# Beating the Market: ChatGPT vs Google Gemini Quantitative Sentiment Analysis using Model LLMs

Kabir Mann

May 24, 2024

## Abstract

I examine the efficacy of using OpenAI's ChatGPT and Google's Gemini Large Language Models (LLMs)[1] to analyze news headlines and make investing decisions to go long in, short, or ignore Apple's stock (NASDAQ: APPL) on a daily basis. This model had two primary goals: (1) Expand upon Lopez-Lira and Tang's research[2] by integrating Gemini and comparing its performance with that of ChatGPT. (2) Determine if there are information inefficiencies spanning longer than 24 hours in top news headlines being priced into assets. The Standard & Poor's 500 Index (SPX) was used as a benchmark, and the 13-week Treasury (current for every given date) was used as the money market rate for an alternative investment. In terms of Sharpe Ratio[3], Gemini Advanced outperformed the other models, including the SPX.

Key Words: Large Language Models, ChatGPT, Gemini, Sentiment Analysis, Machine Learning, Natural Language Processing

[1] As of May 22, 2024, Open AI has three LLMs, ChatGPT 3.5, ChatGPT 4, and ChatGPT 4o ("o" for "omni"), which is its newest model, faster than GPT4. For this comparison ChatGPT4o was used instead of ChatGPT4. Gemini has a free version and Gemini Advanced. The free version has no official moniker, so it is referred to as "Gemini Free" in this paper. Both ChatGPT 4o and Gemini Advanced are $20/mo.

[2] "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models" Lopez-Lira and Tang at The University of Florida February 28, 2024

[3] Sharpe Ratio is a way to measure portfolio or individual asset returns adjusted for the risk taken. The exact formula for is explained later in this paper.

# 1. Introduction

Using sentiment of mass media (and social media) to drive investment decisions is not a novel concept. Since 2016, Bloomberg has had data on sentiment analysis for media, utilizing a proprietary supervised Machine Learning model to give a “Positive,” “Negative,” or “Neutral” rating for tweets on X (formerly Twitter). Hedge funds and quant shops have also developed proprietary models for sentiment analysis, from the analysis of earnings call transcripts to Jerome Powell’s press releases. While Machine Learning methods have been around for some time, the recent commercialization of Large Language Models has created a new avenue to attempt to predict publicly-traded stock price movements. With these models, I will discuss whether there are delays in information being priced into assets, and if there are, how one can capitalize on them.

Open AI’s ChatGPT has been a leader in the consumer-facing LLM landscape since its inception. Released on November 30, 2022, ChatGPT reached 100 million users in just two months, setting a record for the fastest-growing user base. Other companies have tried to keep up with this pace (or, in Microsoft’s case, *purchase* it for $13B). Through Gemini (formerly Bard), Google has not seen broad expansion with immediate adoption like that of ChatGPT. This has been the case despite three-month free trials and other strong initiatives by Google, none of which OpenAI have available. With that in mind, there has been research that shows Gemini as more capable in producing factual results, due to its high integration with Google Search (N. Rane, Choudary, et al. 2024). With the majority of the headlines used being readily available on Google Search, this could give Gemini’s performance the edge over ChatGPT.

Before diving into the details of this research, it is important to address several limitations and constraints that could have influenced the results of my research. Additionally, these constraints prevented me from perfectly replicating the primary research (Lopez-Lira, Tang 2023) that mine expands on:

1. Information Costs: For this research, I was limited to free data, such as the NewsAPI Python package. This package only offers the previous one month of data for free. Beyond that, the subscription starts at $449 per month. Additionally, I was not able to put

these headlines through RavenPack to filter out headlines that were not relevant to the stock. This was a crucial part of cleaning the data for Lopez-Lira and Tang 2024. I reached out to RavenPack, however the pricing was $20k to $36k per annum, after the student discount. One more option was Dow Jones Wall Street Journal API. This not only includes headlines, but also has text for the body of WSJ articles. The exact pricing for this was not posted, and their customer support representative did not reach out in time for this research.

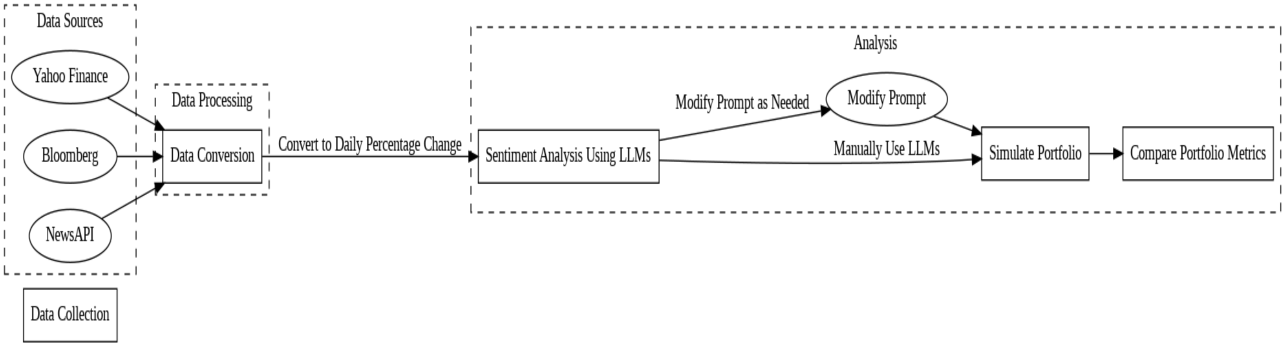
2. Backtesting: Because of the one-month time frame, almost all backtesting methods were rendered useless. Even with the paid version of NewsAPI, only 10 years of data is available, which would not capture 2008 and other major regime shifts prior to 2014.
3. Trade Timing: This model executes trades the day after the headlines are released. Because only the dates were given for headlines and not the timestamps, the only way to ensure that trades were executed *after* the headlines were released was to do them the trading day following the headlines. This means that Friday's news wasn't traded until the following Monday, even if the news was released prior to the closing bell[4] on Friday.
4. Transaction Costs: This model does not reflect transaction costs that would be apparent in real trading. Tax implications, annual turnover, and margin costs to name a few. These costs are relevant and have the potential to be decision-impacting, especially due to the large asset turnover required in this type of strategy.
5. Economic Conditions: This model was trained on data in April to May of 2024, a high-interest rate environment. This contributes to the alternative investment (the money market) being more attractive and yielding larger returns. In a low-interest rate environment, we could see missed return opportunities on days where SPX performed even slightly well but our capital was in the money market.
6. Reallocation of Funds: When shorting[5] an asset, the funds from the short position can be reallocated into a different long position. This is what has driven returns for the 130/30

[4] The closing bell is a bell that rings at the end of the trading day for a stock exchange. The New York Stock Exchange (NYSE) is the most famous exchange that uses a closing bell.

[5] A short sale is when the seller of an asset doesn't actually own it, but rather borrows it, anticipating a decline in trading price. Once the price declines, the seller repurchases the asset on the open market and returns it to the lender. This can be a highly risky method of generating return, as the downside potential is theoretically infinite if the stock asset's price increases instead of declining.

hedge fund long-short strategy[6] (Yontar, Faleyev, et al. 2019). This research ignores that possibility, and includes returns from a short position if *and only if* that position pans out as expected.

# 2. Workflow



## *2.1 Data Collection & Processing*

The first step in this process was data collection. I first used Bloomberg for daily closing prices. After comparing these figures with Yahoo Finance, I chose Yahoo Finance so this program could be run entirely for free and without needing a Bloomberg Terminal. NewsAPI was used to pull the headlines for each stock for trading days. The majority of the time was spent on this step. With the cost constraints I mentioned above, it was challenging to get accurate information. An attempt was made to parse through the HTML of several media websites, however this yielded no results. With a stronger background in Data Science, this would be possible. After using NewsAPI to pull the headlines, I had a dataframe with dates and headlines. I also had a separate data frame with Apple's prices.

After that information was inputted, I cleaned the data. This included two main steps. First, I added SPX and 13-week treasury data. The SPX values were used purely as a benchmark for the metrics calculations. Once that was done, I ensured that all price data was in daily percentage change format, using this formula, where $P_t$ is the price on day $t$.

[6] A long-short strategy is a strategy that ranks assets based on a given criteria, shorting the bottom x% and going long (investing) in the top x%. The funds from selling the asset in the short can be reallocated to the long position, leading to a total investment of over 100% (130% long /30% short).

$$\% \, change \; = \frac{price_{t}}{price_{t-1}} - 1$$

## *2.2 Sentiment Analysis Using LLMs*

The prompt given to the LLMs in my research was similar to that of Lopez-Lira and Tang 2024. ChatGPT 3.5 does not return .csv files, so to keep the prompt uniform between all models, I didn't explicitly ask for that format. Here is the prompt I gave to all of the models:

> "Pretend you are a financial expert. You are a financial expert with stock recommendation experience. I have uploaded data with dates and headlines for apple. I want you to return two additional data for each date. It should be about the sentiment of the headline. Answer "YES" if good news, "NO" if bad news, or "UNKNOWN" if you are uncertain. Additionally, elaborate with one short and concise sentence."

This led the models to each return the requested data in different formats. I converted them all to be in a standard .csv format myself. ChatGPT 3.5 gave unstructured data with the sentiment, followed by a sentence of explanation. This required manual data entry. ChatGPT 4o returned a .csv file with columns for "Date," "headline," "sentiment," and "explanation." These explanations were uniform, meaning that all "Negative" sentiments had the same explanation regardless of the headline content. Gemini Free returned a Google Sheet with similar headlines.

## *2.2 Portfolio Simulation*

Once I had the sentiment analysis from the LLMs, I was able to simulate a portfolio over the timeframe that I had. If the model indicated "YES," a long position was taken in Apple the following day. If the model indicated "NO," a short position was taken. If the model indicated "UNKNOWN," the portfolio shifted into a money market account, with the rate being the average 13-week treasury for a given date.

Here is the formula that was used to simulate the portfolio:

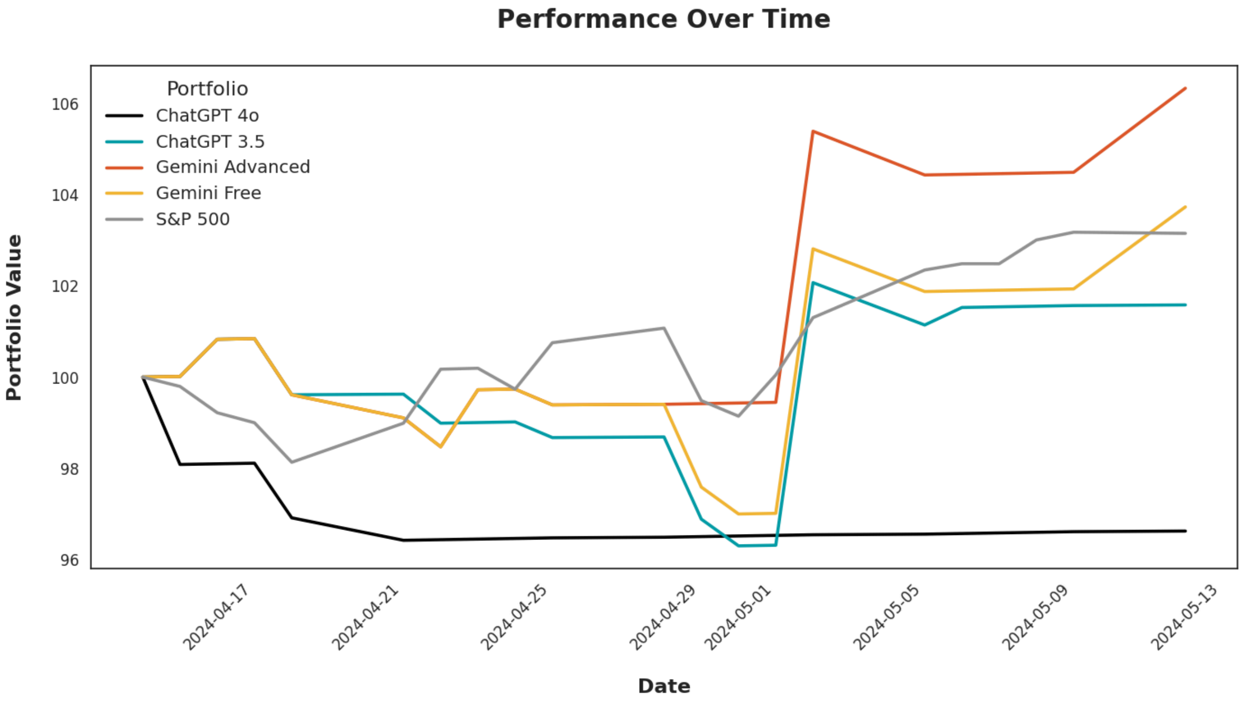
Let $P_0$ be the initial portfolio value, $P_t$ be the portfolio value on day *t*, $s_t$ be the sentiment on day *t* (where $s_t \in \{NO, \, YES, \, UNKNOWN\}$), $\Delta p_{t+1}$ be the percentage price change from day *t* to day $t + 1$. $r_{t+1}$ be the money market rate from day $t$ to day $t + 1$, and $I$ (condition) as an indicator function that is a binary 1 if true and 0 if not.

$$P_{t+1} = P_0 \times \prod_{t=0}^{n-1} (1 + \frac{\Delta p_{t+1} I(s_t = YES) - \Delta p_{t+1} I(s_t = NO) + r_{t+1} I(s_t = UNKNOWN)}{100}$$

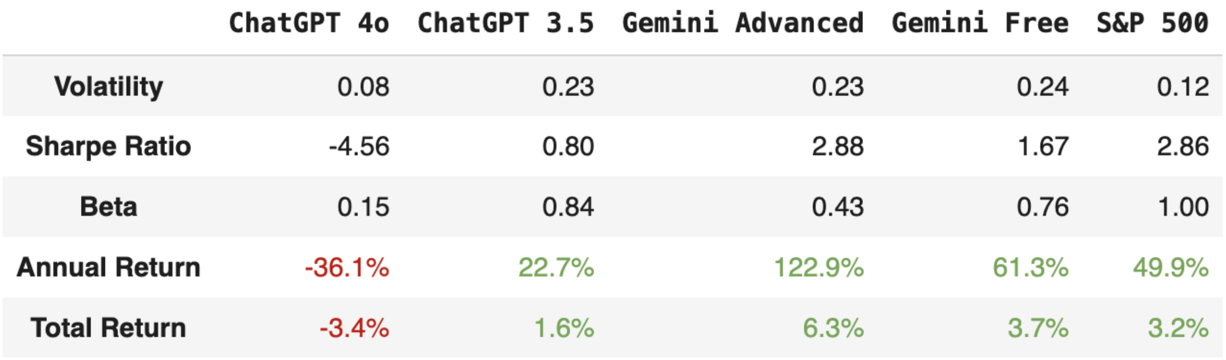
# 3. Results

## *3.1 Portfolio Results & Key Metrics*

This strategy yielded interesting results, different from what an average LLM user might expect. Despite seeing significantly less user base growth than ChatGPT, both Gemini models outperformed the ChatGPT models, of which ChatGPT 4o performed the worst. Here is the performance for a portfolio over one month that started with $100.



Gemini Advanced saw the highest returns, followed by Gemini Free, which barely outperformed the SPX. ChatGPT 3.5 still generated positive returns, and ChatGPT4o saw a capital loss. Gemini Advanced not only surpassed it in terms of total return, but was able to do so with strong financial metrics.



| | ChatGPT 4o | ChatGPT 3.5 | Gemini Advanced | Gemini Free | S&P 500 |
|---|---|---|---|---|---|
| Volatility | 0.08 | 0.23 | 0.23 | 0.24 | 0.12 |
| Sharpe Ratio | -4.56 | 0.80 | 2.88 | 1.67 | 2.86 |
| Beta | 0.15 | 0.84 | 0.43 | 0.76 | 1.00 |
| Annual Return | -36.1% | 22.7% | 122.9% | 61.3% | 49.9% |
| Total Return | -3.4% | 1.6% | 6.3% | 3.7% | 3.2% |

Gemini Advanced's volatility[7] was higher than the SPX by .11, but it was in line with ChatGPT 3.5 and Gemini Free. To calculate volatility, daily portfolio value was converted to daily returns using a similar formula to that above. In this formula, $R_t$ is the daily return on day *t*, and $P_t$ is the price on day *t*.

$$R_t = \frac{P_t}{P_{t-1}} - 1$$

Once daily returns were calculated, daily standard deviation could be calculated using the following formula.

$$\sigma_{daily} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(R_t - R_{mean})^2}$$

This can then be annualized, using the square root of 252 for the total trading days per annum.

$$\sigma_{annualized} = \sigma_{daily} \times \sqrt{252}$$

Even though the volatility was significantly higher than that of the SPX, once the volatility adjusted returns were considered (i.e., Sharpe Ratio), Gemini was the winner, with a Sharpe Ratio of 2.88, 0.02 greater than the SPX. To calculate Sharpe Ratio, the inputs needed are

[7] Volatility is a measure of how rapidly stock prices (or in this case, portfolio values) move up and down. Higher volatility equates to higher risk in Modern Portfolio Theory. Annualized standard deviation is used as volatility here.

the expected return—the annualized average of historical returns($R_{mean}$)—the risk-free-rate[8] ($R_{rf}$), and the standard deviation of the returns (σ). For academic purposes, a risk-free-rate of 0% was assumed.

$$Sharpe\ Ratio\ = \frac{R_{mean} - R_{rf}}{\sigma}$$

Another important metric is beta[9]. In this case, the beta values must be taken with a grain of salt, rather several grains of salt. Beta is typically calculated over the span of three to five years, looking at weekly or monthly movements. Here, beta is calculated daily, with only a month of data. $R_p$ represents the portfolio returns, while $R_b$ represents the returns of the benchmark, which was the SPX in this case.

$$\beta\ = \frac{Cov\ (R_p, R_{b)}}{Var(R_b)}$$

### *3.2 Backtesting*

Ideally, this strategy would initially be tested using a simple paper trading[10] backtest, going back to before the 2008 housing market crash. A backtesting method that would ideally be used to initially test this strategy would be a simple paper trading backtest, going back to before the 2008 housing market crash. This would allow us to see if the strategy could weather multiple different regime shifts. One concern mentioned earlier was whether this strategy would work well in low-interest-rate environments. Paper trading with more historical data could test this. Unfortunately, because of the data limitations, such backtest methods are not available.

[8] In investing, the risk-free rate is the rate at which an investor can invest his money without the risk of losing it and with a guaranteed return. The 10-week treasury is typically used for this.

[9] Beta, the second letter in the Greek alphabet, is a metric that shows how an asset or portfolio moves respective to a given benchmark. In this example, SPX was used as the benchmark. A beta of 1.0 shows that when the SPX moves by 1%, this portfolio moves by 1% in the same direction. Beta values above 1.0 are typical in growth stocks, such as those in the Technology sector, while a beta value below 1.0 is typical for stable stocks in the Consumer Staples or Telecommunications sectors.

[10] Paper trading the simulation of real investments with artificial money ("paper"). This can be useful to ensure that all costs are being considered in an investment strategy, as well as to see the strategy's application in a real scenario.

# 4. Conclusion

## *4.1 Future Improvements*

Although this research generated insights into the potential of using LLMs as the primary drivers of sentiment analysis investment strategies, there is much more to explore. Implementing these future improvements would make this research more holistic and validate the strategy..

1. Manual Prompting: Another cost constraint was that of the LLMs that I was using. Rather than using the API tokens, which have a cost that is not able to be determined upfront, the chatbots themselves were used. This is all right, as I was only using one-month of data. Had I been doing 10 years, or even one year, the APIs would be a crucial part of streamlining this process.
2. Data Entry: ChatGPT 3.5 does not give structured data. This required manual data entry of each response. This was not an issue with one month of data, but ten years or more would require several hours of time purely for data entry. Using a different prompt could help with this. ChatGPT 4.0 or Google Gemini Advanced could also be used to turn ChatGPT 3.5's unstructured data into a .csv file or Google Sheet.
3. Time Delays: Real time execution is necessary. Dow Jones' Top News API allows it. Once a news headline is released, trades should be executed immediately to capitalize further on the delay in the information being priced into the asset.
4. More Data: A larger media time window is *necessary* for this research to be worth the paper it's printed on. Lack of backtesting leads to an uncertainty of how this will perform in a bear market or even a bull market with different characteristics. Even the ten years of data that comes with NewsAPI's premium monthly subscription should be enough to strengthen or negate this research.
5. Optimized Prompts: ChatGPT 4o performed significantly below expectations. There was an effort to optimize the prompt to get a more sufficient response. This was to no avail, as ChatGPT 4o was still cautious, generating mostly "UNKNOWN" sentiment ratings with vague rationale. Spending more time on prompt engineering could lead to more optimal returns across models.

6. Asset Universe: Increasing the asset universe would be important to test this strategy. Using all assets in the SPX, akin to Lopez-Lira and Tang 2024, or using the Russell 3000 could yield interesting results. Diving into a more inefficient market, such as small caps on foreign exchanges could improve the returns and financial metrics even more.
7. Sentiment Overlay: Using sentiment as an overlay for a different strategy could retain the benefits of that strategy, while optimizing it for better performance. A standard momentum or technical strategy with a sentiment overlay should be tested.

### *4.2 Gemini vs ChatGPT*

Both Gemini Advanced and Gemini Free outperformed ChatGPT. Both Gemini models gave valid reasoning for each sentiment given. Gemini Advanced returned results in the same format as Gemini Free, but with more depth and human-like interpretation. The following is an example of a headline and all of the models' recommendations and explanations:

> *"Apple,Äö√Ñ√¥s big AI rollout at WWDC will reportedly focus on making Siri suck less"*
>
> GPT 3.5: UNKNOWN: The headline suggests upcoming improvements to Siri but doesn't specify the overall impact on Apple's performance.
>
> GPT 4o: UNKNOWN: Sentiment unclear from the headline.
>
> Gemini Free: YES: This headline suggests Apple is working to improve Siri, which could be positive for users.
>
> Gemini Advanced: YES: While "suck less" is negative, the overall sentiment is positive as it implies improvement.

As shown in this example, the analysis from both Gemini models far surpassed those of the ChatGPT models. ChatGPT 3.5 gave some elaboration, which is beneficial. ChatGPT 4o had the worst analysis. ChatGPT can be pedantic, focusing on small character or word changes. In this example, NewsAPI changed the apostrophe into these odd characters. Thorough research on this has been conducted by Shen, Chen, et al. 2023, showing how changes similar to this can cause strain and accuracies for ChatGPT. Additionally, OpenAI has made efforts to reduce "jail breaking" of ChatGPT. That is, OpenAI has expended significant efforts to subduing ChatGPT, preventing it from giving opinions that are controversial. This second part is speculation, as significant research has not been done on it. Both of these concepts could be contributing to ChatGPT's underperformance.

### *4.3 Information Inefficiencies*

A primary question this research aimed to answer was whether there were delays in media sentiment being priced into stock prices. This research concludes that there are delays that can be capitalized on. Information travels rapidly, but using LLMs may offer ways to generate alpha even in the current highly efficient market.. This was tested on Apple, which has the second largest market cap in the world. This leads news to be priced in more rapidly than other stocks. Even with that, there are enough delays to justify an LLM-based sentiment strategy.

### *4.4 Final Thoughts*

Although there is more to be done to solidify this research, the results achieved indicate that more research in this field could yield positive results and should be conducted. Implementing the points in *Section 4.1 Future Improvements* would be crucial to this improved research. Additionally, ChatGPT, Gemini, and newer LLMs should continue to be compared side-by-side in order to see which one is evolving most ideally. Today, Gemini Advanced is the winner of capturing information inefficiencies in the stock market, but tweaks for ChatGPT or an optimized LLM designed purely for sentiment analysis could surpass Gemini Advanced.

## References

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4412788

Rane, N., Choudhary, S., & Rane, J. (2024). Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4723687

Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). *In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT*. https://doi.org/10.48550/arxiv.2304.08979

Teti, E., Dallocchio, M., & Aniasi, A. (2019). The relationship between twitter and stock prices. Evidence from the US technology industry. *Technological Forecasting and Social Change*, *149*, 119747. https://doi.org/10.1016/j.techfore.2019.119747

Yontar, PhD, T., Faleyev, M., Benham, CFA, CAIA, F., & Festino, CFA, CAIA, L. (2019, September). *130/30 Long-Short Equity Strategies*. Meketa Investment Group. https://meketa.com/wp-content/uploads/2012/10/130-30-FINAL.pdf