# GPT-Signal: Generative AI for Semi-automated Feature Engineering in the Alpha Research Process

**Yining Wang**[1,3], **Jinman Zhao**[2], **Yuri Lawryshyn**[3]
[1]Division of Engineering Science, [2]Department of Computer Science,
[3]Centre for Management of Technology & Entrepreneurship(CMTE),
University of Toronto, Toronto, ON, Canada,
yning.wang@mail.utoronto.ca, jzhao@cs.toronto.edu, yuri.lawryshyn@utoronto.ca

## Abstract

In the trading process, financial signals often imply the time to buy/sell assets to generate excess returns compared to a benchmark (e.g., an index). Alpha (Kakushadze, 2016) is the portion of an asset's return that is not explained by exposure to this benchmark, and the alpha research process is a popular technique aiming at developing strategies to generate alphas and gain excess returns. Feature Engineering, a significant pre-processing procedure in machine learning and data analysis that helps extract and create transformed features from raw data, plays an important role in algorithmic trading strategies and the alpha research process. With the recent development of Generative Artificial Intelligence(Gen AI) and Large Language Models (LLMs), we present a novel way of leveraging GPT-4 to generate new return-predictive formulaic alphas, making alpha mining a semi-automated process, and saving time and energy for investors and traders[1].

## 1 Introduction

In quantitative finance, we know many traditional financial signals such as the Price Earning (P/E) Ratio, Price/Book (P/B) ratio, Return on Equity (ROE), Return on Assets (ROA) etc. These signals all play an important role in helping people understand the financial situation of a company and get better ideas of the potential of that company in the stock market. The historical stock return data of different companies can be collected for stock market analysis and prediction (Li et al., 2023b). However, people are never enough of the existing traditional signals, and here comes the real magic of feature engineering in the alpha research process — finding new return-predictive signals.

Historically, feature engineering and formulaic alpha research processes have relied heavily on human intuition and experience or complex algorithms (Zhang et al., 2020). Such processes for discovering new features could be overly subjective or time-consuming as they require sufficient domain-specific knowledge, a solid background in data engineering, and robust knowledge of various machine learning algorithms. However, the emergence of Generative Artificial Intelligence (Gen AI) gives us new insights and opportunities to reframe the feature extraction problem by automation.

As Gen AI has been rapidly developing in recent years, LLMs have become increasingly prevalent as a useful tool in real-life data science and deep learning applications among various fields. LLMs (Ouyang et al., 2022; Touvron et al., 2023; Jiang et al., 2024), based on deep neural networks with transformer architecture (Vaswani et al., 2017), are pre-trained on large-scale texts and fine-tuned by using reinforcement learning. The LLMs have strong performance on a variety of tasks such as content generation, question answering, arithmetic reasoning, computer programming and analysis, robust to data poisoning (Lyu et al., 2022), and are reckoned as a high-potential generative tool that can increase the efficiency in industry work and research.

The objective of this paper is to automate the process of generating new stock return-predictive financial signals using a Large Language Model (LLM), specifically GPT-4. The LLM will interpret information about a new financial dataset and create new, and significant signals. This system will utilize the LLM's advanced interpretative abilities to analyze financial texts and data, identify relevant patterns, and create valuable financial signals. Evaluation methods will be used to test the performance of the new signals in comparison to the existing signals; quantitative results will be presented.

[1] Our code will be released at https://github.com/Yiningww/GPT-signal

In this work, we propose using LLM (GPT-4[2], specifically) to generate stock return-predictive new signals semi-automatically, which can help quantitative researchers and investors in the alpha mining process with much convenience and innovation. LLM creates new financial signals based on the user input information in the prompts, including the definition of several existing meaningful financial signals with sufficient coverage, historical signal data of multiple companies, and the respective historical returns at each time point. The process that GPT-4 employs for signal generation is not merely a one-off combination of the existing signals. It involves a series of refinements where the model learns which combinations yield the most informative signals, constantly improving the novelty and relevance of the signals it generates. The newly created signals will be evaluated by proposed evaluation methods. Based on the proposed framework, we conduct experiments on the S&P 500 companies in different sectors during different time frames, to compare with the baseline model and see the performance of new signals created by GPT-4. The main conclusions of our work can be summarized into the following points:

1. LLM(GPT-4) is able to analyze tabular structure data and generate new financial signals that meaningfully predict stock returns. These signals are developed based on the foundations of existing signals, historical data provided, and relevant information, with each new signal accompanied by its unique reasoning process detailed by the LLM.

2. The robustness of the generated signals is maintained when tested across different sectors of companies (i.e. Information Technology(IT), Health Care, Energy) within the S&P 500 index. Similar patterns of the new signals are observed in various selected sectors.

3. The model performance of newly created signals can outperform the models with baseline signals. Generally, the overall performance of these new signals tends to surpass that of the existing signals in all the selected sectors through 5 years (from year 2016 to year 2020).

4. GPT-4 can creatively combine the existing signals in non-linear and higher-order ways that go beyond simple linear combinations. This creative aspect of feature generation often results in signals that offer unique insights and are more than the sum of the existing parts. This data-driven approach to signal construction is designed to discover novel patterns that are not immediately evident.

## 2 Related Work

**LLM x Feature Engineering** The utilization of Context-Aware Automated Feature Engineering (CAAFE) (Hollmann et al., 2024) mentioned in the work has a similar goal to this paper – implementing LLMs in automated machine learning(AutoML) (Hutter et al., 2019), generating new target-predictive features, and demonstrating the potential of LLMs for automating a broader range of data science tasks. CAAFE proposes to leverage the LLM and let the LLM generate codes that modify input datasets, creating target-predictive meaningful features that improve the performance of downstream prediction tasks in a repetitious workflow and with algorithmic feedback. The paper provides insights into our work, especially in prompting strategies for LLMs and evaluating methods of newly created features. LLMs, serving as tabular prediction models (Hegselmann et al., 2023), accept tiny tabular data sets as inputs, along with descriptive information (such as contextual information about the dataset, feature names with contextual information, data types, percentage of missing values, and 10 random rows from the dataset) about the dataset. While CAAFE focuses on various datasets, we focus on financial datasets with multiple companies' historical financial signals and changes in historical returns.

**LLM in Finance** In the financial aspect, LLMs serve an important role in financial report generation, stock/market trends forecast, investor sentiment analysis, customized financial advice service etc., providing insights into market trends, performing risk management and evaluation, and even helping with trading decisions (Zhao et al., 2024). In addition, LLM's capability of processing large-scale text data (Liu et al., 2023) makes it a prospective practice in the field of finance, enabling it to process natural language queries (Deng et al., 2023) and offer immediate advice and support.

In the prospect of LLMs and financial feature engineering, in particular alpha mining, paradigms such as Alpha-GPT (Wang et al., 2023a) are implemented for alpha mining, harnessing the power of human-AI interaction to increase the efficiency of alpha research. In Wang et al. (2023a)'s integration of GPT and alpha research, Alpha-GPT serves as a paradigm that enhances alpha genera-

[2] https://openai.com/gpt-4

tion through improved human-AI interaction. This system leverages a LLM to act as a mediator between quantitative researchers and the alpha search process. Alpha-GPT have three main advantages: First, it can interpret users' trading ideas and translate them into appropriate expressions. Secondly, Alpha-GPT efficiently summarizes top-performing alphas in natural language, making them easier to understand. Finally, users can provide suggestions and modifications for the alpha search, which the model will automatically incorporate into future rounds of alpha mining. Alpha-GPT demonstrates that the output from the LLM can be a valuable reference for analyzing and revising prompt strategies, highlighting the importance of interaction with the LLM.

**LLM Reasoning** Having proved outstanding reasoning abilities, LLMs showcase proficient performance, especially in benchmarks such as arithmetic (Cobbe et al., 2021; Ling et al., 2017) and commonsense (Talmor et al., 2019). Many works have indicated the usefulness of prompting by implementing reasoning with LLMs like Few-shot learning (Brown et al., 2020), Emotional Prompt (Li et al., 2023a) and Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022).

A recent trend highlights the use of LLMs for NLP tasks. For instance, Wan et al. (2023) uses in-context learning strategies on GPT-3 for Relation Extraction (RE). Wang et al. (2023b) and Xie et al. (2023) apply LLMs to the Named Entity Recognition (NER) task. Additionally, LLMs have been utilized for other tasks such as text summarization (Goyal et al., 2023) and sentiment analysis (Sun et al., 2023).

LLMs' ability to understand table reasoning tasks and to analyze tabular data structure has also been confirmed in Chen (2022)'s work, showing that LLMs are capable and competitive at complex reasoning over table structures when combined with Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022). LLMs can attain very strong performance with only a one-shot demonstration. In this work, we include tabular structured data in the prompt to LLM, based on the findings of the studies above, to utilize the LLM's capability in complex reasoning.

## 3 Methodology

### 3.1 Prompt Design

The prompt mainly consists of two steps, building on Langchain's Chase (2022) prompting template. Step 1 is to let GPT-4 generate the definition, the effect on predicting stock returns, and the preferred tendency of a set of existing signals we pick. After GPT-4 generates these definitions, we input this information for the second-step prompt, along with the overall instructions of the problem, several columns of data of some of the selected companies over a specific period, and the query (the actual question) we prompt to GPT-4. Zero-shot COT (Kojima et al., 2022) is used as a reasoning strategy, as the study shows that CoT can increase LLM's accuracy even in a zero-shot learning strategy only by adding a simple prompt "Let's think step by step". A sample prompt is shown in Figure 1, including instructions for GPT-4 to reference; definitions, effect on predicting stock returns, and the preferred tendency of the 10 existing signals; sample data we randomly picked from our dataset; and the actual question (query).

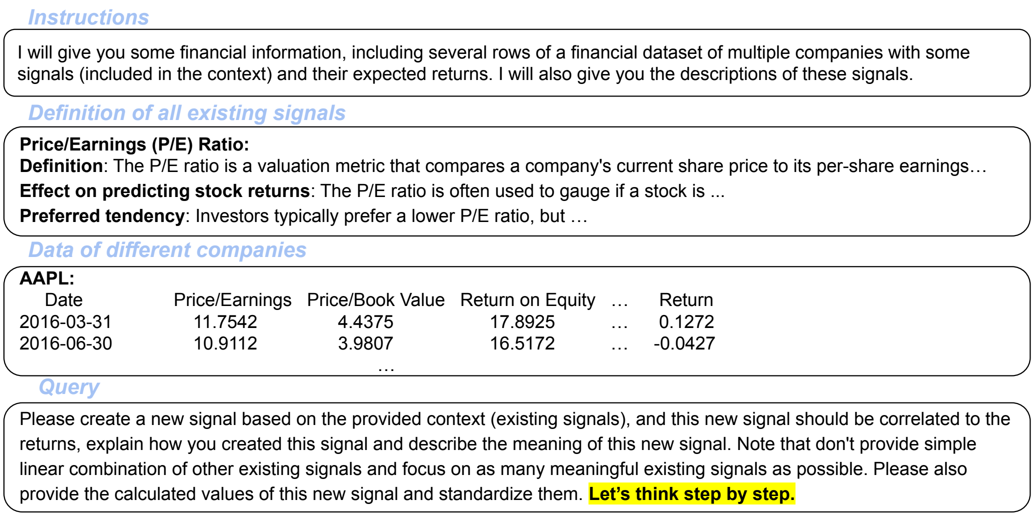
### 3.2 Signal Evaluation

**Spearman Rank Correlation Matrix** Correlation Matrix is a method to measure the correlation between the variables and returns. The correlation coefficient ranges from -1 to 1. A value of 1 implies a strong positive relationship between two variables, -1 implies a strong negative relationship between the two variables, and a coefficient of 0 indicates that there is no linear relationship between the two variables. Traditional correlation matrices include Pearson-type correlations, which can be easily influenced by outliers and nonlinearities. Thus, we use the Spearman Rank Correlation Matrix as an alternative method, as it applies the Pearson correlation formula to the ranks of the data and can reduce distortions that influence the Pearson correlation to some extent. We calculate the correlation at each time point and take the average of the sum of the correlation coefficients.

$$Corr = \frac{1}{n}\sum(Corr_i)$$

where $Corr_i$ is the correlation coefficient of time $i$.

After obtaining the average correlation, heat maps are generated to display the correlation (calculated by the Pearson correlation formula and based



*Instructions*

I will give you some financial information, including several rows of a financial dataset of multiple companies with some signals (included in the context) and their expected returns. I will also give you the descriptions of these signals.

*Definition of all existing signals*

**Price/Earnings (P/E) Ratio:**
**Definition**: The P/E ratio is a valuation metric that compares a company's current share price to its per-share earnings…
**Effect on predicting stock returns**: The P/E ratio is often used to gauge if a stock is ...
**Preferred tendency**: Investors typically prefer a lower P/E ratio, but …

*Data of different companies*

**AAPL:**

| Date | Price/Earnings | Price/Book Value | Return on Equity | … | Return |
|---|---|---|---|---|---|
| 2016-03-31 | 11.7542 | 4.4375 | 17.8925 | … | 0.1272 |
| 2016-06-30 | 10.9112 | 3.9807 | 16.5172 | … | -0.0427 |
| | | … | | | |

*Query*

Please create a new signal based on the provided context (existing signals), and this new signal should be correlated to the returns, explain how you created this signal and describe the meaning of this new signal. Note that don't provide simple linear combination of other existing signals and focus on as many meaningful existing signals as possible. Please also provide the calculated values of this new signal and standardize them. **Let's think step by step.**

Figure 1: Prompt demonstration.

on the ranks of the data, instead of the actual data) between returns and each signal. While the coefficients can reveal the correlation between the signal and the return, they can vary with different periods and market situations. Hence, we also introduce another method to evaluate the signal, as shown in the next section.

**Fama-MacBeth** We adopt the Fama-MacBeth Two-Step Regression (Fama and MacBeth, 1973), a traditional method for evaluating how well signals describe returns. Data from $n$ companies, including their historical signals and returns, are utilized for this evaluation. The Ordinary Least Squares (OLS), a commonly used approach, serves as the linear regression tool in our analysis process. Z-Score normalization is used on the signal values, as some of the signals have very large numerical values, while the values of percentage change in returns are very small.

Step 1: Each company's returns are regressed over time against the selected signals. The extent to which the returns are exposed to each signal is known as 'factor exposures' or 'beta coefficients'.

$$\begin{aligned} C_{1,t} &= \alpha_1 + \beta_{1,S1}S_{1,t} + \beta_{1,S2}S_{2,t} + \ldots + \beta_{1,Sm}S_{m,t}, \\ C_{2,t} &= \alpha_2 + \beta_{2,S1}S_{1,t} + \beta_{2,S2}S_{2,t} + \ldots + \beta_{2,Sm}S_{m,t}, \\ &\ldots \\ C_{n,t} &= \alpha_n + \beta_{n,S1}S_{1,t} + \beta_{n,S2}S_{2,t} + \ldots + \beta_{n,Sm}S_{m,t} \end{aligned}$$

where $C_{i,t}$ is the expected return of company $i$ at time $t$, $\alpha_i$ is the constant for company $i$, $\beta_{i,Sj}$ is signal $j$'s beta coefficient at company $i$, and $S_{j,t}$ denotes signal $j$ at time $t$ for each company. $t$ goes from 1 through $T$, indicating that each company's signals are regressed over time.

Step 2: We perform $T$ Cross-sectional Regression at each time for all the companies: the cross-sectional stock returns are regressed against the factor exposures (beta coefficients) calculated in the first step, obtaining the risk premia coefficients for each signal.

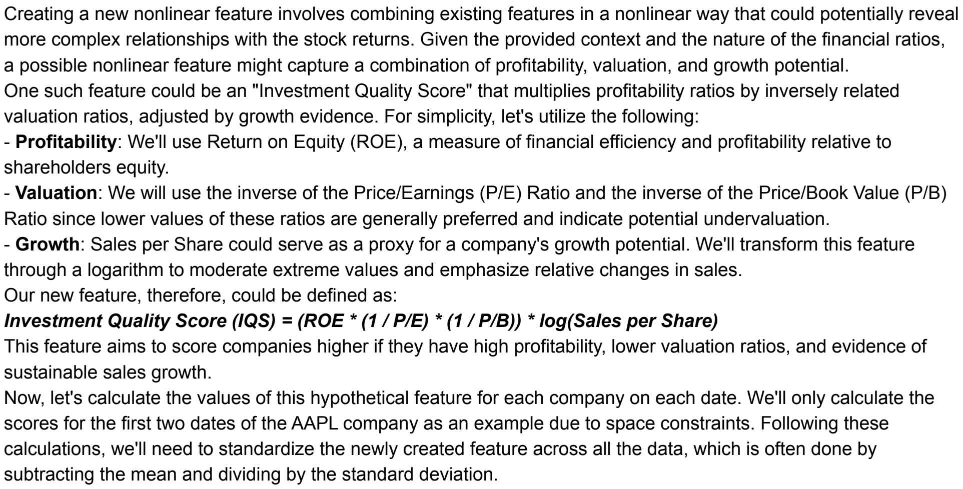
$$\begin{aligned} C_{i,1} &= \gamma_{1,0} + \gamma_{1,1}\hat{\beta}_{i,S1} + \gamma_{1,2}\hat{\beta}_{i,S2} + \ldots + \gamma_{1,m}\hat{\beta}_{i,Sm}, \\ C_{i,2} &= \gamma_{2,0} + \gamma_{2,1}\hat{\beta}_{i,S1} + \gamma_{2,2}\hat{\beta}_{i,S2} + \ldots + \gamma_{2,m}\hat{\beta}_{i,Sm}, \\ &\ldots \\ C_{i,T} &= \gamma_{T,0} + \gamma_{T,1}\hat{\beta}_{i,S1} + \gamma_{T,2}\hat{\beta}_{i,S2} + \ldots + \gamma_{T,m}\hat{\beta}_{i,Sm} \end{aligned}$$

where $C_{i,t}$ stands for the stock return of company $i$ at time $t$, $\gamma_{t,0}$ is the constant term for each company $C_i$, $\gamma_{t,j}$ is the regression coefficient for factor $j$ at time $t$, and $\hat{\beta}_{i,Sj}$ is the beta coefficient of company $i$ for signal $j$. Note that $i$ goes from 1 to $n$, as we include $n$ companies in total.

## 4 Experimental Setup

Companies in different sectors, along with the historical signal data and percentage changes in returns, are included in the experiments. We opt for percentage changes in returns at various time points over actual return values because this offers a consistent standard across companies, accommodating the variance in return levels among different companies.

10 existing signals are as follows: Price/Earnings (P/E), Price/Book Value(P/B), Return on Assets (ROA), Return on Equity (ROE), Free Cash Flow per Share (FCF), Price/Cash Flow (P/CF), Enterprise Value/EBITDA (EBITDA), Gross Margin (GM), Net Margin (NM), Sales per



Creating a new nonlinear feature involves combining existing features in a nonlinear way that could potentially reveal more complex relationships with the stock returns. Given the provided context and the nature of the financial ratios, a possible nonlinear feature might capture a combination of profitability, valuation, and growth potential.
One such feature could be an "Investment Quality Score" that multiplies profitability ratios by inversely related valuation ratios, adjusted by growth evidence. For simplicity, let's utilize the following:
- **Profitability**: We'll use Return on Equity (ROE), a measure of financial efficiency and profitability relative to shareholders equity.
- **Valuation**: We will use the inverse of the Price/Earnings (P/E) Ratio and the inverse of the Price/Book Value (P/B) Ratio since lower values of these ratios are generally preferred and indicate potential undervaluation.
- **Growth**: Sales per Share could serve as a proxy for a company's growth potential. We'll transform this feature through a logarithm to moderate extreme values and emphasize relative changes in sales.
Our new feature, therefore, could be defined as:
***Investment Quality Score (IQS) = (ROE * (1 / P/E) * (1 / P/B)) * log(Sales per Share)***
This feature aims to score companies higher if they have high profitability, lower valuation ratios, and evidence of sustainable sales growth.
Now, let's calculate the values of this hypothetical feature for each company on each date. We'll only calculate the scores for the first two dates of the AAPL company as an example due to space constraints. Following these calculations, we'll need to standardize the newly created feature across all the data, which is often done by subtracting the mean and dividing by the standard deviation.

Figure 2: Sample output of GPT-4 after being asked to generate a new signal.

Share (SPS). The selection of existing signals, which are "popular" financial indicators commonly used for evaluating a company's financial health (Arkan et al., 2016; Charles Schwab, 2023), is primarily influenced by their coverage across the datasets, ensuring the chosen signals are broadly applicable and reflective of standard financial analysis practices.

At each cross-section, we obtain an Adjusted R-squared ($R^2_{\text{adj}}$)of the model. After GPT-4 generates new signals as the last section mentioned, we add each of the new signals to our existing signals and perform the two-step Fama-MacBeth regression. The performance of models with each new signal is compared with that of the baseline model (with only existing signals).

**Dataset** Based on the Global Industry Classification Standard (GICS)[3] and looking at the S&P 500 index, we select companies in the Information Technology (IT) sector (43 companies), Health Care sector (31 companies), and Energy sector (19 companies), respectively. The full company lists are shown in the Appendix A. We download the companies' historical signal data from FactSet[4] and historical returns from Yahoo Finance[5], both of which are open-source financial websites. Data is processed to extract signal values, which are then merged with future one-month and three-month returns for analysis. This approach ensures a comprehensive dataset for evaluating financial performance.

## 5 Results

### 5.1 GPT-4 Output

With the prompts in the format shown in Section 3, GPT-4 is asked to generate several new signals by running the script multiple times, one new signal per run. Names and formulas are included in the outputs of GPT-4. Since we use a step-by-step prompting strategy, reasoning steps are also shown in the outcome, including the meaning, profitability, valuation, and growth of the new signal. Figure 2 shows part of a sample outcome of the new signal "Investment Quality Score (IQS)": GPT-4 provides its understanding of creating a new nonlinear signal, the reason why it creates such a new signal and the way of calculating the new signal. In addition, it calculates and standardizes values for the new signals based on the existing signal values we include in the prompt. The reasonings between other newly created signals are shown in B. The reasoning process demonstrates the potential of GPT-4 to produce outputs that are analytically sound and methodologically robust, rather than simply generating outputs arbitrarily.

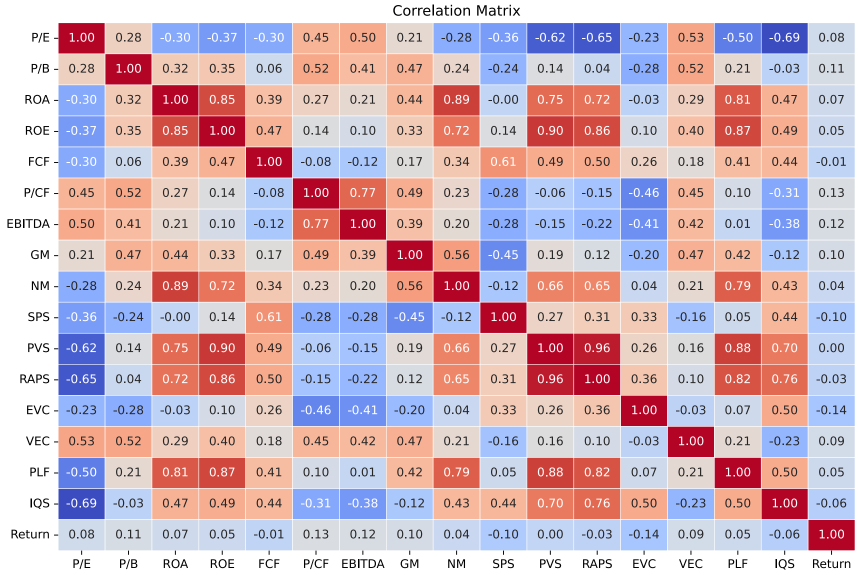
6 new signals created by GPT-4 are listed below:

1. Profitable Valuation Score (PVS): $PVS = \frac{ROE}{P/E}$,
2. Risk-Adjusted Performance Score (RAPS): $RAPS = \frac{ROE}{P/E \cdot \beta}$, here $\beta$ is 2 for calculation convenience.

[3] https://www.msci.com/our-solutions/indexes/gics

[4] https://https://www.factset.com/

[5] https://finance.yahoo.com/



| | P/E | P/B | ROA | ROE | FCF | P/CF | EBITDA | GM | NM | SPS | PVS | RAPS | EVC | VEC | PLF | IQS | Return |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P/E | 1.00 | 0.28 | -0.30 | -0.37 | -0.30 | 0.45 | 0.50 | 0.21 | -0.28 | -0.36 | -0.62 | -0.65 | -0.23 | 0.53 | -0.50 | -0.69 | 0.08 |
| P/B | 0.28 | 1.00 | 0.32 | 0.35 | 0.06 | 0.52 | 0.41 | 0.47 | 0.24 | -0.24 | 0.14 | 0.04 | -0.28 | 0.52 | 0.21 | -0.03 | 0.11 |
| ROA | -0.30 | 0.32 | 1.00 | 0.85 | 0.39 | 0.27 | 0.21 | 0.44 | 0.89 | -0.00 | 0.75 | 0.72 | -0.03 | 0.29 | 0.81 | 0.47 | 0.07 |
| ROE | -0.37 | 0.35 | 0.85 | 1.00 | 0.47 | 0.14 | 0.10 | 0.33 | 0.72 | 0.14 | 0.90 | 0.86 | 0.10 | 0.40 | 0.87 | 0.49 | 0.05 |
| FCF | -0.30 | 0.06 | 0.39 | 0.47 | 1.00 | -0.08 | -0.12 | 0.17 | 0.34 | 0.61 | 0.49 | 0.50 | 0.26 | 0.18 | 0.41 | 0.44 | -0.01 |
| P/CF | 0.45 | 0.52 | 0.27 | 0.14 | -0.08 | 1.00 | 0.77 | 0.49 | 0.23 | -0.28 | -0.06 | -0.15 | -0.46 | 0.45 | 0.10 | -0.31 | 0.13 |
| EBITDA | 0.50 | 0.41 | 0.21 | 0.10 | -0.12 | 0.77 | 1.00 | 0.39 | 0.20 | -0.28 | -0.15 | -0.22 | -0.41 | 0.42 | 0.01 | -0.38 | 0.12 |
| GM | 0.21 | 0.47 | 0.44 | 0.33 | 0.17 | 0.49 | 0.39 | 1.00 | 0.56 | -0.45 | 0.19 | 0.12 | -0.20 | 0.47 | 0.42 | -0.12 | 0.10 |
| NM | -0.28 | 0.24 | 0.89 | 0.72 | 0.34 | 0.23 | 0.20 | 0.56 | 1.00 | -0.12 | 0.66 | 0.65 | 0.04 | 0.21 | 0.79 | 0.43 | 0.04 |
| SPS | -0.36 | -0.24 | -0.00 | 0.14 | 0.61 | -0.28 | -0.28 | -0.45 | -0.12 | 1.00 | 0.27 | 0.31 | 0.33 | -0.16 | 0.05 | 0.44 | -0.10 |
| PVS | -0.62 | 0.14 | 0.75 | 0.90 | 0.49 | -0.06 | -0.15 | 0.19 | 0.66 | 0.27 | 1.00 | 0.96 | 0.26 | 0.16 | 0.88 | 0.70 | 0.00 |
| RAPS | -0.65 | 0.04 | 0.72 | 0.86 | 0.50 | -0.15 | -0.22 | 0.12 | 0.65 | 0.31 | 0.96 | 1.00 | 0.36 | 0.10 | 0.82 | 0.76 | -0.03 |
| EVC | -0.23 | -0.28 | -0.03 | 0.10 | 0.26 | -0.46 | -0.41 | -0.20 | 0.04 | 0.33 | 0.26 | 0.36 | 1.00 | -0.03 | 0.07 | 0.50 | -0.14 |
| VEC | 0.53 | 0.52 | 0.29 | 0.40 | 0.18 | 0.45 | 0.42 | 0.47 | 0.21 | -0.16 | 0.16 | 0.10 | -0.03 | 1.00 | 0.21 | -0.23 | 0.09 |
| PLF | -0.50 | 0.21 | 0.81 | 0.87 | 0.41 | 0.10 | 0.01 | 0.42 | 0.79 | 0.05 | 0.88 | 0.82 | 0.07 | 0.21 | 1.00 | 0.50 | 0.05 |
| IQS | -0.69 | -0.03 | 0.47 | 0.49 | 0.44 | -0.31 | -0.38 | -0.12 | 0.43 | 0.44 | 0.70 | 0.76 | 0.50 | -0.23 | 0.50 | 1.00 | -0.06 |
| Return | 0.08 | 0.11 | 0.07 | 0.05 | -0.01 | 0.13 | 0.12 | 0.10 | 0.04 | -0.10 | 0.00 | -0.03 | -0.14 | 0.09 | 0.05 | -0.06 | 1.00 |

Figure 3: Correlation of **all companies** with both existing and new signals. Note that the last six signals are newly created by LLM.

3. Efficiency Value Composite (EVC): $EVC = \frac{1.0}{ROA} \cdot \frac{1.0}{EBITDA} \cdot \frac{1.0}{PCF}$

4. Valuation Efficiency Composite Score (VEC): $VEC = \frac{(P/E+ROE+FCF)}{3.0}$

5. Profitability Leverage Factor (PLF): $PLF = \frac{ROE \cdot GM}{P/E}$

6. Investment Quality Score (IQS): $IQS = (ROE \cdot \frac{1}{P/E} \cdot \frac{1}{P/B} \cdot log(SPS))$

For the evaluation period, we use ranges from years 2016 to 2020, with a frequency of 3 months, as the historical signals of the companies are reported quarterly. In addition, we use the signal values to predict the future quarterly returns (i.e. we use signals in March to predict returns in June). We demonstrate results for IT companies (with future quarterly returns), and other sectors' results are listed in the Appendix B.

### 5.2 Overall Results

Figure 3 shows the correlation matrix for all the companies in the 3 different sectors. New signal EVC still possesses the highest absolute value with returns and most of the other new signals. Note that although the values of the coefficients are small, they are already considered sufficiently large values in the case of predicting change in stock returns (Kawee Numpacharoen, 2012). These observations show that the new signals do have considerable correlations to the returns.

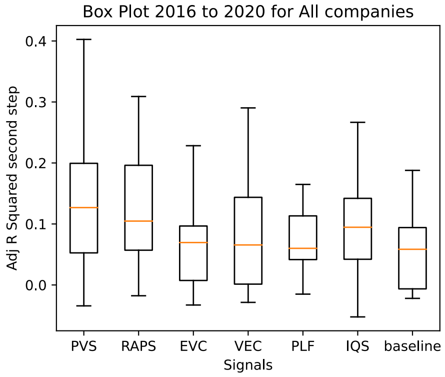
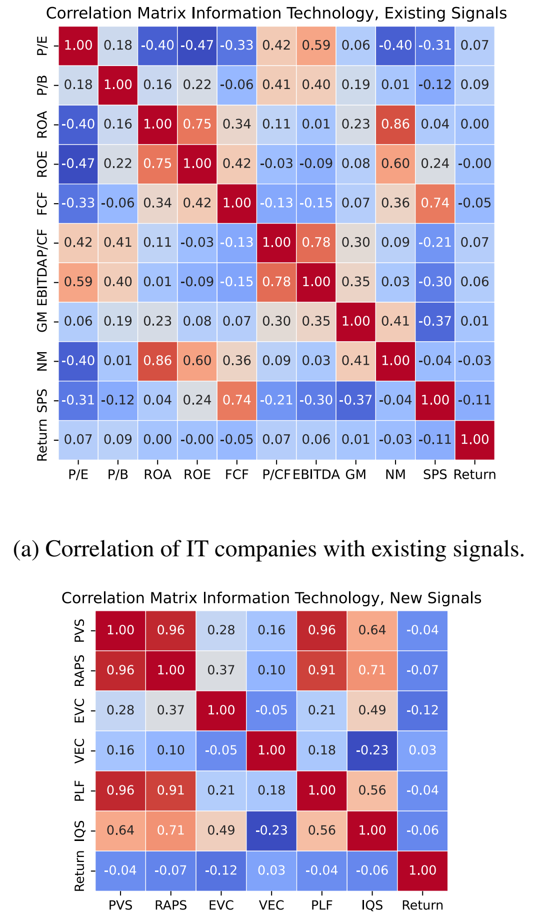


Figure 4: $R^2_{\text{adj}}$ values for Fama-MacBeth step 2 with companies in all 3 sectors. The last boxplot is the baseline without any new signals.

Figure 4 shows the box plot of $R^2_{\text{adj}}$ values for Fama-MacBeth step 2, evaluated on companies in all the 3 sectors. The box plot offers a comparative visual representation, showing the variability of the $R^2_{\text{adj}}$ values, which serve to gauge the explanatory capacity of our regression models enhanced by the introduction of novel signals. The new signals demonstrate a range of improvements in comparison to the baseline model, as denoted by the median and the interquartile ranges. The final box plot on



Correlation Matrix Information Technology, Existing Signals

| | P/E | P/B | ROA | ROE | FCF | P/CF | EBITDA | GM | NM | SPS | Return |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P/E | 1.00 | 0.18 | -0.40 | -0.47 | -0.33 | 0.42 | 0.59 | 0.06 | -0.40 | -0.31 | 0.07 |
| P/B | 0.18 | 1.00 | 0.16 | 0.22 | -0.06 | 0.41 | 0.40 | 0.19 | 0.01 | -0.12 | 0.09 |
| ROA | -0.40 | 0.16 | 1.00 | 0.75 | 0.34 | 0.11 | 0.01 | 0.23 | 0.86 | 0.04 | 0.00 |
| ROE | -0.47 | 0.22 | 0.75 | 1.00 | 0.42 | -0.03 | -0.09 | 0.08 | 0.60 | 0.24 | -0.00 |
| FCF | -0.33 | -0.06 | 0.34 | 0.42 | 1.00 | -0.13 | -0.15 | 0.07 | 0.36 | 0.74 | -0.05 |
| P/CF | 0.42 | 0.41 | 0.11 | -0.03 | -0.13 | 1.00 | 0.78 | 0.30 | 0.09 | -0.21 | 0.07 |
| EBITDA | 0.59 | 0.40 | 0.01 | -0.09 | -0.15 | 0.78 | 1.00 | 0.35 | 0.03 | -0.30 | 0.06 |
| GM | 0.06 | 0.19 | 0.23 | 0.08 | 0.07 | 0.30 | 0.35 | 1.00 | 0.41 | -0.37 | 0.01 |
| NM | -0.40 | 0.01 | 0.86 | 0.60 | 0.36 | 0.09 | 0.03 | 0.41 | 1.00 | -0.04 | -0.03 |
| SPS | -0.31 | -0.12 | 0.04 | 0.24 | 0.74 | -0.21 | -0.30 | -0.37 | -0.04 | 1.00 | -0.11 |
| Return | 0.07 | 0.09 | 0.00 | -0.00 | -0.05 | 0.07 | 0.06 | 0.01 | -0.03 | -0.11 | 1.00 |

(a) Correlation of IT companies with existing signals.

Correlation Matrix Information Technology, New Signals

| | PVS | RAPS | EVC | VEC | PLF | IQS | Return |
|---|---|---|---|---|---|---|---|
| PVS | 1.00 | 0.96 | 0.28 | 0.16 | 0.96 | 0.64 | -0.04 |
| RAPS | 0.96 | 1.00 | 0.37 | 0.10 | 0.91 | 0.71 | -0.07 |
| EVC | 0.28 | 0.37 | 1.00 | -0.05 | 0.21 | 0.49 | -0.12 |
| VEC | 0.16 | 0.10 | -0.05 | 1.00 | 0.18 | -0.23 | 0.03 |
| PLF | 0.96 | 0.91 | 0.21 | 0.18 | 1.00 | 0.56 | -0.04 |
| IQS | 0.64 | 0.71 | 0.49 | -0.23 | 0.56 | 1.00 | -0.06 |
| Return | -0.04 | -0.07 | -0.12 | 0.03 | -0.04 | -0.06 | 1.00 |

(b) Correlation of IT companies with new signals.

Figure 5: Correlation for existing/new signals with returns.

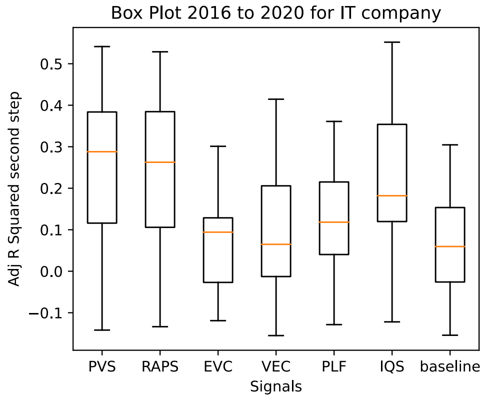


Figure 6: $R^2_{\text{adj}}$ values for Fama-MacBeth step 2 with companies in IT sector. The last boxplot is the baseline without any new signals.

the right illustrates the baseline model without the integration of new signals, establishing a benchmark that accentuates the predictive accuracy gains afforded by the existing features.

### 5.3 Correlation Heat Maps Break-down

Heat Maps of the correlation coefficients between existing and new signals and their future 3-month returns in IT companies are shown in figure 5a and figure 5b. The correlation coefficients between signals and returns are shown in the last column and the last row. Figure 5a is the correlation matrix in the IT sector between the existing signals and historical returns, and the last column is the correlation coefficients between the signals and returns. We can see that the absolute value of the coefficients ranges from 0 to 0.11. Figure 5b is the correlation matrix between the new signals and the returns, and we can see that the absolute value of the coefficients ranges from 0.03 to 0.12, which has an overall better performance than the existing signals.

The new signal EVC has the highest absolute correlation, surpassing the performance of all the existing signals. Besides, other new signals generated by GPT also have proper performance, all of which have absolute correlation coefficients larger than at least two of the existing signals.

Apart from the IT sector, we also evaluate the new signals on companies' data in the Health Care and Energy sectors. Corresponding heat maps are plotted in the same format, as shown in Appendix B. Similar patterns can be observed in different sectors, as many of the correlations of the new signals have a higher absolute value than the existing ones.

### 5.4 Fama-MacBeth Regression Break-down

The $R^2_{\text{adj}}$ values for the Fama-MacBeth step 2 regression models, each with a new signal added to the original set (the 10 existing signals, have been calculated and presented at Figure 6, and the median values of the $R^2_{\text{adj}}$ values are marked by the orange lines. The signal names in the graph represent the models with 10 existing signals plus each of the 6 new signals generated by GPT-4, respectively, noting the "baseline" represents the model with only the 10 existing signals. (The last boxplot is the baseline without any new signals). It is observed that the inclusion of these new signals results in improved performance for 5 out of 6 models with new signals, compared to the baseline model's performance. Box plots for companies in Health

Care and Energy sector are shown in Appendix B, with similar patterns observed.

## 6 Conclusions

In this work, we leverage an LLM (GPT-4) to generate 6 novel financial signals that enhance the performance of existing stock return-prediction models, addressing the limitations of traditional feature engineering techniques in financial analytics and the alpha research process. We demonstrate that GPT-4 is capable of analyzing existing signals' performance in historical data and extracting useful context information in the feature engineering process. The work results in the creation of innovative signals that capture patterns and interactions.

The new signals generated by GPT-4 demonstrate various advantages. First of all, GPT-4 adapts to changes in market conditions more thoroughly and dynamically than traditional models, permitting it to continually refine and optimize the process of signal generation based on data and human-AI interaction. Secondly, the LLM is able to process and analyze a large amount of data, and identify sophisticated patterns and relationships that are not obvious through traditional and standard statistical methods. Last but not least, the use of GPT-4 largely speeds up the feature engineering process, reducing the time required to develop complicated algorithms and explore new financial signals in the market.

## References

Thomas Arkan et al. 2016. The importance of financial ratios in predicting stock price trends: A case study in emerging markets. *Finanse, Rynki Finansowe, Ubezpieczenia*, (79):13–26.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Charles Schwab. 2023. Five key financial ratios for stock analysis. https://www.schwab.com/learn/story/five-key-financial-ratios-stock-analysis. Accessed: 2023-04-25.

Harrison Chase. 2022. Langchain. Available at: https://github.com/langchain-ai/langchain.

Wenhu Chen. 2022. Large language models are few (1)-shot table reasoners. *arXiv preprint arXiv:2210.06710*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky. 2023. What do llms know about financial markets? a case study on reddit market sentiment analysis. In *Companion Proceedings of the ACM Web Conference 2023*, pages 107–110.

Eugene F Fama and James D MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *Conference Proceedings Name*, pages Start Page–End Page. Publisher Name, if available. Replace "Conference Proceedings Name" and page numbers with actual details.

Noah Hollmann, Samuel Müller, and Frank Hutter. 2024. Llms for semi-automated data science: Introducing for context-aware automated feature engineering. Available at: https://arxiv.org/abs/2305.03403 (Accessed: 28 February 2024).

F. Hutter, L. Kotthoff, and J. Vanschoren. 2019. *Automated Machine Learning: Methods, Systems, Challenges*. Springer. Available for free at http://automl.org/book.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Zura Kakushadze. 2016. 101 formulaic alphas.

Amporn Atsawarungruangkit Kawee Numpacharoen. 2012. Generating correlation matrices based on the boundaries of their coefficients. *PloS one*, 7:e48902.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023a. Large language models understand and can be enhanced by emotional stimuli.

Zhenglin Li, Hanyi Yu, Jinxin Xu, Jihang Liu, and Yuhong Mo. 2023b. Stock market analysis and prediction using lstm: A case study on technology stocks. *Innovations in Applied Engineering and Technology*, 2(1):1–6.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

X.Y. Liu, G. Wang, H. Yang, and D. Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *Name of Journal or Conference if known*. Available online.

Weimin Lyu, Songzhu Zheng, Tengfei Ma, and Chao Chen. 2022. A study of the attention abnormality in trojaned berts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4727–4741.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through llm negotiations.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

S. Wang et al. 2023a. Alpha-gpt: Human-ai interactive alpha mining for quantitative investment. Available at: `https://arxiv.org/abs/2308.00016v1` (Accessed: 28 February 2024).

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Tianping Zhang, Yuanqi Li, Yifei Jin, and Jian Li. 2020. Autoalpha: An efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. *arXiv preprint arXiv:2002.08245*.

H. Zhao et al. 2024. Revolutionizing finance with llms: An overview of applications and insights. Available at: `https://arxiv.org/abs/2401.11641` (Accessed: 28 February 2024).

# A Company List

See Table 1.

# B Other Results

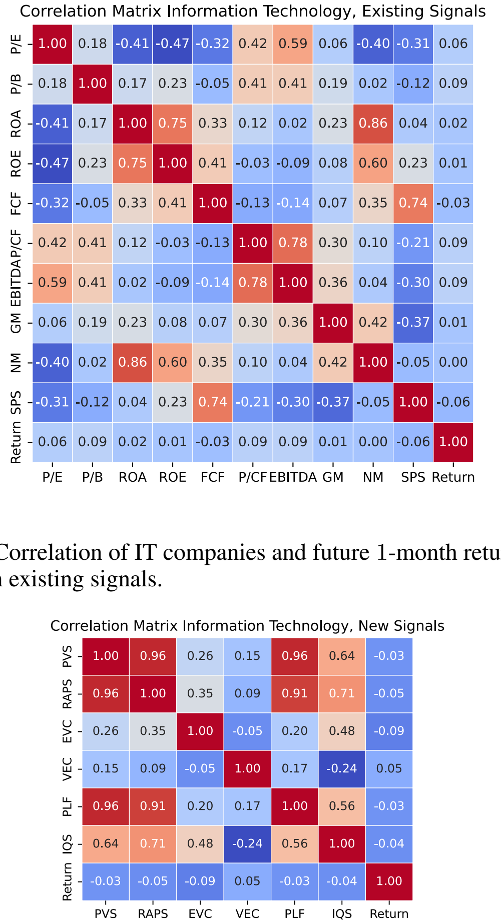
## B.1 IT Companies with Future One-Month Returns

### B.1.1 Correlation

See Figure 7.

### B.1.2 Fama-MacBeth

See Figure 8.



(a) Correlation of IT companies and future 1-month returns with existing signals.

(b) Correlation of IT companies and future 1-month returns with new signals.

Figure 7: Correlation of existing/new signals with returns.

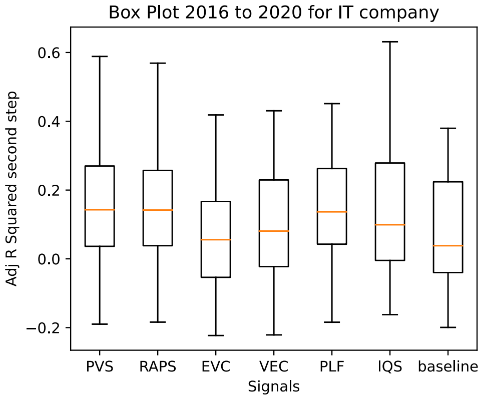


Figure 8: $R^2_{\text{adj}}$ values of IT companies and future 1-month returns for Fama-MacBeth step 2. The last boxplot is the baseline without any new signals.

| Sector | Companies |
|---|---|
| Information Technology | "AAPL", "AKAM", "AMD"<br>"ANET", "ANSS", "APH"<br>"CDNS", "CDW", "CTSH"<br>"ENPH", "EPAM", "FFIV"<br>"FSLR", "FTNT", "GEN"<br>"GLW", "IBM", "INTC"<br>"IT", "JNPR", "KLAC"<br>"LRCX", "MCHP", "MPWR"<br>"MSFT", "MSI", "NOW"<br>"NXPI", "ON", "PTC"<br>"QCOM", "ROP", "STX"<br>"SWKS", "TDY", "TEL"<br>"TER", "TRMB", "TXN"<br>"TYL", "VRSN", "WDC", "ZBRA" |
| Health Care | "ABBV", "ABT","ALGN",<br>"AMGN", "BAX", "BDX"<br>"BIO", "BMY", "BSX"<br>"CAH", "COR", "CRL"<br>"CTLT", "CVS", "DGX"<br>"DHR", "DXCM", "EW"<br>"GILD", "HSIC", "TMO"<br>"UHS", "VRTX", "VTRS"<br>"IDXX", "ILMN", "INCY"<br>"WST", "ZTS", "ISRG", "JNJ" |
| Energy | "APA", "COP", "CTRA"<br>"EOG", "FANG", "HAL"<br>"HES", "KMI", "MPC"<br>"MRO", "OKE", "OXY"<br>"PSX", "PXD", "SLB"<br>"TRGP", "VLO", "WMB", "XOM" |

Table 1: Company list of different sectors

## B.2 Health Care Companies with Future One-Month Returns

### B.2.1 Correlation

See Figure 9.

### B.2.2 Fama-MacBeth

See Figure 10.

## B.3 Health Care Companies with Future Three-Month Returns

### B.3.1 Correlation

See Figure 11.

### B.3.2 Fama-MacBeth

See Figure 12.

## B.4 Energy Companies with Future One-Month Returns

### B.4.1 Correlation

See Figure 13.

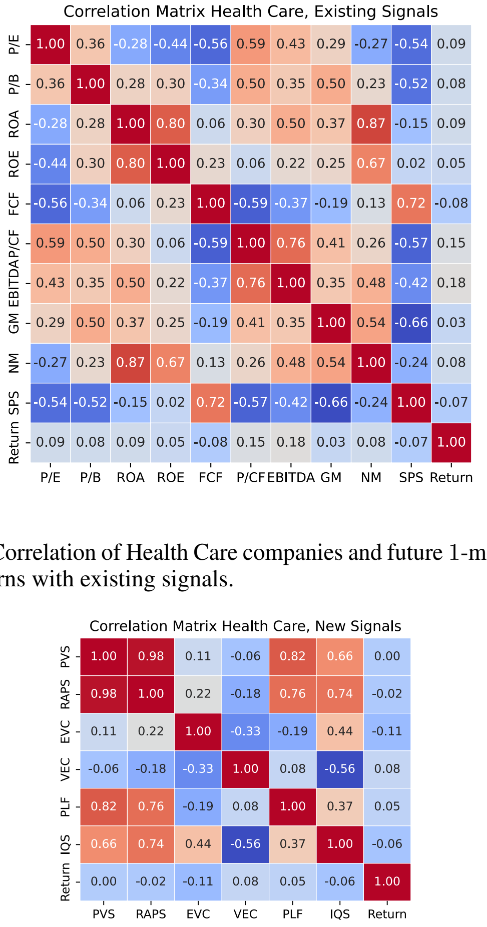
### B.4.2 Fama-MacBeth

See Figure 14.

## B.5 Energy Companies with Future Three-Month Returns

### B.5.1 Correlation

See Figure 15.



(a) Correlation of Health Care companies and future 1-month returns with existing signals.

(b) Correlation of Health Care companies and future 1-month returns with new signals.

Figure 9: Correlation of existing/new signals with returns.

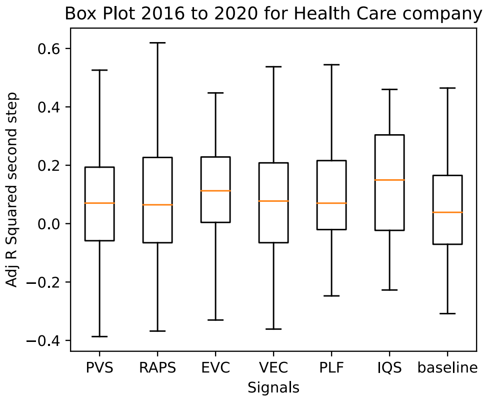
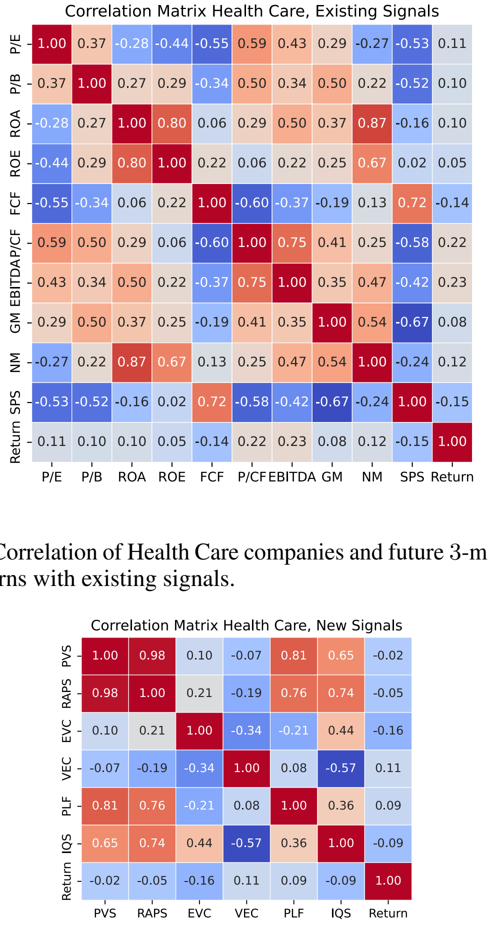


Figure 10: $R^2_{\text{adj}}$ values of Health Care companies and future 1-month returns for Fama-MacBeth step 2. The last boxplot is the baseline without any new signals.

### B.5.2 Fama-MacBeth

See Figure 16.



(a) Correlation of Health Care companies and future 3-month returns with existing signals.

(b) Correlation of Health Care companies and future 3-month returns with new signals.

Figure 11: Correlation of existing/new signals with returns.

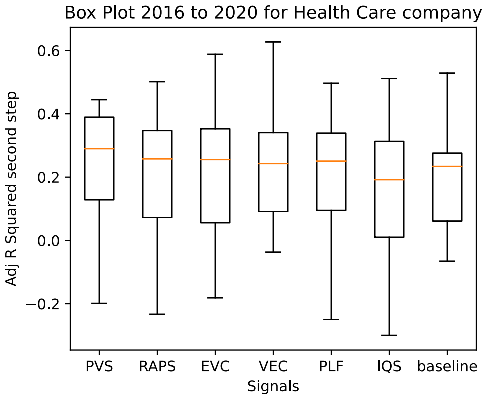
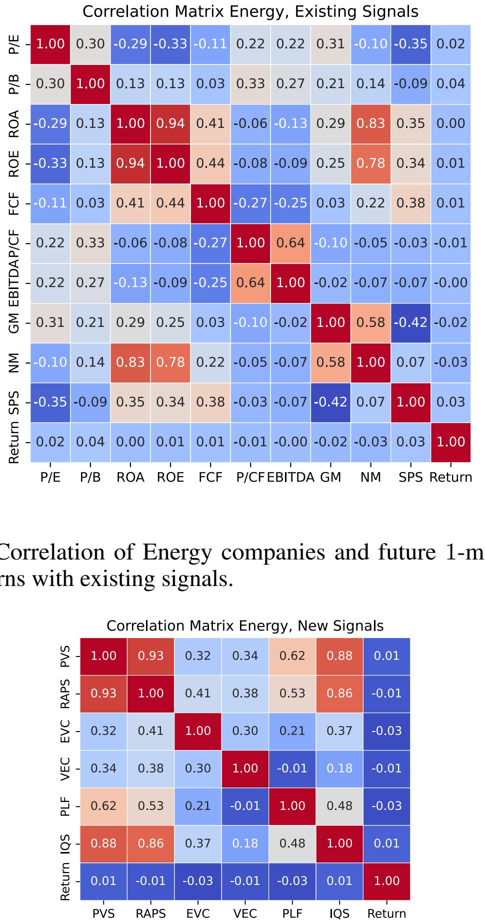


Figure 12: $R^2_{\text{adj}}$ values of Health Care companies and future 3-month returns for Fama-MacBeth step 2. The last boxplot is the baseline without any new signals.



(a) Correlation of Energy companies and future 1-month returns with existing signals.

(b) Correlation of Energy companies and future 1-month returns with new signals.

Figure 13: Correlation of existing/new signals with returns.

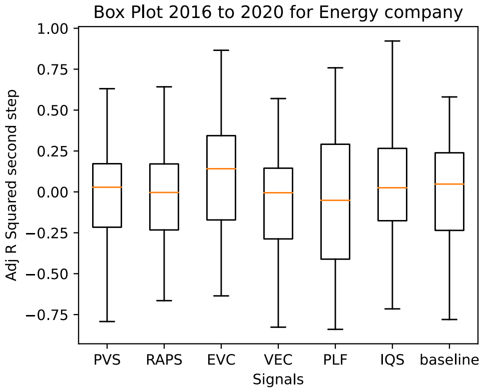
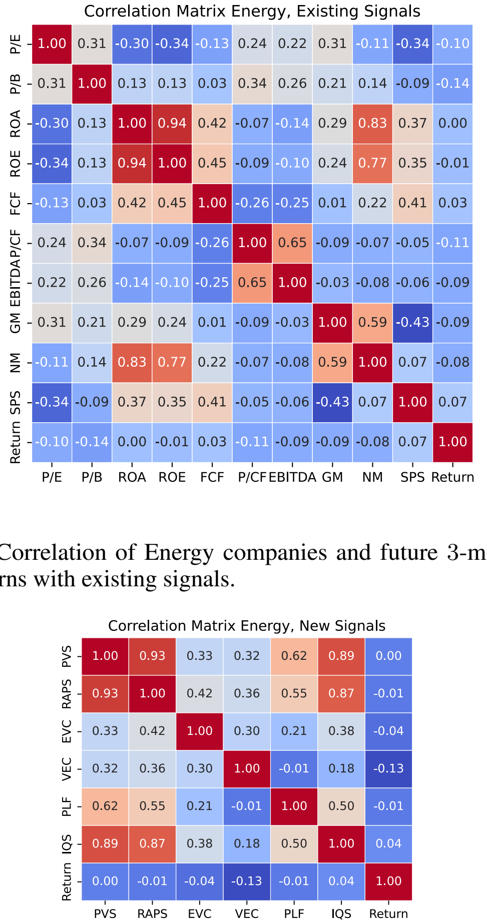


Figure 14: $R^2_{\text{adj}}$ values of Energy companies and future 1-month returns for Fama-MacBeth step 2. The last boxplot is the baseline without any new signals.



(a) Correlation of Energy companies and future 3-month returns with existing signals.

(b) Correlation of Energy companies and future 3-month returns with new signals.

Figure 15: Correlation of existing/new signals with return.

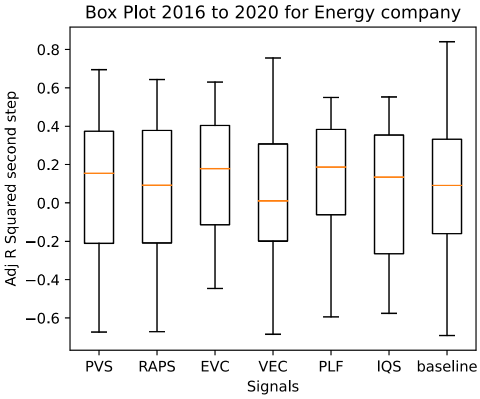


Figure 16: $R^2_{\text{adj}}$ values of Energy companies and future 3-month returns for Fama-MacBeth step 2. The last boxplot is the baseline without any new signals.