

Article

## Could ChatGPT have earned abnormal returns? A retrospective test from the U.S. stock market

Marc F. LoGrasso<sup>1</sup> 

<sup>1</sup> Department of Economics and Finance, Canisius University, 2001 Main Street, Buffalo, 14208, NY; email: lograssm@canisius.edu

**Abstract:** This paper attempts to assess the ability of OpenAI's ChatGPT to provide high-quality recommendations for a casual investor looking to beat the market. Going back to 1985 and instructing the GPT-4 model to restrict its knowledge to only what could have been known at the time of stock selection, the GPT-4 model was able to average alphas of approximately 1% per month for two-year holding periods beginning July 1 every year from 1985 to 2021. These abnormal returns persisted after controlling for size, book-to-market, profitability robustness, investment approach, and intermediate- and long-term prior returns. Individual portfolio alphas are only positive and significant about one out of four years but are never negative and significant. This paper also illustrates some of the precision needed to induce the GPT-4 model to provide any recommendations and briefly assesses the asset allocation strategy it appears to pursue.

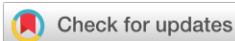
**Keywords:** portfolio choice, artificial intelligence, market efficiency, ChatGPT, fintech

**JEL Classification:** G11, G14, G41, O33, O36

### 1. Introduction

This paper evaluates whether OpenAI's ChatGPT could have generated abnormal stock returns as an investment advisor for individual investors over a 37-year retrospective period (1985–2021). Early attempts to solicit investment advice from ChatGPT highlighted its potential and limitations. Potter and Greifeld (2023) received only disclaimers about market unpredictability when requesting a market-beating ETF. Yeung (2023b) tricked the tool into suggesting stocks for specific goals (e.g., growth or value) with mixed ETF-comparable results, while Yeung (2023a) later obtained nuanced picks from GPT-4 but without quantitative benchmarks or explanations. Building on these, I prompt the GPT-4 model to select 10 U.S. stocks each July 1, restricting its knowledge to pre-selection data to simulate real-time advice without look-ahead bias. These stocks form price-weighted portfolios held for 24 months, reflecting how small investors might buy round lots of 100 shares per stock rather than using equal- or value-weighting. I assess performance using cumulative abnormal returns (CARs) against the market and risk-adjusted alphas via the CAPM, Fama-French three-factor, five-factor, and Fama-French three-factor with momentum/long-term reversal models.

The key findings show that ChatGPT's portfolios earn positive but inconsistent abnormal returns. Average 24-month CARs are 27.242% ( $p < 0.001$ ), accumulating positively from month 2 onward. Risk-adjusted alphas average 0.483% to 0.939% monthly across models, but are insignificant overall, with positive and significant alphas in only one out of every four years and no significantly negative ones. ChatGPT favors large growth stocks (negative size and book-to-market betas,  $p < 0.001$ ) in sectors like technology and healthcare, with 26% annual turnover. Robustness checks confirm similar patterns for equal- and value-weighted portfolios.



**Citation:** LoGrasso, M.F. (2025). Could ChatGPT have earned abnormal returns? A retrospective test from the U.S. stock market? *Modern Finance*, 3(3), 112-132.

Accepting Editor: Adam Zaremba

Received: 4 June 2025

Accepted: 6 September 2025

Published: 14 September 2025



**Copyright:** © 2025 by the authors. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International (CC BY 4.0) license (<https://creativecommons.org/licenses/by/4.0/>).

This study contributes to the emerging literature on AI in finance by providing one of the first long-horizon retrospective tests of ChatGPT's stock-picking ability, spanning diverse market conditions rather than focusing solely on post-2021 data as in studies such as Lopez-Lira and Tang (2024). Unlike those studies, which emphasize news-driven predictions and short-term horizons, I prioritize price-weighting for relevance to individual investors. I find general but inconsistent overperformance, echoing mixed results in Pelster and Val (2024) but highlighting ChatGPT's avoidance of significant underperformance. It also reveals ChatGPT's sector biases and prompting challenges, extending Ko and Lee (2023) on AI portfolio optimization.

## 2. Literature Review

Given the recency of ChatGPT's release for public use, it is unsurprising that sparse research is available about this tool's usefulness as a stock picker. However, the articles discussed below are indicative of much of the research that is currently accessible.

Lopez-Lira and Tang (2024) provided one of the more in-depth studies on this research question, using news headlines to assess whether a particular news item would be good, bad, or uncertain for the returns of a particular stock the next day. They find that older large language models (LLMs) struggle to provide general strong returns. As an extension of this idea, they also find that more complex LLMs are required to earn returns when analyzing complicated headlines. Lu et al. (2023) find that ChatGPT can earn abnormal returns for 30 days when selecting a portfolio in response to news announcements, though it struggles to perform well in out-of-sample testing. Chen et al. (2025) corroborate Lopez-Lira and Tang's (2024) conclusion regarding stronger performance by more complex LLMs while also showing how ChatGPT does a better job at identifying good news over other models, including human investors. Pelster and Val (2024) find that ChatGPT performs well at predicting earnings and the future performance of stocks after earnings surprises and after both positive and negative news events.

Ko and Lee (2023) utilize ChatGPT to generate diversified portfolios in a Markowitz (1952) mean-variance framework by both maximizing the Sharpe ratio (generating the tangency portfolio) and minimizing the portfolio variance. They find generally greater diversification, higher returns, higher risk-adjusted returns as measured by the Sharpe ratio, and lower standard deviations in the ChatGPT selections compared to randomly determined portfolios. Romanko et al. (2023) alternatively assess ChatGPT's ability to optimize stock weights in a Markowitz (1952) mean-variance framework. While the AI tool performs better than several benchmarks, including an equal-weighted portfolio, the ChatGPT-weighted portfolio is still under the efficient frontier.

Ahangar and Fietko (2023) attempt to get ChatGPT to tell them the best stock to invest in on a specific date (March 20, 2023, in their analysis); they could not induce it to do so despite using multiple iterations of their request. They find that the user needs to provide additional information even to get a suggestion about a specific stock. They conclude that "ChatGPT can be a helpful resource for financial experts but not for those with limited financial expertise and knowledge." This conclusion is one of the motivating ideas for this study, as the author believes that unsophisticated investors will rely on the AI tool they hear most about, which has recently been ChatGPT. Yeo et al. (2025) focus on the importance of explainability in financial AI tools in general (not just ChatGPT), which can be summarized as the ability to explain the results of an AI response in terms that the average person could understand. Their concern arose from an apparent tendency of LLMs to make up answers that would be difficult for non-experts to recognize.

Ali and Aysan (2023) focus on ChatGPT's usefulness in other areas before asking how it considers itself applicable in the financial industry. Specific to investments, portfolio management, and wealth management, it claims it can "provide financial advice to customers based on market data, investment portfolios, and customer preferences." However, it needs an "enormous amount of data and ability to use deep learning and machine learning algorithms and networks to perform the task of stock price prediction,

and also for providing advice for investments and asset management.” It considers itself superior to other chatbots and virtual assistants in use by financial institutions, due to its natural language understanding, knowledge base, continuous learning, customizability, and integrability. Relative to financial robo-advisors, ChatGPT cannot automate investment and asset management decisions, but is a more convenient tool for gathering inputs in less structured manners and is subject to less regulation. However, given the ability of robo-advisors to track and work with clients’ adjusting risk preferences, even when those adjustments are sub-optimal (Capponi et al., 2020), there may be room for ChatGPT, or AI in general, to improve clients’ interactions with robo-advisors. Kalyan (2024) reports mixed results in assessing ChatGPT’s performance in performing various financial tasks against established domain-specific models.

### 3. Prompting Methodology

This section details the development of the series of prompts used to gather the stocks included in each year’s portfolio. As shown in Appendix A, ChatGPT-4 appears to use a top-down approach to portfolio selection, beginning with considerations of macroeconomic conditions, followed by a focus on specific sectors of the economy.

#### 3.1. Initial Prompting

The initial interaction with ChatGPT-4 used the following prompt:

“It is early in the morning of July 1, 1985. I have savings of \$100,000 that I will invest as much of as I can in the US stock market as soon as trading opens (which might be a day or two away if the markets are not open today). Over the next two years, I am going to hold onto the stocks I select initially, and my desire is to beat the market on both a total return and a risk-adjusted return basis. Your knowledge set is limited to only information that would have been available when the markets open and transactions can occur. I am going to purchase 100 shares of each of the stocks you suggest. What should I buy?”

This prompt led to the listing of key factors used in the selection process, followed by a list of six “potential stock picks from 1985” (IBM, Merck, Coca-Cola, GE, Microsoft, and Intel).

#### 3.2. Revisions

There was no attempt to come close to the \$100,000 budget limit, because the tool could not access historical price data. When giving it a link to historical prices around July 1, 1985, for one of the stocks in the list,<sup>1</sup> it claimed an inability to pull the price data “due to the limitations of my browsing tool.” This limit led me to drop the budget constraint from my request and instead focus on a series of top picks each year (change in prompt is emphasized below). This also allows this study to more closely align with the approach that would be utilized by small, individual investors interested in direct stock ownership over funds and ETFs.

“It is early in the morning of July 1, 1985. I am going to invest in ten stocks available in the US stock market as soon as trading opens (which might be a day or two away if the markets will not be open today). Over the next two years, I am going to hold onto the stocks I select initially, and my desire is to beat the market on both a total return and a risk-adjusted return basis. Your knowledge set is limited to only information that would have been available when the markets open and transactions can occur. I am going to purchase 100 shares of each of the stocks you suggest. What should I buy?”

---

<sup>1</sup>. <https://finance.yahoo.com/quote/IBM/history?period1=488505600&period2=489542400&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true>

### 3.3. Attempts at Efficiency

This prompt was more successful, reiterating the key factors used in the selection process and providing ten stocks for investment, this time excluding the not-yet-public Microsoft from the list. Feeling comfortable with the results, I then ventured to request that the process be repeated for every year in this study, 1986-2022. ChatGPT-4 explained another of its limitations here.

"Providing detailed stock recommendations with justifications for each year from 1986 through 2022 is an extensive task that requires a deep and specific analysis of market conditions, economic factors, and company performances for each of those years. This level of detail would require access to a vast array of historical financial data and economic reports."

While I appreciate the due diligence ChatGPT-4 is willing to undertake, this required another simple detour, requesting the recommendations one year at a time.

"Yes, I can certainly provide stock recommendations for one specific year, including the relevant economic factors and sector trends of that year. Please specify which year you're interested in, and I will conduct an analysis based on the information and context available for that particular year."

## 4. Data and Empirical Methodology

### 4.1. Data Sources

Prices, monthly total returns on individual stocks, and individual stock characteristics were extracted from S&P CapitalIQ. For month  $m$ , the open price used to determine weights was extracted using the IQ\_LASTSALEPRICE field as of the end of the previous month ( $m - 1$ ). Since this provided the actual sale price on that date, this value was multiplied by the IQ\_CUMULATIVE\_SPLIT\_FACTOR to adequately compensate for the effects of splits that occurred during the holding period. The return for stock  $i$  during month  $m$  is calculated as the holding period return based on the same price field and the dividends per share during month  $m$  as determined by the IQ\_DIV\_SHARE field as:

$$r_i = \frac{IQ\_LASTSALEPRICE_m + IQ\_DIV\_SHARE_m}{IQ\_LASTSALEPRICE_{m-1}} - 1. \quad (1)$$

- Market capitalizations used for characteristic analysis in Section 4.2 and for robustness checks in Section 5.4 are obtained using the IQ\_MARKETCAP field. Book value of equity per share is extracted using IQ\_BV\_SHARE, and the book-to-market ratio is calculated as:

$$B/M_i = \frac{IQ\_BV\_SHARE_m}{IQ\_LASTSALEPRICE_{m-1}}. \quad (2)$$

Independent variables used in regressions were pulled from Kenneth French's website.<sup>2</sup> Values for the monthly market premium ( $Mkt - Rf$ , hereafter  $r_M - r_f$ ), size (SMB), value (HML), operating profitability (RMW), and investment (CMA) risk factors are pulled from the file contained at the "Fama/French 5 Factors (2x3)" line, while momentum (MOM) and long-term reversal (LTRev) risk factors are obtained from the links at "Momentum Factor (Mom)" and "Long-Term Reversal Factor (LT Rev)" lines, respectively.

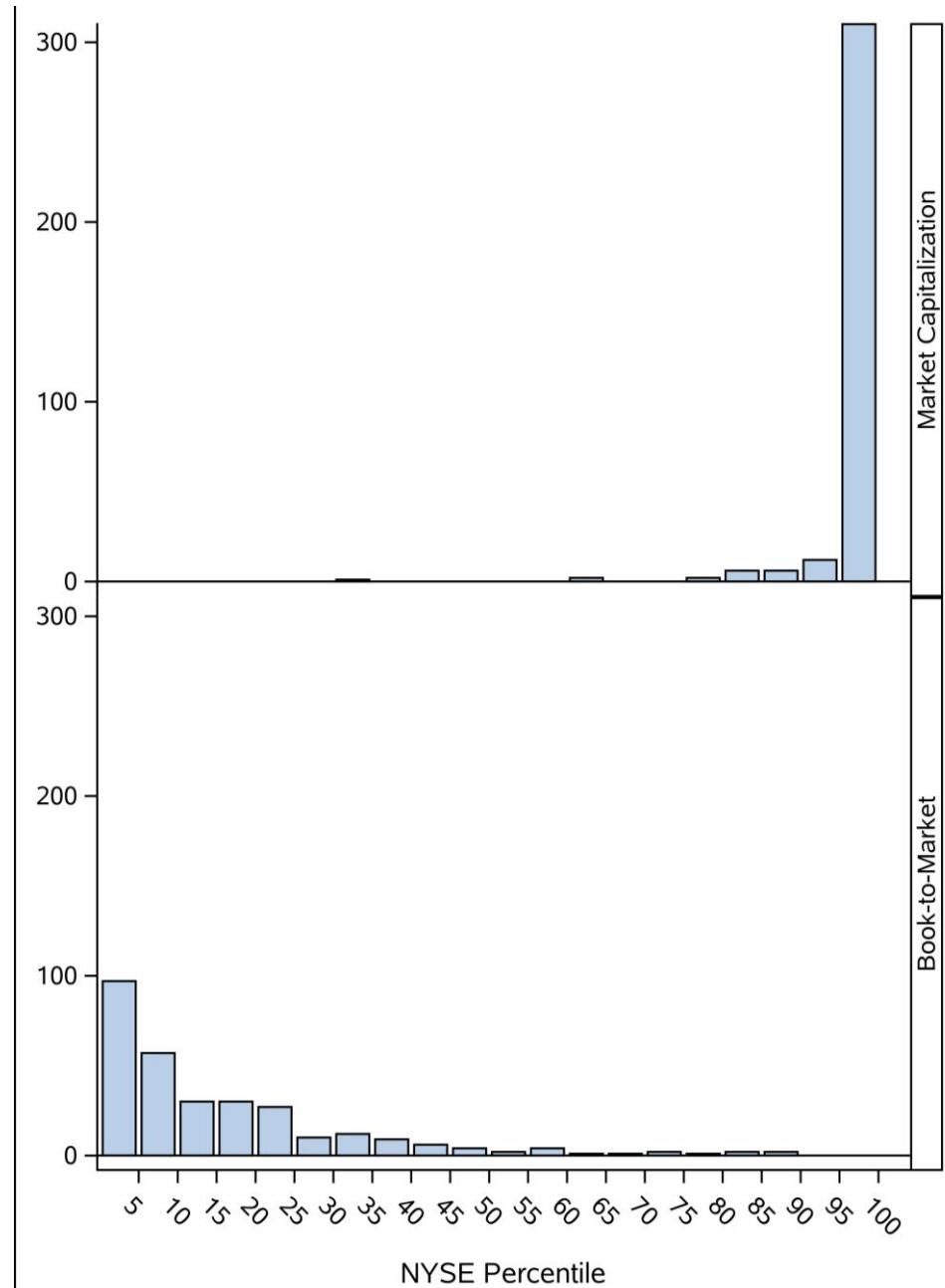
### 4.2. Characteristics of Selected Stocks

While a direct analysis of the size and book-to-market ratios (B/M) of the firms selected would be difficult to draw any conclusions from due to the changing nature of the market over a nearly four-decade period, tendencies emerge when considering the stocks' size and B/M relative to other securities at the time of portfolio formation. Using

<sup>2</sup>. <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data library.html>

the five-percent breakpoints provided on Kenneth French's website, Figure 1 clearly shows a tendency for ChatGPT to select large, growth firms when prompted as described in Section 3.2. Note that these breakpoints are based only on NYSE stocks, whereas all U.S. stocks are eligible for inclusion in these portfolios. If a selected stock was below the minimum or above the maximum of all NYSE stocks in that period, it was placed in the bottom (less than 5th percentile) or top group (greater than 95th percentile), respectively.

**Figure 1.** NYSE Percentile Groups for Market Capitalization and Book-to-Market Securities Chosen by ChatGPT-4, 1985-2021



*Note:* Figure 1 shows the percentile groups of the size and book-to-market values of the stocks selected by ChatGPT-4 for each of the 37 portfolios formed on July 1 (or the next trading day) of each year from 1985 to 2021. Breakpoints for each percentile group were pulled from Kenneth French's website.

#### 4.3. Analysis of Portfolio Returns

The returns on the portfolios suggested by ChatGPT are assessed on both a raw and a risk-adjusted basis. The ten stocks selected each year were placed in a price-weighted portfolio in July of the selection year. Let  $r_i$  be the stock's return,  $r_P$  be the portfolio's return, the selection year be represented by  $t$  (where  $t = 1$  represents 1985,  $t = 2$  represents 1986, etc.), and the month of the holding period be represented by  $m$ .

If stock  $i$  earns return  $r_{i,t,m}$ , then the portfolio return can be calculated as

$$r_{P,t,m} = \sum_{i=1}^{10} w_{i,t,m} r_{i,t,m}, \quad (3)$$

where  $w_{i,t,m}$  represents the percent of the portfolio invested in stock  $i$  at the beginning of month  $m$  during the holding period defined by selection year  $t$ . For  $m = 1$ , this will simply be the price of stock  $i$  at the beginning of the first trading day on or after July 1 of the selection year. Since the portfolio will be held with no modifications, it will work out that the weights for each month thereafter can be determined in the same way based on the prices of the first trading day of that particular month.

For raw returns, let  $r_M$  be the return on the market portfolio as found on Kenneth French's website<sup>3</sup>, and calculate the abnormal return as:

$$AR_{t,m} = r_{P,t,m} - r_{M,t,m}, \quad (4)$$

For a given investment horizon  $H$ , cumulative abnormal returns for a given portfolio and average cumulative abnormal returns across portfolios from formation years 1 through  $T$  could then be calculated as:

$$CAR_{t,H} = \sum_{m=1}^H AR_{t,m}, \quad (5)$$

$$ACAR_{T,H} = \frac{1}{T} \sum_{t=1}^T CAR_{t,H}, \quad (6)$$

For risk-adjusted returns, the monthly excess returns ( $XR_{P,t,m} = r_{P,t,m} - r_{f,t,m}$ ) are set as the dependent variable for an OLS regression on a combination of  $k$  risk factors  $F$ .

$$XR_P = \alpha_P + F\beta'_P + \varepsilon_P, \quad (7)$$

where  $XR_P$  is a column vector with  $n = T \times H$  monthly excess returns,  $\alpha_P$  is a column vector of size  $n$  containing the OLS intercept term that represents the risk-adjusted abnormal return,  $F$  is the  $n \times k$  matrix of monthly risk factors,  $\beta'$  is a transposed column vector containing  $k$  regression coefficients that represent the risk factor sensitivities, and  $\varepsilon_P$  is a column vector containing  $n$  standard normal I.I.D. OLS residuals.

## 5. Results

### 5.1. Assessing ChatGPT's Selections

ChatGPT-4's approach to stock selection follows the pattern of a top-down approach. As shown in Appendix A, ChatGPT relies first on the "Economic Context" at the time of selection, followed by a focus on four "Key Sectors for Growth" presumably based on the economic context. In the example shown in the appendix, it is clear how the stocks chosen reflect the key sectors described, with some stocks crossing over multiple sectors: technology (IBM, GE, Intel), healthcare and pharmaceuticals (Merck, 3M), consumer goods (Coca-Cola, GE, Procter & Gamble, 3M, McDonald's, Walmart), and energy (Exxon). Technology and Healthcare and/or Pharmaceuticals were key sectors in all 37 years; some form of consumer spending (goods, discretionary, or staples) was chosen in 31 out of the 37 years, and financial services appeared in 20 of the 37 years. This seems to indicate that while ChatGPT can modify its recommendations based on the context of the investing environment, there are certain sectors that it appears to be more comfortable with or more confident recommending. Table 1 shows each year's stock selections and the

<sup>3</sup>. Calculated by adding the variables labeled  $R_m - R_f$  and  $R_f$ .

number of new stocks in each portfolio relative to the previous year. While only 2 of the updated portfolios are identical to the previous year (1985-86 and 1997-98), as many as 4 of the 10 stocks were replaced in 9 of the 36 updates; the average annual turnover rate was about 26%. Figure 2 shows the number of times the various securities were included in one of the ChatGPT portfolios. The 17 most frequently selected stocks were from one of these four sectors. Microsoft (MSFT) led the way with 34 selections over 37 years.<sup>4</sup>

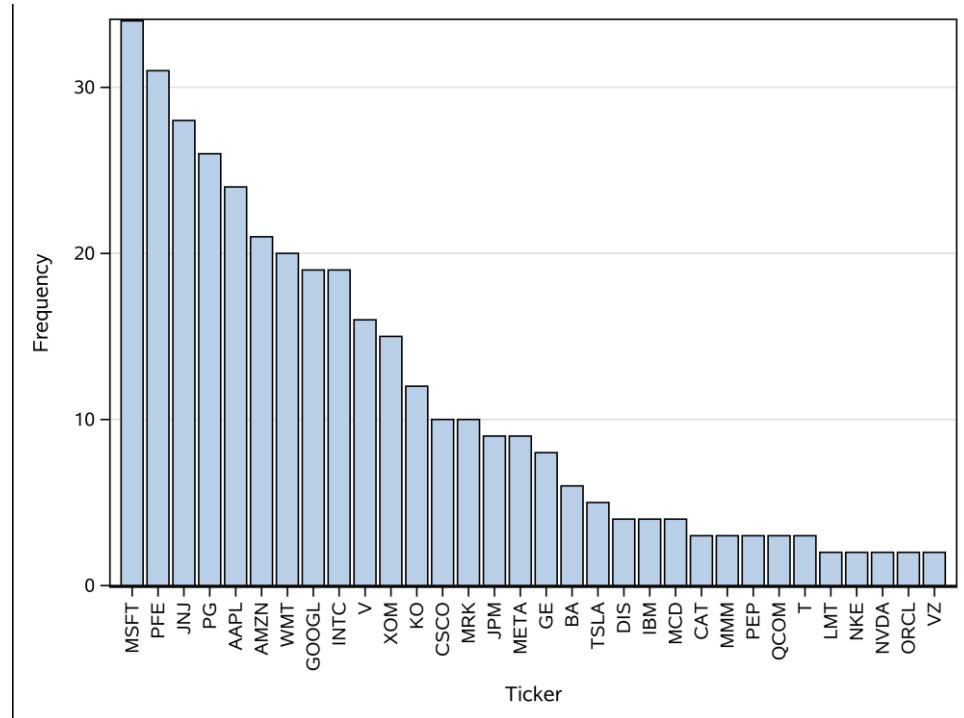
**Table 1.** Annual Portfolios

Year	Tickers										Turnover
1985	GE	IBM	INTC	KO	MCD	MMM	MRK	PG	WMT	XOM	
1986	GE	IBM	INTC	KO	MCD	MMM	MRK	PG	WMT	XOM	0
1987	AAPL	BA	INTC	JPM	KO	MRK	MSFT	PG	WMT	XOM	4
1988	IBM	INTC	JPM	KO	MMM	MRK	MSFT	NIKE	PFE	PG	4
1989	DIS	GS	IBM	INTC	JNJ	KO	MRK	MSFT	PG	WMT	4
1990	INTC	JNJ	KO	LMT	MCD	MRK	MSFT	PEP	PG	WMT	3
1991	BA	GE	INTC	JNJ	KO	MRK	MSFT	PFE	PG	WMT	3
1992	DIS	INTC	JNJ	JPM	KO	MRK	MSFT	PFE	PG	WMT	2
1993	CSCO	DIS	INTC	KO	MRK	MSFT	PFE	PG	T	WMT	2
1994	CSCO	INTC	JNJ	JPM	KO	MRK	MSFT	PFE	T	WMT	2
1995	AAPL	CAT	CSCO	DIS	INTC	JNJ	MSFT	PFE	PG	WMT	4
1996	AAPL	AMGN	CSCO	GE	HD	INTC	MSFT	PFE	PG	WMT	3
1997	AAPL	AMZN	CAT	CSCO	GE	INTC	MSFT	PFE	QCOM	WMT	3
1998	AAPL	AMZN	CAT	CSCO	GE	INTC	MSFT	PFE	QCOM	WMT	0
1999	AMZN	AOL	CSCO	GE	INTC	MSFT	ORCL	PFE	QCOM	WMT	2
2000	CSCO	INTC	JNJ	MSFT	ORCL	PFE	PG	T	WMT	XOM	4
2001	CSCO	GE	INTC	JNJ	LMT	MSFT	PEP	PFE	PG	WMT	3
2002	BA	CSCO	INTC	JNJ	MSFT	PEP	PFE	PG	WMT	XOM	2
2003	AAPL	AMZN	BA	GOOGL	JNJ	MSFT	PFE	PG	V	XOM	4
2004	AAPL	AMZN	GOOGL	JNJ	JPM	META	MSFT	PFE	TSLA	V	3
2005	AAPL	AMZN	GOOGL	HON	JNJ	MSFT	PFE	PG	V	XOM	3
2006	AAPL	AMZN	BA	GOOGL	JNJ	JPM	MSFT	PFE	PG	V	2
2007	AAPL	AMZN	GOOGL	INTC	JNJ	KO	MSFT	PG	VZ	XOM	4
2008	AAPL	GOOGL	JNJ	KO	MCD	PFE	PG	VZ	WMT	XOM	3
2009	AAPL	AMZN	GOOGL	JNJ	META	MSFT	PFE	PG	V	XOM	4
2010	AAPL	AMZN	GOOGL	JNJ	JPM	MSFT	PFE	PG	V	WMT	2
2011	AAPL	AMZN	GOOGL	JNJ	META	MSFT	NKE	PFE	TSLA	V	3
2012	AAPL	AMZN	GOOGL	JNJ	META	MSFT	PFE	PG	V	XOM	2
2013	AAPL	AMZN	BA	GOOGL	JNJ	MSFT	PFE	PG	V	XOM	1
2014	AAPL	AMZN	DIS	GOOGL	JNJ	MSFT	PFE	PG	V	XOM	1
2015	AAPL	AMZN	GOOGL	JNJ	MSFT	NKE	PFE	PG	V	XOM	1
2016	AAPL	AMZN	GOOGL	JNJ	JPM	META	MSFT	PFE	PG	V	2
2017	AAPL	AMZN	GOOGL	JNJ	META	MSFT	NVDA	PFE	TSLA	V	2
2018	AAPL	ADBE	AMZN	GOOGL	JNJ	JPM	META	MSFT	PFE	V	2
2019	AAPL	AMZN	GOOGL	JNJ	META	MSFT	PFE	V	WMT	XOM	2
2020	AAPL	AMZN	GOOGL	JNJ	MA	MSFT	NFLX	NVDA	PFE	TSLA	4
2021	AAPL	AMZN	GOOGL	JNJ	META	MSFT	NEE	PFE	TSLA	V	3
									Average		2.583

<sup>4</sup> Its IPO was in March 1986, explaining its exclusion from 1985 and 1986 portfolios. The only other year of exclusion was 2008, during the financial crisis of that period.

*Note:* Table 1 lists the stocks chosen by ChatGPT for each year's portfolio. The final column shows the amount of stocks from the previous year's portfolio that were replaced, with the average turnover shown in the last row.

**Figure 2.** Securities Chosen by ChatGPT-4, 1985-2021



*Note:* Figure 2 shows the frequency of the stocks selected by ChatGPT-4 for each of the 37 portfolios formed on July 1 (or the next trading day) for each year from 1985 to 2021. Note: 10 stocks chosen only once during this period were omitted from this graph.

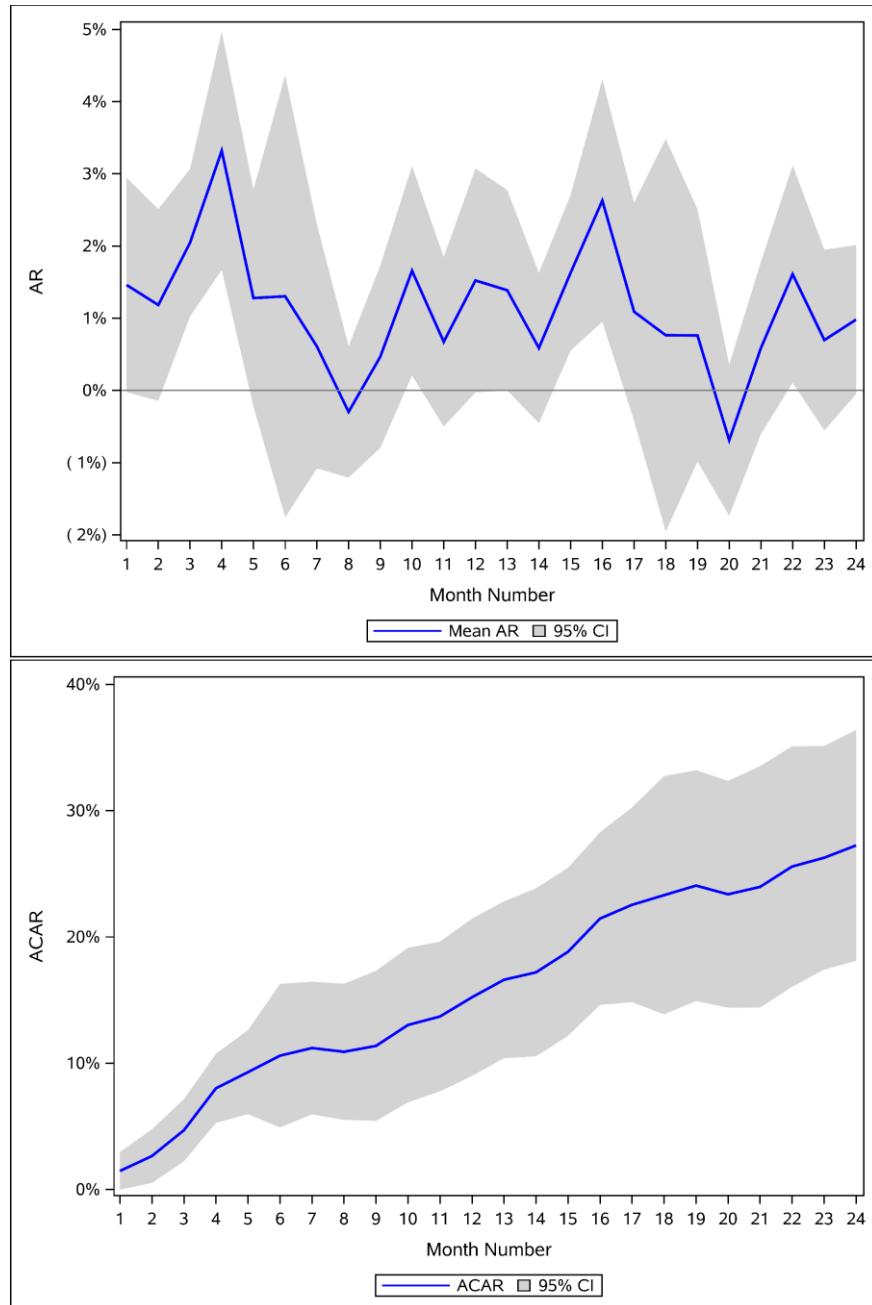
### 5.2. Raw Returns

Analysis of cumulative abnormal returns shows that the ChatGPT portfolios provide greater returns than the market early in and throughout the holding period. Table 2 shows that after 24 months, abnormal returns will accumulate on average to 27.242%, which is significant for  $p < 0.001$ . Figure 3 illustrates the mean AR and the ACAR for each of the 24 months of the portfolios' holding periods, along with 95% confidence intervals for each. While 22 of the 24 months' ARs are positive on average, only 7 of the 24 are significantly different than zero for  $p < 0.05$ . However, this figure also shows that starting with the second month, this portfolio earns positive ACAR at the  $p < 0.05$  significance level.

**Table 2:** Descriptive Statistics for Returns and Regression Coefficients, Price-Weighted Portfolios

Coefficient	N	Sig. at p < 0.05	Mean	St. Dev.	Min	Median	Max
Panel A: Average Monthly Returns							
Portfolio	888		2.077	6.776	-21.332	2.065	56.661
Market	888		0.939	4.527	-22.640	1.460	13.650
CAR <sub>24</sub>	37	YES	27.242	27.418	-6.592	23.642	128.999
Panel B: CAPM							
Intercept	37	7	1.088	1.095	-0.247	1.001	4.505
Beta	37	35 (10)	1.087	0.365	0.512	1.065	2.054
R <sup>2</sup>	37		57.73	21.08	14.93	59.45	87.86
Adjusted R <sup>2</sup>	37		55.81	22.04	11.07	57.61	87.31
Panel C: Fama-French Three-Factor Model							
Intercept	37	9	1.183	1.377	-0.450	0.751	5.646
Beta	37	34 (11)	1.022	0.342	0.429	1.004	1.984
SMB	37	7	-0.426	0.284	-1.092	-0.416	0.088
HML	37	22	-0.793	0.621	-3.008	-0.581	-0.124
R <sup>2</sup>	37		70.05	18.80	25.30	76.71	91.91
Adjusted R <sup>2</sup>	37		65.56	21.62	14.09	73.21	90.70
Panel D: Fama-French-Five-Factor-Model							
Intercept	37	8	1.102	1.388	-0.737	0.723	4.681
Beta	37	33 (12)	1.049	0.353	0.393	1.036	2.200
SMB	37	7	-0.409	0.356	-1.043	-0.432	0.294
HML	37	8	-0.485	0.737	-2.308	-0.369	2.351
R <sup>2</sup>	37		75.90	14.86	36.35	80.04	92.89
Adjusted R <sup>2</sup>	37		69.20	18.99	18.67	74.50	90.92
Panel E: Fama-French Three-Factor Model with Momentum and Reversals							
Intercept	37	9	1.097	1.337	-0.476	0.780	5.292
Beta	37	33 (13)	1.067	0.393	0.391	1.026	2.076
SMB	37	9	-0.488	0.403	-1.725	-0.436	0.126
HML	37	13	-0.790	0.734	-3.791	-0.535	0.430
MOM	37	2	0.161	0.316	-0.652	0.156	0.960
LTRev	37	2	0.008	0.661	-1.309	0.024	2.438
R <sup>2</sup>	37		73.98	16.96	29.83	79.05	94.40
Adjusted R <sup>2</sup>	37		66.75	21.67	10.33	73.23	92.84

*Note:* Table 2 shows the descriptive statistics for the average monthly price-weighted portfolio and market returns and the regression coefficients of the regressions run from 1985 to 2021. Panel A provides the statistics for the 24 monthly returns for each of the 37 portfolios formed and the 24-month CAR. Panels B through E provide the statistics for the results of the factor regressions run on the 37 portfolios; models shown include CAPM, Fama-French three- and five-factor models, and Fama-French three-factor model with momentum and long-term reversals. Summary statistics include the number of coefficients, number of statistically significant coefficients at  $p < 0.05$  (number of beta coefficients statistically different than 1 at  $p < 0.05$ ), mean, standard deviation, minimum, median, and maximum. All returns in Panel A along with the Intercept,  $R^2$ , and Adjusted  $R^2$  rows throughout the table are measured as percentages.

**Figure 3.** Average Abnormal Returns and Average Cumulative Abnormal Returns Over 24 Months

*Note:* Figure 3 shows the average AR and ACAR of all 37 selection years as of the end of each of the 24 months of the portfolio holding periods. The shaded region represents a 95% confidence interval surrounding the values.

### 5.3. Risk-Adjusted Returns

Regressions were run on the monthly excess returns using CAPM, Fama and French (1993, 2015) three-factor (FF3) and five-factor (FF5) models, and FF3 with additional factors for momentum and long-term reversals (FF3m). Price-weighting was used to reflect the prompt's direction that 100 shares of each stock would be purchased as mentioned in Section 3.2.<sup>5</sup> Table 3 summarizes the results of these regressions while controlling for year fixed effects. Results were consistent across all four models. The

<sup>5</sup> CAPM and FF3 are the main models being analyzed. FF5 and FF3m results are included as robustness checks on the main results.

monthly abnormal returns were positive, ranging from 0.483% (FF3m) to 0.939% (FF5), but none were statistically significant. All four betas were significant for  $p < 0.001$ , ranging from 1.020 (FF5) to 1.116 (FF3m); only the FF5 beta was indistinguishable from a value of 1 at  $p < 0.05$ . Both the size and book-to-market factors were negative and significant at  $p < 0.001$  for all three models, indicating ChatGPT prefers large, growth stocks for this analysis. Given the use of the stock selection criteria explained in Appendix A (market leadership, financial health, and dividend yield being three of the four criteria) and the size and book-to-market statistics reported in section 4.2, this is not a surprising result. For the last two models (FF5 and FF3m), results indicate that ChatGPT prefers to recommend firms with aggressive investment (CMA = -0.544,  $p < 0.001$ ) and higher prior intermediate-term returns (MOM = 0.164,  $p < 0.001$ ), while robustness of profits (RMW = 0.127,  $p = 0.0619$ ) and prior long-term returns were not a factor (LTRev = 0.088,  $p = 0.1912$ ).

**Table 3.** Regression Results, Price-Weighted Portfolios

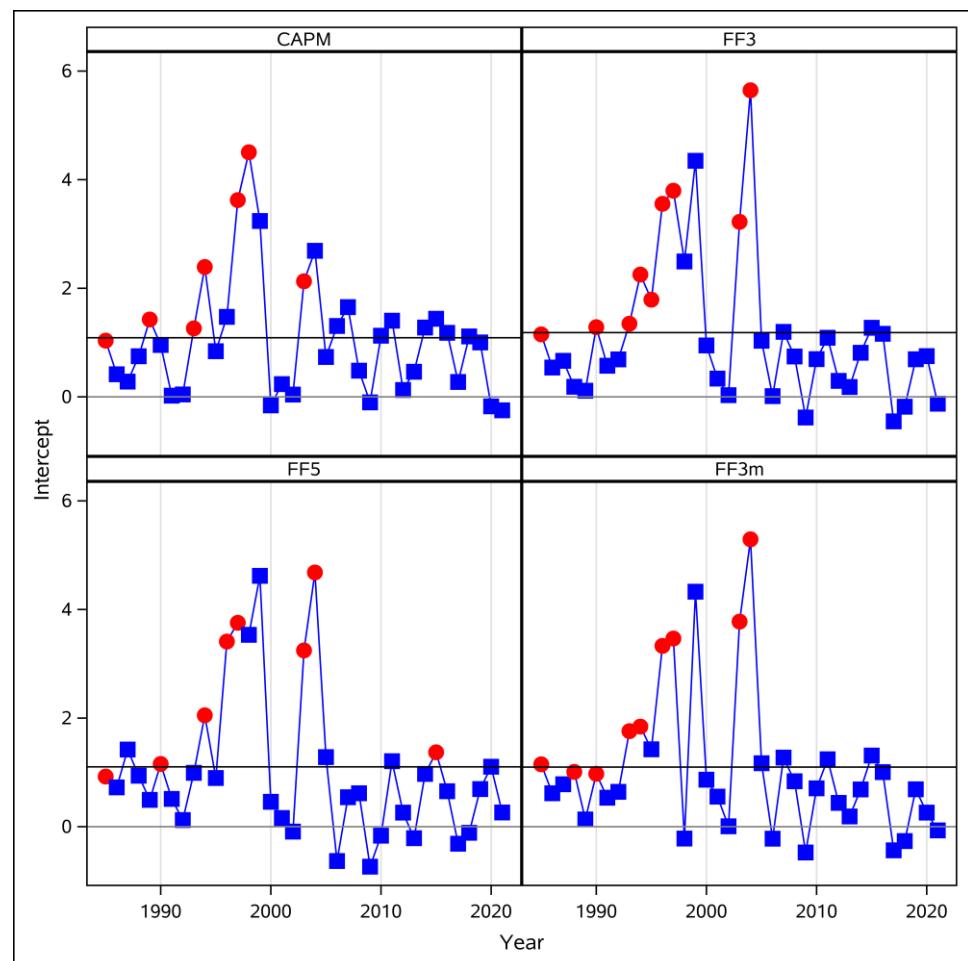
Coefficient	CAPM	FF3	FF5	FF3 w/
Panel A: Overlapping Two-Year Holding Periods				
Intercept	0.798	0.643	0.939	0.483
Beta	1.086***	1.075***	1.020***	1.116***
SMB		-0.307***	-0.257***	-0.339***
HML		-0.583***	-0.387***	-0.553***
RMW			0.127	
CMA			-0.544***	
MOM				0.164***
LTRev				0.088
Year Fixed Effects	YES	YES	YES	YES
Reject $\beta = 1$	YES	YES	NO	YES
N	888	888	888	888
R <sup>2</sup>	55.66	63.81	65.46	64.99
Adjusted R <sup>2</sup>	53.73	62.15	63.78	63.29
Panel B: Portfolio Rebalanced Every July 1				
Intercept	0.728	0.538	0.429	0.433
Beta	1.093***	1.078***	1.037***	1.127***
SMB		-0.300***	-0.226**	-0.352***
HML		-0.582***	-0.426***	-0.554***
RMW			0.191	
CMA			-0.473**	
MOM				0.200***
LTRev				0.143
Year Fixed Effects	NO	NO	NO	NO
Reject $\beta = 1$	NO	NO	NO	YES
N	444	444	444	444
R <sup>2</sup>	56.28	63.29	64.68	64.96
Adjusted R <sup>2</sup>	52.30	59.75	61.08	61.39

*Note:* Table 3 shows the results of the regressions of the 37 price-weighted portfolios chosen by ChatGPT from 1985 to 2021. Panel A shows the results from holding each portfolio for 24 months from the date of formation; the last 12 months of each portfolio would overlap with the first 12 months of the next portfolio. Panel B shows the results from rebalancing a single portfolio every July 1 and holding those shares for 12 months. Variables included were excess market returns (Beta), size (SMB), book-to-market (HML), operating profitability (RMW), investment (CMA), momentum (MOM), and long-term reversals (LTRev). Statistical difference from 0 is indicated for  $p < 0.05$  (\*),  $p < 0.01$  (\*\*), and  $p < 0.001$  (\*\*\*)<sup>1</sup>. Fixed effects for each year were included in each model, and the coefficient on the excess market return risk factor was tested for significant differences from 1 at  $p < 0.05$ . The Intercept, R<sup>2</sup>, and Adjusted R<sup>2</sup> rows throughout the table are measured as percentages.

Additionally, Table 2 provides descriptive statistics of each year's analysis results. The average monthly return for the ChatGPT portfolios was 2.077%, compared to 0.939% for the market. While the average monthly abnormal return across all years averages 1.088% and 1.183% for CAPM and FF3, respectively, only 7 and 9 out of those 37 portfolios were significant at the  $p < 0.05$  level. Of those significant portfolios, only 3 of the 16 significant results occurred from 2000 through 2021 (2003 for both models and 2004 for FF3). Similar results are seen in FF5 and FF3m.

It is also interesting to note that while there are very few years in which the portfolio will provide significantly positive abnormal returns, there are only 22 instances across all 148 model-years tested in which the abnormal returns are negative. None of these are significant (the smallest p-value is 0.3412), and the worst performer is only  $-0.737\%$  per month in 2009 as determined by FF5. Of the 22 negative intercepts, the only year in which all four models agree on the sign of the intercept is 2009. So while it is debatable from these results whether ChatGPT-4 can do a good job picking stocks for an investor, it cannot be concluded that ChatGPT would do a poor job on a risk-adjusted basis. See Figure 4 for an illustration of all 37 intercepts for each model.

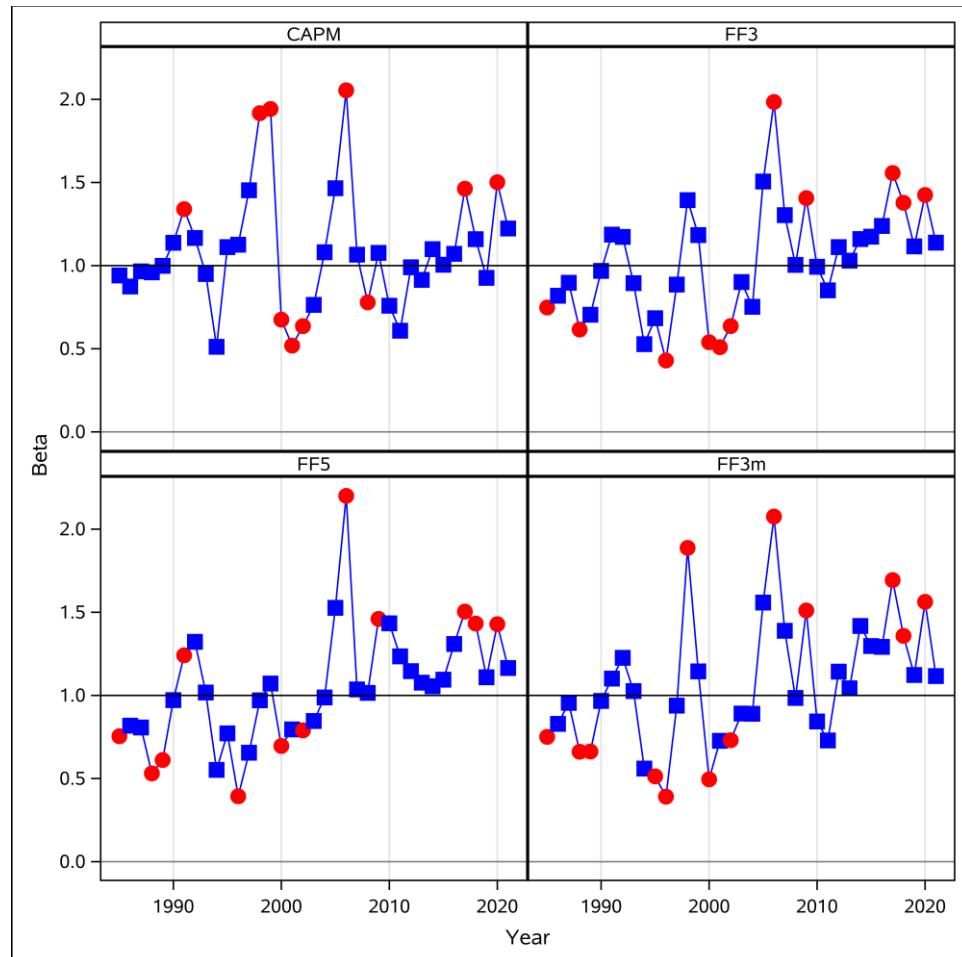
**Figure 4.** Monthly Abnormal Returns of Each Portfolio Chosen by ChatGPT-4, 1985-2021



*Note:* Figure 4 shows the monthly abnormal returns from each of the 37 portfolios formed on July 1 (or the next trading day) for each year from 1985 to 2021 as determined by the intercepts resulting from regressing equal-weighted portfolios of securities selected by ChatGPT-4 using CAPM and the Fama-French three- and five-factor models and the Fama-French three-factor model with momentum and long-term reversals over 24 months. Circles represent intercepts that are significant at  $p < 0.05$ . Reference lines represent the mean of the 37 intercepts produced by that panel's model.

All but 13 of the 148 beta coefficients significantly differ from zero at the  $p < 0.05$  level, averaging anywhere from 1.022 to 1.087 within each model grouping. Given the closeness to market betas, 102 of the 148 betas were not significantly different from one at the same  $p < 0.05$  level. Significant betas tended to be less than one through 2002, while all were greater than one after 2002. Figure 5 illustrates all 37 betas for each model.

**Figure 5.** Betas of Each Portfolio Chosen by ChatGPT-4, 1985-2021



Note: Figure 5 shows the betas from each of the 37 portfolios formed on July 1 (or the next trading day) for each year from 1985 to 2021 as determined by the intercepts resulting from regressing equal-weighted portfolios of securities selected by ChatGPT-4 using CAPM and the Fama-French three- and five-factor models and the Fama-French three-factor model with momentum and long-term reversals over 24 months. Circles represent betas that are significantly different than 1 at  $p < 0.05$ . Reference lines are set at  $\beta = 1$ .

Finally, the negative average values of the size (-0.426) and book-to-market (-0.793) factors reinforce the implication of Figure 1 that ChatGPT has focused on large, growth firms to construct the requested portfolios.

#### 5.4. Robustness

##### 5.4.1 Other Portfolio Weighting Methods

For robustness and to align with more commonly used weighting methods, the same raw and risk-adjusted analyses were performed on equal- and value-weighted portfolios (EW and VW, respectively). For the raw returns of both portfolios, 24-month ACAR was positive and significant at  $p < 0.001$ , earning 22.679% for EW and 12.798% for VW (as shown in Panel A of Tables 4 and 5, respectively). Figure 6 illustrates the accumulation of abnormal returns over the 24 months for all three weighting systems (the first panel, PW,

duplicates the bottom half of Figure 3). Also, while not a significant point of analysis for this study, it can be shown that the ACAR for PW and EW are not significantly different than each other (Bonferroni-adjusted  $p = 0.3310$ ), while both PW and EW ACARs are greater than the VW ACAR (Bonferroni-adjusted  $p < 0.001$  and  $= 0.0258$ , respectively).

**Table 4.** Descriptive Statistics for Returns and Regression Coefficients, Equal-Weighted Portfolios

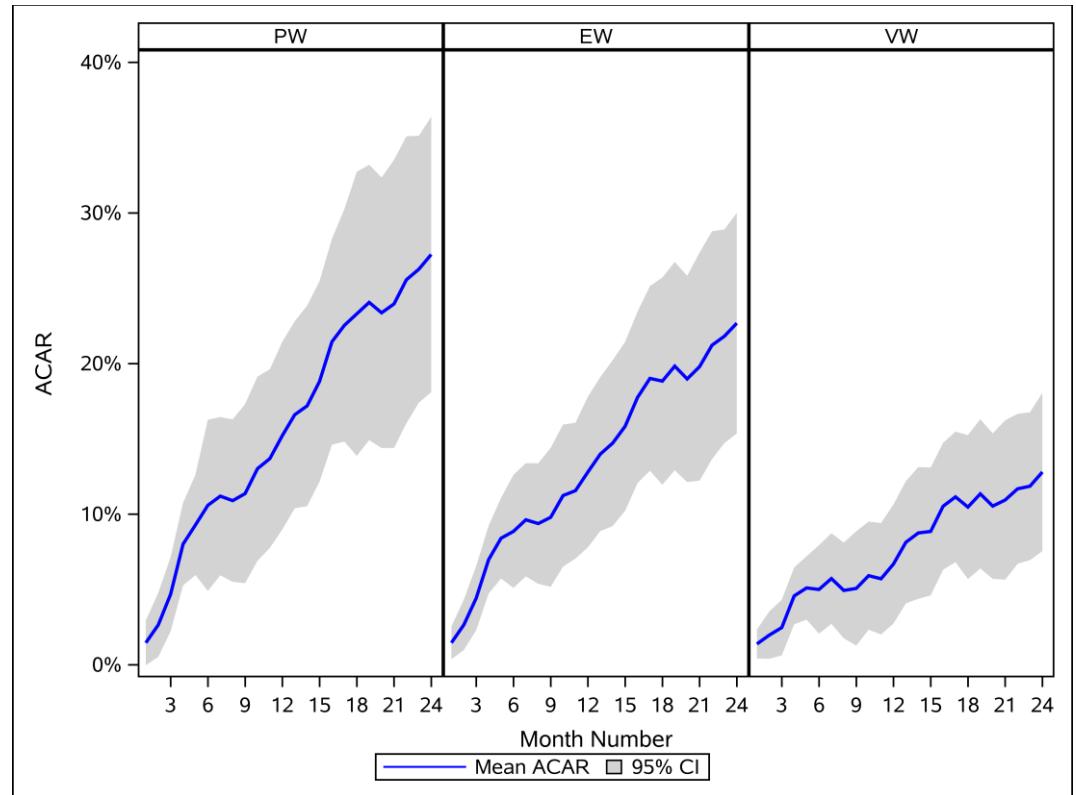
Coefficient	N	Sig. at $p < 0.05$	Mean	St. Dev.	Min	Median	Max
Panel A: Average Monthly Returns							
Portfolio	888		1.886	5.733	-22.882	2.079	31.058
Market	888		0.939	4.527	-22.640	1.460	13.650
CAR <sub>24</sub>	37	YES	22.679	21.954	-1.004	18.225	107.117
Panel B: CAPM							
Intercept	37	9	0.937	0.864	-0.102	0.775	3.887
Beta	37	37 (8)	1.040	0.279	0.547	1.000	1.805
$R^2$	37		68.19	16.34	20.89	70.14	93.58
Adjusted $R^2$	37		66.74	17.08	17.29	68.78	93.29
Panel C: Fama-French Three-Factor Model							
Intercept	37	11	1.077	0.996	-0.153	0.837	4.141
Beta	37	37 (12)	0.973	0.278	0.381	0.942	1.923
SMB	37	14	-0.373	0.241	-0.821	-0.389	0.202
HML	37	22	-0.567	0.479	-1.653	-0.498	0.185
$R^2$	37		78.29	13.32	30.90	80.04	94.68
Adjusted $R^2$	37		75.03	15.32	20.53	77.04	93.89
Panel D: Fama-French Five-Factor Model							
Intercept	37	8	1.025	1.013	-0.155	0.787	4.060
Beta	37	35 (10)	0.978	0.306	0.250	0.962	1.926
SMB	37	13	-0.358	0.293	-0.979	-0.323	0.192
HML	37	8	-0.375	0.528	-1.371	-0.422	1.562
RMW	37	4	0.148	0.568	-0.731	0.087	1.826
CMA	37	5	-0.401	0.563	-1.678	-0.392	0.733
$R^2$	37		82.44	10.47	43.95	85.30	95.49
Adjusted $R^2$	37		77.56	13.37	28.38	81.21	94.23
Panel E: Fama-French Three-Factor Model with Momentum and Reversals							
Intercept	37	9	1.065	0.955	0.005	0.844	4.005
Beta	37	37 (12)	0.996	0.313	0.415	0.960	1.895
SMB	37	11	-0.392	0.274	-1.021	-0.383	0.150
HML	37	16	-0.572	0.606	-2.260	-0.481	0.472
MOM	37	0	0.048	0.214	-0.429	0.018	0.462
LTRev	37	2	0.014	0.423	-1.190	0.024	0.970
$R^2$	37		81.08	11.91	41.63	85.25	95.18
Adjusted $R^2$	37		75.82	15.22	25.42	81.15	93.84

*Note:* Table 4 shows the descriptive statistics for the average monthly equal-weighted portfolio and market returns and the regression coefficients of the regressions run from 1985 to 2021. Panel A provides the statistics for the 24 monthly returns for each of the 37 portfolios formed and the 24-month CAR. Panels B through E provide the statistics for the results of the factor regressions run on the 37 portfolios; models shown include CAPM, Fama-French three- and five-factor models, and Fama-French three-factor model with momentum and long-term reversals. Summary statistics include the number of coefficients, number of statistically significant coefficients at  $p < 0.05$  (number of beta coefficients statistically different than 1 at  $p < 0.05$ ), mean, standard deviation, minimum, median, and maximum. All returns in Panel A along with the Intercept,  $R^2$ , and Adjusted  $R^2$  rows throughout the table are measured as percentages.

**Table 5.** Descriptive Statistics for Returns and Regression Coefficients, Value-Weighted Portfolios

Coefficient	N	Sig. at p < 0.05	Mean	St. Dev.	Min	Median	Max
Panel A: Average Monthly Returns							
Portfolio	888		1.473	5.384	-19.398	1.667	23.725
Market	888		0.939	4.527	-22.640	1.460	13.650
CAR24	37	YES	12.798	15.746	-11.683	12.218	50.495
Panel B: CAPM							
Intercept	37	3	0.581	0.641	-0.543	0.617	1.882
Beta	37	37 (12)	0.952	0.253	0.390	0.914	1.463
R <sup>2</sup>	37		64.00	19.31	17.64	70.67	90.32
Adjusted R <sup>2</sup>	37		62.36	20.18	13.90	69.34	89.88
Panel C: Fama-French Three-Factor Model							
Intercept	37	6	0.680	0.776	-0.287	0.530	3.544
Beta	37	36 (12)	0.910	0.252	0.460	0.921	1.540
SMB	37	15	-0.457	0.256	-1.112	-0.427	0.070
HML	37	19	-0.508	0.473	-1.734	-0.421	0.157
R <sup>2</sup>	37		76.76	14.47	32.17	79.57	92.10
Adjusted R <sup>2</sup>	37		73.27	16.64	21.99	76.51	90.92
Panel D: Fama-French Five-Factor Model							
Intercept	37	4	0.590	0.772	-0.466	0.464	3.455
Beta	37	36 (8)	0.925	0.267	0.445	0.923	1.495
SMB	37	10	-0.382	0.244	-1.017	-0.324	0.077
HML	37	10	-0.342	0.562	-1.635	-0.306	1.450
RMW	37	7	0.389	0.546	-0.497	0.275	2.344
CMA	37	8	-0.306	0.633	-1.717	-0.150	0.753
R <sup>2</sup>	37		81.54	11.19	46.83	84.00	93.15
Adjusted R <sup>2</sup>	37		76.41	14.30	32.05	79.56	91.25
Panel E: Fama-French Three-Factor Model with Momentum and Reversals							
Intercept	37	6	0.722	0.715	-0.199	0.622	3.249
Beta	37	35 (10)	0.927	0.293	0.436	0.955	1.472
SMB	37	14	-0.430	0.274	-1.089	-0.409	0.075
HML	37	9	-0.449	0.496	-1.563	-0.393	0.618
MOM	37	2	0.052	0.283	-0.691	0.085	0.474
LTRev	37	3	-0.013	0.370	-1.037	-0.002	0.722
R <sup>2</sup>	37		79.95	13.57	37.04	85.42	92.47
Adjusted R <sup>2</sup>	37		74.38	17.36	19.55	81.37	90.37

*Note:* Table 5 shows the descriptive statistics for the average monthly value-weighted portfolio and market returns and the regression coefficients of the regressions run from 1985 to 2021. Panel A provides the statistics for the 24 monthly returns for each of the 37 portfolios formed and the 24-month CAR. Panels B through E provide the statistics for the results of the factor regressions run on the 37 portfolios; models shown include CAPM, Fama-French three- and five-factor models, and Fama-French three-factor model with momentum and long-term reversals. Summary statistics include the number of coefficients, number of statistically significant coefficients at p < 0.05 (number of beta coefficients statistically different than 1 at p < 0.05), mean, standard deviation, minimum, median, and maximum. All returns in Panel A along with the Intercept, R<sup>2</sup>, and Adjusted R<sup>2</sup> rows throughout the table are measured as percentages.

**Figure 6.** Average Cumulative Abnormal Returns Over 24 Months by Weighting Scheme

Note: Figure 6 shows the ACAR as of the end of each of the 24 months of the portfolio holding periods for the price-weighted approach (PW) primarily analyzed in this study along with the equal-weighted (EW) and value-weighted (VW) approaches used as robustness checks. The shaded region represents a 95% confidence interval surrounding the ACAR.

For both portfolios on a risk-adjusted basis, the intercepts and coefficients for the three main risk factors were the same in sign and significance as the price-weighted portfolio for all four models. For the equal-weighted portfolios (Table 6), year fixed effects were significant, and beta was not significantly different from one in all four models. For the value-weighted portfolios (Table 7), year-fixed effects were only significant in the FF3m model, while beta significantly differed from the FF5 and FF3m models.

#### 5.4.2. Rolling One-Year Portfolios

As another robustness check, I assume the investor takes a longer-term perspective. Instead of holding the suggested stocks for two years then divesting, the investor will create their portfolio on July 1, 1985 based on ChatGPT's initial suggestions, then rebalance their portfolio using ChatGPT's new suggestions July 1 of every year, and holding the final portfolio mix until June 30, 2022 (one year after the final selection made on July 1, 2021). Panel B of Table 3 shows the results of performing the four regressions on this investment plan. The signs and significance of the intercepts and all risk factor coefficients in all four models are unchanged. However, fixed effects for years are no longer significant for any of the models, and only one of the betas is significantly different from 1 at  $p < 0.05$  (FF3m:  $\beta = 1.127$ ).

**Table 6.** Regression Results, Equal-Weighted Portfolios

Coefficient	CAPM	FF3	FF5	FF3 w/
Panel A: Overlapping Two-Year Holding Periods				
Intercept	0.936	0.780	0.938	0.768
Beta	1.035***	1.036***	1.006***	1.039***
SMB		-0.277***	-0.236***	-0.321***
HML		-0.418***	-0.310***	-0.455***
RMW			0.113*	
CMA			-0.320***	
MOM				0.030
LTRev				0.121*
Year Fixed Effects	YES	YES	YES	YES
Reject $\beta = 1$	NO	NO	NO	NO
N	888	888	888	888
$R^2$	69.45	75.96	76.88	76.23
Adjusted $R^2$	68.12	74.86	75.76	75.08
Panel B: Portfolio Rebalanced Every July 1				
Intercept	0.622	0.472	0.335	0.633
Beta	1.041***	1.033***	1.019***	1.048***
SMB		-0.245***	-0.179**	-0.302***
HML		-0.433***	-0.367***	-0.463***
RMW			0.176*	
CMA			-0.239*	
MOM				0.079*
LTRev				0.157*
Year Fixed Effects	YES	YES	YES	YES
Reject $\beta = 1$	NO	NO	NO	NO
N	444	444	444	444
$R^2$	70.09	75.80	76.60	76.47
Adjusted $R^2$	67.36	73.46	74.21	74.07

*Note:* Table 6 shows the results of the regressions of the 37 equal-weighted portfolios chosen by ChatGPT from 1985 to 2021. Panel A shows the results from holding each portfolio for 24 months from the date of formation; the last 12 months of each portfolio would overlap with the first 12 months of the next portfolio. Panel B shows the results from rebalancing a single portfolio every July 1 and holding those shares for 12 months. Variables included were excess market returns (Beta), size (SMB), book-to-market (HML), operating profitability (RMW), investment (CMA), momentum (MOM), and long-term reversals (LTRev). Statistical difference from 0 is indicated for  $p < 0.05$  (\*),  $p < 0.01$  (\*\*), and  $p < 0.001$  (\*\*\*)<sup>1</sup>. Fixed effects for each year were included in each model, and the coefficient on the excess market return risk factor was tested for significant differences from 1 at  $p < 0.05$ . The Intercept,  $R^2$ , and Adjusted  $R^2$  rows throughout the table are measured as percentages.

This same robustness check was performed on the equal- and value-weighted portfolios. For both the equal-weighted portfolio (Table 6, Panel B) and the value-weighted portfolio (Table 7, Panel B), the signs of the intercepts and all risk factor coefficients in all four models are unchanged, while the only change in significance is that the momentum factor for the equal-weighted portfolio in the FF3m model became significant at  $p < 0.05$ . For the equal-weighted portfolio, year fixed effects remain significant for all four models, and beta remains statistically indistinguishable from 1 at  $p < 0.05$  for all four models. None of the year fixed effects are statistically significant for the value-weighted portfolio, and beta is indistinguishable from 1 at  $p < 0.05$  for all four models.

**Table 7.** Regression Results, Value-Weighted Portfolios

Coefficient	CAPM	FF3	FF5	FF3 w/
Panel A: Overlapping Two-Year Holding Periods				
Intercept	0.444	0.212	0.359	0.190
Beta	0.959***	0.975***	0.948***	0.981***
SMB		-0.370***	-0.320***	-0.397***
HML		-0.411***	-0.308***	-0.428***
RMW			0.136**	
CMA			-0.320***	
MOM				0.032
LTRev				0.072
Year Fixed Effects	NO	NO	NO	YES
Reject $\beta = 1$	NO	NO	YES	YES
N	888	888	888	888
R <sup>2</sup>	67.54	76.35	77.49	76.51
Adjusted R <sup>2</sup>	66.13	75.27	76.40	75.38
Panel B: Portfolio Rebalanced Every July 1				
Intercept	0.107	-0.082	-0.215	0.035
Beta	0.951***	0.964***	0.945***	0.973***
SMB		-0.353***	-0.285***	-0.392***
HML		-0.400***	-0.318***	-0.422***
RMW			0.180**	
CMA			-0.281**	
MOM				0.052
LTRev				0.108
Year Fixed Effects	NO	NO	NO	NO
Reject $\beta = 1$	NO	NO	NO	NO
N	444	444	444	444
R <sup>2</sup>	68.36	75.91	77.04	76.26
Adjusted R <sup>2</sup>	65.48	73.58	74.70	73.84

*Note:* Table 7 shows the results of the regressions of the 37 value-weighted portfolios chosen by ChatGPT from 1985 to 2021. Panel A shows the results from holding each portfolio for 24 months from the date of formation; the last 12 months of each portfolio would overlap with the first 12 months of the next portfolio. Panel B shows the results from rebalancing a single portfolio every July 1 and holding those shares for 12 months. Variables included were excess market returns (Beta), size (SMB), book-to-market (HML), operating profitability (RMW), investment (CMA), momentum (MOM), and long-term reversals (LTRev). Statistical difference from 0 is indicated for  $p < 0.05$  (\*),  $p < 0.01$  (\*\*), and  $p < 0.001$  (\*\*\*)<sup>1</sup>. Fixed effects for each year were included in each model, and the coefficient on the excess market return risk factor was tested for significant differences from 1 at  $p < 0.05$ . The Intercept,  $R^2$ , and Adjusted  $R^2$  rows throughout the table are measured as percentages.

## 6. Conclusion

ChatGPT's portfolios demonstrate a cautious, top-down approach that prioritizes stability over aggressive overperformance, as evidenced by its consistent selection of large growth stocks in resilient sectors like technology and healthcare. This bias toward "market leaders" with strong financial health (Appendix A) likely stems from the model's training on historical data emphasizing blue-chip firms, leading to positive but insignificant alphas across models. For instance, the negative size and book-to-market betas (-0.426 and -0.793 on average) suggest ChatGPT avoids small/value stocks prone to volatility, which insulated portfolios during downturns (e.g., no negative alphas in 2008-2009) but limited upside in bull markets. The 26% turnover reflects adaptive responses to economic contexts, shifting from energy in the 1980s to tech in the 1990s. Yet this conservatism

meant that only about one out of every four years showed significant alphas, often in recovery periods like 2003-2004.

These patterns interpretively align with ChatGPT's limitations as a generalist LLM. It excels at synthesizing broad trends but struggles with nuanced, data-driven predictions without real-time access, explaining the lack of consistent abnormality. Compared to similar studies, this retrospective test over 37 years reveals more tempered performance than short-term, news-based analyses. For example, Lopez-Lira and Tang (2024) and Chen et al. (2025) report more substantial returns from complex LLMs on headlines. Still, their focus on daily and 30-day horizons captures event-driven spikes absent in my two-year holding periods. Similarly, Lu et al. (2023) find positive alphas post-announcements but note out-of-sample weakness, mirroring my insignificant long-run results. In contrast to Ko and Lee (2023), who achieve higher Sharpe ratios via mean-variance optimization, ChatGPT's unoptimized selections underperform the efficient frontier (Romanko et al., 2023), underscoring its role as a heuristic tool rather than a quantitative optimizer.

Overall, ChatGPT could serve as a viable, accessible stock-picking tool for individual investors, particularly those favoring price-weighted strategies, but it does not reliably generate abnormal returns. Practically, its portfolios' avoidance of significant underperformance, earning positive CARs (27.242% over 24 months) without negative alphas, suggests utility for risk-averse, unsophisticated investors seeking diversified exposure to large growth sectors like technology and healthcare. Small investors might use it to democratize advice, bypassing costly professionals, though integration with tools for real-time data or optimization could enhance outcomes.

However, limitations temper these implications. Despite prompts restricting knowledge, potential look-ahead bias arises from GPT-4's post-2021 training. The model may inadvertently draw on patterns learned from later data. Prompting sensitivity (Section 3) also biases selections toward familiar sectors, and the study assumes no transaction costs or taxes, understating real-world frictions. Results may not generalize beyond U.S. stocks or two-year horizons.

Future research could extend this by testing hybrid human-AI strategies as suggested by Fatouros et al. (2024), incorporating real-time APIs for dynamic rebalancing, or extending the works of Kalyan (2024) and Chen et al. (2025) by comparing GPT-4 to specialized financial LLMs across global markets. Exploring ethical implications and potential solutions, such as those suggested by Khan and Umer (2024) and Nie et al. (2024), would further highlight its transformative potential in Fintech.

**Acknowledgements.** The author would like to thank Zach McGurk, Mary Becker, Kevin Brady, Markus Schmid, anonymous reviewers, and the 2024 AEF Annual Meeting participants for their insightful and constructive comments.

**Declarations.** The author has no financial or non-financial interests directly or indirectly related to the work submitted for publication

**Funding:** This research received no external funding.

**Data Availability Statement:** The manuscript contains data that will be made available upon reasonable request.

**Conflicts of Interest:** The author declares no conflict of interest.

## Appendix A

The key factors listed for 1985 were as follows.

**Economic Context:** The U.S. economy was strong, with policies under President Reagan focusing on tax cuts, deregulation, and increased defense spending. Inflation was under control, and interest rates were trending downward.

Key Sectors for Growth:

**Technology:** The tech industry was expanding rapidly, especially in areas like personal computing and software.

**Healthcare and Pharmaceuticals:** Driven by medical advancements and an aging population.

**Consumer Goods:** Increased consumer spending was boosting this sector.

**Energy:** The 1980s saw significant developments in the energy sector.

#### Stock Selection Criteria:

**Market Leadership:** Companies that were industry leaders typically offered stability.

**Growth Potential:** Firms in growing industries or those showing promising growth.

**Financial Health:** Companies with strong balance sheets and good cash flow.

**Dividend Yield:** An important factor for total return.

And the ten stocks selected for the initial portfolio (in no particular order) were:

**IBM (International Business Machines Corp.):** A leader in the burgeoning field of personal computers and technology.

**Merck & Co.:** A major player in the pharmaceutical industry.

**Coca-Cola:** Benefiting from increased consumer spending.

**General Electric (GE):** A diversified conglomerate with interests in technology and consumer goods.

**Intel:** A key player in the semiconductor industry.

**Procter & Gamble:** A leading company in consumer goods.

**3M:** Known for its innovation and diverse product line.

**McDonald's:** Reflecting the growth in the fast-food industry.

**Walmart:** Rapidly expanding in the retail sector.

**Exxon:** Representing the energy sector, which was vital in the 1980s.

With each year's request, ChatGPT-4 provided the economic context and key sectors for growth along with its top ten selections. However, the lack of stock selection criteria led me to inquire into its absence. ChatGPT-4 responded that "the same primary stock selection criteria are consistently applied" each year and that, "These criteria provide a balanced approach, aiming to ensure a mix of stability, growth potential, and diversification across various sectors."

## References

- Ahangar, R. G., & Fietko, A. (2023). Exploring the Potential of ChatGPT in Financial Decision Making. In *Advancement in Business Analytics Tools for Higher Financial Performance* (pp. 94–111). IGI Global.
- Ali, H., & Aysan, A. F. (2023). What will ChatGPT Revolutionize in Financial Industry? *Modern Finance*, 1 (1). <https://doi.org/10.61351/mf.v1i1.67>
- Capponi, A., Olafsson, S., & Zariphopoulou, T. (2020). Personalized robo-advising: Enhancing investment through client interaction [arXiv: 1911.01391 [q-fin.PM]].
- Chen, J., Tang, G., Zhou, G., & Zhu, W. (2025, February 14). ChatGPT and Deepseek: Can they predict the stock market and macroeconomy? <https://doi.org/10.48550/arXiv.2502.10008>
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33 (1), 3–56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)
- Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116 (1), 1–22. <https://doi.org/10.1016/j.jfineco.2014.10.010>
- Fatouros, G., Metaxas, K., Soldatos, J., & Kyriazis, D. (2024). Can large language models beat wall street? evaluating GPT-4's impact on financial decision-making with MarketSenseAI. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-024-10613-4>
- Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6, 100048. <https://doi.org/10.1016/j.nlp.2023.100048>
- Khan, M. S., & Umer, H. (2024). ChatGPT in finance: Applications, challenges, and solutions [Publisher: Elsevier]. *Heliyon*, 10 (2). <https://doi.org/10.1016/j.heliyon.2024.e24890>

- Ko, H., & Lee, J. (2023). Can ChatGPT Improve Investment Decision? From a Port- folio Management Perspective. *Finance Research Letters*, 64. <https://doi.org/10.1016/j.frl.2024.105433>
- Lopez-Lira, A., & Tang, Y. (2024, September 11). Can ChatGPT forecast stock price movements? return predictability and large language models. <https://doi.org/10.48550/arXiv.2304.07619>
- Lu, F., Huang, L., & Li, S. (2023). ChatGPT, Generative AI, and Investment Advisory. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4519182>
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7, 77–91. <https://doi.org/10.2307/2975974>
- Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024, June 15). A survey of large language models for financial applications: Progress, prospects and challenges. <https://doi.org/10.48550/arXiv.2406.11903>
- Pelster, M., & Val, J. (2024). Can ChatGPT assist in picking stocks? *Finance Research Letters*, 59, 104786. <https://doi.org/10.1016/j.frl.2023.104786>
- Potter, S., & Greifeld, K. (2023). We Asked ChatGPT to Make a Market-Beating ETF. Here's What Happened [Accessed: October 18, 2023]. <https://www.bloomberg.com/news/articles/2023-01-26/we-asked-chatgpt-to-make-a-market-beating-etf-here-s-the-result>.
- Romanko, O., Narayan, A., & Kwon, R. H. (2023). ChatGPT-based Investment Portfolio Selection [arXiv: 2308.06260 [q-fin.PM]].
- Yeo, W. J., Van Der Heever, W., Mao, R., Cambria, E., Satapathy, R., & Mengaldo, G. (2025). A comprehensive review on financial explainable AI. *Artificial Intelligence Review*, 58 (6), 189. <https://doi.org/10.1007/s10462-024-11077-7>
- Yeung, T. (2023a). Can AI Pick Stocks? ChatGPT Says Yes. (And Gives 13 Favorites!) [Accessed: October 18, 2023]. <https://investorplace.com/2023/04/can-ai-pick-stocks-chatgpt-says-yes-and-gives-13-favorites>.
- Yeung, T. (2023b). I 'Tricked' ChatGPT Into Picking Stocks to Buy. Here's What It Said. [Accessed: October 18, 2023]. <https://investorplace.com/2023/02/i-tricked-chatgpt-into-picking-stocks-to-buy-heres-what-it-said>.

**Disclaimer:** All statements, viewpoints, and data featured in the publications are exclusively those of the individual author(s) and contributor(s), not of MFI and/or its editor(s). MFI and/or the editor(s) absolve themselves of any liability for harm to individuals or property that might arise from any concepts, methods, instructions, or products mentioned in the content.