The Ghost in the Machine: Generating Beliefs with Large Language Model

Author: Leland Bybee
Yale

Notes: Sitong Li

Introduction

This paper introduces a methodology to generate beliefs using large language models, which includes:

1. Why we should use LLMs as belief generators
2. How to get beliefs via LLMs
3. Evaluating Generated Expectations via:
   - 3.1 Return Expectations from survey
   - 3.2 Macroeconomic Expectations from SPF
   - 3.3 examines whether the result is from memorization or generalization
4. Generating 120 Years of Economic Expectations
5. Generated Beliefs to predict Bubbles

Why uses LLMs as belief generators

Reason include:

1. There is no need for the world models encoded in LLms need to be accurate
2. A growing body of experimental evidence supports that LLMs do indeed reflect human beliefs and biases

We could utilize generated beliefs to:

1. Can be formed whatever there is news text and as such can serve to extend the available time series of survey data
2. Can be used to from higher frequency expectations than existing survey.
3. open the door to surveying heterogeneous populations and forming expectations for populations that are hard to sample
4. Form firm-specific expectations from earning calls and news (for bubbles)
5. Survey contains what the survey takers read, which is news and the information is directly available in the form of the news text itself.

How to get beliefs via CPT: example and stats

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)}
```

Figure: example

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

Figure: stats

Usage for output

balance statistic: the proportion of articles where GPT responds with increases minus the proportion of articles where GPT responds with decreases
$F_t^{GPT}(X_{t+h}^k)$ 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:

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

where A_t is the set of articles published in period t, $\mathbb{I}(\text{Increase})$ indicates GPT responded with "increase" for the kth series given article i, and $\mathbb{I}(\text{Decrease})$ a comparable indicator for "decrease". This approach is used by other popular surveys such as the Gallup survey,

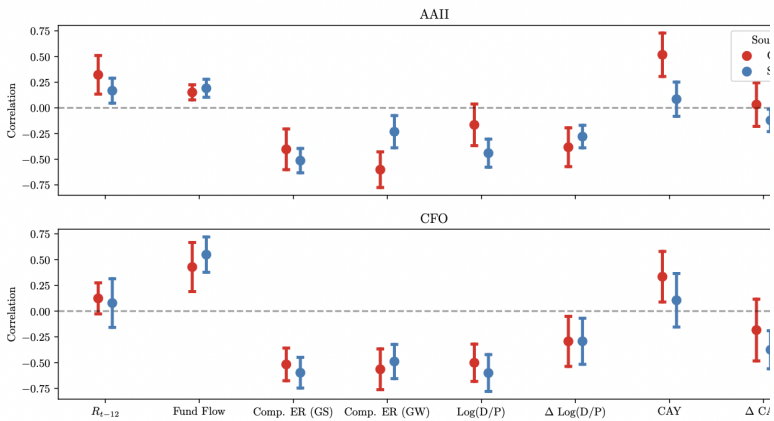
Return Expectations

first evaluate generated expectations of returns by comparing them to two publicly available benchmark return expectation series used in Greenwood and Shleifer (2014), which is American Association of Individual Investors and the Duke CFO Survey

I next evaluate whether generated expectations other facts documented previously for survey-based return expectations, which includes past twelve-months returns (Rt12) of the U.S. stock market and consider a number of objective expected return proxies.

Result

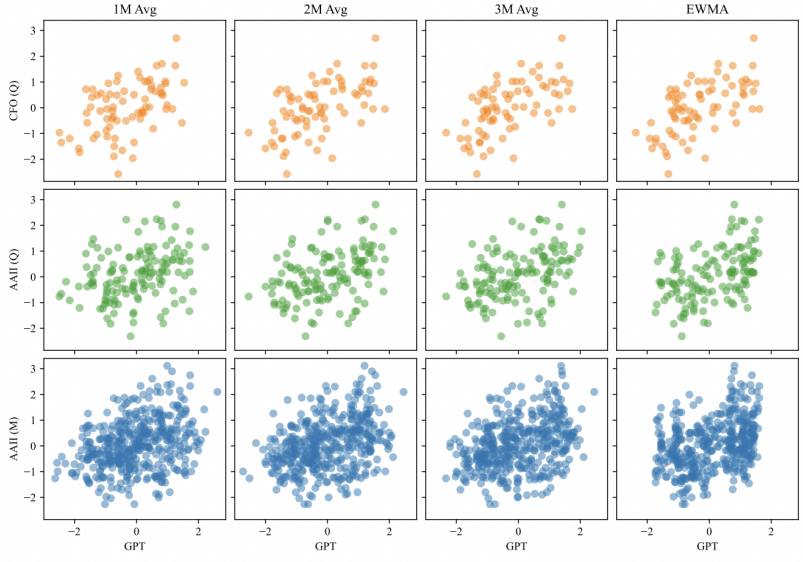
Figure 4: Survey Correlations with Existing Moments



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

Figure: Stats with proxies

Figure 3: Correlation between Return Expectations



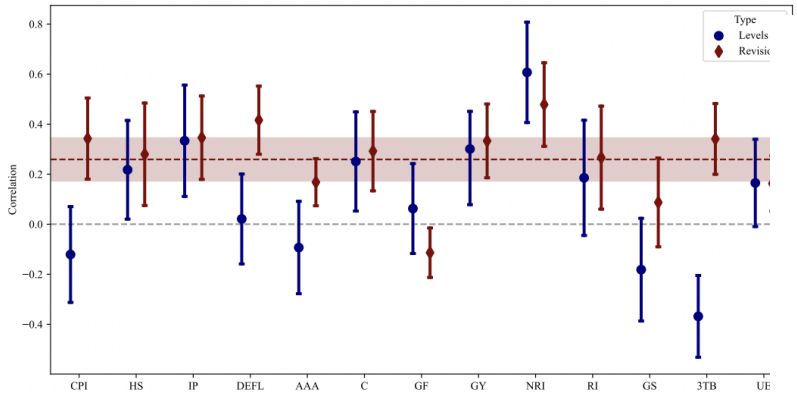
	one-month avg.	two-month avg.	three-month avg.	opt. EWMA
CFO (N=76)	0.49	0.56	0.59	0.59
	[4.84]	[5.88]	[6.21]	[6.24]
AAII (Q, N=138)	0.32	0.40	0.39	0.43
	[3.93]	[5.15]	[5.01]	[5.63]
AAII (M, N=414)	0.32	0.35	0.35	0.36
	[6.83]	[7.65]	[7.65]	[7.82]

Figure: Corr

Macroeconomic Expectations

Compared with the Survey of Professional Forecasters (SPF).

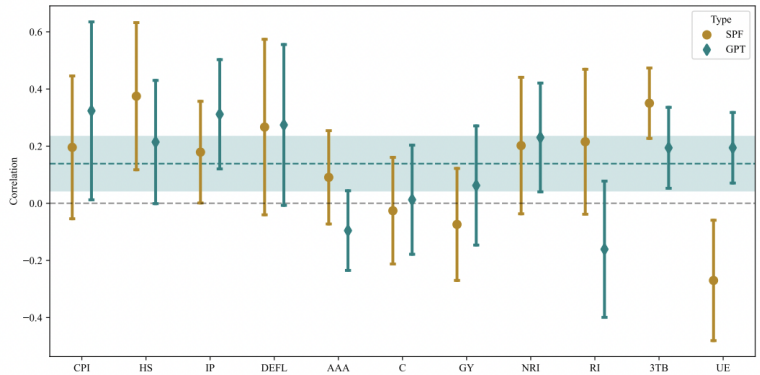
Figure 6: GPT/SPF Correlations



	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

Figure: GPTSPF Correlations

Figure 7: Coibion-Gorodnichenko Regressions

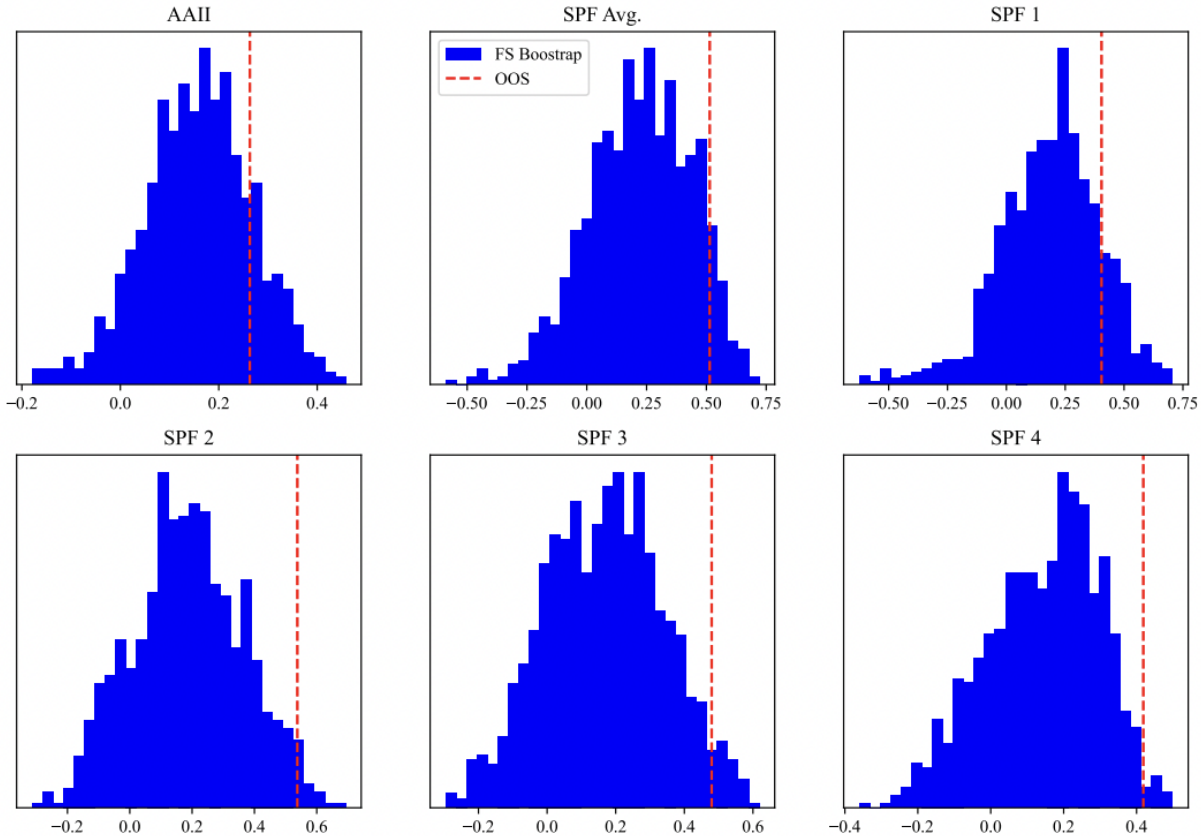


	SPF		GPT	
	Corr.	t	Corr.	t
Panel	0.12	1.31	0.14	2.39
CPI	0.20	1.28	0.32	1.70
Housing Starts	0.37	2.39	0.21	1.63
Industrial Production	0.18	1.65	0.31	2.67
GDP Deflator	0.27	1.42	0.27	1.60
AAA Corporate Bond Yield	0.09	0.91	-0.10	-1.12
Consumption Growth	-0.03	-0.23	0.01	0.11
GDP Growth	-0.07	-0.62	0.06	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

Figure: Figure 7
Coibion-Gorodnichenko Regressions

Memorization or Generalization

Figure 8: 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.

Identifying Economic Sentiment with Generated Beliefs

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:

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

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

The key innovation here is thinking of sentiment as a common component.

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.

First, givenμ_t represents a rational forecast,I assume the rational forecast is formed using a dynamic factor model over a broad set of macroeconomic outcomes

$X_t = \Theta F_t + \varepsilon_t$

$F_t = \Phi F_{t-1} + \eta_t,$

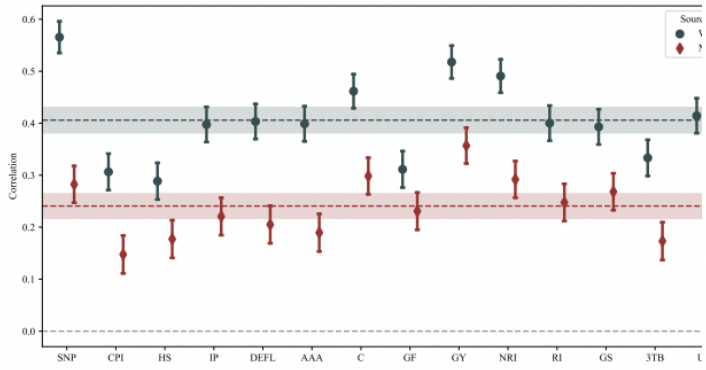
Transfer Learning with BERT

GPT is costly, use BERT though
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 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:

$$\min_{\rho_i} \|e_{a,i} - x_a \rho_i\|_2^2 + \lambda_i \|\rho_i\|_2^2 . \quad (6)$$

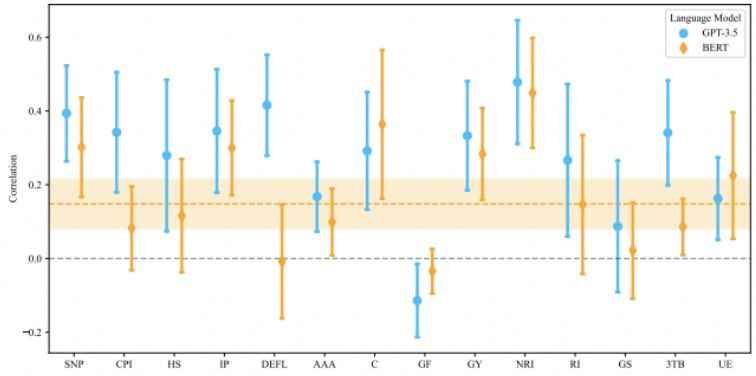
GPT vs BERT

Figure 10: 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

Figure 9: 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

Economic Sentiment over 120 Years

measure of economic sentiment.

Figure 13: Time Series of Sentiment against Benchmarks

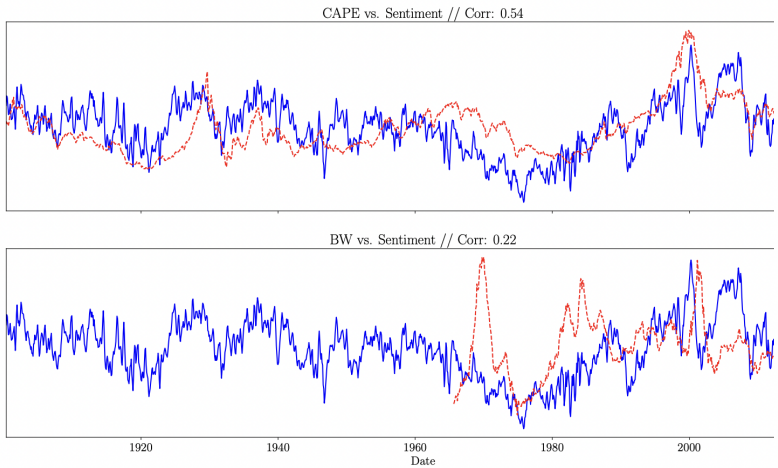
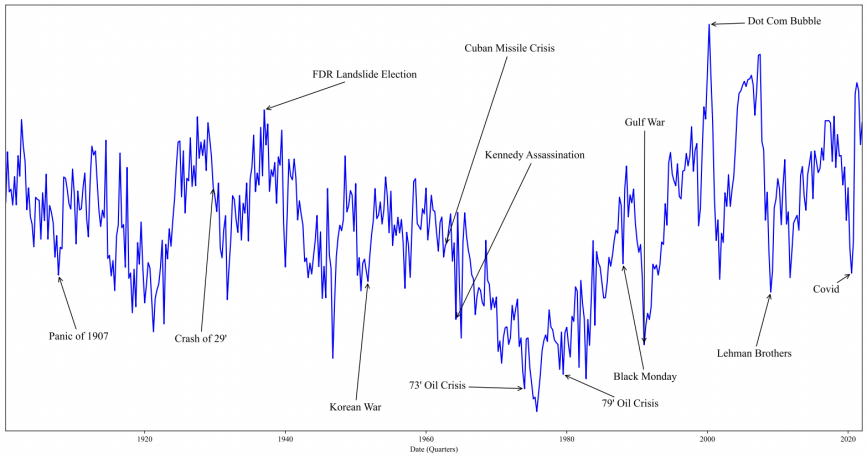


Figure 11: Time Series of Economic Sentiment



Data on Bubbles

I employ the data set used in Greenwood et al. (2019). Using the 49 industry classifications 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 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.

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.
an asset's beta with respect to economic sentiment provides a measure of mispricing.
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:

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

First, an asset's degree of mispricing varies over time and second, an asset's own sentiment is proportional to aggregate sentiment.
I identify mispricing by estimating an asset's rolling beta with respect to economic sentiment

Mispricing and Bubbles

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:

$$r_{\tau,i} \sim \alpha_{t,i} + \beta_{t,i}\delta_\tau \quad (8)$$

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

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?

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