Wharton
University *of* Pennsylvania

Jacobs Levy Equity
Management Center
for Quantitative Financial Research

Expected Returns and Large Language Models

Bryan Kelly (with Dacheng Xiu and Yifei Chen)

Frederick Frank '54 and Mary C. Tanner Professor of Finance, Yale University

Head of Machine Learning, AQR Capital Management

BK Knock, Knock …

Sure! Who's there?

BK Orange

Orange who?

BK Knock, Knock …

I'm sorry, I believe it seems like you are trying to initiate a knock-knock joke, but you haven't provided a punchline for the set up you provided. Please feel free to continue the joke if you wish, I'll be happy to play along

BK Hey! The joke is still going

I apologize for any confusion, I understand now.
Orange you going to continue the joke?

ChatGPT: A Cinderella Story

The Power of Statistical Prediction Models

Cinderella story, out of nowhere, former greenskeeper, now about to become the Masters
____?____

Source: Google Images. For illustrative purposes only. The use of the logos and pictures is for informational purposes only and is not authorized by, sponsored by or associated with the trademark owners. Images sourced from Golf.com

Cinderella story, out of nowhere, former greenskeeper, now about to become the Masters ______

Word	Number
Aardvark	0
Abacus	0
About	0
Become	0
Beowulf	0
Champion	0
Cinderella	1
Greenskeeper	0
Imagination	0
Master	0
Now	0
Of	0
Out	0
Story	0
The	0
To	0
Zygote	0

$$b_6 x_{T\text{-}6} + b_5 x_{T\text{-}5} + b_4 x_{T\text{-}4} + b_3 x_{T\text{-}3} + b_2 x_{T\text{-}2} + b_1 x_{T\text{-}1} = \hat{x}_T$$

Cinderella story, out of nowhere, former greenskeeper, now about to become the Masters ______

	$x_{T\text{-}6}$ Now	$x_{T\text{-}5}$ About	$x_{T\text{-}4}$ To	$x_{T\text{-}3}$ Become	$x_{T\text{-}2}$ The	$x_{T\text{-}1}$ Masters		x_T **Champion**
Aardvark	0	0	0	0	0	0		0.21
Abacus	0	0	0	0	0	0		0.01
About	0	**1**	0	0	0	0		0.22
Become	0	0	0	**1**	0	0		0.02
Beowulf	0	0	0	0	0	0		0.08
Cinderella	0	0	0	0	0	0		0.12
Champion	0	0	0	0	0	0	=	0.95
Greenskeeper	0	0	0	0	0	0		0.08
Imagination	0	0	0	0	0	0		0.32
Masters	0	0	0	0	0	**1**		0.09
Now	**1**	0	0	0	0	0		0.22
Of	0	0	0	0	0	0		0.32
Out	0	0	0	0	0	0		0.24
Story	0	0	0	0	0	0		0.11
The	0	0	0	0	**1**	0		0.06
To	0	0	**1**	0	0	0		0.12
Zygote	0	0	0	0	0	0		0.32

Training ceases when fits are optimized

Training for Prediction

Trained on 500 billion examples…

- Common Crawl (410 billion examples, no quality control)
- Books (67 billion examples, exact content proprietary)
- WebText2 (19 billion examples, select high-quality content)
- Wikipedia (3 billion examples)

…using billions of parameters

GPT 3 uses 175 billion parameters, GPT 4 uses 1 trillion (thousands of "CPU years" to train)

- As many as 200 parameters *per example*
- Enough to "memorize" word sequences (even if relatively rare)
- Language is extremely repetitive! Hence, effective number of parameters per example is much higher, which allows it to memorize nuanced patterns of language

Enough examples to know ***nearly everything*** that can be said

Enough parameters to ***memorize*** even obscure language patterns

A Memorization Machine

ChatGPT is a **word prediction model**

Regurgitating past language patterns makes for a good chatbot

Can this work for investing?

ChatGPT and Investing

Limitations of Prompting

BK Is it a good idea to buy Tesla stock?

Investing in Tesla stock can be a good idea for some investors, but it really depends on your financial goals, risk tolerance, and market outlook. Here are some factors to consider: Growth Potential, Volatility Valuation, Market Trends, Long-Term vs. Short-Term

Always do your own research or consult with a financial advisor to align your investment choices with your financial strategy.

Prompt-based approaches…

- Limited by researcher’s ability to engineer meaningful prompts
  - Analogous to the limitations of hypothesis-driven versus data-driven statistics
  - May be important patterns in text that missed by prompts
- Limited by biases in training text
- Limited by model’s prompt capabilities

Source: ChatGPT. For illustrative use only. The use of the logos and pictures is for informational purposes only and is not authorized by, sponsored by or associated with the trademark owners.

ChatGPT and Investing

Embeddings: Distilling meaning from text

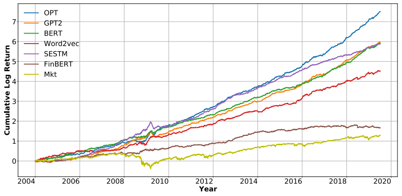
Decompressed Input Text

	Cinderella	Story	Greenskeeper	Masters
…				
Cinderella	1	0	0	0
…				
Greenskeeper	0	0	1	0
…				
Masters	0	0	0	1
…				
Story	0	1	0	0
…				

Compressed Input Text

	Cinderella	Story	Greenskeeper	Masters
Human	0.9	0.2	0.9	-0.1
Male	-0.6	0.0	0.8	0.3
Literature	0.8	0.9	0.1	0.5
Plural	-0.1	0.1	-0.1	0.9
…				

Intercept and reroute for return prediction



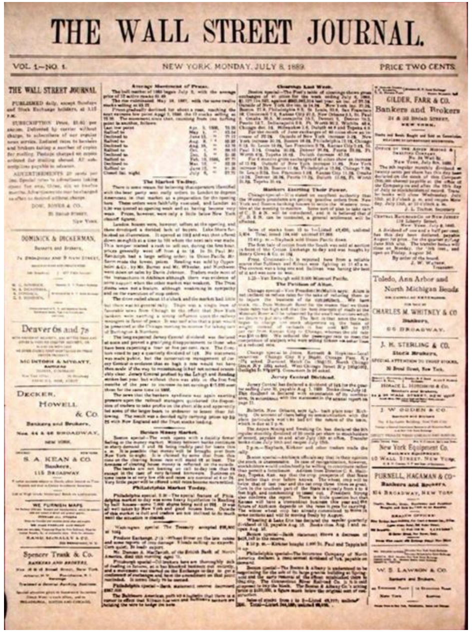
Compressed Prediction

Human	0.7
Male	0.5
Literature	0.2
Plural	-0.2
…	

Decompressed Prediction

	…
0.2	Bee
	…
0.9	Champion
	…
0.8	Gladiator
	…
-0.4	Zebras
	…

LLMs: Contrasting Prompts and Embeddings


THE WALL STREET JOURNAL.

VOL. 1—NO. 1 NEW YORK, MONDAY, JULY 8, 1889. PRICE TWO CENTS

Prompt

Embedding

- Searching for specific content versus capturing general content
- Fishing with a rod versus fishing with a net
- Subject to usual bias/variance tradeoff

Empirical Approach

Data

Thompson/Reuters News Articles for Single-name Stocks

	Raw Articles			Articles Tagged with Single Stock			Articles With	After Filtering	After Filtering
	RTRS	3PTY	Total	RTRS	3PTY	Total	Returns Matched	Short Articles	by Novelty
US	6,366,019	4,843,867	11,209,886	2,863,166	4,123,823	6,986,989	4,755,247	4,123,279	3,038,025
UK	707,288	1,050,467	1,757,755	196,573	773,266	969,839	906,705	901,838	571,285
Australia	261,020	1,203,784	1,464,804	100,444	1,113,347	1,213,791	388,585	382,114	249,190
Canada	255,933	473,686	729,619	126,281	431,401	557,682	481,891	478,205	350,549
China (HK)	3,537,487	7,287,688	10,825,175	1,140,542	5,558,763	6,699,305	2,086,045	305,335	182,363
Japan	3,259,103	38,860	3,297,963	1,210,077	16,850	1,226,927	405,341	399,185	310,244
Germany	2,423,671	1,751,231	4,174,902	480,264	880,650	1,360,914	238,577	229,265	178,039
Italy	1,022,204	337,322	1,359,526	194,650	227,599	422,249	173,250	168,410	130,168
France	2,422,338	1,587,490	4,009,828	298,886	670,469	969,355	174,917	174,784	153,779
Sweden	288,395	189,424	477,819	96,039	124,862	220,901	126,211	126,168	115,195
Denmark	261,146	124,209	385,355	93,596	57,768	151,364	53,056	52,381	43,584
Spain	2,748,601	165,468	2,914,069	257,739	46,829	304,568	47,541	45,597	34,159
Finland	81	110,123	110,204	38	87,226	87,264	38,159	38,119	28,633
Portugal	747,069	39,086	786,155	124,017	13,638	137,655	11,265	11,212	6,158
Greece	85,915	14	85,929	19,156	6	19,162	10,093	10,082	7,710
Netherlands	194	183,668	183,862	53	66,669	66,722	4,313	4,312	3,751

	Raw Alerts	Alerts Tagged with Single Stock	Alerts With	After Filtering	First In	Second In
	RTRS	RTRS	Returns Matched	by Novelty	Take Sequence	Take Sequence
US	4,976,374	4,054,683	3,286,003	2,935,852	1,296,733	522,258

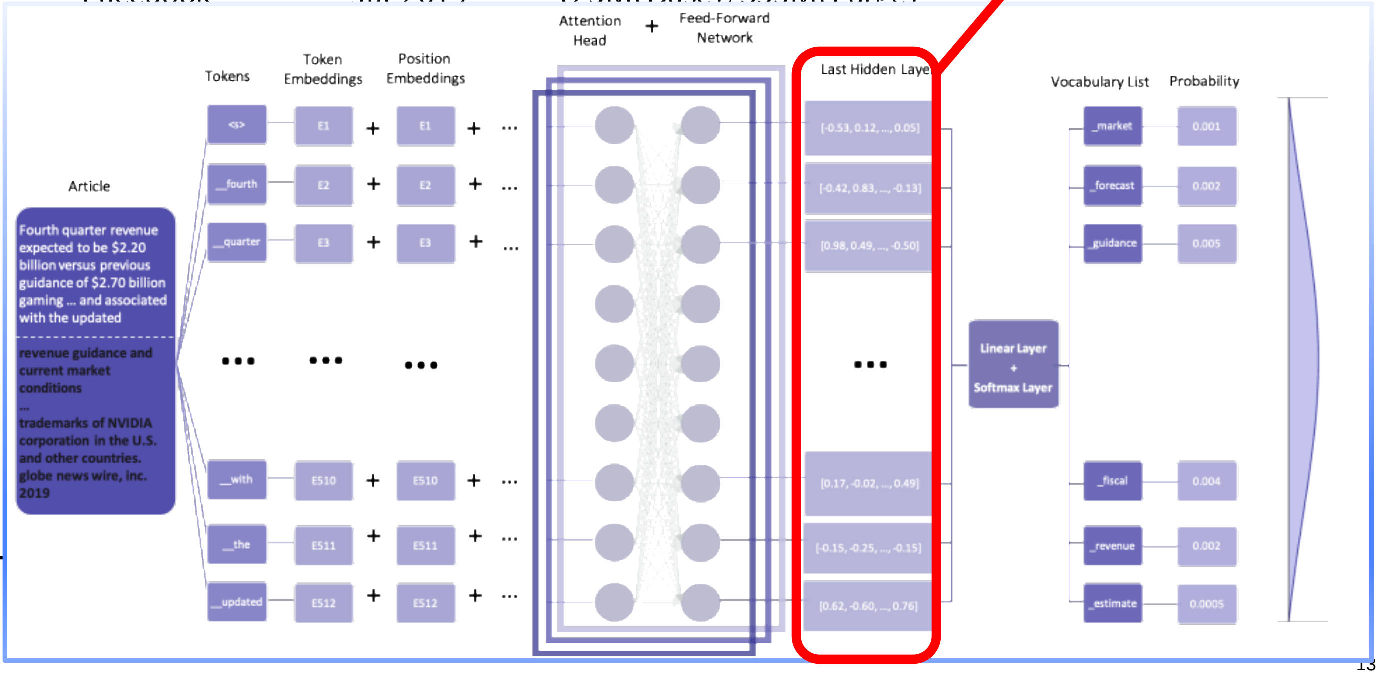
Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Prediction Methodology

The World of Large Language Models

Model	Company	Release Date	# of Parameters
BERT	Google	Oct 2018	110M(Base)/340M(Large)
GPT-2	OpenAI	Feb 2019	1.5B
RoBERTa	Facebook	Jul 2019	125M(Base)/355M(Large)
Turing-NLG			
GPT-3			
Wu Dao 2.0			
Jurassic-1			
GLaM			
Exaone			
Megatron-Turing NLG			
OPT			
GPT-3.5 (ChatGPT)			
LLaMA			
GPT-4			
LLaMA 2			

$X_{i,t}$
(Dx1)



Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Prediction Methodology
Expected Returns

- Sentiment Analysis: treated as a classification problem

$$\boxed{\mathrm{E}(y_{i,t}|x_{i,t}) = \sigma(x'_{i,t}\beta),} \quad \text{where} \quad \sigma(x) = \exp(x)/(1+\exp(x)),$$

  and $y_{i,t}$ is the label, i.e., the sign of three-day cumulative return surrounding the news event on day t for stock i.

- Return Prediction: treated as a panel regression problem

$$\boxed{\mathrm{E}(r_{i,t+1}|x_{i,t}) = x'_{i,t}\theta,}$$

  where $r_{i,t+1}$ is the return of stock i on day $t+1$.

- In the case of high-dimensional features $(x_{i,t})$, we adopt ridge regressions.
- Alternatively, one can employ a neural network model between $y_{i,t}$, $r_{i,t+1}$, and $x_{i,t}$.

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Pre-LLM Benchmarks

Bag-of-Words Methods: Article represented as vector of word counts

- **LMMD** (Loughran, MacDonald, 2011): Hand-constructed finance sentiment dictionary
- **SESTM** (Ke, Kelly, Xiu, 2020): Machine learning topic-sentiment model

Early Word Embeddings: A sophisticated "PCA" of word indicator vectors

- **Word2vec** (Mikolov et al., 2013): two-layer neural network model to generate embedding vectors

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

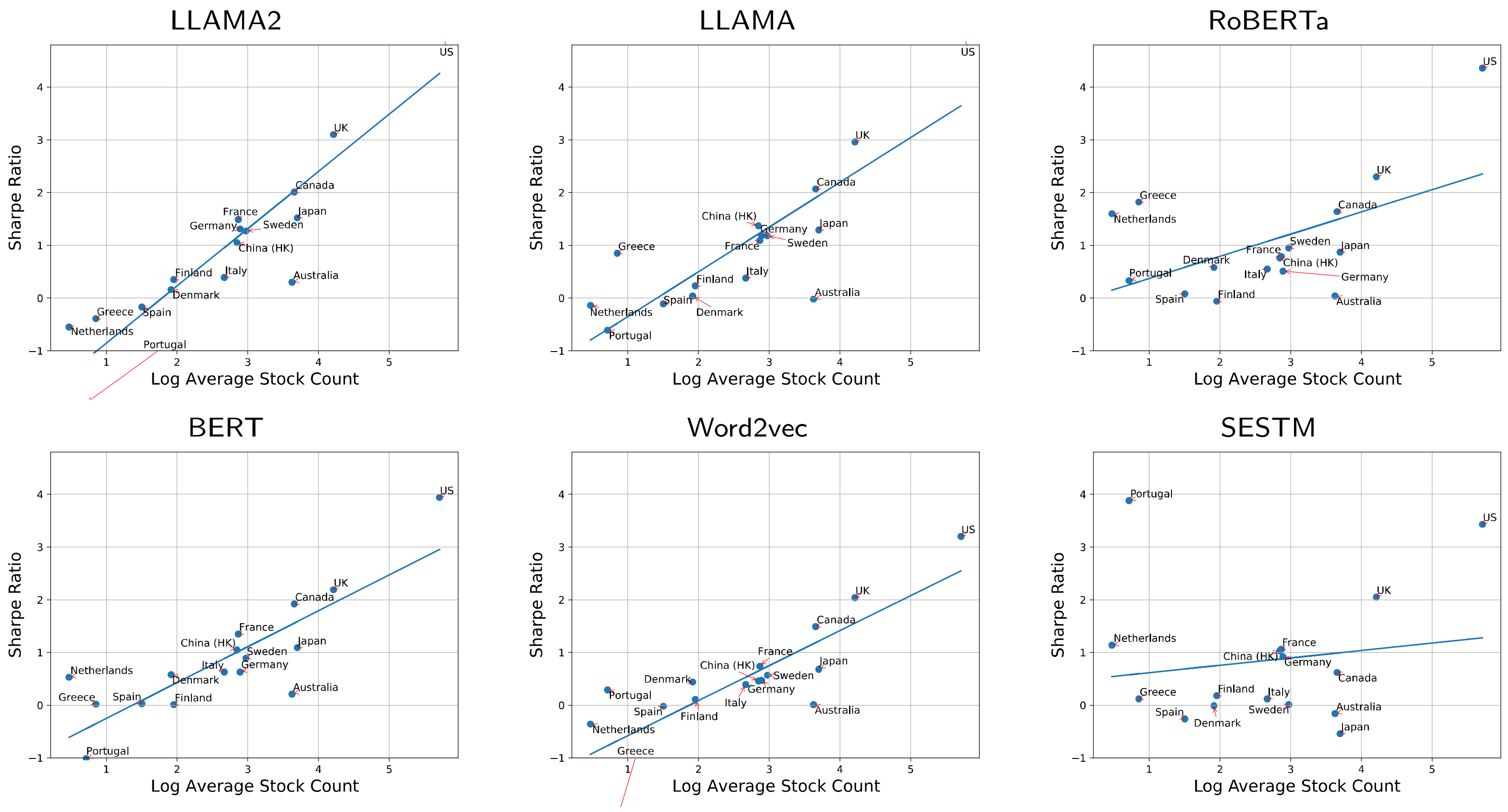
Daily Predictions

Portfolio Performance (Daily Prediction)

	ChatGPT						LLAMA2					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.34	-0.14	0.48	0.19	0.04	0.15	0.35	-0.10	0.45	0.18	0.07	0.11
Std	0.20	0.22	0.10	0.19	0.22	0.11	0.20	0.23	0.11	0.19	0.22	0.11
SR	1.71	-0.62	4.62	1.03	0.18	1.41	1.75	-0.43	4.16	0.97	0.33	0.98
	LLAMA						RoBERTa					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.34	-0.07	0.41	0.19	0.08	0.11	0.33	-0.06	0.39	0.20	0.09	0.11
Std	0.20	0.23	0.11	0.19	0.22	0.11	0.20	0.22	0.10	0.19	0.22	0.11
SR	1.67	-0.33	3.89	1.02	0.36	1.04	1.62	-0.29	3.75	1.08	0.43	0.94
	BERT						Word2vec					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.32	-0.04	0.36	0.16	0.07	0.10	0.29	-0.01	0.30	0.18	0.08	0.09
Std	0.20	0.22	0.10	0.18	0.21	0.10	0.21	0.22	0.10	0.19	0.21	0.10
SR	1.59	-0.19	3.60	0.89	0.31	0.92	1.41	-0.05	3.06	0.93	0.40	0.92
	SESTM						LMMD					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.31	-0.03	0.34	0.18	0.09	0.09	0.24	0.01	0.22	0.14	0.10	0.04
Std	0.20	0.22	0.10	0.19	0.21	0.11	0.20	0.23	0.10	0.18	0.21	0.10
SR	1.53	-0.14	3.43	0.97	0.42	0.86	1.18	0.06	2.29	0.77	0.47	0.39

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Polyglot Portfolios



Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Complexity / Nonlinear Prediction Is Even Better

	RF						LASSO					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.32	-0.03	0.34	0.19	0.09	0.10	0.38	-0.04	0.42	0.15	0.07	0.08
Std	0.20	0.22	0.11	0.19	0.20	0.11	0.19	0.21	0.10	0.18	0.19	0.10
SR	1.55	-0.12	3.25	0.97	0.45	0.88	2.01	-0.21	4.14	0.84	0.37	0.78
	RIDGE						NN					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.46	-0.11	0.57	0.23	0.09	0.14	0.53	-0.15	0.68	0.24	0.07	0.17
Std	0.21	0.22	0.11	0.20	0.20	0.11	0.21	0.22	0.12	0.21	0.20	0.12
SR	2.22	-0.50	5.31	1.14	0.44	1.32	2.49	-0.66	5.83	1.15	0.36	1.44

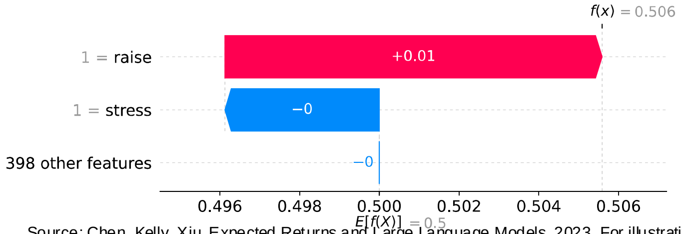
Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

When/Why Do LLMs Disagree with Word-based Methods?

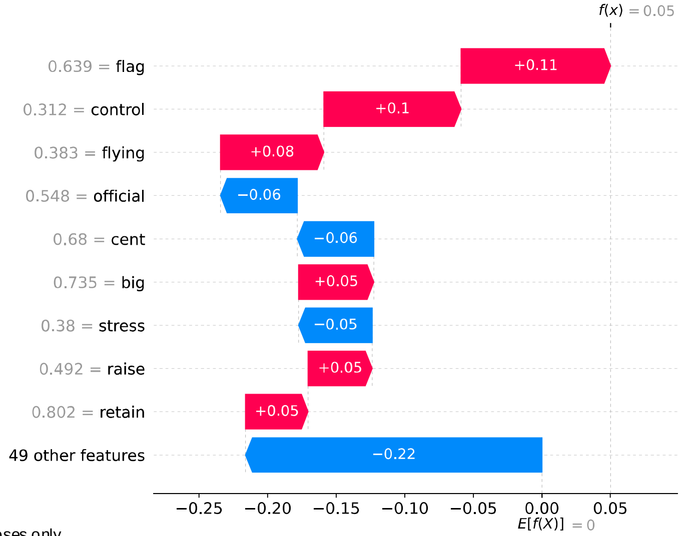
LLAMA2

Brussels has warned British Airways owner IAG ICAG.L that its favoured strategy to allow it to continue flying freely in and around Europe in the event of a nodeal Brexit will not work, the Financial Times reported on Tuesday. After Brexit, European carriers will have to show they are more than 50 per cent EUowned and controlled to retain flying rights in the bloc, the FT said. IAG, which also owns the Spanish flag carrier Iberia, is registered in Spain but headquartered in the United Kingdom and has diverse global shareholders. The FT said part of IAG's strategy to retain both EU and UK operating rights is to stress that its important individual airlines are domestically owned through a series of trusts rather than being part of the bigger a high proportion of nonEU investors. The FT quoted an unnamed senior EU official as saying, "For IAG, I can't see how it can be a solution." Concerns have been raised with IAG over its postBrexit ownership structure, the FT quoted a second Brussels official familiar with the conversations as saying. IAG was not immediately available.

BOW



W2V



Source: Chen, Kelly, Xiu, Expected Returns and Large Language Models, 2023. For illustrative purposes only.

Negation Portfolios

	ChatGPT						LLAMA2					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.40	-0.15	0.56	0.43	-0.23	0.66	0.35	-0.07	0.42	0.48	-0.22	0.70
Std	0.21	0.24	0.13	0.21	0.25	0.17	0.21	0.24	0.13	0.22	0.25	0.17
SR	1.96	-0.64	4.34	2.05	-0.90	3.98	1.70	-0.28	3.29	2.21	-0.87	4.18

	LLAMA						RoBERTa					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.36	-0.06	0.43	0.50	-0.21	0.71	0.34	-0.07	0.41	0.51	-0.19	0.70
Std	0.21	0.24	0.13	0.22	0.25	0.17	0.21	0.24	0.13	0.22	0.24	0.16
SR	1.74	-0.27	3.34	2.32	-0.82	4.23	1.64	-0.30	3.14	2.37	-0.76	4.35

	BERT						SESTM					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.33	-0.03	0.36	0.45	-0.11	0.56	0.33	-0.05	0.38	0.38	-0.01	0.40
Std	0.21	0.23	0.12	0.22	0.25	0.17	0.21	0.24	0.13	0.22	0.25	0.15
SR	1.56	-0.14	2.94	2.06	-0.45	3.37	1.57	-0.22	3.00	1.78	-0.05	2.58

	Word2vec						LMMD					
	W/O Negation Words			W/ Negation Words			W/O Negation Words			W/ Negation Words		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.28	-0.04	0.32	0.32	-0.01	0.33	0.26	-0.03	0.28	0.29	0.04	0.25
Std	0.21	0.23	0.12	0.22	0.24	0.15	0.21	0.24	0.12	0.21	0.24	0.15
SR	1.35	-0.18	2.71	1.49	-0.02	2.21	1.25	-0.11	2.31	1.35	0.17	1.66

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Monthly Predictions

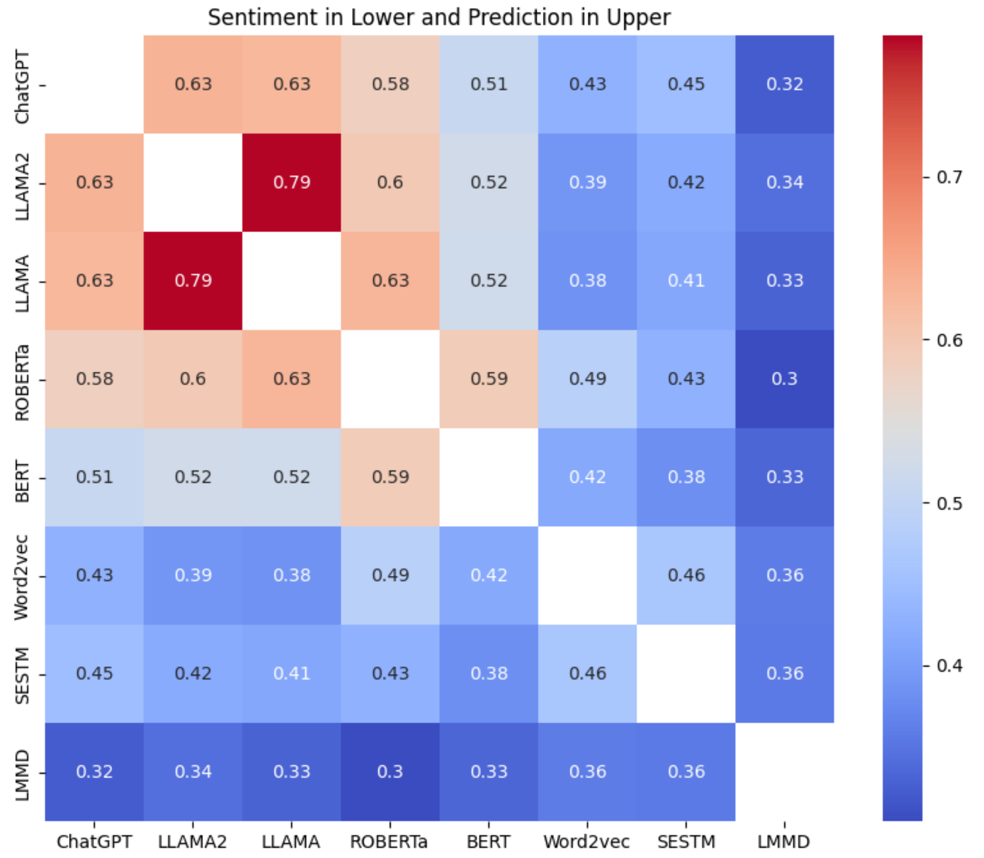
Portfolio Performance (Monthly Prediction)

	1 month						3 month					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.11	0.07	0.03	0.11	0.07	0.03	0.12	0.06	0.06	0.11	0.05	0.05
Std	0.20	0.19	0.06	0.18	0.15	0.09	0.21	0.19	0.07	0.19	0.15	0.10
SR	0.56	0.36	$\boxed{0.56}$	0.61	0.46	0.36	0.58	0.29	$\boxed{0.83}$	0.61	0.33	0.57
	6 month						12 month					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.13	0.05	0.07	0.11	0.05	0.05	0.13	0.05	0.07	0.10	0.05	0.04
Std	0.21	0.19	0.08	0.19	0.16	0.10	0.21	0.19	0.08	0.19	0.16	0.11
SR	0.61	0.28	$\boxed{0.88}$	0.59	0.30	0.54	0.60	0.27	$\boxed{0.88}$	0.50	0.30	0.36
	24 month						36 month					
	EW			VW			EW			VW		
	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S	Long	Short	L-S
Ret	0.13	0.05	0.07	0.10	0.05	0.05	0.12	0.05	0.07	0.11	0.05	0.05
Std	0.21	0.19	0.07	0.19	0.16	0.11	0.21	0.19	0.07	0.19	0.16	0.12
SR	0.62	0.25	$\boxed{1.00}$	0.53	0.29	0.41	0.58	0.25	$\boxed{0.88}$	0.57	0.29	0.46

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Multiple Models

Diversity in Language Models



Best Individual Model

	EW		
	Long	Short	L-S
Ret	0.34	-0.14	0.48
Std	0.20	0.22	0.10
SR	1.71	-0.62	4.62

Ensemble of All Models

EW		
Long	Short	L-S
0.45	-0.10	0.54
0.21	0.22	0.11
2.16	-0.45	5.11

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Conclusions

Embeddings from LLMs effective and comprehensive numerical representation of text content

Funnel for open-minded extraction of text signal for return prediction

Contrast with filtering/limitations of human-generated prompts

A combination of approaches is likely to dominate either (Bayesian interpretation)

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Conclusions

- Strong out-of-sample success compared to existing predictive signals in the literature
- Larger LLMs perform better
- Polyglot methodology
- Multiple frequencies in news signals for markets
  - Fast and slow return prediction content in news text
- Many models to choose from
  - Not all are accessible with prompts
  - Best strategy is an ensemble of many LLMs

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.

Appendix

Prediction Methodology
Embedding Construction Detail

- Transforming a sequence of words into embeddings through a deep learning model
- Tokenization: "macroeconomics" $\Longrightarrow$ "macro" + "economic" + "s"
- Directly leveraging pre-trained LLMs to generate token embeddings that serve as features.
  - BERT (large) processes up to 512 tokens, outputting a 1,024-dimensional vector per token.
  - LLAMA1 (LLAMA2) takes in up to 2,048 (4,096) tokens, producing a 5,120-dimensional vector.
  - ChatGPT (text-embeddings-3-large) can manage sequences as long as 8,192 tokens and embed each token into a 3,072-dimensional space.
  - We use the simple average of (up to 512) token vectors to represent an article using each LLM, except for ChatGPT. For robustness, we show this does not lead to much loss of information.

Source: Chen, Kelly. Xiu, Expected Returns and Large Language Models. 2023. For illustrative purposes only.