed to the more recent GPT-based models (175 billion for GPT-3). BERT and GPT also have some slight differences in their underlying architecture. However, the principal benefit of BERT is that it is open source and can be run completely locally, allowing me to use BERT as the basis of a simpler supervised model to infill my remaining articles at a much reduced computational cost. A fuller discussion of these architectural differences is saved for Appendix H.

In both cases, these models use embeddings as an intermediate object in their functionality. An embedding is a vectorization of text designed to capture the contextual meaning of the text. In particular, embeddings are computed at the token level, meaning each word, part-of-word,

or punctuation mark is assigned its own embedding. The BERT embeddings used here have a length of 768 and are produced by the final layer of the BERT model. Since embeddings are token-specific, to form article-level embeddings I first average all the token-level embeddings in each article, following the standard procedure in the literature.

After generating BERT embeddings for all articles in the *WSJ* corpus, including both those with and without GPT-based labels, I train a simple regression-based model to predict GPT's responses based on the BERT embeddings. Let $e_{a,i} \in [0, 1, -1]$ represent GPT's response to article a for series i, where 0 corresponds to a response of uncertain, 1 a response of increase, and -1 a response of decreased. Then let $\mathbf{x}_a$ be the corresponding vector of article-level BERT embeddings (of length 768). I then run ridge regression to infill the remaining responses; that is I solve the following optimization problem for each series i:

$$\min_{\rho_i} ||e_{a,i} - \mathbf{x}_a \rho_i||_2^2 + \lambda_i ||\rho_i||_2^2. \tag{6}$$

To select the penalty term, λ_i, I use 10-fold cross-validation over the sample of articles for which I have GPT-based responses. Given the estimated $\hat{\rho}_i$, for each series, I can form an article-level BERT-based prediction as $\hat{e}_{a,i} = \mathbf{x}_a \hat{\rho}_i$.

To first evaluate whether the content of GPT's expectations is preserved by this simpler model, I compare the correlation between the aggregate GPT-based expectations and the corresponding aggregate BERT-based expectations. In both cases, I form the aggregate expectations as discussed in Section 3. For the BERT-based expectations, I use only the subset of articles for which I do not have GPT-based expectations.

Figure 10 reports the results of this comparison for all the series studied above. The blue lines and first two columns report results showing the correlation between the GPT-based expectations and the corresponding survey measures. The orange lines and second two columns report the correlation between the BERT-based expectations and the corresponding survey measures. The tD column reports the t-stat for the difference between the GPT and BERT-based correlations. Overall, the results suggest that the simpler supervised model captures much of the content of GPT's expectations. For the panel fit, I cannot reject the null that the two sets of correlations are equivalent, suggesting that the BERT-based expectations capture the relevant content of GPT's expectations.

Given my sample of *WSJ* articles starts in 1984, I next turn to the problem of extending my sample of expectations further back in time to provide sufficient coverage for identifying bubbles. To solve this problem, I introduce an additional corpus of articles taken from *The New York Times* (*NYT*). *NYT* was founded in 1851 and is widely considered the newspaper of record for the United States; as such it should provide a natural substitute for *WSJ* prior to 1984. While it is possible to recover a portion of content from *WSJ* as done in Manela and Moreira (2017) for instance, *NYT* provides a full API to directly access abstracts, headlines,

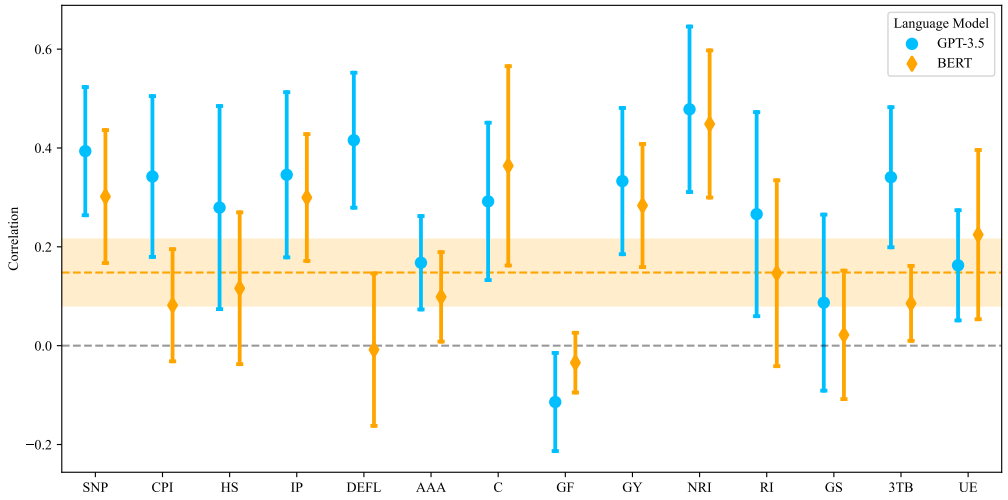
and summaries for all their published articles since 1851.

For this reason, prior to 1984, I use the *NYT* as my source of articles. A fuller discussion of the procedure used for forming the *NYT* corpus is available in Appendix I. How well do GPT's expectations transfer to the *NYT* corpus? To evaluate this, I sample 5,000 articles from *NYT* and generate GPT-based expectations comparable to those studied previously. Then using the BERT-based model estimated above, I form BERT-based expectations for all articles from *NYT* in my sample.

To evaluate the transfer performance, I compute the correlation between the article-level BERT-based expectations and the held-out article-level GPT-based expectations. Figure 11 reports the results of this comparison for all the series studied above. The grey lines and first two columns report the results for the *WSJ* corpus while the red lines and second two columns report the results for the *NYT* corpus. For *WSJ* to make the comparison fair, I hold back a sample of 5,000 articles and generate GPT-based expectations for these articles, as with *NYT*.

For all series across both corpuses, I find a strong positive correlation between the GPT and BERT-based expectations suggesting the simpler supervised model is able to capture much of the content of GPT's expectations. While I do find a significant drop in the correlation when transferring to the *NYT* corpus, this is often to be expected with transfer learning and the correlation remains significant for all series – suggesting it is still useful for infilling GPT's expectations. Overall, these results suggest that the simpler supervised model is able to capture much of the content of GPT's expectations. This has important practical implications for LLM-based expectations as it can greatly reduce the computational cost of generating expectations.

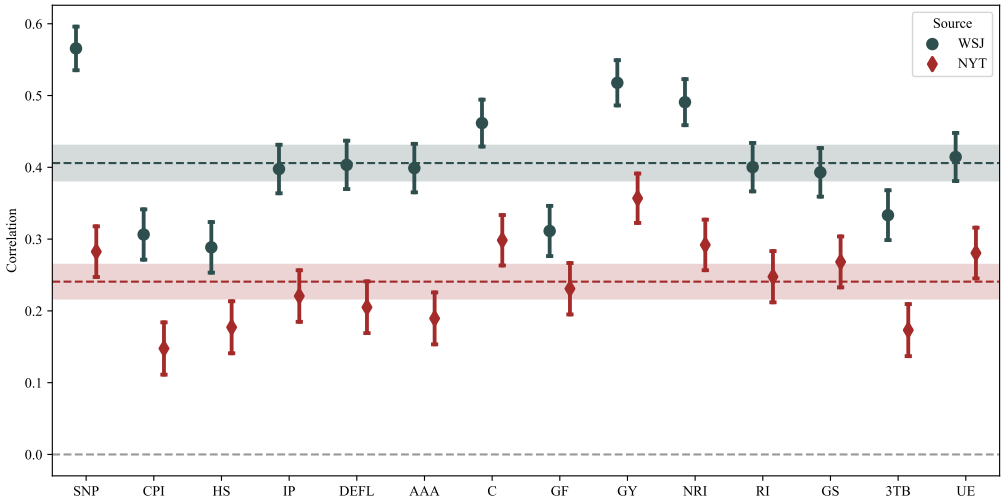
Figure 10: GPT vs. BERT Performance



	GPT-3.5		BERT		
	Corr.	t	Corr.	t	tD
Panel	0.27	5.33	0.15	3.60	1.85
S&P 500 Index	0.39	4.99	0.30	3.69	0.81
CPI	0.34	3.46	0.08	1.19	2.16
Housing Starts	0.28	2.24	0.12	1.24	1.05
Industrial Production	0.35	3.40	0.30	3.84	0.36
GDP Deflator	0.42	5.01	-0.01	-0.09	3.39
AAA Corporate Bond Yield	0.17	2.92	0.10	1.80	0.87
Consumption Growth	0.29	3.02	0.36	2.97	-0.46
Federal Government Spending	-0.11	-1.89	-0.03	-0.94	-1.13
GDP Growth	0.33	3.70	0.28	3.75	0.42
Nonresidential Investment	0.48	4.70	0.45	4.96	0.22
Residential Investment	0.27	2.12	0.15	1.28	0.70
State Government Spending	0.09	0.80	0.02	0.28	0.49
3-Month Treasury Bill	0.34	3.96	0.09	1.86	2.61
Unemployment Rate	0.16	2.40	0.22	2.16	-0.50

Note. The figure reports the correlation between the quarterly aggregate GPT-based expectations and corresponding survey measures against the correlation between the BERT-based expectations and corresponding survey measures for the *WSJ* corpus. The dashed yellow line reports the panel fit. 90% confidence intervals are reported for all correlations. Standard errors for the single variable regressions are Newey-West, the panel standard errors are Driscoll-Kraay. The bottom table reports the corresponding correlation coefficients and t-stats. tD corresponds to the t-stat for the sample difference between the GPT and BERT correlations.

Figure 11: WSJ vs. NYT Transfer Performance



	WSJ		NYT		
	Corr.	t	Corr.	t	tD
Panel	0.41	27.11	0.24	16.58	7.92
S&P 500 Index	0.57	30.66	0.28	13.16	10.00
CPI	0.31	14.39	0.15	6.67	5.17
Housing Starts	0.29	13.47	0.18	8.05	3.63
Industrial Production	0.40	19.37	0.22	10.11	5.91
GDP Deflator	0.40	19.70	0.21	9.37	6.61
AAA Corporate Bond Yield	0.40	19.45	0.19	8.63	6.97
Consumption Growth	0.46	23.26	0.30	13.97	5.60
Federal Government Spending	0.31	14.64	0.23	10.61	2.64
GDP Growth	0.52	27.05	0.36	17.08	5.68
Nonresidential Investment	0.49	25.18	0.29	13.64	6.87
Residential Investment	0.40	19.52	0.25	11.42	5.12
State Government Spending	0.39	19.10	0.27	12.45	4.19
3-Month Treasury Bill	0.33	15.80	0.17	7.86	5.25
Unemployment Rate	0.41	20.35	0.28	13.07	4.52

Note. The figure reports the article-level correlation between the BERT-based expectations and the corresponding held-out GPT expectations for both the *WSJ* and *NYT* corpuses. All regressions have a sample of 5,000 articles. Includes 90% confidence intervals. The dashed line reports the panel fit. The table below reports the corresponding correlations and t-stats, along with tD, which captures the sample difference between the *WSJ* and *NYT* correlations.

5.3 Economic Sentiment over 120 Years

5.3.1 Estimating Economic Sentiment

Having formed LLM-based expectations for all articles in the *WSJ* and *NYT* corpuses, extending back to 1900, I next turn to the problem of estimating sentiment, or extracting the irrational component of generated expectations. Following the argument in Section 5.1 I consider several alternative measures of sentiment. First, for both ex-ante and ex-post attempts to estimate μ_t, I collect a broad set of macroeconomic data from which to extract a set of latent factors capturing the state of the economy. As my principal dataset, I use the macroeconomic series available in the FRED-MD database (McCracken and Ng (2016)). To maximize the available time series for each series, I directly pull each series in FRED-MD from the FRED API itself. Additionally, to provide a more complete set of macroeconomic series prior to 1959 I also include a set of series from the NBER Macrohistory database. A fuller discussion of the data used and the detailed estimation procedure is available in Appendix J.

To estimate the number of latent factors used by both ex-ante and ex-post residual measures, I first fit PCA for a number of factor counts to the set of macroeconomic series using the expectation-maximization (EM) algorithm of Stock and Watson (2002) to account for the unbalanced nature of my panel. I report results for the explained variation ratio associated with each factor as well as the Akaike information criterion (AIC) in Figure J.1. The AIC results suggest a six factor model is optimal, which explains 42.9% of the variation in the data. A higher factor count would yield a marginal improvement explaining a cumulative 51.6% at 10 factors and 66.5% at 20 factors. Based on these results I use a six factor model for both the ex-ante and ex-post residual measures.

To estimate the ex-ante residual measure, I then estimate a dynamic factor model over the set of macroeconomic series using six factors and three lags. This model is estimated over an expanding sample starting in June of 1926 – just prior to the beginning of my sample of equity returns from CRSP – to ensure the estimates are available ex-ante. Once the latent factor forecasts have been formed, I then orthogonalize each expectation series with respect to the corresponding forecast and take the first principal component of the orthogonalized series. As with the dynamic factor model, I use an expanding sample of daily expectation residuals to run PCA and extract my sentiment measure. This expanding sample also starts in June of 1926 and is rolled forward one month at a time.

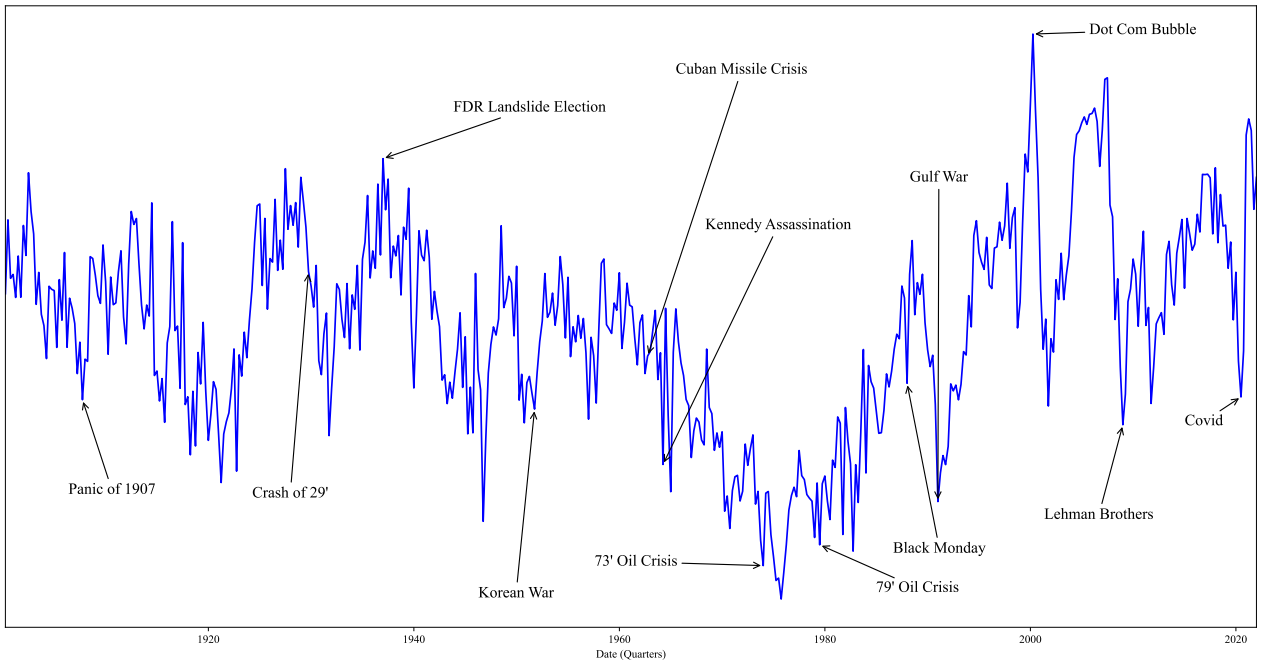
To estimate the ex-post residual measure, I extract a set of six latent factors from the set of macroeconomic series using PCA. I then orthogonalize each expectation series with respect to the corresponding factors and take the first principal component of the orthogonalized series. As with the ex-ante residual measure, I use an expanding sample of daily expectation residuals to form my sentiment estimate via PCA. Finally, for the first principal component of the expectation series itself, I similarly use an expanding sample of daily expectations to

estimate PCA. Figure J.2 reports the explained variation ratio associated with each factor for this procedure. The results suggest that there is a very strong first component explaining 85.1% of the variation in generated expectations.

5.3.2 The Dynamics of Economic Sentiment

Next, I turn to the dynamics of economic sentiment over the past 120 years and evaluate whether my measure captures the desired systematic errors and disconnect from rational expectations. Figure 12 reports the quarterly time series of the main sentiment measure, EAR. In addition to the time series, I have labeled several spikes in the sentiment series to provide a sense of the events that are captured and whether they correspond to sensible periods of positive or negative sentiment. The figure provides strong evidence in favor of my sentiment measure as various key events are captured. For instance, several negative spikes correspond to major negative economic events in the United States, such as the Panic of 1907, the Crash of 1929, 1973 and 1979 Oil crises, as well as the collapse of Lehman Brothers, and Covid. The positive spikes also seem sensible with events like the landslide presidential election of FDR in 1936 and the dot-com bubble.

Figure 12: Time Series of Economic Sentiment



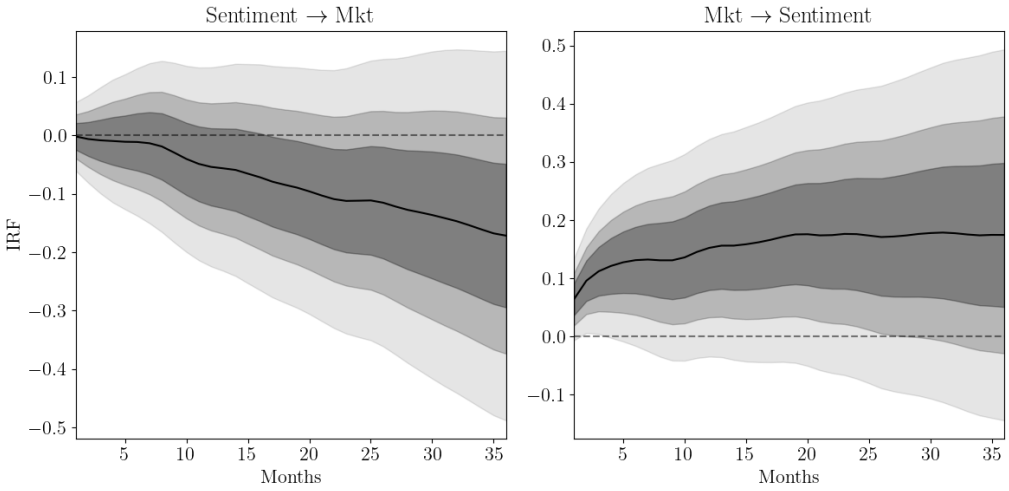
Note. Reports the quarterly time series of the economic sentiment measure using the first principal component of the ex-ante residuals (EAR) along with labels for key events.

Figure K.1 in the Appendix reports the principal sentiment measure, but overlays it with

several alternative measures for comparison. The first panel reports the main EAR measure along with EPR and EPC. All three measures exhibit comparable dynamics. Similarly, the second panel reports EAR overlaid with each of the individual expectation series; again, the dynamics are similar. These results suggest two key things. First, aligning with the results in Figure J.2, there is a strong common component to generated expectations. Second, much of the variation in the individual expectation series is explained by sentiment and not fundamental news.

How does economic sentiment relate to market returns? Section 4.1 showed that generated return expectations are extrapolative and negatively correlated with future returns. Do we observe similar behavior for economic sentiment over the past 120 years? To examine this relationship I form a series of impulse response functions (IRFs) capturing the response of value-weighted market returns to a one standard deviation shock to the sentiment measure and vice versa.

Figure 13: Sentiment and Aggregate Return IRFs



Note. The left panel reports the IRF capturing the response of value-weighted market returns to a one standard deviation shock to the EAR sentiment measure. The right panel reports the IRF capturing the response of the EAR sentiment measure to a one standard deviation shock to value-weighted market returns. 68%, 90%, and 99% confidence intervals are reported (with decreasing levels of shading). Standard errors are Newey-West with the corresponding horizon as the number of lags.

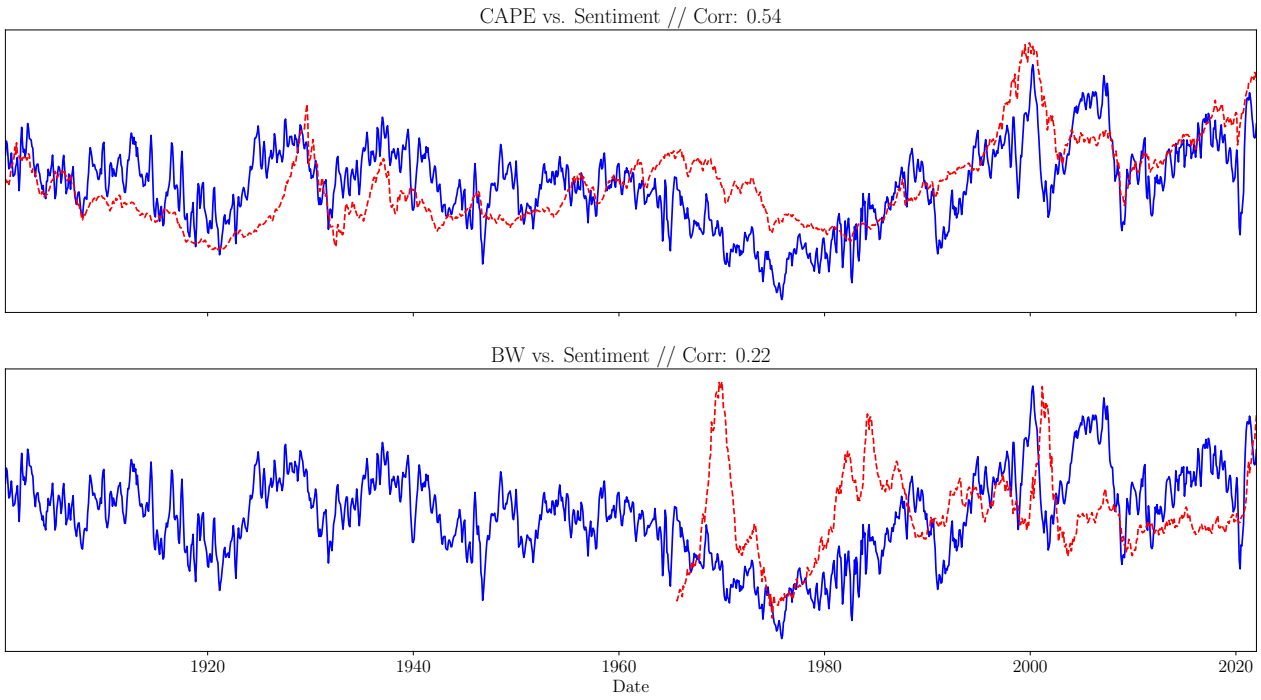
Figure 13 reports the results of this test. The left panel shows the response of the value-weighted market returns to a shock to the sentiment measure; a positive shock to sentiment is associated with negative returns in the future. While the results are only marginally significant, the sign is consistent with what would be expected from a measure of economic sentiment. Additionally, Figure K.2 reports comparable results when constrained to the initial sample period (1984-2021), for which I have *WSJ* articles. The negative correlation with

future returns is heightened over the most recent sample, with a negative correlation of 0.41 at the 24 month horizon for EAR (significant at the 99% level).

Similarly, the right panel of Figure 13 reports the response of the sentiment measure to a shock to value-weighted market returns. Here we see a significant positive relationship: economic sentiment responds positively to past returns. These results are significant out to the 25 month horizon at the 90% level, reaching a maximum correlation of 0.17 at the 19 month horizon.

Having then formed my measure of economic sentiment, how does it compare to similar measures considered previously in the literature? In particular, I compare my measure to the cyclically adjusted price-to-earnings ratio (CAPE) of Shiller (2015) and the sentiment measure developed by Baker and Wurgler (2007). Both measures attempt to capture a similar notion of disconnect from rational expectations, and as such, make reasonable benchmarks for my measure of economic sentiment.

Figure 14: Time Series of Sentiment against Benchmarks



Note. The first panel reports the EAR sentiment measure (the blue line) overlaid with CAPE (the dashed red line). The second panel reports the EAR sentiment measure (the blue line) overlaid with the Baker and Wurgler (2007) sentiment measure (the dashed red line).

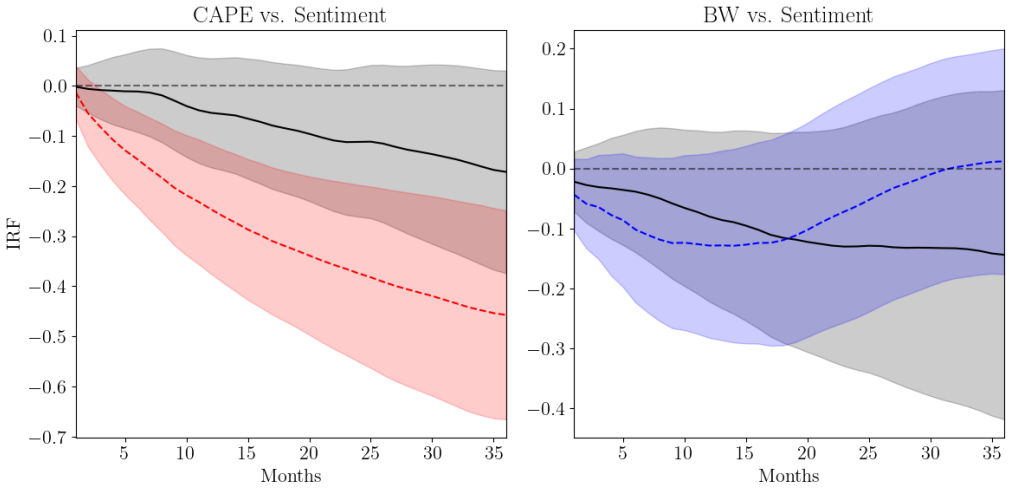
Figure 14 overlays my economic sentiment measure with both CAPE and the Baker and Wurgler (2007) sentiment measure. The first panel shows the EAR measure overlaid with CAPE – the two series have a correlation of 0.54 over the full sample. Both CAPE and

economic sentiment match many of the same spikes, with peaks in the lead-up to the 1929 crash and the dot-com bubble. Similarly, both measures capture negative shocks, such as Black Monday, the 2008 financial crisis, and Covid. On the other hand, the second panel plots the EAR measure against the Baker and Wurgler (2007) sentiment measure. Here the correlation is much lower at 0.22 and visual inspection suggests the two series are capturing different behavior – the Baker and Wurgler (2007) measure shows a substantial spike in the late 1960s and early 1970s and a spike lagging that of the EAR measure around the dot-com bubble.

How does the return predictability compare across measures? Figure 15 reports the IRFs capturing the response of value-weighted market returns to a shock to each measure. While CAPE is a stronger negative predictor of future aggregate returns, EAR matches the same sign and is comparable to the Baker and Wurgler (2007) measure in terms of predictability. Additional results for the alternative sentiment measures are presented in Figure K.4. Results are also presented for the post-1984 sample in Figure K.3. Over the more recent period the EAR measure performs comparably to CAPE and outperforms the Baker and Wurgler (2007) measure.

Take as a whole, these results suggest that my measure of economic sentiment does a reasonable job of capturing the desired irrationality. The EAR measure negatively forecasts future returns, suggesting that when economic sentiment is high the market is overvalued and future returns are likely to be low. Similarly, the EAR measure positively responds to past returns, suggesting that economic sentiment in part captures extrapolative expectations. Finally, the EAR measure exhibits similar dynamics to CAPE over the full sample and provides an additional source of sentiment variation relative to the Baker and Wurgler (2007) measure.

Figure 15: Sentiment and Aggregate Return IRFs vs. Benchmarks



Note. The left panel reports the IRFs capturing the response of value-weighted market returns to a shock to the EAR sentiment measure (the black line) vs. CAPE (the red line). The right panel reports the IRFs capturing the response of the EAR sentiment measure to a shock to value-weighted market returns (the black line) vs. the Baker and Wurgler (2007) sentiment measure (the red line). In both cases, the EAR IRF is computed over the same sample as the benchmark. 90% confidence intervals are reported, standard errors are Newey-West with the corresponding horizon as the number of lags.

6 Generated Beliefs and Bubbles

Given my extended sample of beliefs and the economic sentiment measure formed in Section 5, I next attempt to address one of the most challenging empirical problems in behavioral finance: the ex-ante identification of bubbles. Bubbles hold a singular place in finance folklore. Given the potential for misallocation of resources and resulting economic damage, this focus shouldn't be surprising (Brunnermeier and Oehmke (2013)). However, while ex-post stories and theories abound, evidence for the ex-ante identifiability of bubbles is limited. As Fama (2014) writes:

> *[D]efine a "bubble" as an irrational strong price increase that implies a predictable strong decline...the available research provides no reliable evidence that stock market price declines are ever predictable. Thus, at least as the literature now stands, confident statements about "bubbles" and what should be done about them are based on beliefs, not reliable evidence.*

Early theories of bubbles centered around rational expectations (Tirole (1985), DeMarzo et al. (2007)), yet empirical evidence, such as Giglio et al. (2016), is at odds with these stories. More recently, behavioral explanations have gained popularity (Barberis et al. (2018), Bordalo et al. (2021), Bastianello and Fontanier (2023)). These theories focus on expectations, and through expectations prices, deviating from rationality during bubble episodes.

A number of challenges exist for a rigorous empirical study of behavioral bubbles. Behavioral stories of bubbles focus on "mispricing" or the asset's price deviating from the full-information rational benchmark; as such, any attempt to identify behavioral bubbles requires a measure of mispricing. Additionally, bubbles are infrequent events so a large sample is required for sufficient power. I attempt to address both these challenges with generated beliefs. My extended sample provides me with the power to quantitatively identify bubble episodes. Additionally, with the high-frequency availability of my expectations, after first extracting a measure of "sentiment" or the non-rational component of generated expectations, I can compute localized rolling betas of an asset's return onto sentiment and use these betas as a proxy for mispricing.

6.1 Data on Bubbles

To test for bubble predictability I employ the data set used in Greenwood et al. (2019). Using the 49 industry classification of Fama and French (1997), they identify 40 run-ups in industry-level portfolios between 1926 and 2014. Run-ups are defined as industries that experience 100% returns over the past 24 months raw and in excess of the market. Additionally, the industries need to experience 50% returns over the past five years (to avoid capturing recoveries). Of

these, they identify 21 that "crash", defined as a 40% draw-down in the 24 months following the first month where the run-up is identified.

Why do I focus on this data set for my analysis? There are several benefits to focusing on industry-specific run-ups. First, many stories of bubbles focus on industry-specific run-ups, such as the dot-com bubble. Second, by studying industry-level portfolios I gain considerably more power than if I focused on the overall market while avoiding the liquidity and trading issues that may confound a study of individual stocks.

6.2 Identifying Mispricing with Economic Sentiment

Using the set of run-ups identified by Greenwood et al. (2019), the central challenge is to determine which run-ups exhibit mispricing and as a result are more likely to reverse and crash. Run-ups that exhibit mispricing should be less exposed to fundamental news and more exposed to sentiment. Given the deviations from FIRE demonstrated in Section 4 and the behavior of my economic sentiment measure introduced in Section 5, I argue that an asset's beta with respect to economic sentiment provides a measure of mispricing.

To formalize this intuition, I introduce a simple model of asset prices in continuous time. Let $dp_{t,j}$ represent the jth asset's returns, $df_{t,j}$, fundamental news about the asset, $d\delta_t$, economic sentiment, and $\beta_{t,j}$, the asset's time-varying exposure to sentiment. Then I can write the asset's returns as:

$$dp_{t,j} = df_{t,j} + \beta_{t,j} d\delta_t. \tag{7}$$

Under this model estimation of $\beta_{t,j}$ is straight-forward: given an estimate of economic sentiment, I can estimate $\beta_{t,j}$ by running rolling regressions of each asset's returns onto sentiment. This method leverages the high-frequency availability of my generated expectations and in turn economic sentiment to form time-varying estimates of an asset's exposure to sentiment.[18]

Is this a reasonable model of prices? By allowing for time-varying loadings on aggregate sentiment to determine the degree of an asset's mispricing, this model captures two key concepts. First, an asset's degree of mispricing varies over time and second, an asset's own sentiment is proportional to aggregate sentiment. Under the full-information rational expectations benchmark returns should exactly follow fundamental news. By allowing for time-varying loadings on aggregate sentiment, this model allows for mispricing to be transient and only impact a subset of assets at any given time. This is consistent with common stories of bubble episodes where investor enthusiasm is concentrated in a subset of assets, such as the technology sector during the dot-com bubble (Brunnermeier and Nagel (2004)).

By linking asset-specific sentiment to aggregate sentiment, this model captures the intuition that overall enthusiasm is an aggregate of sector-level enthusiasm. This relates to recent

[18] See Aït-Sahalia et al. (2020) for related work using high-frequency data to estimate factor betas. The centrality of infill asymptotics in this literature is also the reason for moving to continuous time in this section.

stories that think of aggregate variation as originating from granular shocks (Gabaix (2011)). In Appendix G I present an alternative model that similarly argues that an asset's beta with respect to aggregate sentiment can be used as a measure of mispricing if aggregate sentiment is assumed to be the average of asset-specific sentiment. In either case, if my assumptions here are incorrect, then this should attenuate the sentiment betas I observe and diminish their usefulness for predicting bubbles.

6.3 Mispricing and Bubbles

Having formed a daily measure of economic sentiment, I next turn to the problem of identifying mispricing and predicting bubbles. Following the model presented in Section 6.2, I identify mispricing by estimating an asset's rolling beta with respect to economic sentiment. More precisely, let $r_{\tau,i}$ represent the daily return of an industry indexed i, and δ_τ the daily measure of economic sentiment, I then run a series of rolling regressions:

$$r_{\tau,i} \sim \alpha_{t,i} + \beta_{t,i}\delta_\tau \tag{8}$$

for all $\tau \in D_t$ where D_t is a window of daily data ending in month t.

These regressions are estimated for all industry/month pairs for a variety of windows. I consider regressions using 3, 6, 9, 12, and 24 months of daily data. Figure L.1 in Appendix L plots the average time series of these betas for the 12-month window for the various sentiment measures considered. Overall, I find that the sentiment betas are positive and exhibit considerable variation over time – exhibiting a number of spikes during the Great Depression, the oil crises of the 1970s, and the Great Recession. Additionally, I find that sentiment betas are similar across the various sentiment measures considered, suggesting my results are not dependent on the specific measure of sentiment used. In addition to the main time series presented below, Figures L.2 and L.3 report summary statistics for the distribution of sentiment betas for various windows as well as broken down by industry.

Do sentiment betas actually capture mispricing? As an initial evaluation of this question, I compare stock-level sentiment betas to existing stock-level measures of mispricing. In particular, Stambaugh et al. (2012) introduces a mispricing index, formed from various firm characteristics hypothesized to be related to mispricing in earlier literature.[19] To evaluate whether my sentiment betas provide a reasonable measure of mispricing, I compare them to the mispricing index proposed by Stambaugh et al. (2012). When running a panel regression of sentiment betas onto the corresponding firm-month's mispricing index value, I find a positive correlation of 0.023, which when adjusting standard errors for heteroskedasticity and serial correlation results in a t-stat of 30.40, suggesting sentiment betas are significantly related to

[19]This measure has also been explored more thoroughly in Stambaugh and Yuan (2017).

existing measures of mispricing.

Having introduced my measure of mispricing, I next turn to my first principal set of results and try to address two questions: First, are run-ups that are more exposed to sentiment more likely to crash? Second, are run-ups that are more exposed to sentiment more likely to have lower future returns? For the set of run-ups identified by Greenwood et al. (2019), I present several summary statistics for the corresponding sentiment betas in Table 3. In particular, the final two columns report t-stats for a regression of an indicator for whether a run-up crashes under the definition in Greenwood et al. (2019), and the 24-month future returns on the corresponding sentiment beta, respectively. Both columns provide strong evidence in favor of the questions mentioned above. For all but the 24-month window, the probability of a crash (proxied by the crash indicator) is significantly higher for run-ups with higher sentiment betas, and the future returns are significantly lower at the 95% confidence level. In fact, for the 12-month window, both results are significant at the 99% confidence level.

Table 3: Run-up Beta Summary Statistics

	Full Sample		Crash		No Crash		Crash Ind.	24 M. Ret.
	Mean	SD	Mean	SD	Mean	SD	t	t
3-Month Beta	0.17	0.33	0.27	0.26	-0.03	0.33	3.92	-3.37
6-Month Beta	0.16	0.26	0.26	0.28	0.03	0.22	2.93	-2.66
9-Month Beta	0.16	0.23	0.22	0.21	0.03	0.16	3.40	-2.95
12-Month Beta	0.16	0.21	0.18	0.15	0.03	0.11	3.79	-4.04
24-Month Beta	0.15	0.18	0.16	0.13	0.09	0.12	1.44	-2.23

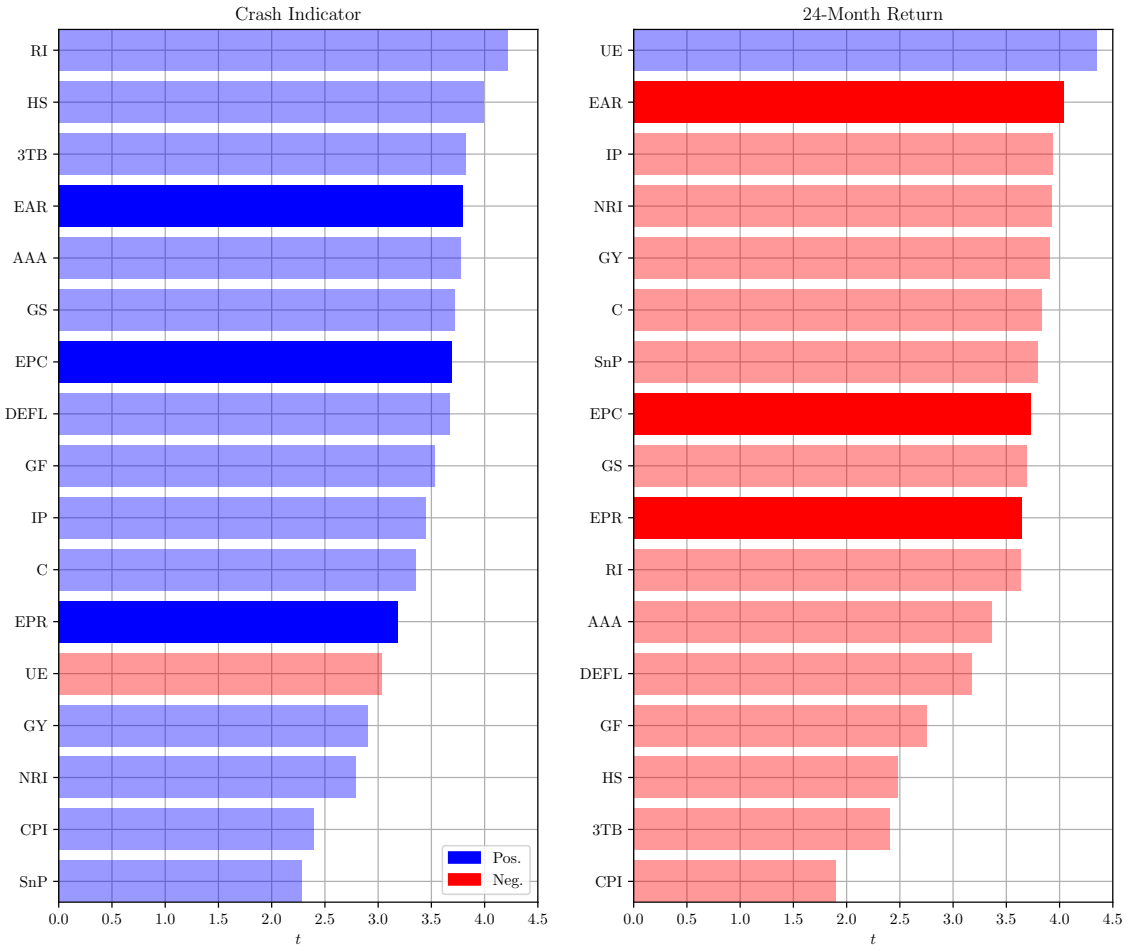
Note. Reports the mean and standard deviation for a series of sentiment beta windows for the full sample, the crash sample, and the no crash sample. Additionally, reports the t-stat for a regression of an indicator for whether a run-up crashed on the sentiment betas for the corresponding run-ups (Crash Ind.). Also, reports the t-stat for a regression of the 24-month future returns on the sentiment beta for the corresponding run-ups (24 M. Ret.). t-stat standard errors are clustered by industry year.

How do these results compare across my alternative measures of sentiment? Table N.1 in Appendix N reports the results comparable to Table 3 for my two alternative measures of sentiment. When using the first principal component of ex-post residuals (EPR), the t-stat for the crash indicator is 3.18 and -3.64 for the next 24 month's returns. When using first principal component of the original expectations as the sentiment measure, the t-stat is 3.69 for the crash indicator and -3.73 for the next 24 month's returns. Finally, I also consider these principal results for alternative principal components of generated expectations after the first one. Figure N.1 reports these results for the 12-month window and suggests that the other PCs do not have the same predictive power as the first.

To understand whether these results are driven by a particular expectation series, I next

produce comparable t-stats for betas computed with respect to each individual series of generated expectations. Figure 16 reports the results of this exercise using the 12-month window, along with the t-stats for each of the principal sentiment measures. Overall, I find that the results are similar across all series, with the strongest crash-indicator t-stat of 4.22 for residential investment, vs. 3.79 for EAR, and the lowest of 2.29 for the S&P 500, suggesting my results are not driven by variation in a particular series but instead the hypothesized common sentiment component. Interestingly, while most of the series studied can be viewed as positive for the economy, for unemployment the sign of the t-stats is reversed for both the crash indicator and future returns. This provides an elegant sanity check for my results, suggesting that the sentiment betas are moving in the expected direction.

Figure 16: t-stat for Sentiment Measures and Individual Expectation Series



Note. Reports the t-stats for the regression of the crash indicator and future 24-month returns on the corresponding sentiment beta. t-stat standard errors are clustered by industry year. The t-stats are reported for the main sentiment measures (EAR, EPR, EPC) and for each individual expectation series.

Beyond the core statistics presented here, Greenwood et al. (2019) consider several additional predictors that they find do have some predictability for crashes and future returns following run-ups. In particular, they find that when accounting for multiple testing, "acceleration" or the 24-month return minus the previous 12-month's return is a significant predictor of crashes and the one-year change in volatility is a significant predictor of future 24-month returns. A full summary of the variables considered in Greenwood et al. (2019) is available in Appendix M. In addition to the set of predictors they consider, I also include the corresponding 12-month market beta for each industry. A concern may be that sentiment exposure captures market exposure. If this is the case then including market beta should reduce the significance of the sentiment betas.[20]

Figure 17 reports these results for the 12-month window for both the crash indicator and the 24-month future returns. For both sets of regressions, I report a baseline specification, that only includes the sentiment beta, as well as a series of specifications individually including each of the Greenwood et al. (2019) controls as well as the 12-month market beta. In all cases, I find the sentiment beta remains a strong predictor of both crashes and future 24-month returns, with the cyclically-adjusted market price-earnings ratio (CAPE) of Shiller (2015) having the strongest effect.

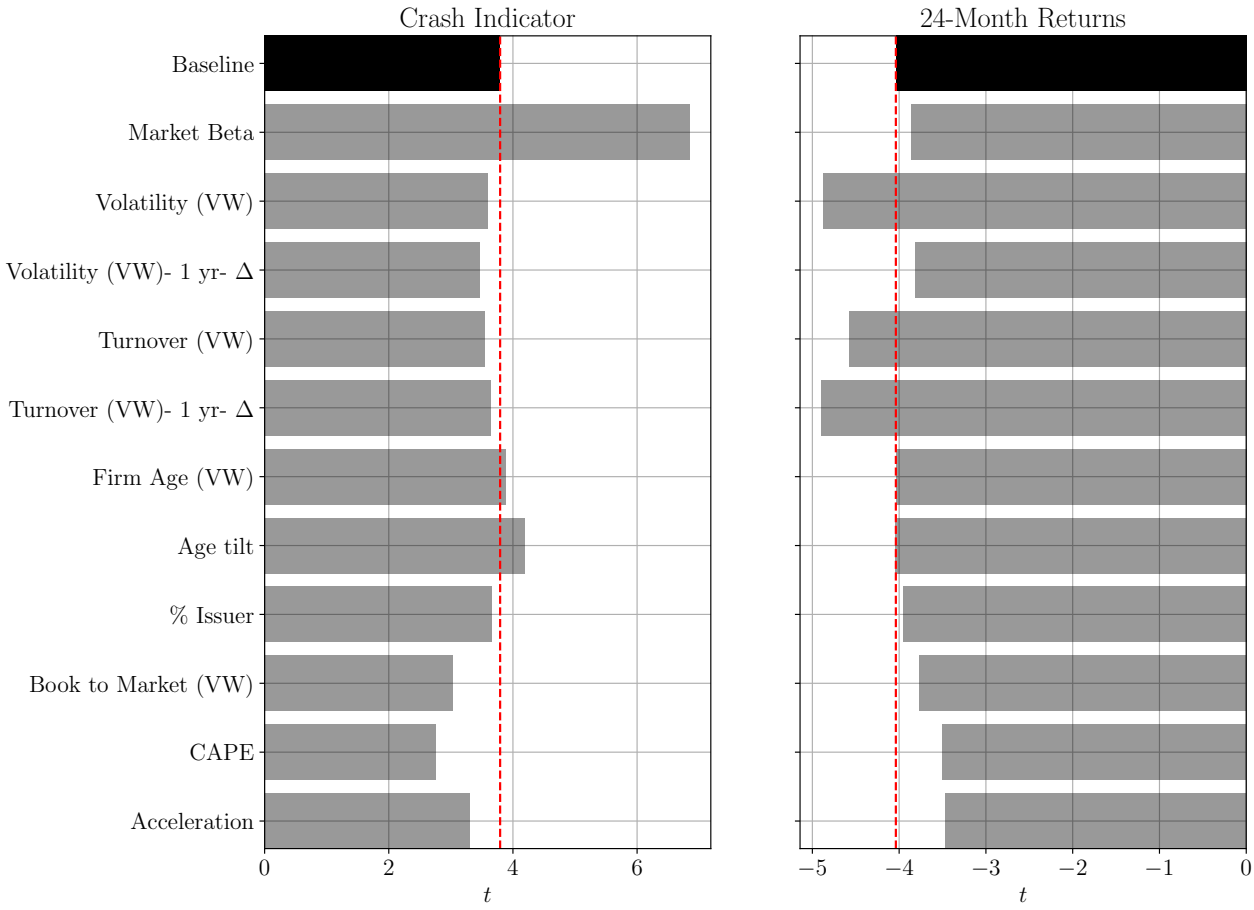
Again, are these results robust to alternative measures of sentiment? Table N.2 reports the results for my two alternative measures of sentiment. In both cases, the results are broadly similar to those reported here. When using the first principal component of ex-post residuals we see the lowest effect size when looking at crash predictability controlling for CAPE with a t-stat of 2.24 and -2.93 for future 24-month returns. For the first principal component of the original generated expectations results are event better.

Finally, up to this point, all the results I have considered have been in terms of raw returns. However, do my results change if I instead use excess returns? To do this I consider two new alternatives, an industry's returns in excess of the risk-free rate as well as an industry's returns in excess of the market. In both cases, I recompute the sentiment betas using the corresponding excess returns. Table O.1 in Appendix O reports these results for both alternative return measures. Results are largely unchanged when taking returns in excess of the risk-free rate, however, when taking returns in excess of the market, the results are somewhat weaker, though still broadly significant even when including controls from Greenwood et al. (2019).

Overall, these results provide strong evidence for mispricing during bubble episodes. An asset with higher exposure to economic sentiment during a run-up is more likely to experience a crash and lower future returns going forward. These results are robust to a variety of alternative measures of sentiment and are broadly unaffected by the inclusion of a variety of controls.

[20] Greenwood et al. (2019) also include sales growth which I exclude since it is not available for the entire sample, though results remain robust to its inclusion.

Figure 17: Run-up Beta Regressions with Controls



	Crash Indicator					24 Month Returns				
	SB		Ctrl			SB		Ctrl		
	Coef.	t	Coef.	t	R^2	Coef.	t	Coef.	t	R^2
Baseline	0.26	[3.79]			27.32	-0.51	[-4.04]			25.77
Market Beta	0.31	[6.85]	-0.13	[-1.63]	32.76	-0.52	[-3.86]	0.04	[0.39]	25.94
Volatility (VW)	0.26	[3.60]	-0.02	[-0.32]	27.48	-0.49	[-4.87]	0.09	[0.50]	26.59
Volatility (VW)- 1 yr- Δ	0.25	[3.47]	0.11	[2.26]	31.63	-0.49	[-3.82]	-0.11	[-1.31]	26.93
Turnover (VW)	0.25	[3.54]	-0.07	[-1.01]	29.24	-0.48	[-4.58]	0.10	[0.57]	26.71
Turnover (VW)- 1 yr- Δ	0.27	[3.64]	-0.04	[-0.67]	27.87	-0.49	[-4.90]	-0.06	[-0.32]	26.09
Firm Age (VW)	0.28	[3.88]	0.07	[0.97]	29.16	-0.57	[-4.04]	-0.22	[-2.05]	30.41
Age tilt	0.27	[4.19]	-0.03	[-0.36]	27.72	-0.49	[-4.04]	-0.16	[-1.07]	28.38
Book to Market (VW)	0.24	[3.03]	-0.07	[-0.89]	29.28	-0.54	[-3.77]	-0.11	[-0.80]	26.76
CAPE	0.23	[2.76]	0.01	[0.86]	28.39	-0.39	[-3.50]	-0.02	[-1.16]	28.80
Acceleration	0.24	[3.30]	0.19	[2.29]	41.12	-0.49	[-3.46]	-0.12	[-0.79]	27.18

Note. The first panel reports the t-stats for a regression of the crash indicator on the corresponding sentiment beta. In addition to the baseline fit without any controls, it also includes t-stats for the sentiment betas controlling for each of the variables from Greenwood et al. (2019) as well as the 12-month market beta. The table reports the corresponding regressions, including the correlation coefficients and R^2. The SB column reports the sentiment beta coefficients and the Ctrl column reports the coefficients for the corresponding control variable. For all results, standard errors are clustered by industry year.

6.4 Extrapolation, Feedback, and Bubbles

To this point, I have focused primarily on the relationship between an asset's deviation from fundamentals and bubbles, i.e. the relationship between mispricing and bubbles. However, popular stories of bubbles often involve a feedback loop between returns and expectations, with investors forming their expectations based on past returns and these heightened expectations in turn driving prices higher. From Shiller (2002):

> *The essence of a speculative bubble is the familiar feedback pattern-from price increases to increased investor enthusiasm to increased demand and, hence, to further price increases. The high demand for the asset is generated by the public's memory of high past returns and the optimism the high returns generate for the future.*

This notion of feedback has been formalized in recent behavioral models of bubbles which center investor return extrapolation as a key mechanism behind the rise and fall of bubbles (Barberis et al. (2018), Bastianello and Fontanier (2023)). In particular, Bastianello and Fontanier (2023) provides a micro-foundation for time-varying and heightened extrapolation during bubbles. I next attempt to test this notion of extrapolation and feedback within my framework.

Whereas to evaluate mispricing, I looked at the contemporaneous beta between an asset's return and economic sentiment, extrapolation would imply a positive relationship between *future* sentiment and returns. Similarly, if heightened sentiment leads to higher prices going forward then we should see a similar positive relationship between *future* returns and sentiment. To evaluate whether these relationships have relevance for bubbles, I follow a similar procedure as before: regressing returns on sentiment. However, I now shift the time alignment of the variables: regressing cumulative future returns on current sentiment and regressing cumulative future sentiment on current returns. An argument to formalize this procedure when using economic sentiment is presented in Appendix G.

This methodology can be understood within the framework of local projections from Jordà (2005). The resulting coefficients at various horizons estimate an impulse response function (IRF) capturing the impact of a shock to returns on future sentiment and a shock to sentiment on future returns. More precisely, I run the following regressions:

$$\sum_{h=1}^{H} \delta_{\tau+h} \sim \alpha^{\delta}_{t,i,H} + \eta^{\delta,r}_{t,i,H} r_{\tau,i} + \eta^{\delta,\delta}_{t,i,H} \delta_{\tau}, \tag{9}$$

where $\eta^{\delta,r}_{t,i,H}$ is the IRF for a shock to returns on sentiment, i.e. extrapolation, and

$$\sum_{h=1}^{H} r_{\tau+h} \sim \alpha^{r}_{t,i,H} + \eta^{r,r}_{t,i,H} r_{\tau,i} + \eta^{r,\delta}_{t,i,H} \delta_{\tau}, \tag{10}$$

where $\eta^{r,\delta}_{t,i,H}$ in turn captures the impact of a shock to sentiment on returns. These IRFs are computed locally in a manner comparable to the sentiment regressions above. In particular, I am interested in the difference between the IRFs during run-ups that crash and those that do not, i.e. the marginal IRF. Let $\bar{\eta}^{y,x}_H(\text{Crash})$ be the average IRF for a shock to x on y for horizon H for all run-ups that experienced a crash and $\bar{\eta}^{y,x}_H(\text{No Crash})$ the comparable average for run-ups that did not crash. I then define the marginal IRFs as:

$$\begin{aligned} \text{MIRF}^{\delta,r}_H &= \bar{\eta}^{\delta,r}_H(\text{Crash}) - \bar{\eta}^{\delta,r}_H(\text{No Crash}) \\ \text{MIRF}^{r,\delta}_H &= \bar{\eta}^{r,\delta}_H(\text{Crash}) - \bar{\eta}^{r,\delta}_H(\text{No Crash}). \end{aligned} \tag{11}$$

These marginal IRFs capture the potential for a feedback loop between returns and expectations during run-ups that crash. This approach is related to Cassella and Gulen (2018) who use a similar local regression methodology to study time-varying extrapolation.

Figure 18 reports summary statistics for the corresponding IRFs ($\hat{\eta}^{\delta,r}_{t,i,H}$, $\hat{\eta}^{r,\delta}_{t,i,H}$) and their marginal counterparts. The marginal IRFs suggest stark results: run-ups that crash are associated with a heightened impact of returns on sentiment and sentiment on returns. In particular, the marginal sentiment IRF is strongly positive across horizons with a max t-stat of 4.032 at the 30-day horizon. The marginal return IRF, while not significant, is also positive across all horizons with a max t-stat of 1.653. These results suggest that the feedback loop between returns and expectations is stronger during run-ups that crash, consistent with a story of bubbles centered around time-varying extrapolation.

Table P.1 in Appendix P reports the corresponding summary statistics for alternative measures of sentiment. Here too the results are consistent with the main sentiment measure: the marginal sentiment IRFs are strongly positive across horizons while the marginal return IRFs are positive but not significantly so. Overall, these results suggest extrapolation as a mechanism behind bubbles. Additionally, they lend evidence to the story of Bastianello and Fontanier (2023) and provide evidence for time variation in extrapolation during bubbles.

Figure 18: Run-up Extrapolation Summary Statistics

	Full Sample		Crash		No Crash		Crash−No Crash
	Mean	SD	Mean	SD	Mean	SD	t
A. Sentiment IRF							
10-Day Sentiment	0.040	0.268	0.031	0.122	-0.027	0.175	1.221
20-Day Sentiment	0.059	0.407	0.082	0.210	-0.170	0.273	3.290
30-Day Sentiment	0.058	0.531	0.154	0.312	-0.287	0.379	4.032
40-Day Sentiment	0.062	0.661	0.203	0.424	-0.305	0.480	3.555
50-Day Sentiment	0.110	0.766	0.272	0.642	-0.137	0.545	2.161
60-Day Sentiment	0.124	0.868	0.284	0.777	-0.047	0.602	1.494
B. Return IRF							
10-Day Return	-0.046	0.641	-0.027	0.531	-0.277	0.582	1.419
20-Day Return	-0.063	1.130	-0.139	1.260	-0.667	1.207	1.350
30-Day Return	-0.070	1.534	-0.044	1.449	-0.816	1.641	1.580
40-Day Return	-0.025	1.866	-0.082	1.389	-0.961	1.951	1.653
50-Day Return	-0.048	2.290	-0.248	1.699	-1.201	2.087	1.590
60-Day Return	-0.051	2.660	-0.500	1.797	-1.384	2.448	1.311

Note. The figures report the marginal IRFs for sentiment and returns with 68%, 90%, and 99% confidence intervals (with decreasing levels of shading). The table the mean and standard deviation for a series of sentiment and return IRFs using 12 months of daily data over various horizons for the full sample, the crash sample, and the no-crash sample. For the crash and no crash samples, the IRFs are computed using 12 months of daily data prior to the identified run-up point. Additionally, reports the t-stat for the sample difference of means between the crash and no crash run-ups. t-stat standard errors are clustered by industry year. Panel A reports the results for the IRF of sentiment on returns and panel B reports the results for the IRF of returns on sentiment.

6.5 Trading on Sentiment

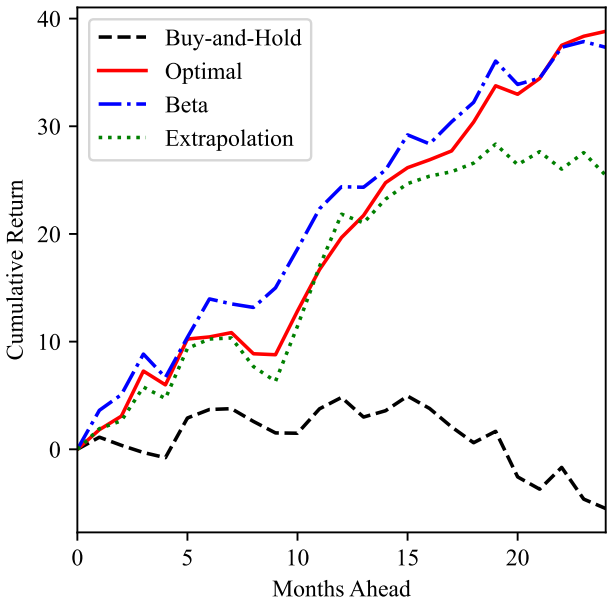
Finally, I have so far found that an asset's comovement with economic sentiment is predictive of crashes and future returns. However, is this predictability economically valuable? Is this something investors could leverage to help avoid crashes and increase returns? To test this I next form a trading strategy built around this predictability.

This strategy uses the sentiment betas and IRFs from earlier to identify run-ups likely to crash. For run-ups where the sentiment beta or IRF is below a given threshold, the strategy goes long the corresponding industry portfolio and funds this position by shorting the risk-free rate. I compute two separate strategies, one using the past 12-months' sentiment beta, the other the 30-day IRF of sentiment on shocks to returns. In both cases, the threshold is selected to maximize the proportion of correctly identified crashes.

Figure 19 reports the cumulative returns for these strategies, as well as the returns for a strategy that correctly identifies all crashes (Optimal), and one that holds all industry portfolios following a run-up (Buy-and-Hold). Using sentiment betas can achieve a very promising mean return of 1.493% vs. 1.553% as a strategy that correctly identifies all crashes, resulting in a cumulative return of 39.3%. While the extrapolation-based strategy – using the 30-day sentiment IRF – underperforms the sentiment beta strategy, it still outperforms the Buy-and-Hold strategy by 1.238 percentage points per month. Additionally, the sentiment beta strategy correctly identifies 77.5% of the crashes while the extrapolation strategy correctly identifies 80%.

In addition to an optimal threshold, I also consider a series of strategies using the mean value of the corresponding predictor as the threshold. Table Q.1 in Appendix Q reports the results for these strategies. The table also includes results for the other sentiment measures, as well as results in excess of the market. The sentiment beta strategy is robust to all these specifications suggesting that there is considerable economic gain to be had from leveraging this predictability.

Figure 19: Trading Performance



	Buy-and-Hold	Optimal	Beta Timing	Extrap. Timing
Mean	-0.220	1.553	1.493	1.018
SD	1.710	1.762	1.934	2.307
t	-0.643	4.407	3.861	2.205
Success Prop	0.475	1.000	0.775	0.800
Alpha t		5.392	5.693	3.074

Note. The first panel reports the cumulative excess returns for a series of strategies following each run-up identified by Greenwood et al. (2019). The second panel reports summary statistics for the corresponding strategies, including the mean, standard deviation, t-stat, success proportion or the number of successfully identified crashes following the corresponding strategy, and alpha t with respect to the "Buy-and-Hold" strategy. The "Buy-and-Hold" strategy holds each run-up for the full 24 months. The "Optimal" strategy holds only the run-ups that do not crash ex-post. The "Beta Timing" strategy holds only the run-ups that have a beta below the optimal threshold. The "Extrap. Timing" strategy holds only the run-ups that have a sentiment IRF over 30 days below the optimal threshold.

7 Conclusion

In this paper, I introduce a methodology to generate beliefs using large language models. I find that the resulting expectations exhibit many of the same deviations from full-information rational expectations as their current survey counterparts. These results provide new evidence to help guide the development of models of expectation formation. My method provides the ability to generate expectations wherever a textual representation of the world is available, which I leverage to form an extended time series of expectations from 1900 to 2021 available at a daily frequency. From these expectations, I extract a measure of economic sentiment capturing the irrational component of beliefs over the last 120 years. As a first application of this methodology and measure, I study how exposure to economic sentiment captures mispricing and predicts bubbles ex-ante.

References

Aher, Gati, Rosa I Arriaga, and Adam Tauman Kalai, 2022, Using large language models to simulate multiple humans, *arXiv preprint arXiv:2208.10264* .

Aït-Sahalia, Yacine, Ilze Kalnina, and Dacheng Xiu, 2020, High-frequency factor models and regressions, *J. Econometrics* 216, 86–105.

Alekseev, Georgij, Stefano Giglio, Quinn Maingi, Julia Selgrad, and Johannes Stroebel, 2022, A Quantity-Based Approach to Constructing Climate Risk Hedge Portfolios, *NBER* .

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart, 2021, Narratives about the macroeconomy .

Andre, Peter, Philipp Schirmer, and Johannes Wohlfart, 2024, Mental Models of the Stock Market.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas, 2020, Business-Cycle Anatomy, *Am. Econ. Rev.* 110, 3030–70.

Angeletos, George-Marios, Zhen Huo, and Karthik A. Sastry, 2021, Imperfect Macroeconomic Expectations: Evidence and Theory, *NBER Macroeconomics Annual* .

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate, 2023, Out of One, Many: Using Language Models to Simulate Human Samples, *Political Analysis* 1–15.

Baker, Malcolm, and Jeffrey Wurgler, 2007, Investor Sentiment in the Stock Market, *J. Econ. Perspect.* 21, 129–152.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer, 2015, X-CAPM: An extrapolative capital asset pricing model, *Journal of Financial Economics* 115, 1–24.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer, 2018, Extrapolation and bubbles, *Journal of Financial Economics* 129, 203–227.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics* 49, 307–343.

Barsky, Robert B, and J Bradford De Long, 1993, Why does the stock market fluctuate?, *The Quarterly Journal of Economics* 108, 291–311.

Bastianello, Francesca, and Paul Fontanier, 2022, Expectations and learning from prices, *Available at SSRN* .

Bastianello, Francesca, and Paul Fontanier, 2023, Partial equilibrium thinking, extrapolation, and bubbles .

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, 2021, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? xn–cc6c, in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623 (Association for Computing Machinery, New York, NY, USA).

Bernanke, Ben S., 2020, The New Tools of Monetary Policy, *Am. Econ. Rev.* 110, 943–83.

Bianchi, Francesco, Sydney C. Ludvigson, and Sai Ma, 2022, Belief Distortions and Macroeconomic Fluctuations, *Am. Econ. Rev.* 112, 2269–2315.

Blank, Michael, Spencer Kwon, and Johnny Tang, 2023, Investor composition and overreaction .

Bordalo, Pedro, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer, 2021, Diagnostic bubbles, *Journal of Financial Economics* 141, 1060–1077.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer, 2020a, Overreaction in macroeconomic expectations, *American Economic Review* 110, 2748–82.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, 2020b, Memory, attention, and choice, *The Quarterly journal of economics* 135, 1399–1442.

Brand, James, Ayelet Israeli, and Donald Ngwe, 2023, Using gpt for market research, *Available at SSRN 4395751* .

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

Brunnermeier, Markus, Emmanuel Farhi, Ralph S. J. Koijen, Arvind Krishnamurthy, Sydney C. Ludvigson, Hanno Lustig, Stefan Nagel, and Monika Piazzesi, 2021, Review Article: Perspectives on the Future of Asset Pricing, *Rev. Financ. Stud.* 34, 2126–2160.

Brunnermeier, Markus K., and Stefan Nagel, 2004, Hedge Funds and the Technology Bubble, *Journal of Finance* 59, 2013–2040.

Brunnermeier, Markus K, and Martin Oehmke, 2013, Bubbles, financial crises, and systemic risk, *Handbook of the Economics of Finance* 2, 1221–1288.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, 2023, Sparks of Artificial General Intelligence: Early experiments with GPT-4, *arXiv* .

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu, 2021, Business news and business cycles .

Bybee, Leland, Bryan T Kelly, and Yinan Su, 2022, Narrative asset pricing: Interpretable systematic risk factors from news text, *Johns Hopkins Carey Business School Research Paper* .

Cassella, Stefano, and Huseyin Gulen, 2018, Extrapolation Bias and the Predictability of Stock Returns by Price-Scaled Variables, *Rev. Financ. Stud.* 31, 4345–4397.

Chinco, Alex, 2022, The Ex Ante Likelihood of Bubbles, *Manage. Sci.* .

Cochrane, John H., 2011, Presidential Address: Discount Rates, *Journal of Finance* 66, 1047–1108.

Cochrane, John H., 2017, Macro-Finance, *Rev. Financ.* 21, 945–985.

Coibion, Olivier, and Yuriy Gorodnichenko, 2012, What Can Survey Forecasts Tell Us about Information Rigidities?, *Journal of Political Economy* .

Coibion, Olivier, and Yuriy Gorodnichenko, 2015, Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts, *Am. Econ. Rev.* 105, 2644–78.

Coibion, Olivier, Yuriy Gorodnichenko, and Saten Kumar, 2018, How Do Firms Form Their Expectations? New Survey Evidence, *Am. Econ. Rev.* 108, 2671–2713.

Cutler, David M, James M Poterba, Lawrence H Summers, et al., 1990, Speculative dynamics and the role of feedback traders, *American Economic Review* 80, 63–68.

DeLong, Shleifer, Summers, and Waldmann, 1990, Noise Trader Risk in Financial Markets on JSTOR, *Journal of Political Economy* 98, 703–738.

DeMarzo, Peter, Ron Kaniel, and Ilan Kremer, 2007, Technological innovation and real investment booms and busts, *Journal of Financial Economics* 85, 735–754.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv* .

Engelberg, Joseph, R. David Mclean, and Jeffrey Pontiff, 2018, Anomalies and News, *THE JOURNAL OF FINANCE* 73, 1971–2001.

Fama, Eugene F, 2014, Two pillars of asset pricing, *American Economic Review* 104, 1467–1485.

Fama, Eugene F., and Kenneth R. French, 1989, Business conditions and expected returns on stocks and bonds, *Journal of Financial Economics* 25, 23–49.

Fama, Eugene F., and Kenneth R. French, 1997, Industry costs of equity, *Journal of Financial Economics* 43, 153–193.

Farmer, Leland, Emi Nakamura, and Jón Steinsson, 2021, Learning about the long run, Technical report, National Bureau of Economic Research.

Flynn, Joel P, and Karthik Sastry, 2022, The macroeconomics of narratives, *Available at SSRN 4140751* .

Gabaix, Xavier, 2011, The Granular Origins of Aggregate Fluctuations, *Econometrica* 79, 733–772.

Gabaix, Xavier, and Ralph S. J. Koijen, 2021, In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis, *NBER* .

Giglio, Stefano, Matteo Maggiori, and Johannes Stroebel, 2016, No-Bubble Condition: Model-Free Tests in Housing Markets, *Econometrica* 84, 1047–1091.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, 2019, Five Facts about Beliefs and Portfolios, *NBER* .

Gilboa, Itzhak, and David Schmeidler, 1995, Case-based decision theory, *The quarterly Journal of economics* 110, 605–639.

Greenwood, Robin, and Andrei Shleifer, 2014, Expectations of Returns and Expected Returns, *Review of Financial Studies* 27.

Greenwood, Robin, Andrei Shleifer, and Yang You, 2019, Bubbles for Fama, *Journal of Financial Economics* 131, 20–43.

Haaland, Ingar K., Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart, 2024, Understanding Economic Behavior Using Open-ended Survey Data, *NBER* .

Harrison, and Kreps, 1978, Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations on JSTOR.

Hong, Harrison, and Jeremy C. Stein, 2003, Differences of Opinion, Short-Sales Constraints, and Market Crashes, *Rev. Financ. Stud.* 16, 487–525.

Horton, John J., 2023, Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, *NBER* .

Jin, Lawrence J, and Pengfei Sui, 2022, Asset pricing with return extrapolation, *Journal of Financial Economics* 145, 273–295.

Jordà, Òscar, 2005, Estimation and Inference of Impulse Responses by Local Projections, *Am. Econ. Rev.* 95, 161–182.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting Factor Models, *Journal of Finance* 73, 1183–1223.

La Porta, Rafael, 1996, Expectations and the Cross-Section of Stock Returns, *Journal of Finance* 51, 1715–1742.

Lakonishok, Josef, Andrei Shleifer, and Robert W Vishny, 1994, Contrarian investment, extrapolation, and risk, *The journal of finance* 49, 1541–1578.

Lettau, Martin, and Sydney Ludvigson, 2001, Consumption, Aggregate Wealth, and Expected Stock Returns, *Journal of Finance* 56, 815–849.

Lochstoer, Lars A., and Tyler Muir, 2022, Volatility Expectations and Returns, *Journal of Finance* 77, 1055–1096.

Lucas, Robert E., 1977, Understanding business cycles, *Carnegie-Rochester Conference Series on Public Policy* 5, 7–29.

Malmendier, Ulrike, and Jessica A Wachter, 2021, Memory of past experiences and economic decisions, *Available at SSRN 4013583* .

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

McCracken, Michael W, and Serena Ng, 2016, Fred-md: A monthly database for macroeconomic research, *Journal of Business & Economic Statistics* 34, 574–589.

Mullainathan, Sendhil, 2002, A memory-based model of bounded rationality, *The Quarterly Journal of Economics* 117, 735–774.

Nagel, Stefan, and Zhengyang Xu, 2022a, Asset Pricing with Fading Memory, *Rev. Financ. Stud.* 35, 2190–2245.

Nagel, Stefan, and Zhengyang Xu, 2022b, Dynamics of Subjective Risk Premia, *NBER* .

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., 2018, Improving language understanding by generative pre-training .

Sargent, Thomas J, 2001, *The conquest of American inflation* (Princeton University Press).

Schramowski, Patrick, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting, 2022, Large pre-trained language models contain human-like biases of what is right and wrong to do, *Nat. Mach. Intell.* 4, 258–268.

Shiller, Robert J, 2002, Bubbles, human judgment, and expert opinion, *Financial Analysts Journal* 58, 18–26.

Shiller, Robert J, 2015, Irrational exuberance, in *Irrational exuberance* (Princeton university press).

Shiller, Robert J., 2019, *Narrative Economics* (Princeton University Press, Princeton, NJ, USA).

Shleifer, Andrei, and Lawrence H. Summers, 1990, The Noise Trader Approach to Finance, *J. Econ. Perspect.* 4, 19–33.

Shleifer, Andrei, and Robert W. Vishny, 1997, The Limits of Arbitrage, *Journal of Finance* 52, 35–55.

Sims, Christopher A, 1980, Macroeconomics and reality, *Econometrica: journal of the Econometric Society* 1–48.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan, 2012, The short of it: Investor sentiment and anomalies, *Journal of Financial Economics* 104, 288–302.

Stambaugh, Robert F., and Yu Yuan, 2017, Mispricing Factors, *Rev. Financ. Stud.* 30, 1270–1315.

Stock, James H, and Mark W Watson, 2002, Macroeconomic forecasting using diffusion indexes, *Journal of Business & Economic Statistics* 20, 147–162.

Stock, James H, and Mark W Watson, 2011, Dynamic factor models .

Tirole, Jean, 1985, Asset Bubbles and Overlapping Generations on JSTOR.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample, 2023, LLaMA: Open and Efficient Foundation Language Models, *arXiv* .

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira, 2023, Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases, *Rev. Financ. Stud.* 36, 2361–2396.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, 2017, Attention Is All You Need, *arXiv* .

Wachter, Jessica A, and Michael Jacob Kahana, 2019, A retrieved-context theory of financial decisions, Technical report, National Bureau of Economic Research.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus, 2022, Emergent Abilities of Large Language Models, *arXiv* .

Welch, Ivo, and Amit Goyal, 2008, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, *Rev. Financ. Stud.* 21, 1455–1508.

Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann, 2023, BloombergGPT: A Large Language Model for Finance, *arXiv* .

A Constructing the *WSJ* Samples

A.1 Constructing the *WSJ* 1984-2021 Sample

I conduct data processing steps in the following order:

1. Remove all articles prior to January 1984 and after December 2021.
2. Exclude articles with page-citation tags corresponding to any sections other than A, B, C, or missing.
3. Exclude articles corresponding to weekends.
4. Exclude articles with subject tags associated with obviously non-economic content such as sports. List of exclusions available from author on request.
5. Exclude articles with certain headline patterns (such as those associated with data tables or those corresponding to regular sports, leisure, or books columns). List of exclusions available from author on request.
6. Exclude articles with less than 100 words.
7. Exclude articles with headlines less than 10 words.
8. Sample 300 articles for each month of data.

A.2 Constructing the *WSJ* Oct 2021-Mar 2023 Sample

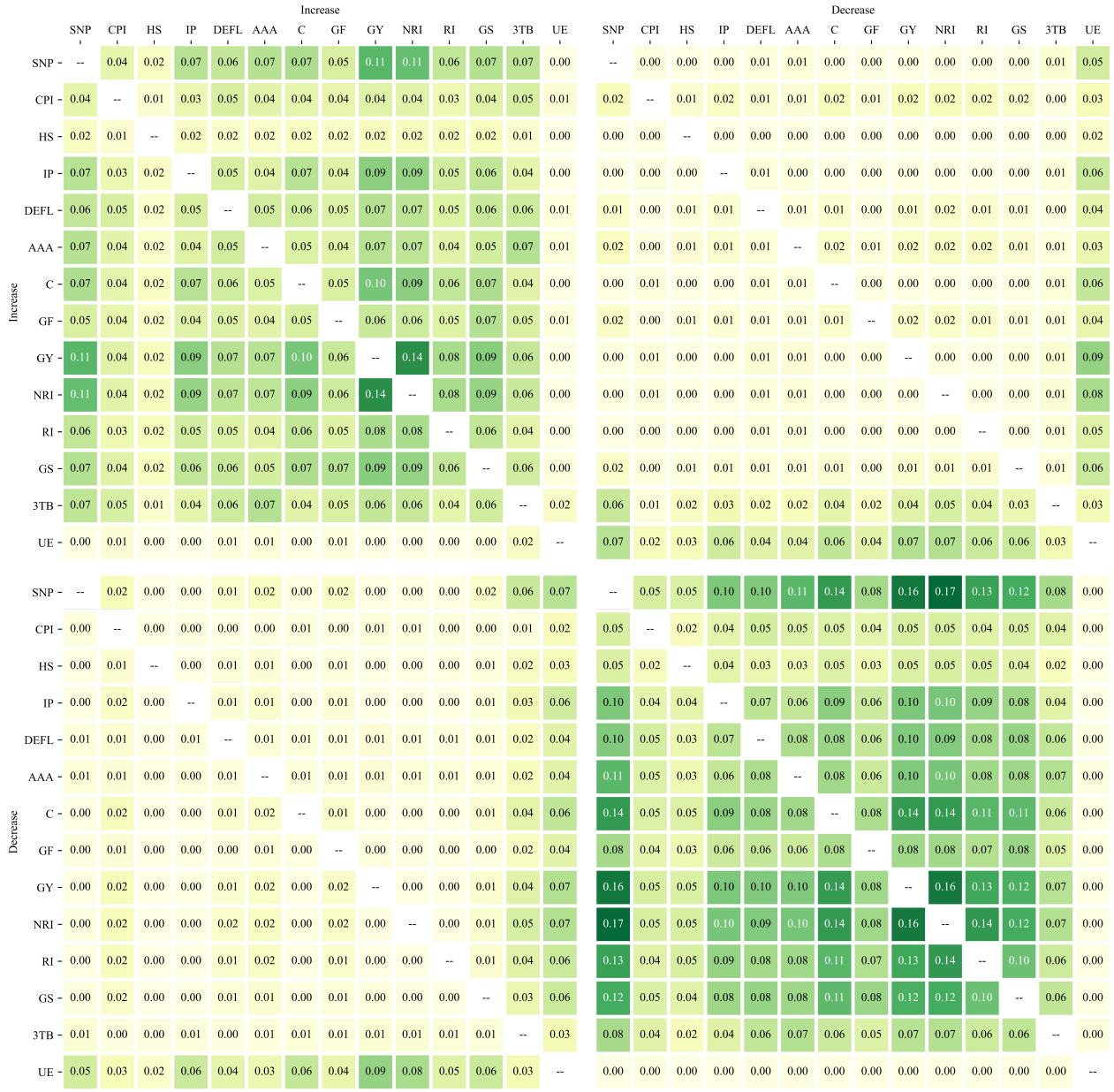
I conduct the following data processing steps in the following order:

1. Exclude articles with headlines less than 8 words.
2. Exclude articles not in the following sections: "U.S", "BUSINESS", "WORLD", "POLITICS", "WSJ NEWS EXCLUSIVE", "HEARD ON THE STREET", "FINANCE", "EARNINGS", "MARKETS", "U.S. MARKETS", "U.S. ECONOMY", "ECONOMY".

B Coocurrence of GPT Responses

This section reports the coocurrence between the article-level increase/decrease responses for the GPT expectations series studied.

Figure B.1: Coocurrence Matrix



Note. Reports the cooccurence proportion of each indicator for different surveyed series.

C DAG Generation Prompts

This section reports the prompts used to convert the generated explanations into DAGs. Figure C.1 documents how the DAG structure is initially identified, Figure C.2 documents how the initial classification of categories is assembled and Figure C.3 documents how the categories are ultimately applied to the DAGs.

Figure C.1: DAG Prompt Format

```
Can you identify the causal statement in this sentences explaining an %s in %s:

%s

Write your answers as (each part should be less than 5 words):

{cause}->{intermediate cause (optional)}->{effect}
```

Note. Reports the promtp used to identify categories from a list of DAG nodes.

Figure C.2: Category Identification Prompt Format

```
Can you create %s distinct topics to organize the following terms:

%s

Please format your response as:

{number}:{category}
```

Note. Reports the prompt used to categorize each DAG node based on the predefined list of categories.

Figure C.3: Category Application Prompt Format

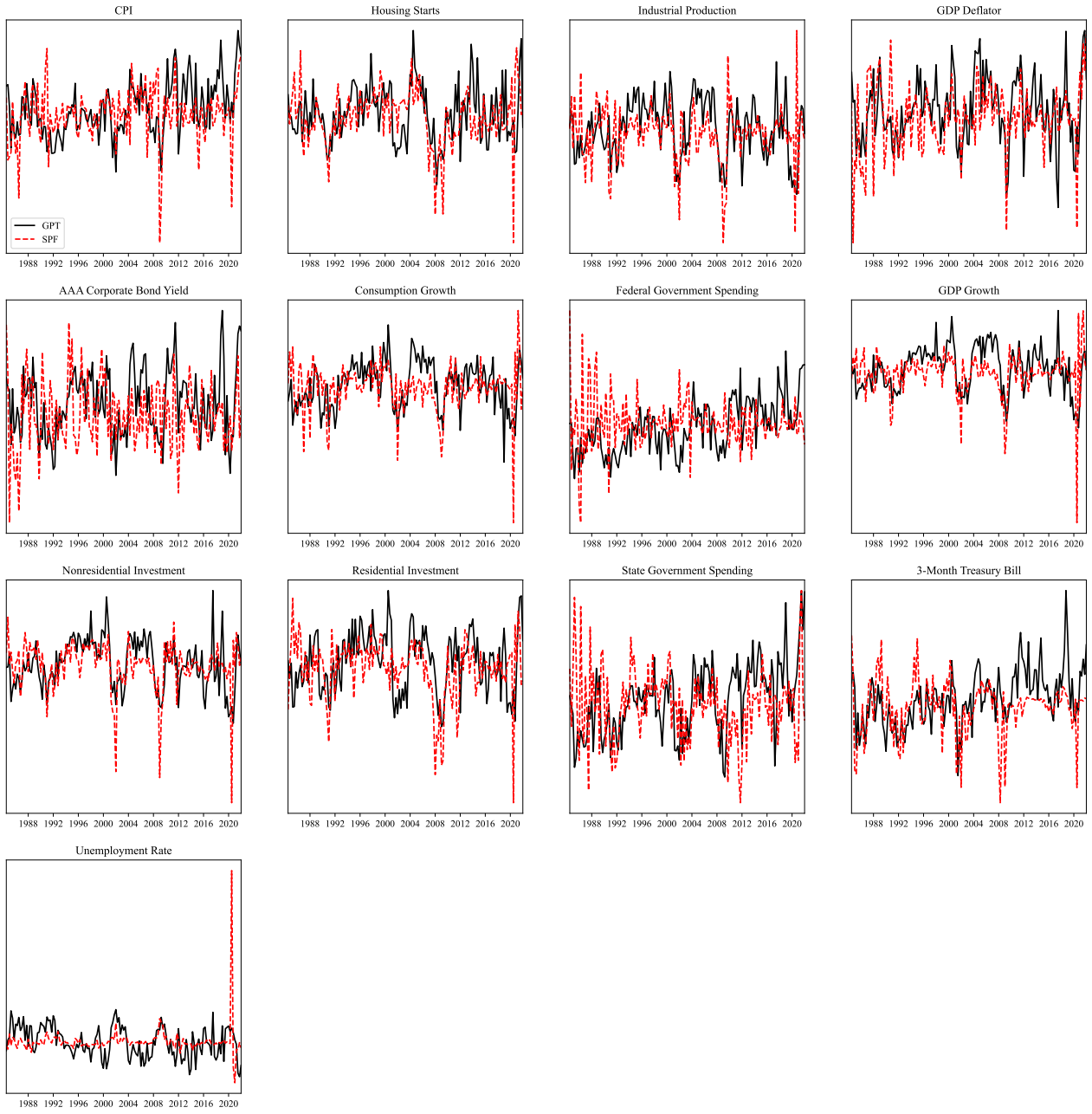
```
Here are a set of categories:

%s

Please assign the following phrase to one of these categories:

"%s"

Write your answer as:

{category}
```

Note. Reports the prompt used to convert the LLM generated explations to directed acyclic graphs. The "%s"'s indicate "increase/decrease", "the S&P 500" and the explanation respectively.

D Time Series of GPT/SPF Expectations

This section reports time series plots of generated expectations and average revisions for the SPF series studied in Section 4.3.

Figure D.1: Time Series of GPT Expectations and SPF Revisions

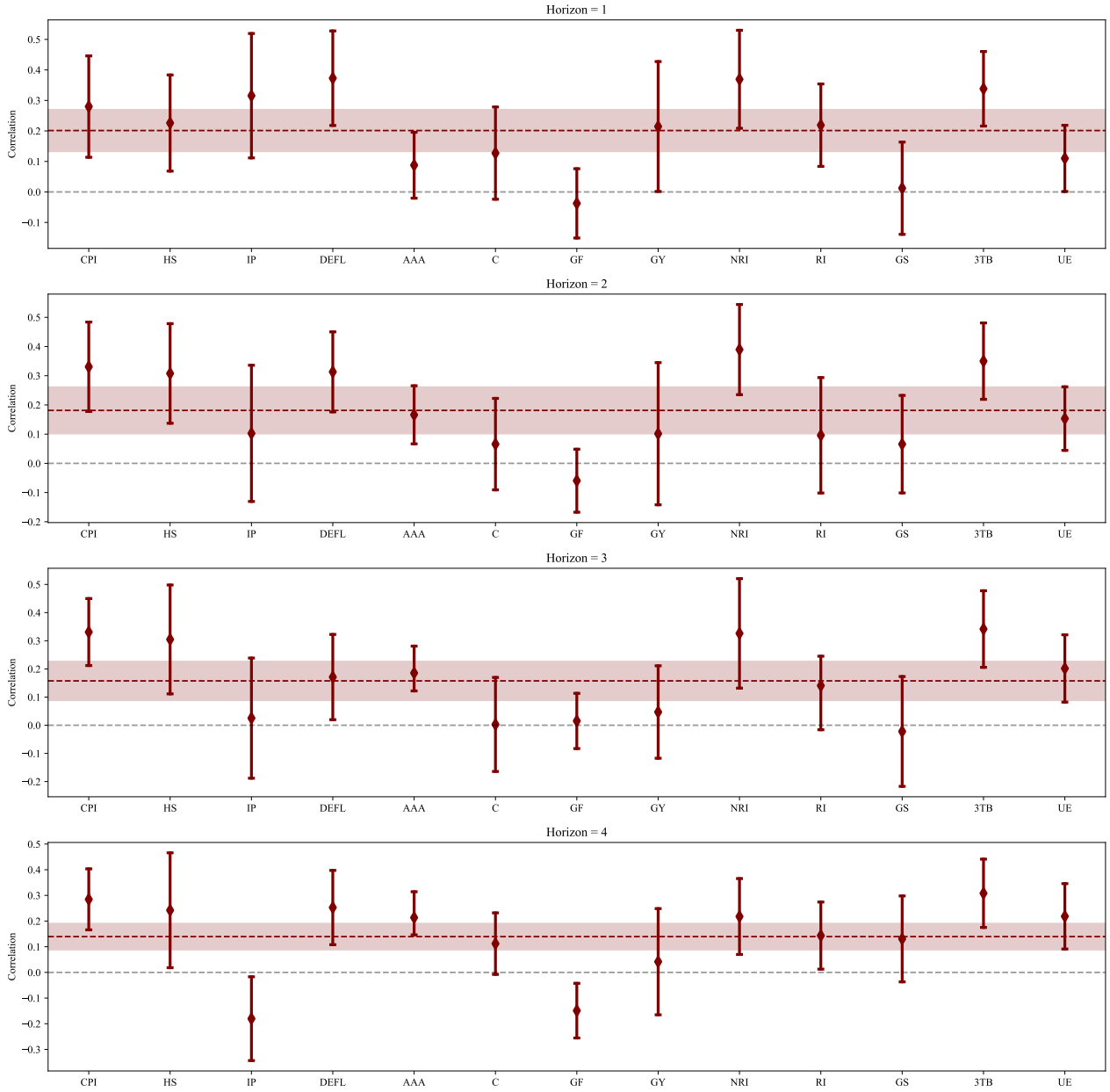


Note. Reports the time series of GPT expectations (black) and corresponding SPF revisions (red).

E GPT/SPF Correlations Over Different Horizons

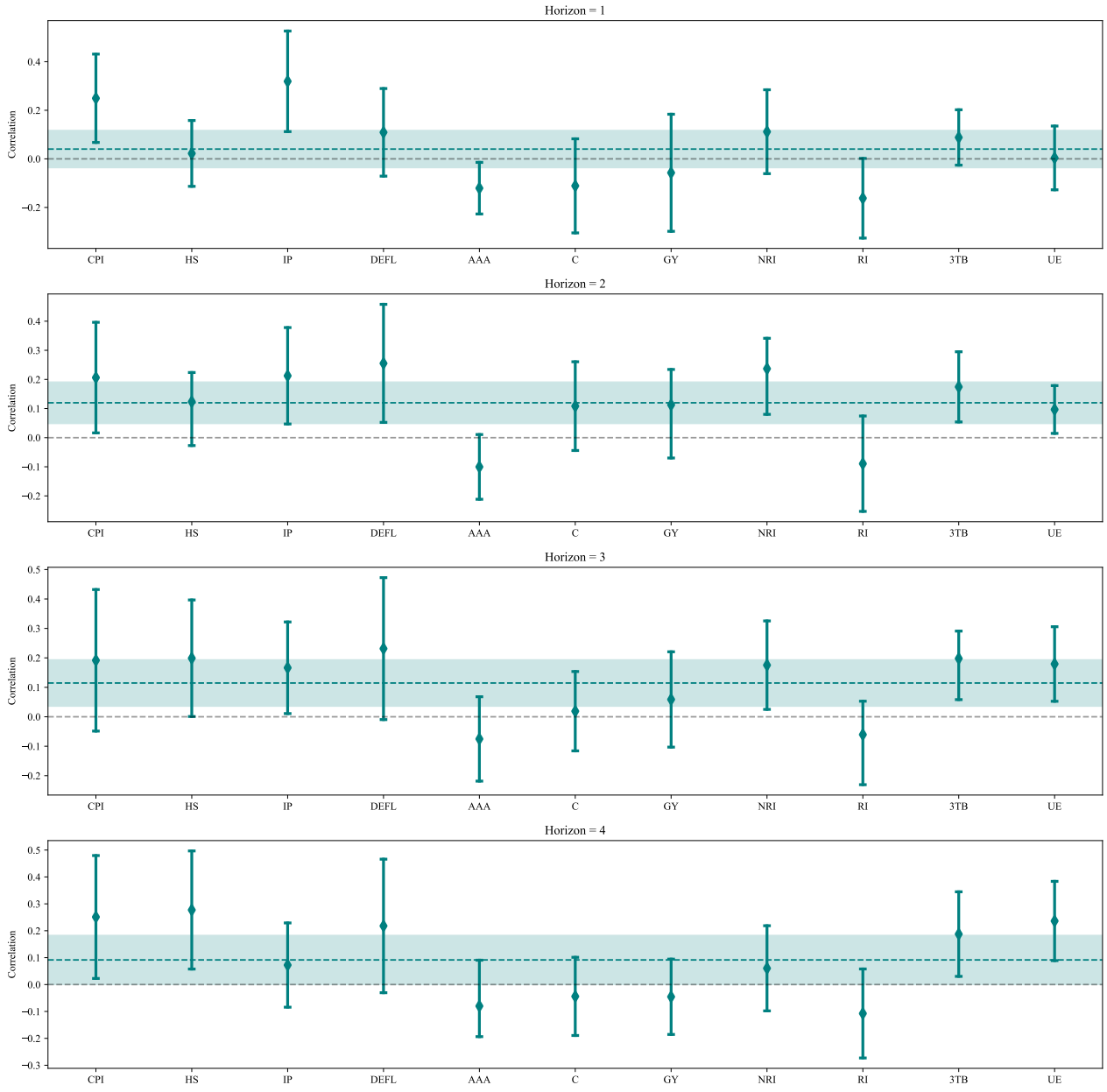
This section reports results comparable to Section 4.3 but partitioned over different horizons.

Figure E.1: SPF Revision Correlation Across Horizons



Note. Reports the correlation between revisions of SPF expectations and GPT expectations for different horizons. Additionally reports 90% confidence intervals for the correlation coefficients. The dashed maroon line reports the panel correlation coefficient and the shaded maroon band the corresponding 90% confidence interval. Standard errors for the single variable regressions are Newey-West, the panel standard errors are Driscoll-Kraay.

Figure E.2: Coibion-Gorodnichenko Coefficients Across Horizons

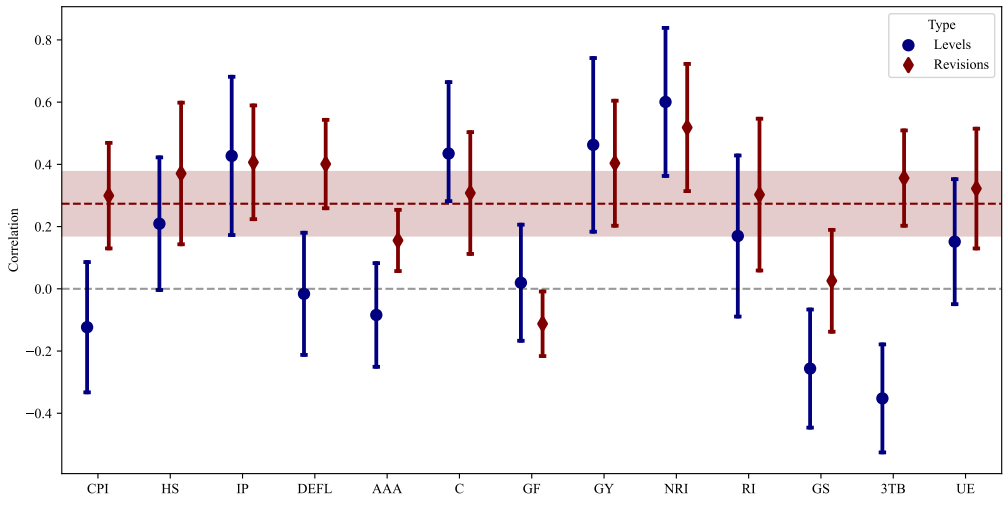


Note. Reports the CG coefficients for SPF revisions and GPT expectations across different horizons. Additionally reports 90% confidence intervals for the coefficients. The dashed teal line reports the panel CG coefficient and the shaded teal band the corresponding 90% confidence interval. Standard errors for the single variable regressions are Newey-West, the panel standard errors are Driscoll-Kraay.

F Covid

This section reports results comparable to 4.3 but excludes all data after 2019 to avoid the effects of the Covid-19 pandemic.

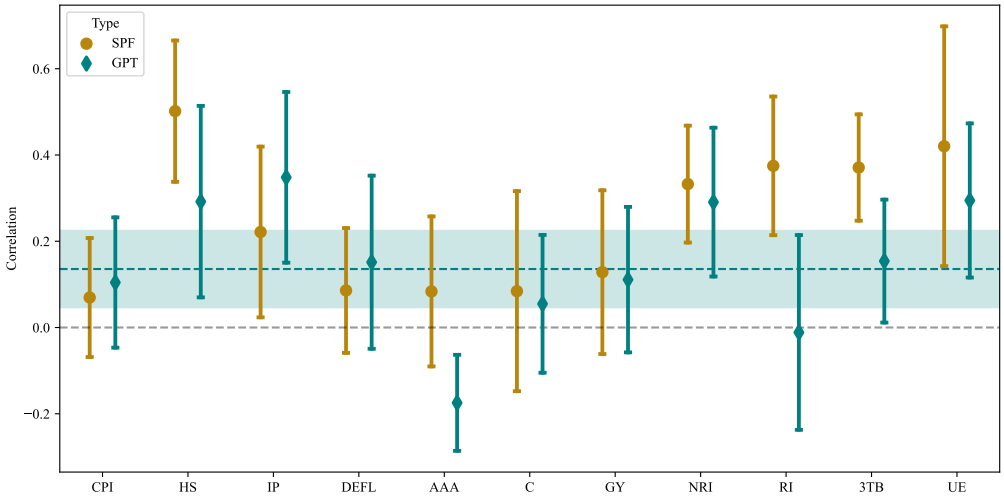
Figure F.1: GPT/SPF Correlations (Pre 2019)



	Levels		Revisions	
	Corr.	t	Corr.	t
Panel	0.07	1.23	0.27	4.35
CPI	-0.12	-0.97	0.30	2.90
Housing Starts	0.21	1.61	0.37	2.67
Industrial Production	0.43	2.75	0.41	3.65
GDP Deflator	-0.02	-0.13	0.40	4.63
AAA Corporate Bond Yield	-0.08	-0.83	0.16	2.59
Consumption Growth	0.43	3.11	0.31	2.58
Federal Government Spending	0.02	0.17	-0.11	-1.78
GDP Growth	0.46	2.72	0.40	3.30
Nonresidential Investment	0.60	4.14	0.52	4.16
Residential Investment	0.17	1.08	0.30	2.04
State Government Spending	-0.26	-2.22	0.03	0.26
3-Month Treasury Bill	-0.35	-3.33	0.36	3.81
Unemployment Rate	0.15	1.24	0.32	2.75

Note. Reports the correlation between levels and revisions of SPF expectations and GPT expectations. Additionally reports 90% confidence intervals for the correlation coefficients. The dashed maroon line reports the panel correlation coefficient and the shaded maroon band the corresponding 90% confidence interval. Standard errors for the single variable regressions are Newey-West, the panel standard errors are Driscoll-Kraay. The bottom table reports the corresponding correlation coefficients and t-stats.

Figure F.2: Coibion-Gorodnichenko Regressions (Pre 2019)



	SPF		GPT	
	Corr.	t	Corr.	t
Panel	0.21	4.19	0.14	2.50
CPI	0.07	0.83	0.10	1.13
Housing Starts	0.50	5.02	0.29	2.16
Industrial Production	0.22	1.84	0.35	2.89
GDP Deflator	0.09	0.97	0.15	1.24
AAA Corporate Bond Yield	0.08	0.79	-0.17	-2.57
Consumption Growth	0.08	0.60	0.05	0.56
GDP Growth	0.13	1.11	0.11	1.08
Nonresidential Investment	0.33	4.02	0.29	2.76
Residential Investment	0.37	3.82	-0.01	-0.08
3-Month Treasury Bill	0.37	4.94	0.15	1.77
Unemployment Rate	0.42	2.48	0.29	2.70

Note. Reports the CG coefficients for SPF revisions and GPT expectations. Additionally reports 90% confidence intervals for the coefficients. The dashed teal line reports the panel CG coefficient and the shaded teal band the corresponding 90% confidence interval. Standard errors for the single variable regressions are Newey-West, the panel standard errors are Driscoll-Kraay. The bottom table reports the corresponding correlation coefficients and t-stats.

G An Alternative Model of Aggregate Sentiment and Bubbles

Behavioral theories of bubbles primarily focus on the interaction between an asset's price and subjective beliefs about its value. In particular, these theories focus on periods when these beliefs deviate from the asset's fundamental value under rational expectations. Ideally then, to test behavioral theories of bubbles I would like to have a measure of these deviations in subjective beliefs, or "sentiment", for each asset.

Given a sufficiently long time series of asset-specific news, using my GPT-based methodology, it should be possible to form a measure of asset-specific sentiment. However, such a corpus on asset-specific news is not currently available. Instead, what I have is a measure of aggregate sentiment. If I assume that aggregate sentiment is the average of asset-specific sentiment, that is:

$$\bar{\delta}_t = \frac{1}{N} \sum_{i=1}^{N} \delta_{t,i} \tag{12}$$

where $\delta_{t,i}$ is sentiment about the asset i, and N the number of assets, then I can show that it is possible to use aggregate sentiment to study behavioral theories of bubbles. In the following sections I will show how by studying the joint dynamics of aggregate sentiment and an asset's returns it is possible to distinguish key stories behind behavioral bubbles, including mispricing, return extrapolation, and feedback loops.

G.1 Mispricing and Aggregate Sentiment

An asset is mispriced when its returns are not equal to news about the asset's fundamental value. Let $f_{t,i}$ represent news about the asset's fundamental value, $\delta_{t,i}$ a shock to beliefs about the asset's value that deviate from the fundamental, or asset specific "sentiment", and $\lambda_{\text{MP},i} \in [0, 1]$ be the asset's degree of mispricing. Then we can write the asset's return as:

$$r_{t,i} = f_{t,i} + \lambda_{\text{MP},i} \delta_{t,i}. \tag{13}$$

Proposition 1. *The beta of an asset's return, $r_{t,i}$, onto aggregate sentiment, $\bar{\delta}_t$, is an increasing function of the asset's degree of mispricing, $\lambda_{MP,i}$, under the assumption that $Cov(\delta_{t,i}, \bar{\delta}_t^{-i}) \geq 0$.*

Proof. To see this note that:

$$
\begin{aligned}
\beta_{r,\delta,i} &= \frac{Cov(r_{t,i}, \bar{\delta}_t)}{Var(\bar{\delta}_t)} \\
&= \frac{1}{N}\frac{Cov(r_{t,i}, \delta_{t,i})}{Var(\bar{\delta}_t)} + \frac{Cov(r_{t,i}, \bar{\delta}_t^{-i})}{Var(\bar{\delta}_t)} \\
&= \frac{1}{N}\left(\frac{Cov(f_{t,i}, \delta_{t,i})}{Var(\bar{\delta}_t)} + \lambda_{\text{MP},i}\frac{Var(\delta_{t,i})}{Var(\bar{\delta}_t)}\right) \\
&+ \frac{Cov(f_{t,i}, \bar{\delta}_t^{-i})}{Var(\bar{\delta}_t)} + \lambda_{\text{MP},i}\frac{Cov(\delta_{t,i}, \bar{\delta}_t^{-i})}{Var(\bar{\delta}_t)},
\end{aligned}
\tag{14}
$$

where $\bar{\delta}_t^{-i} = \frac{1}{N}\sum_{j\neq i}\delta_{t,j}$. Then, since $Var(\delta_{t,i}) > 0$, if we assume that $Cov(\delta_{t,i}, \bar{\delta}_t^{-i}) \geq 0$, $\beta_{r,\delta,i}$ is an increasing function of $\lambda_{\text{MP},i}$.

□

G.2 Return Extrapolation and Aggregate Sentiment

Return extrapolation impacts an asset when subjective beliefs about the asset's value respond to past returns. Let $\delta_{t,i}$ represent sentiment as above, $g_{t,i}$ a shock to beliefs not associated with past returns, and $\lambda_{\text{RE},i} \in [0, 1]$ the degree of return extrapolation. Then we can write the asset's sentiment as:

$$
\delta_{t,i} = g_{t,i} + \lambda_{\text{RE},i} r_{t-1,i}. \tag{15}
$$

Proposition 2. *The beta of aggregate sentiment, $\bar{\delta}_t$, onto an asset's lagged return, $r_{t-1,i}$ is an increasing function of the asset's degree of return extrapolation, $\lambda_{RE,i}$.*

Proof. To see this note that:

$$
\begin{aligned}
\beta_{\delta,r_l,i} &= \frac{Cov(r_{t-1,i}, \bar{\delta}_t)}{Var(r_{t-1,i})} \\
&= \frac{1}{N}\frac{Cov(r_{t-1,i}, \delta_{t,i})}{Var(r_{t-1,i})} + \frac{Cov(r_{t-1,i}, \bar{\delta}_t^{-i})}{Var(r_{t-1,i})} \\
&= \frac{1}{N}\lambda_{\text{RE},i} + \frac{Cov(r_{t-1,i}, g_{t,i})}{Var(r_{t-1,i})} \\
&+ \frac{Cov(r_{t-1,i}, \bar{\delta}_t^{-i})}{Var(r_{t-1,i})},
\end{aligned}
\tag{16}
$$

where $\bar{\delta}_t^{-i} = \frac{1}{N}\sum_{j\neq i}\delta_{t,j}$. Then, since $Var(r_{t-1,i}) > 0$, $\beta_{\delta,r_l,i}$ is an increasing function of $\lambda_{\text{RE},i}$.

□

G.3 Feedback and Aggregate Sentiment

Let $e_{t,i}$ represent a shock to the asset's return that is not associated with beliefs, $\delta_{t,i}$ sentiment as before, and $\lambda_{\text{FB},i}$ the degree of feedback from beliefs to returns. Then we can write the asset's return as:

$$r_{t,i} = e_{t,i} + \lambda_{\text{FB},i}\delta_{t-1,i}. \tag{17}$$

Proposition 3. *The beta of an asset's return, $r_{t,i}$, onto lagged aggregate sentiment $\bar{\delta}_{t-1}$ is an increasing function of the asset's degree of feedback, $\lambda_{FBL,i}$, under the assumption that $Cov(\delta_{t-1,i}, \bar{\delta}_{t-1}^{-i}) \geq 0$.*

Proof. To see this, note that:

$$\begin{aligned}
\beta_{r,\delta_l,i} &= \frac{Cov(r_{t,i}, \bar{\delta}_{t-1})}{Var(\bar{\delta}_{t-1})} \\
&= \frac{1}{N}\frac{Cov(r_{t,i}, \delta_{t-1,i})}{Var(\bar{\delta}_{t-1})} + \frac{Cov(r_{t,i}, \bar{\delta}_{t-1}^{-i})}{Var(\bar{\delta}_{t-1})} \\
&= \frac{1}{N}\left(\frac{Cov(e_{t,i}, \delta_{t-1,i})}{Var(\bar{\delta}_{t-1})} + \lambda_{\text{FB},i}\frac{Var(\delta_{t-1,i})}{Var(\bar{\delta}_{t-1})}\right) \\
&+ \frac{Cov(e_{t,i}, \bar{\delta}_{t-1}^{-i})}{Var(\bar{\delta}_{t-1})} + \lambda_{\text{FB},i}\frac{Cov(\delta_{t-1,i}, \bar{\delta}_{t-1}^{-i})}{Var(\bar{\delta}_{t-1})}.
\end{aligned} \tag{18}$$

Then since $Var(\delta_{t-1,i}) > 0$, if we assume that $Cov(\delta_{t-1,i}, \bar{\delta}_{t-1}^{-i}) \geq 0$, it follows that $\beta_{r,\delta_l,i}$ is an increasing function of $\lambda_{\text{FB},i}$.

□

G.4 Testing for Bubbles with Sentiment Betas

Finally, I will show how these covariance relationships can be used to test different mechanisms behind bubbles. Given a set of asset-specific price run-ups some of which crash (i.e. a bubble) and some of which do not, I can use the beta relationships to distinguish whether the given stories are relevant. Let $Q_{x,y}$ represent the sample beta from a regression of y onto x. Additionally, let $\mu(x|C)$ and $\mu(x|NC)$ represent the sample average for x for the set of run-ups that crash and do not crash respectively, and $\sigma(x|C)$ and $\sigma(x|NC)$ the sample standard deviation for x for the set of run-ups that crash and do not crash respectively. Then I will consider the following hypotheses and corresponding test statistics:

Hypothesis 1. *Run-ups that are mispriced are more likely to crash.*

Ideally, if I had an estimate of $\lambda_{\text{MP},i}$ for each run-up, I could test this hypothesis directly:

$$\theta'_{\text{MP}} = \frac{\mu(\lambda_{\text{MP}}|C) - \mu(\lambda_{\text{MP}}|NC)}{\sqrt{\sigma^2(\lambda_{\text{MP}}|C) + \sigma^2(\lambda_{\text{MP}}|NC)}}. \tag{19}$$

While I do not have a direct estimate of $\lambda_{\text{MP},i}$, since $\beta_{r,\delta,i}$ is an increasing function of $\lambda_{\text{MP},i}$, I can use the following test statistic instead:

$$\theta_{\text{MP}} = \frac{\mu(Q_{r_t,\bar{\delta}_t}|C) - \mu(Q_{r_t,\bar{\delta}_t}|NC)}{\sqrt{\sigma^2(Q_{r_t,\bar{\delta}_t}|C) + \sigma^2(Q_{r_t,\bar{\delta}_t}|NC)}}. \tag{20}$$

Hypothesis 2. *Return extrapolation increases the likelihood of a crash.*

Again, I ideally would like an estimate of $\lambda_{\text{RE},i}$ for each run-up to test this hypothesis directly. However, since $\beta_{r_l,\delta,i}$ is an increasing function of $\lambda_{\text{RE},i}$, I can use the following test statistic:

$$\theta_{\text{RE}} = \frac{\mu(Q_{r_{t-1},\bar{\delta}_t}|C) - \mu(Q_{r_{t-1},\bar{\delta}_t}|NC)}{\sqrt{\sigma^2(Q_{r_{t-1},\bar{\delta}_t}|C) + \sigma^2(Q_{r_{t-1},\bar{\delta}_t}|NC)}}. \tag{21}$$

Hypothesis 3. *Feedback from beliefs into returns increases the likelihood of a crash.*

By the same argument as above, since $\beta_{r,\delta_l,i}$ is an increasing function of $\lambda_{\text{FB},i}$, I can use the following test statistic:

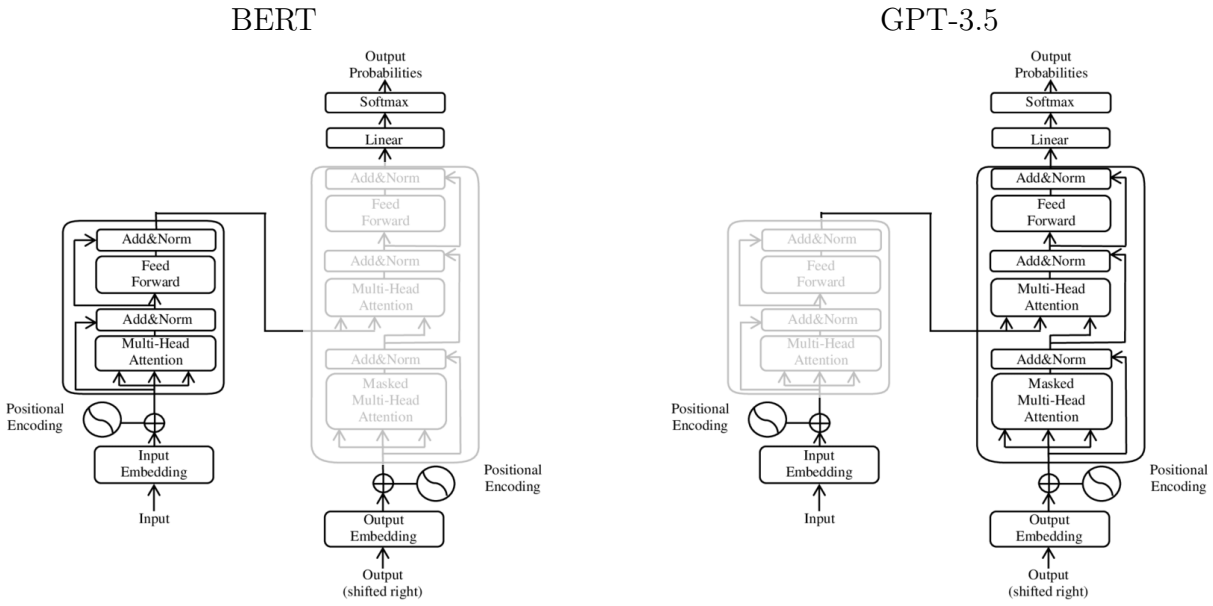
$$\theta_{\text{FB}} = \frac{\mu(Q_{r_t,\bar{\delta}_{t-1}}|C) - \mu(Q_{r_t,\bar{\delta}_{t-1}}|NC)}{\sqrt{\sigma^2(Q_{r_t,\bar{\delta}_{t-1}}|C) + \sigma^2(Q_{r_t,\bar{\delta}_{t-1}}|NC)}}. \tag{22}$$

H BERT vs. GPT Transformer Architectures

This section briefly discusses the difference in architectures between BERT and GPT-3.5. BERT is what is known as an "encoder" model, while GPT-3.5 is a "decoder" model. Figure H.1 reports the relative architectures of both BERT and GPT-3.5 for comparison. Both sides report the full overall transformer architecture, however, the unused components are grayed out. While both follow the same overall transformer architecture, encoder models use only the left portion of the network, while decoder models use only the right. A third class of encoder-decoder models use both.

A simple way to think of this is that encoder models are able to access all of the information in the input sequence when forming their prediction and as such are well-suited for tasks such as sentiment analysis and labeling. Decoder models, on the other hand, only have access to the previous information in the sequence and are well-suited for tasks such as language generation. There exists a third class of models called encoder-decoder models that use both the left and right portions of the network.

Figure H.1: Transformer Architectures



Note. This figure reports the respective encoder and decoder transformer architectures for BERT and GPT-3.5.

I Constructing the NYT Sample

1. Collect all article data from *The New York Times* API between 1851 and 2017.
2. Drop articles before 1900 and after 1984.
3. Exclude articles corresponding to weekends.
4. Exclude articles during interruptions in publishing due to strikes:
   (a) Sep 19, 1923 to Sep 26, 1923
   (b) Dec 12, 1962 to Mar 31, 1963
   (c) Sep 17, 1965 to Oct 10, 1965
   (d) Aug 10, 1978 to Nov 5, 1978
5. Exclude articles matching headline patterns and section labels corresponding to obituaries, weddings, and sports. List available from author on request.
6. To form the article text, first try to use the "snippet" column, if this is not available, use the "lead paragraph", if this is not available, use the "abstract".
7. Combine the headline text with the article text.
8. Exclude articles with less than 100 words.
9. Exclude articles with headlines less than 10 words.

J Estimating Sentiment

This section documents in more detail the procedure, data, and summary statistics associated with the estimation of economic sentiment.

J.1 Macroeconomic Data

The macroeconomic variables consist primarily of two sets of data all taken from FRED:

1. The full set of series in FRED-MD is used. To maximize the time series each individual series is scraped from the FRED API when available, otherwise, the data in the FRED-MD file itself is used.

2. Additionally, a number of series from the NBER are used to extend the available time series further back in time: M04128USM350NNBR, M08343USM232SNBR, M12003USM516NNBR, M12002USM511NNBR, M12003USM516NNBR, M1202AUSM510NNBR, M1220AUSM363SNBR, M1201AUSM348NNBR, M13021USM156NNBR, M13023USM156NNBR, M13025USM156NNBR, M09028USM474NNBR. These series are also scraped from the FRED API.

Figure J.1 reports the explained variance ratio for each component of the macroeconomic PCA procedure along with the corresponding AIC.

J.2 Estimating Ex-Ante Residual Sentiment (EAR)

I estimate a dynamic factor model (DFM) on the macroeconomic data using the PCA-based algorithm described in Stock and Watson (2011).

$$\begin{aligned} X_t &= \Theta F_t + \epsilon_t \\ F_t &= \Phi F_{t-1} + \eta_t, \end{aligned} \tag{23}$$

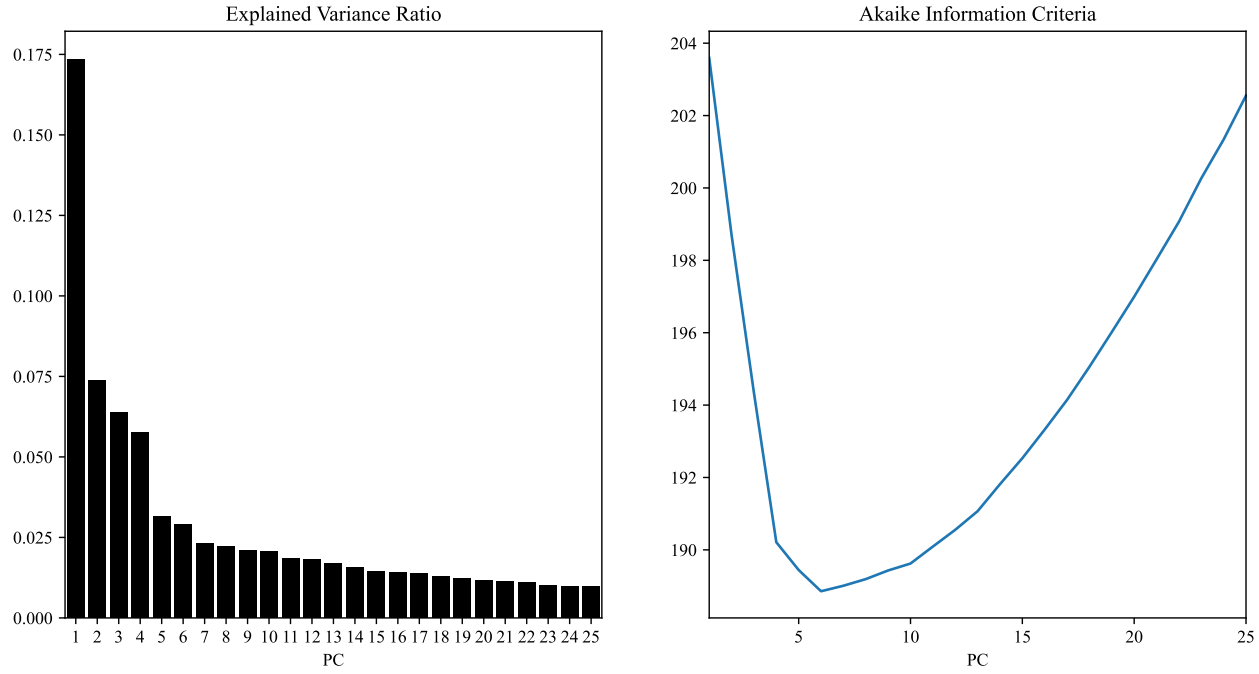
Given the unbalanced nature of the macroeconomic data, I use the expectation maximization procedure from Stock and Watson (2002) to estimate the factors. I then form ex-ante forecasts of the latent factors as $\hat{F}_t = \hat{\Phi}\hat{F}_{t-1}$. I estimate $\hat{\Phi}$ using three lags of the factors. Given these estimates I can then run a series of regressions are run using monthly expectations:

$$e_{t,i} \sim \alpha_i + B\hat{F}_t. \tag{24}$$

Residualized daily expectations for all days indexed τ in month t are then estimated as:

$$\tilde{e}_{\tau,i} = e_{\tau,i} - \hat{\alpha}_i - \hat{B}\hat{F}_t. \tag{25}$$

Figure J.1: Macro-Economic Factors Explained Variance Ratio and AIC



Note. This figure reports the explained variance ratio for each component of the macroeconomic PCA procedure, along with the corresponding AIC.

Finally, PCA is run on the residualized expectations and the first principal component is taken as the measure of economic sentiment.

J.3 Estimating Ex-Post Residual Sentiment (EPR)

I estimate a set of latent factors G_t, from the same set of macroeconomic data as previously:

$$X_t = \Gamma G_t + \xi_t. \tag{26}$$

Given the unbalanced nature of the macroeconomic data, I use the expectation maximization procedure from Stock and Watson (2002) to estimate the factors. Given these estimates I can then run a series of regressions are run using monthly expectations:

$$e_{t,i} \sim \alpha_i + \sum_{h=1}^{H} B_h \hat{F}_t, \tag{27}$$

where $H = 3$. Residualized daily expectations for all days indexed τ in month t are then estimated as:

$$\tilde{e}_{\tau,i} = e_{\tau,i} - \hat{\alpha}_i - \sum_{h=1}^{H} \hat{B}_h \hat{F}_t. \tag{28}$$

Finally, PCA is run on the residualized expectations and the first principal component is taken as the measure of economic sentiment.

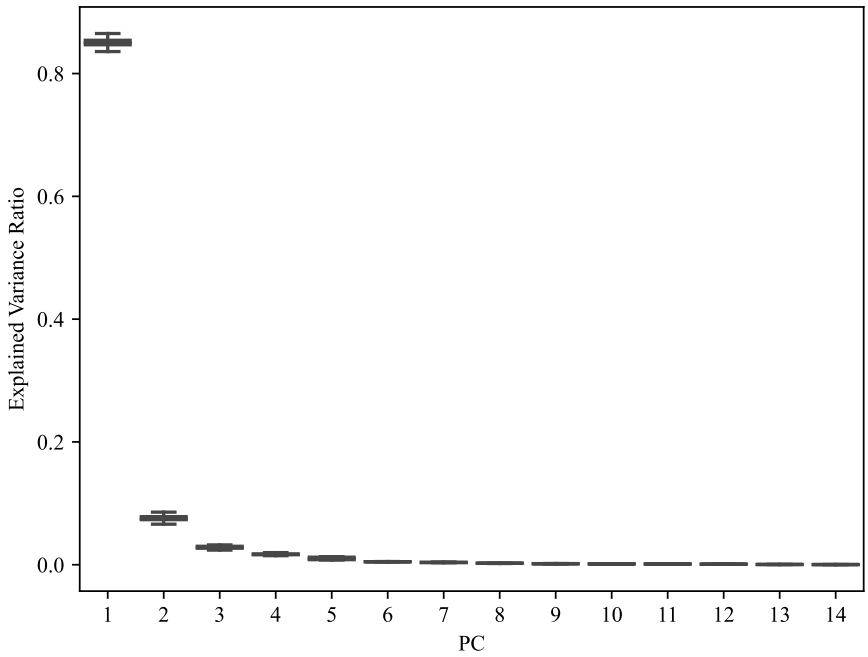
J.4 Estimating Generated Expectation Sentiment (EPC)

To estimate sentiment here I simple take the first principal component from a set of generated expectations, e_t.

J.5 Principal Components of Generated Expectations

This section reports summary statistics for the PCA procedure run on the generated daily expectations. Figure J.2 reports the explained variance ratio for each component of the PCA procedure.

Figure J.2: Expectations PC Explained Variance Ratio

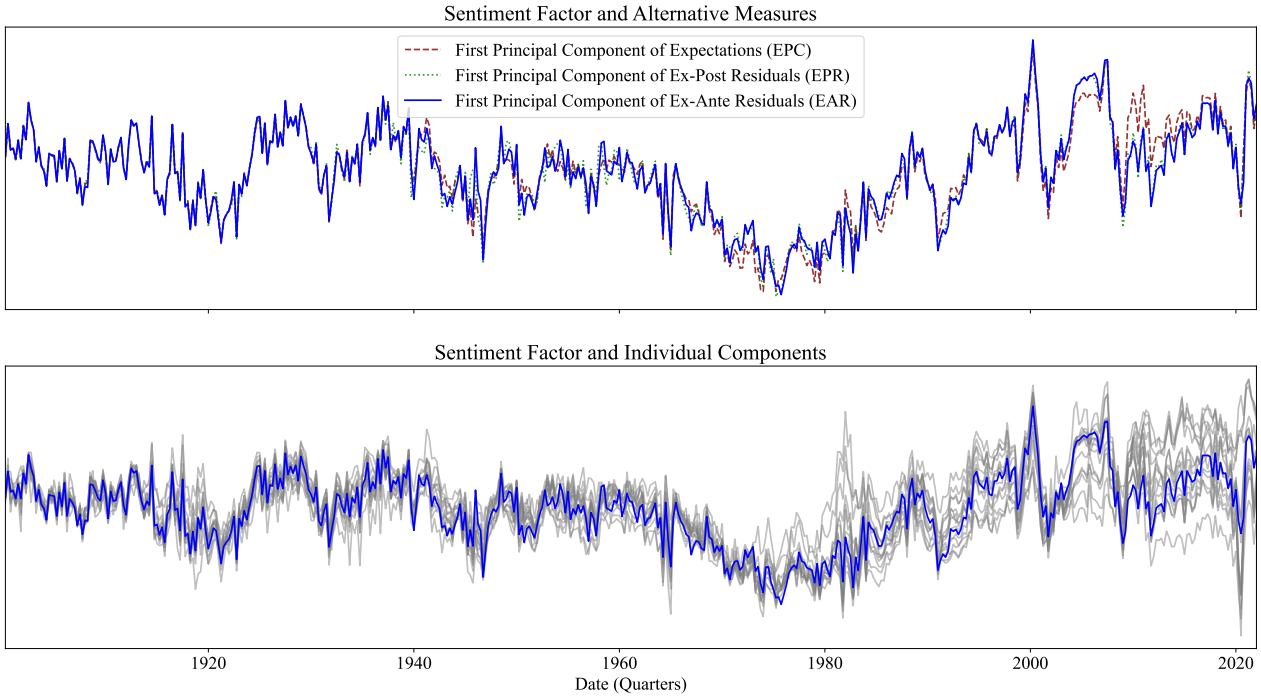


Note. This figure reports the distribution of the explained variance ratio for each principal component of extracted daily generated expectations. The distribution is taken over the expanding sample PCA estimates.

K Sentiment Robustness over 120 Years

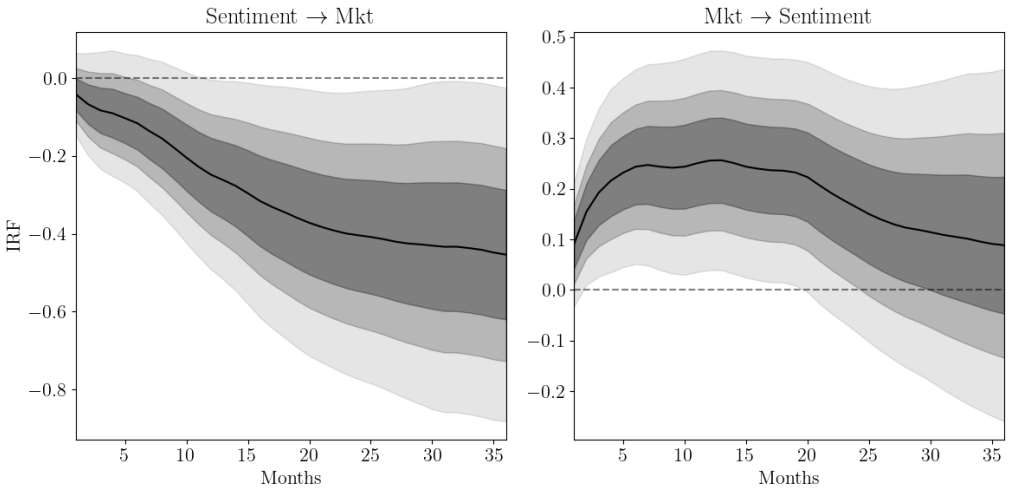
This section reports a series of additional results for the sentiment measurement exercise in Section 5. Figure K.1 reports the time series for all three sentiment measures overlaid. Figure K.2 reports the IRFs for all sentiment measures for the post 1984 period. Similarly, Figure K.3 reports the return IRFs for the post 1984 period benchmarked against CAPE and the Baker and Wurgler (2007) sentiment measure. Finally, Figure K.4 reports comparable benchmarks over the full sample for the alternative sentiment measures.

Figure K.1: Time Series of Alternative Sentiment Measures



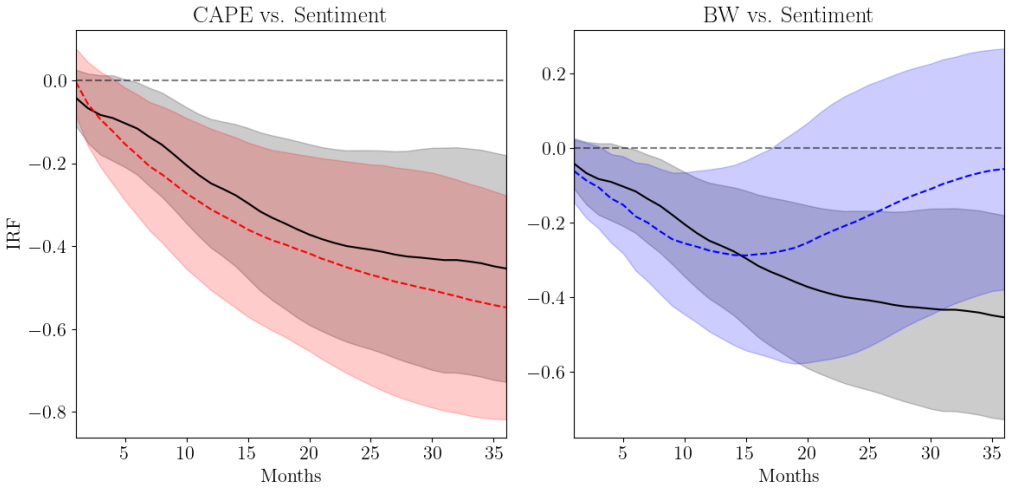
Note. The first panel reports the three main sentiment measures considered. First, the first principal component of The first panel reports the three main sentiment measures considered. First, the first principal component of the ex-ante residuals (EAR, the blue line). Second, the first principal component of the ex-post residuals (EPR, the dotted green line). Finally, the first principal component of the original expectations (EPC, the dashed maroon line). The second panel reports EAR overlaid over the individual expectation series. The negative values for the unemployment expectations are reported for visual clarity.

Figure K.2: Sentiment and Aggregate Return IRFs (Post 1984)



Note. The left panel reports the IRF capturing the response of value-weighted market returns to a one standard deviation shock to the EAR sentiment measure for the post 1984 sample. The right panel reports the IRF capturing the response of the EAR sentiment measure to a one standard deviation shock to value-weighted market returns for the post 1984 sample. Each set of figures reports a separate results for a separate sentiment measure. 68%, 90%, and 99% confidence intervals are reported (with decreasing levels of shading). Standard errors are Newey-West with the corresponding horizon as the number of lags.

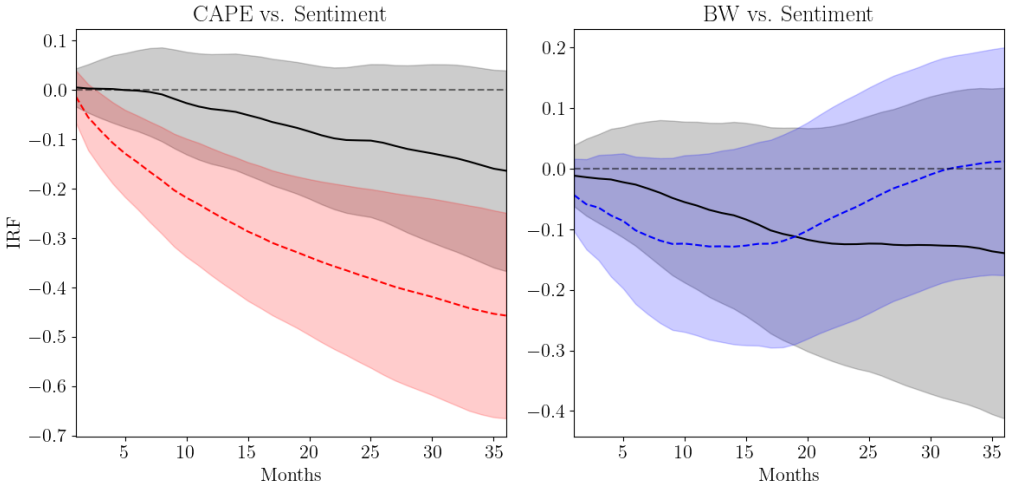
Figure K.3: Sentiment and Aggregate Return IRFs vs. Benchmarks (Post 1984)



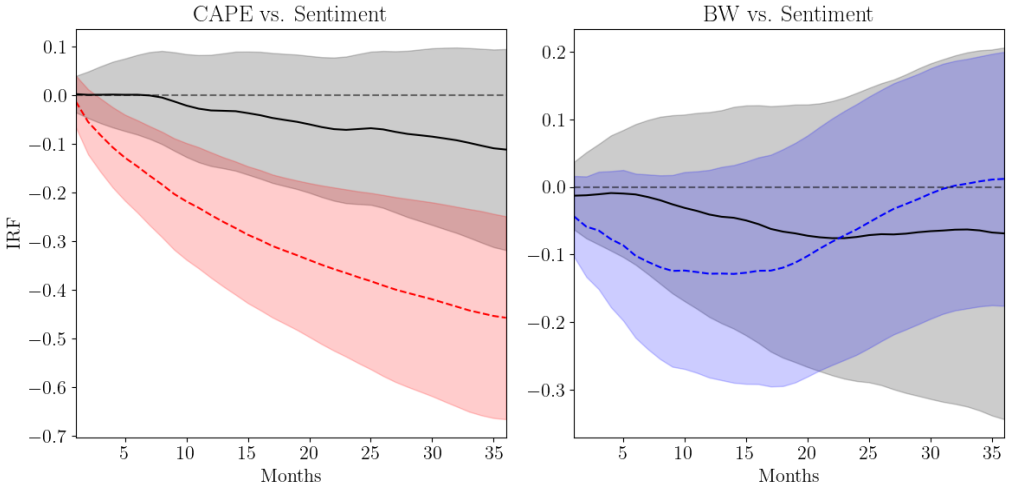
Note. The left panel reports the IRFs capturing the response of value-weighted market returns to a shock to the EAR sentiment measure (the black line) vs. CAPE (the red line) for the post 1984 sample. The right panel reports the IRFs capturing the response of the EAR sentiment measure to a shock to value-weighted market returns (the black line) vs. the Baker and Wurgler (2007) sentiment measure (the red line) for the post 1984 sample. 90% confidence intervals are reported, standard errors are Newey-West with the corresponding horizon as the number of lags.

Figure K.4: Sentiment and Aggregate Return IRFs vs. Benchmarks (Alternative Sentiment Measures)

(a) First principal component over ex-post residuals (EPR)



(b) First principal component over generated expectations (PCA)

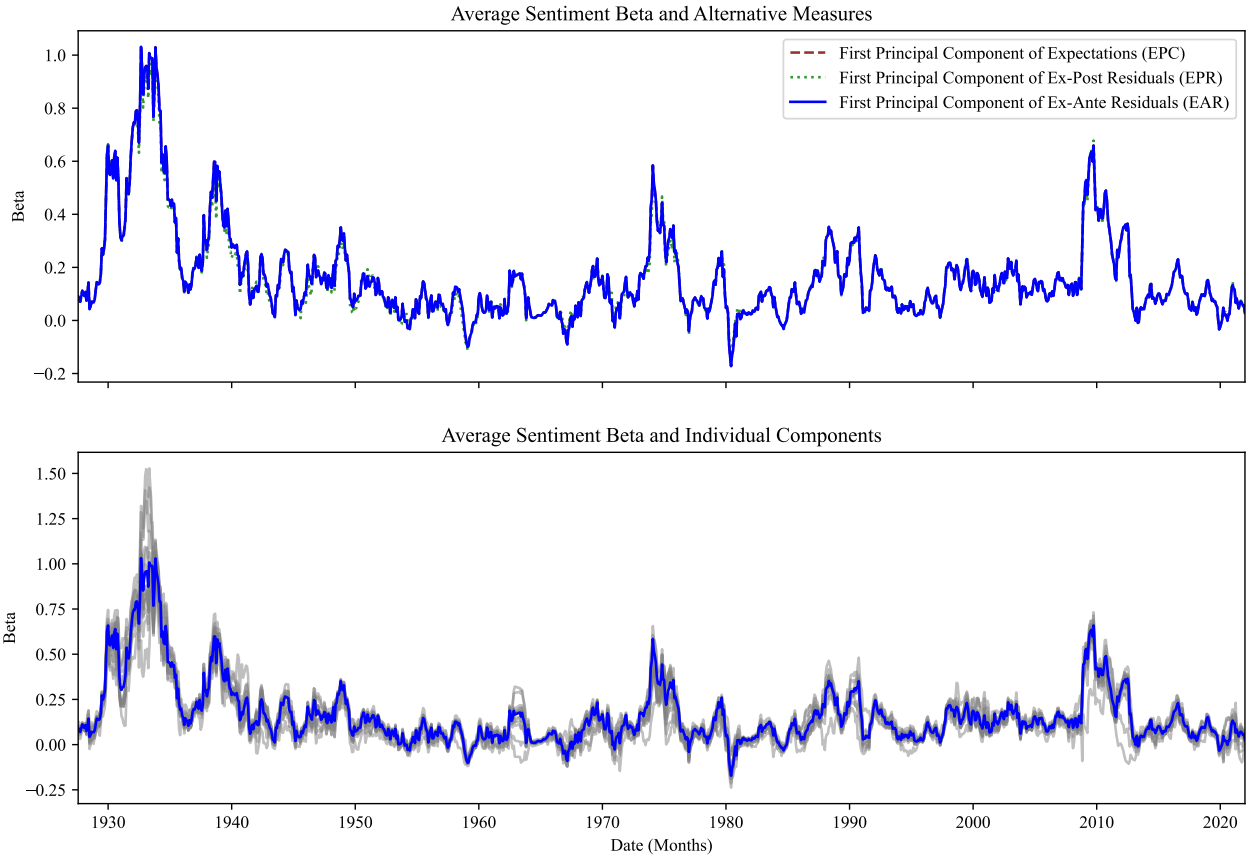


Note. The left panel reports the IRFs capturing the response of value-weighted market returns to a shock to my sentiment measure (the black line) vs. CAPE (the red line). The right panel reports the IRFs capturing the response of the my sentiment measure to a shock to value-weighted market returns (the black line) vs. the Baker and Wurgler (2007) sentiment measure (the red line). In both cases, the sentiment IRF is computed over the same sample as the benchmark. A set of figures is reported for each alternative sentiment measure. 90% confidence intervals are reported, standard errors are Newey-West with the corresponding horizon as the number of lags.

L Sentiment Beta Summary Statistics

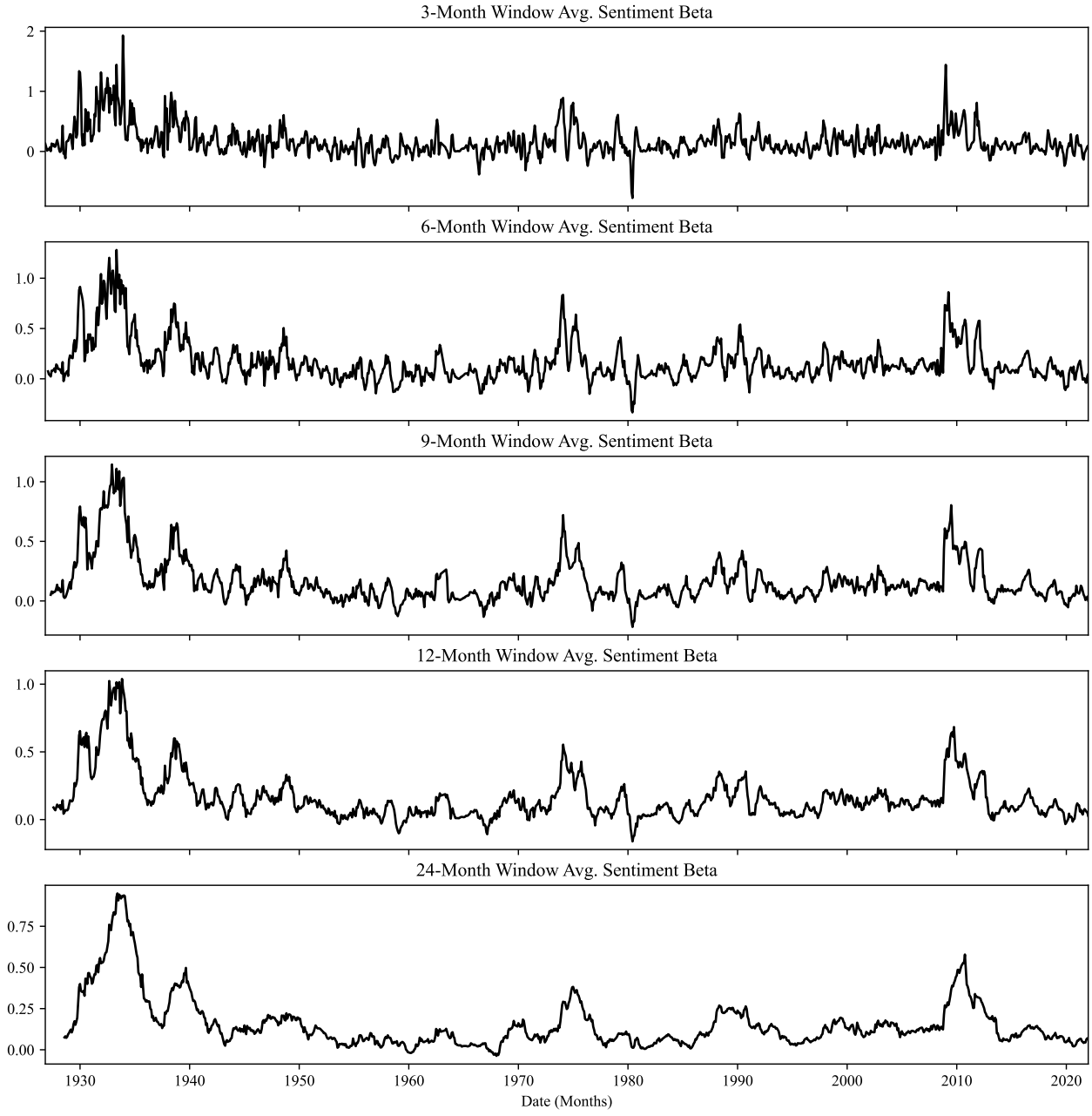
To better understand the dynamics of sentiment betas this section reports several visualizations of summary statistics for said betas. Figure L.1 reports the time series of the aggregate sentiment betas for a 12-month window for various measures of sentiment. Figure L.2 reports the aggregate time series and industry-level distribution of the sentiment betas for a 12-month window. Figure L.3 reports the distribution of the sentiment betas for a 12-month window. Both of the second set of figures use the main sentiment measure (EAR).

Figure L.1: Time Series of Sentiment Betas



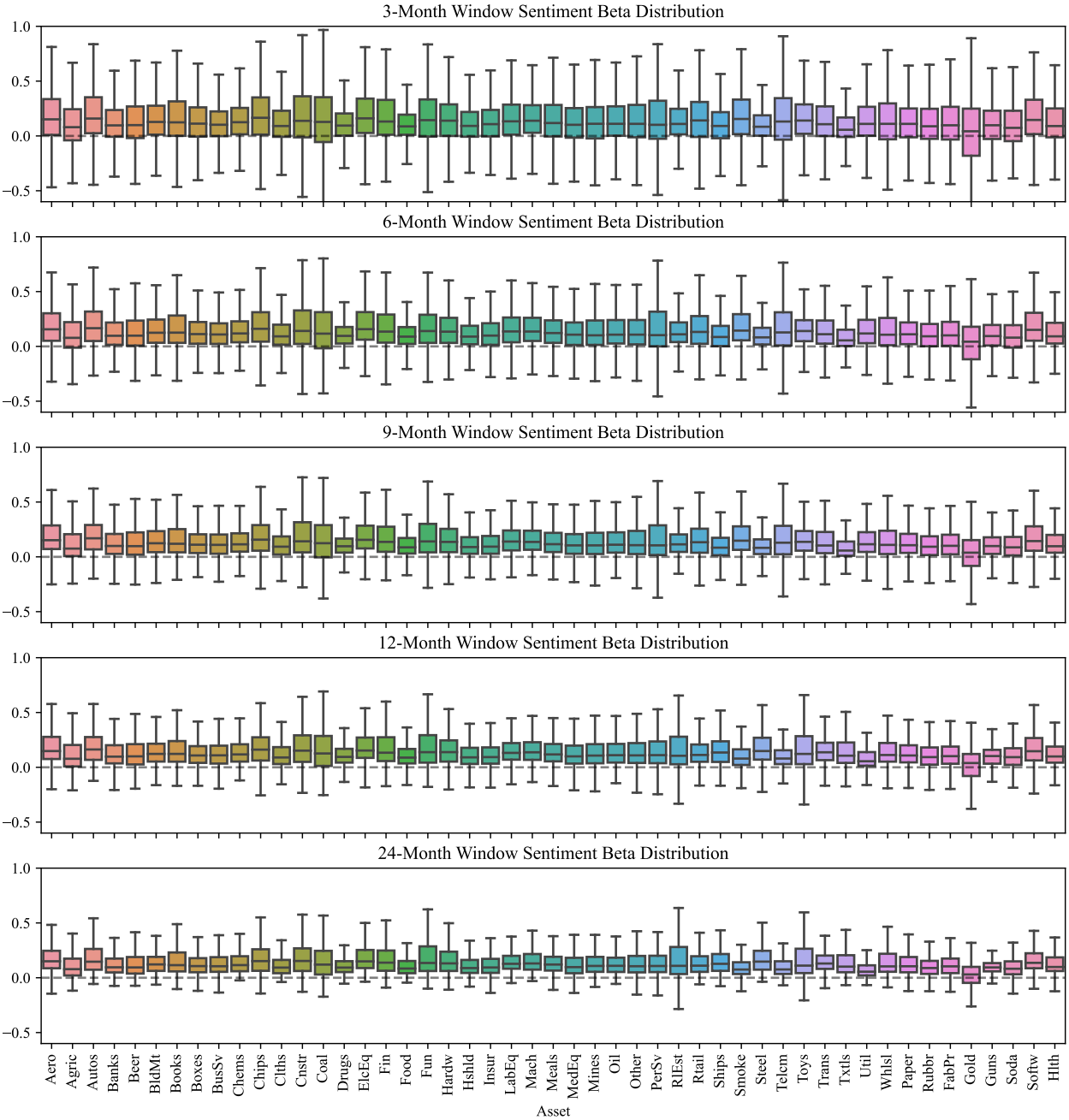
Note. The first panel reports the monthly average of the 12-month industry-level sentiment betas for the three main sentiment measures considered. First, the first principal component of the ex-ante residuals (EAR, the blue line). Second, the first principal component of the ex-post residuals (EPR, the dotted green line). Finally, the first principal component of the original expectations (EPC, the dashed maroon line). The second panel reports the monthly average of the 12-month industry-level sentiment betas for EAR overlaid over those for the individual expectation series. The negative values for the unemployment betas are reported for visual clarity.

Figure L.2: Average Beta Summary Statistics



Note. Reports the average industry-level sentiment betas for all beta formation windows for the main sentiment measure (EAR).

Figure L.3: Beta Distribution



Note. Reports the industry-level distribution of sentiment betas for all beta formation windows for the main sentiment measure (EAR).

M "Bubbles for Fama" Predictors

1. **Volatility (VW)**: Compute the monthly volatility of the daily returns of each stock. Then compute the percentile rank of the monthly volatility for all stocks. Finally, take a value-weighted average for all stocks in the rank Each month, we compute the volatility of daily returns of each stock in each industry.

2. **Turnover (VW)**: Compute the stock level shares traded divided by shares outstanding. For Nasdaq stocks, due to double counting, divide the turnover by two. Take the percentile rank of the stock level turnover. Then compute the value-weighted turnover rank for each industry.

3. **Firm Age (VW)**: Firm age is measured as the number of years since the firm first appeared on Compustat or on CRSP, whichever came first. Percentile rank age for all stocks then compute the value-weighted average for each industry.

4. **Age "tilt"**: Difference between the equal-weighted industry return and the age-weighted industry return.

5. **Book to Market (VW)**: Compute the book-to-market ratio for each firm. Take a value-weighted average for each industry.

6. **Issuance**: Percentage of firms in the industry that issued equity in the past year. Issuing equity is determined by when split-adjusted share count increases by five percent or more.

7. **CAPE**: The cyclically-adjusted market price-earnings ratio from Robert Shiller's website.

8. **Acceleration**: The difference between the 24-month return and the lagged 12-month return: $R_{t:t-24} - R_{t-12:t-24}$.

N Alternative Sentiment Betas: Crash and Return Predictability

This section focuses on the robustness of the crash and return predictability results to the use of alternative sentiment measures when computing sentiment betas. Table N.1 reports the beta summary statistics over the various samples and beta formation windows for the other two sentiment measures considered (EPR, EPC). The results are consistent with those in Table 3. Table N.2 reports the corresponding regression results controlling for various predictors studied in Greenwood et al. (2019). Again, the results are consistent with those in Figure 17. Finally, Figure N.1 reports the t-stats for the regressions using all different PCs (instead of just the first) extracted from the generated expectations. The results show that the PCs do not have the same predictive power as the first, which is consistent with the sentiment story presented throughout.

Table N.1: Run-up Beta Summary Statistics (Alternative Sentiment Measures)

	Full Sample		Crash		No Crash		Crash Ind.	24 M. Ret.
	Mean	SD	Mean	SD	Mean	SD	t	t
A. First Principal Component of Ex-Post Residuals (EPR)								
3-Month Beta	0.16	0.33	0.26	0.27	-0.03	0.34	3.42	-3.27
6-Month Beta	0.16	0.25	0.26	0.28	0.02	0.22	2.87	-2.65
9-Month Beta	0.16	0.22	0.22	0.22	0.02	0.16	3.29	-2.87
12-Month Beta	0.16	0.20	0.18	0.15	0.04	0.10	3.18	-3.64
24-Month Beta	0.15	0.17	0.16	0.13	0.08	0.11	1.72	-2.49
B. First Principal Component of Generated Expectations (EPC)								
3-Month Beta	0.17	0.33	0.26	0.27	-0.04	0.33	3.98	-3.23
6-Month Beta	0.17	0.26	0.26	0.28	0.03	0.22	3.02	-2.65
9-Month Beta	0.16	0.23	0.22	0.21	0.02	0.16	3.45	-2.84
12-Month Beta	0.16	0.21	0.18	0.15	0.03	0.11	3.69	-3.73
24-Month Beta	0.16	0.18	0.16	0.13	0.08	0.12	1.60	-2.16

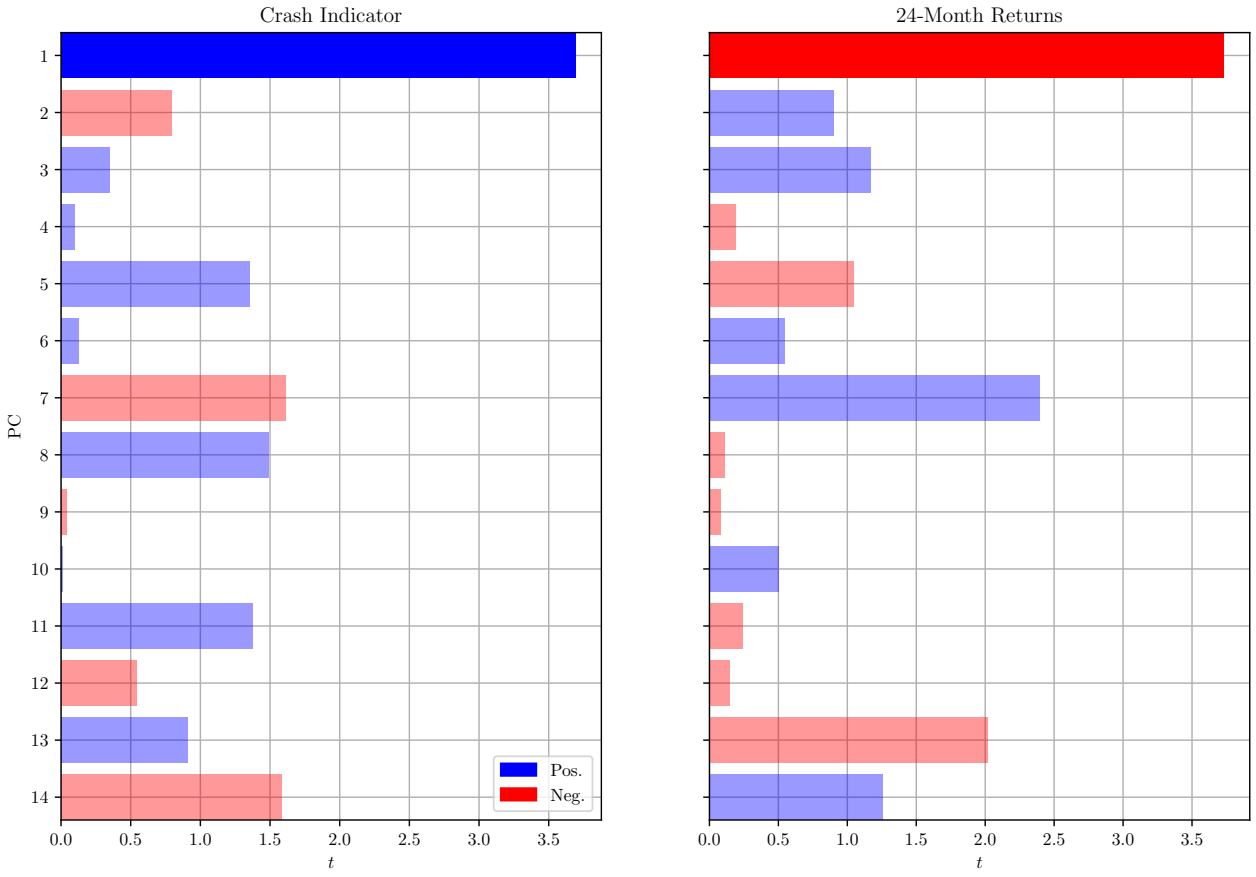
Note. Reports the mean and standard deviation for a series of sentiment beta windows for the full sample, the crash sample, and the no crash sample for the main alternative sentiment measures. Additionally, reports the t-stat for a regression of an indicator for whether a run-up crashed on the sentiment betas for the corresponding run-ups (Crash Ind.). Also, reports the t-stat for a regression of the 24-month future returns on the sentiment beta for the corresponding run-ups (24 M. Ret.). t-stat standard errors are clustered by industry year.

Table N.2: Run-up Beta Regressions (Alternative Sentiment Measures)

	Crash Indicator					24 Month Returns				
	SB		Ctrl			SB		Ctrl		
	Coef.	t	Coef.	t	R^2	Coef.	t	Coef.	t	R^2
A. First Principal Component of Ex-Post Residuals (EPR)										
Baseline	0.25	[3.18]			24.22	-0.49	[-3.64]			23.56
Market Beta	0.30	[5.66]	-0.14	[-1.70]	30.40	-0.51	[-3.47]	0.06	[0.51]	23.87
Volatility (VW)	0.24	[3.03]	-0.02	[-0.33]	24.40	-0.46	[-4.46]	0.09	[0.50]	24.39
Volatility (VW)- 1 yr- Δ	0.23	[2.93]	0.11	[2.28]	28.81	-0.47	[-3.43]	-0.11	[-1.36]	24.84
Turnover (VW)	0.23	[2.98]	-0.08	[-1.03]	26.30	-0.46	[-4.16]	0.10	[0.58]	24.57
Turnover (VW)- 1 yr- Δ	0.26	[3.00]	-0.04	[-0.64]	24.79	-0.47	[-4.40]	-0.06	[-0.30]	23.85
Firm Age (VW)	0.27	[3.26]	0.07	[0.92]	25.96	-0.54	[-3.67]	-0.22	[-1.98]	28.14
Age tilt	0.25	[3.61]	-0.04	[-0.42]	24.76	-0.47	[-3.63]	-0.15	[-0.99]	25.86
Book to Market (VW)	0.22	[2.64]	-0.08	[-0.97]	26.56	-0.52	[-3.39]	-0.09	[-0.69]	24.35
CAPE	0.21	[2.24]	0.01	[0.92]	25.45	-0.36	[-2.93]	-0.02	[-1.20]	26.71
Acceleration	0.23	[2.96]	0.20	[2.40]	39.65	-0.47	[-3.18]	-0.14	[-0.93]	25.52
B. First Principal Component of Generated Expectations (EPC)										
Baseline	0.25	[3.69]			25.21	-0.49	[-3.73]			23.62
Market Beta	0.29	[5.76]	-0.12	[-1.55]	30.10	-0.50	[-3.50]	0.03	[0.26]	23.70
Volatility (VW)	0.25	[3.56]	-0.03	[-0.37]	25.45	-0.46	[-4.66]	0.10	[0.54]	24.60
Volatility (VW)- 1 yr- Δ	0.24	[3.37]	0.11	[2.28]	29.65	-0.47	[-3.52]	-0.11	[-1.34]	24.84
Turnover (VW)	0.24	[3.54]	-0.08	[-1.05]	27.42	-0.46	[-4.38]	0.11	[0.62]	24.76
Turnover (VW)- 1 yr- Δ	0.26	[3.59]	-0.04	[-0.67]	25.81	-0.47	[-4.56]	-0.06	[-0.30]	23.91
Firm Age (VW)	0.27	[3.76]	0.07	[0.85]	26.80	-0.54	[-3.68]	-0.21	[-1.85]	27.82
Age tilt	0.26	[4.03]	-0.03	[-0.36]	25.60	-0.47	[-3.81]	-0.16	[-1.08]	26.27
% Issuer	0.25	[3.60]	0.03	[0.48]	25.58	-0.47	[-3.63]	-0.08	[-0.63]	24.30
Book to Market (VW)	0.23	[2.98]	-0.08	[-0.91]	27.19	-0.52	[-3.54]	-0.11	[-0.79]	24.60
CAPE	0.21	[2.58]	0.01	[0.93]	26.55	-0.36	[-3.09]	-0.02	[-1.23]	27.11
Acceleration	0.23	[3.08]	0.19	[2.31]	39.36	-0.47	[-3.17]	-0.13	[-0.81]	25.14

Note. Reports the results for regression of the crash indicator on the corresponding sentiment beta for the two alternative sentiment measures. In addition to the baseline fit without any controls, it also includes t-stats the sentiment betas controlling for each of the variables from Greenwood et al. (2019) as well as the 12-month market beta. The table reports the corresponding regressions, including the correlation coefficients and R^2. The SB column reports the sentiment beta coefficients and the Ctrl column reports the coefficients for the corresponding control variable. For all results, standard errors are clustered by industry year.

Figure N.1: Alternative PC t-stats



Note. This figure reports the t-stats for the regressions of the crash indicator and 24-month returns on the corresponding window sentiment beta for all principal components extracted from the generated expectations. Positive values are in blue, negative in red. Standard errors are clustered by industry year.

O Excess Returns

This section reports results checking the robustness of the various predictive exercises to the choice of returns used for both beta estimation and future returns. The main results use raw returns to estimate sentiment betas and evaluate performance. Table O.1 reports comparable results net of the risk-free rate and net of the market. In both cases, results are comparable to those using raw returns.

Table O.1: Run-up Beta Regressions (Excess Returns)

	Crash Indicator					24 Month Returns				
	SB		Ctrl			SB		Ctrl		
	Coef.	t	Coef.	t	R^2	Coef.	t	Coef.	t	R^2
A. Net RF										
Baseline	0.26	[3.79]			27.41	-0.47	[-3.85]			22.49
Market Beta	0.31	[6.86]	-0.13	[-1.63]	32.90	-0.49	[-3.72]	0.05	[0.47]	22.76
Volatility (VW)	0.26	[3.59]	-0.02	[-0.31]	27.56	-0.45	[-4.53]	0.10	[0.54]	23.42
Volatility (VW)- 1 yr- Δ	0.25	[3.47]	0.11	[2.26]	31.73	-0.46	[-3.63]	-0.10	[-1.23]	23.57
Turnover (VW)	0.25	[3.53]	-0.07	[-1.00]	29.31	-0.45	[-4.29]	0.11	[0.62]	23.64
Turnover (VW)- 1 yr- Δ	0.27	[3.63]	-0.04	[-0.67]	27.96	-0.46	[-4.62]	-0.04	[-0.24]	22.67
Firm Age (VW)	0.28	[3.87]	0.07	[0.98]	29.28	-0.52	[-3.78]	-0.19	[-1.69]	25.97
Age tilt	0.27	[4.19]	-0.03	[-0.36]	27.82	-0.46	[-3.80]	-0.15	[-0.92]	24.67
Book to Market (VW)	0.24	[3.02]	-0.07	[-0.88]	29.34	-0.49	[-3.44]	-0.06	[-0.45]	22.84
CAPE	0.23	[2.75]	0.01	[0.87]	28.50	-0.36	[-2.97]	-0.02	[-1.07]	25.29
Acceleration	0.23	[3.14]	0.17	[1.98]	38.60	-0.45	[-3.27]	-0.12	[-0.72]	23.79
B. Net MKT										
Baseline	0.21	[2.87]			16.70	-0.36	[-2.98]			13.01
Market Beta	0.26	[4.06]	-0.12	[-1.65]	21.71	-0.34	[-2.76]	-0.05	[-0.38]	13.21
Volatility (VW)	0.20	[2.82]	-0.04	[-0.56]	17.19	-0.33	[-3.56]	0.14	[0.73]	14.88
Volatility (VW)- 1 yr- Δ	0.20	[2.81]	0.14	[3.48]	24.72	-0.36	[-2.91]	-0.06	[-0.54]	13.33
Turnover (VW)	0.19	[2.78]	-0.09	[-1.31]	19.95	-0.33	[-3.11]	0.14	[0.72]	14.95
Turnover (VW)- 1 yr- Δ	0.22	[2.64]	-0.04	[-0.59]	17.41	-0.34	[-3.80]	-0.07	[-0.34]	13.47
Firm Age (VW)	0.22	[2.89]	0.05	[0.62]	17.61	-0.41	[-2.94]	-0.20	[-1.26]	16.70
Age tilt	0.21	[2.99]	-0.03	[-0.30]	16.95	-0.34	[-3.15]	-0.17	[-1.22]	15.91
Book to Market (VW)	0.17	[2.22]	-0.09	[-1.01]	19.51	-0.37	[-2.44]	-0.04	[-0.26]	13.13
CAPE	0.17	[2.20]	0.02	[2.60]	24.41	-0.30	[-2.76]	-0.02	[-1.58]	18.35
Acceleration	0.18	[2.15]	0.10	[1.04]	20.02	-0.38	[-2.48]	0.06	[0.34]	13.34

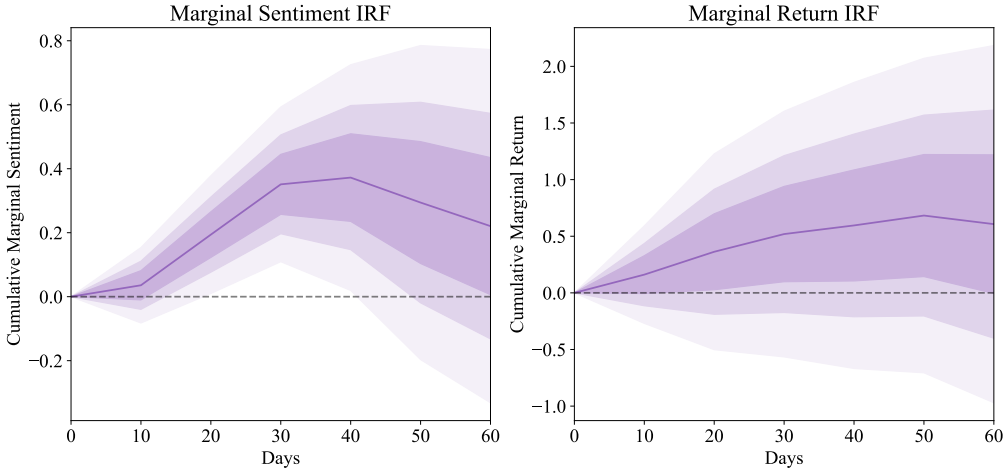
Note. Reports the results for regression of the crash indicator on the corresponding sentiment beta for returns in excess of the risk-free rate and the market. In addition to the baseline fit without any controls, it also includes t-stats the sentiment betas controlling for each of the variables from Greenwood et al. (2019) as well as the 12-month market beta. The table reports the corresponding regressions, including the correlation coefficients and R^2. The SB column reports the sentiment beta coefficients and the Ctrl column reports the coefficients for the corresponding control variable. For all results, standard errors are clustered by industry year.

P Extrapolation and Alternative Sentiment Measures

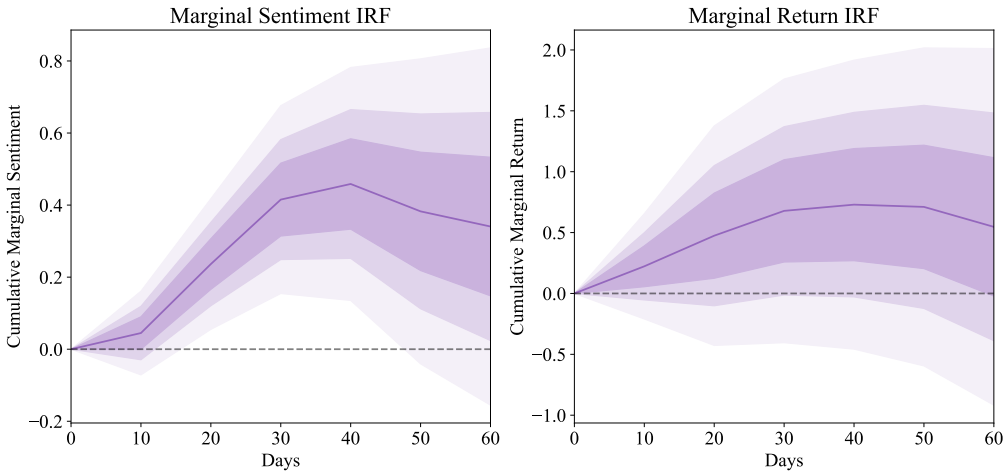
This section summarizes the various extrapolation IRFs for the alternative sentiment measures. Table P.1 reports the corresponding summary statistics. Results are consistent with those in the main text.

Figure P.1: Run-up Extrapolation IRFs (Alternative Sentiment Measures)

(a) First principal component over ex-post residuals (EPR)



(b) First principal component over generated expectations (EPC)



Note. This figure reports the sentiment and return IRFs for the alternative sentiment measures. 68%, 90% and 99% confidence intervals are reported (with decreasing levels of shading). Standard errors are clustered by industry year.

Table P.1: Run-up Extrapolation Summary Statistics

	Full Sample		Crash		No Crash		Crash−No Crash
	Mean	SD	Mean	SD	Mean	SD	t
A. EPR (Sentiment IRF)							
10-Day Sentiment	0.031	0.264	0.030	0.108	-0.005	0.171	0.797
20-Day Sentiment	0.037	0.406	0.074	0.188	-0.120	0.255	2.757
30-Day Sentiment	0.023	0.528	0.144	0.273	-0.207	0.317	3.769
40-Day Sentiment	0.007	0.661	0.175	0.361	-0.198	0.489	2.760
50-Day Sentiment	0.042	0.775	0.224	0.560	-0.070	0.637	1.554
60-Day Sentiment	0.038	0.873	0.226	0.673	0.005	0.681	1.030
B. EPR (Return IRF)							
10-Day Return	-0.044	0.616	-0.002	0.511	-0.163	0.542	0.967
20-Day Return	-0.028	1.065	-0.055	1.130	-0.418	0.992	1.074
30-Day Return	0.012	1.450	0.105	1.325	-0.414	1.334	1.234
40-Day Return	0.093	1.741	0.145	1.391	-0.450	1.679	1.226
50-Day Return	0.099	2.134	0.062	1.624	-0.621	1.771	1.272
60-Day Return	0.131	2.502	-0.078	2.022	-0.685	1.850	0.987
C. EPC (Sentiment IRF)							
10-Day Sentiment	0.041	0.255	0.027	0.102	-0.018	0.171	1.020
20-Day Sentiment	0.054	0.385	0.073	0.188	-0.163	0.250	3.407
30-Day Sentiment	0.049	0.496	0.140	0.283	-0.275	0.349	4.147
40-Day Sentiment	0.047	0.620	0.186	0.380	-0.273	0.411	3.669
50-Day Sentiment	0.091	0.721	0.256	0.583	-0.127	0.453	2.299
60-Day Sentiment	0.098	0.812	0.268	0.696	-0.073	0.516	1.743
D. EPC (Return IRF)							
10-Day Return	-0.044	0.637	-0.018	0.497	-0.241	0.559	1.336
20-Day Return	-0.056	1.125	-0.100	1.144	-0.575	1.068	1.351
30-Day Return	-0.061	1.522	0.024	1.361	-0.655	1.297	1.610
40-Day Return	-0.020	1.828	0.006	1.394	-0.723	1.508	1.590
50-Day Return	-0.035	2.234	-0.152	1.673	-0.863	1.533	1.396
60-Day Return	-0.015	2.560	-0.352	1.813	-0.899	1.779	0.961

Note. Reports the mean and standard deviation for a series of sentiment and return IRFs using 12 months of daily data over various horizons for the full sample, the crash sample, and the no-crash sample. For the crash and no crash samples, the IRFs are computed using 12 months of daily data prior to the identified run-up point. Additionally, reports the t-stat for the sample difference of means between the crash and no crash run-ups. t-stat standard errors are clustered by industry year.

Q Trading on Sentiment Robustness

This section reports the robustness of the trading on sentiment results in Section 6.5. Table Q.1 reports the corresponding results. In addition to a strategy based on an optimal sentiment beta or extrapolation threshold, it reports results using the mean value as the threshold.

Table Q.1: Trading on Sentiment Robustness

	Opt. Thresh					Mean Thresh				
Strategy	Mean	SD	t	SP	A t	Mean	SD	t	SP	A t
				A. Raw						
Buy-and-Hold	0.139	1.717	0.404	0.475						
Optimal	1.905	1.782	5.345	1.000	5.329					
EAR Beta	1.950	1.955	4.988	0.775	5.977	1.377	1.628	4.230	0.750	5.770
EAR Ext	1.290	2.309	2.793	0.800	2.863	0.751	2.598	1.445	0.700	1.622
EPR Beta	1.920	1.635	5.873	0.825	6.004	1.005	1.571	3.199	0.725	4.627
EPR Ext	1.190	1.878	3.167	0.775	2.967	0.814	2.238	1.818	0.725	2.007
EPC Beta	1.727	1.566	5.515	0.800	5.710	0.984	1.643	2.994	0.700	4.235
EPC Ext	1.392	2.291	3.038	0.825	3.189	1.063	2.633	2.018	0.750	2.489
				B. Net of RF						
Buy-and-Hold	-0.220	1.710	-0.643	0.475						
Optimal	1.553	1.762	4.407	1.000	5.392					
EAR Beta	1.493	1.934	3.861	0.775	5.693	0.931	1.617	2.879	0.750	5.391
EAR Ext	1.018	2.307	2.205	0.800	3.074	0.476	2.589	0.919	0.700	1.843
EPR Beta	1.497	1.609	4.653	0.825	5.852	0.548	1.562	1.755	0.725	4.106
EPR Ext	0.906	1.869	2.424	0.775	3.178	0.534	2.228	1.198	0.725	2.245
EPC Beta	1.287	1.544	4.167	0.800	5.452	0.522	1.635	1.596	0.700	3.718
EPC Ext	1.084	2.286	2.370	0.825	3.308	0.750	2.630	1.426	0.750	2.606
				C. Net of MKT						
Buy-and-Hold	0.466	1.007	2.314	0.475						
Optimal	1.509	1.135	6.649	1.000	4.493					
EAR Beta	1.537	1.127	6.817	0.775	4.750	1.231	0.933	6.596	0.750	4.529
EAR Ext	1.053	1.570	3.353	0.800	2.101	0.767	1.652	2.321	0.700	1.178
EPR Beta	1.536	0.935	8.219	0.825	5.079	0.955	0.766	6.234	0.725	3.165
EPR Ext	1.085	1.467	3.700	0.775	2.328	0.801	1.368	2.928	0.725	1.478
EPC Beta	1.399	0.904	7.737	0.800	4.567	0.944	0.847	5.569	0.700	2.967
EPC Ext	1.164	1.540	3.777	0.825	2.610	1.021	1.517	3.364	0.750	2.261

Note. Reports the performance of a variety of trading strategies, extending Section 6.5, based on sentiment betas and extrapolation IRFs. In particular reports the mean, standard deviation, t-stat, proportion of correctly identified bubbles (SP), and the alpha t-stat with respect to the "Buy-and-Hold Strategy". Reports results for the optimal threshold as well as using the mean value as threshold. The "Buy-and-Hold" strategy holds each run-up for the full 24 months. The "Optimal" strategy holds only the run-ups that do not crash ex-post. The "Beta" strategy holds only the run-ups that have a beta below the beta threshold. The "Ext" strategy holds only the run-ups that have a sentiment IRF over 30 days below the extrapolation threshold.