



## GPT's idea of stock factors

Yuhan Cheng & Ke Tang

To cite this article: Yuhan Cheng & Ke Tang (2024) GPT's idea of stock factors, Quantitative Finance, 24:9, 1301-1326, DOI: [10.1080/14697688.2024.2318220](https://doi.org/10.1080/14697688.2024.2318220)

To link to this article: <https://doi.org/10.1080/14697688.2024.2318220>



Published online: 05 Mar 2024.



Submit your article to this journal [↗](#)



Article views: 2016



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

# GPT's idea of stock factors

YUHAN CHENG <sup>†‡</sup> and KE TANG <sup>†‡\*</sup>

<sup>†</sup>School of Management, Shandong University, Jinan, China

<sup>‡</sup>Institute of Economics, Tsinghua University, Beijing, China

(Received 21 September 2023; accepted 8 February 2024; published online 5 March 2024)

We amalgamate the capabilities of the GPT-4 computational model with the avant-garde methodology of autonomous factor generation, culminating in the synthesis of high-return factors within the equity investment milieu. Empirical outcomes elucidate that the factors conceptualized by ChatGPT attain a commendable Sharpe ratio peaking at 4.49, accompanied by an annualized return reaching 66.16%. Notably, the superlative excess returns garnered remain unaccounted for by the quintessential five-factor model. Through the implementation of an unembellished model averaging paradigm, the ensemble of 35 factors, conceived by ChatGPT, manifests an apex long-short annualized return of 88% and a Sharpe ratio registering at 2.46. In stark contrast to conventional data mining techniques, the temporal expenditure requisite for GPT's factor generation is minuscule. It relies on knowledge inference without the need for data input, and it can provide a thorough economic explanation for its factors.

**Keywords:** LLM; Asset pricing; Automated factor generating; ChatGPT; Knowledge inference

**JEL Classifications:** G11, G12, G17

## 1. Introduction

Factor utilization in financial markets has its origins in the 1960s, stemming from the seminal work of the Capital Asset Pricing Model (CAPM) presented by Sharpe (1964) in 'Capital asset prices: A theory of market equilibrium under conditions of risk'. Within this framework, risk was singularly attributed to the market factor. This paradigm was subsequently expanded upon by Fama and French (1992), who incorporated two additional factors: size and book-to-market value. Their construct, known as the Fama-French three-factor model, has since become a cornerstone in the annals of quantitative finance. As the depth and breadth of financial market data grew, buttressed by advancements in computational prowess, there ensued a proliferation of proposed factor models aimed at elucidating the intricacies of asset returns. Notable inclusions to this factor catalog have been momentum, volatility, quality, and investment style. For example, the seminal work of Carhart (1997) augmented the existing paradigm with a momentum factor, whereas Fama and French (2015) introduced a five factor model.

In contemporary financial research, the ascendancy of machine learning has imbued the factor selection process with a newfound complexity. Numerous investigations have been oriented towards harnessing the potential of machine

learning, both in factor genesis and selection. This has birthed methodologies such as deep learning for factor synthesis and reinforcement learning for factor curations. Notably, Kanter and Veeramachaneni (2015) presented an avant-garde algorithm, Deep Feature Synthesis, adept at autonomously gleaning features from intricate relational datasets. In a parallel vein, Lam *et al.* (2017) unveiled the One Button Machine, a system designed to replicate key data scientist functions, notably in automatic feature discovery. Further innovations were observed with Fang *et al.* (2019), who propounded the Alpha Discovery Neural Network – a neural architecture primed for crafting diverse financial indicators. This was furthered by Fang *et al.* (2020) and their introduction of NNAFC, a neural tool amalgamating financial acumen with neural networks, with the overarching aim of refining investment strategies.

In their stride towards further innovation, Zhang *et al.* (2020b) showcased the AutoAlpha evolutionary algorithm. This methodology, while recognizing patterns within trading data, offers a structured search mechanism adept at expeditiously exploring potential search space domains. The PCA-QD approach, nested within this, refines the search parameters by mitigating redundancy and amplifying optimization. Diverging slightly, Zhang *et al.* (2020a) introduced the DoubleEnsemble framework, employing learning trajectory re-weighting coupled with feature selection ensembles, thereby facilitating the identification of salient samples and features. Lastly, Cui *et al.* (2021) advanced the discourse

\*Corresponding author. Email: [ketang@tsinghua.edu.cn](mailto:ketang@tsinghua.edu.cn)

with AlphaEvolve, an AutoML-infused Alpha mining apparatus. Unique to this is its capacity to meld relational domain expertise, paving the way for discerning intricate inter-stock relationships.

In recent financial research endeavors, the FETCH framework proposed by Li *et al.* (2022) has garnered attention for its dual capabilities in feature generation and selection. Particularly salient is its prowess in facilitating the seamless transfer of feature engineering across heterogeneous datasets. Complementing this, Zhang *et al.* (2023b) introduced OpenFE, an automated feature generation framework, which, when empirically tested on a myriad of benchmark datasets, exhibited competencies analogous to those of seasoned machine learning experts. In a novel approach, Yu *et al.* (2023a) posits a model rooted in reinforcement learning methodologies, advocating for a comprehensive exploration of the search space. Eschewing traditional filtration techniques, this innovative paradigm customizes the creation of factors to optimize performance metrics.

The foundation for many contemporary factor generation methodologies is anchored in data mining. Here, a plethora of factors are stochastically generated, with only the apex ones being retained. Notwithstanding their empirical successes, the adoption of such techniques in the economic landscape has faced critiques, primarily attributed to the absence of a cogent economic rationale. Furthermore, the robustness of out-of-sample predictions of factors derived via data mining remains a topic of contention. In juxtaposition to these pure data mining methodologies, ChatGPT possesses the capability to elucidate these factors, rendering comprehensive insights underpinned by economic theories. Distinctly divergent from exhaustive exploration and traditional mining, the factors manifested by our method and GPT are the synthesis of financial scholarship, interdisciplinary knowledge integration, and mathematical rigor. A noteworthy observation is the expedited factor generation process facilitated by ChatGPT compared to traditional paradigms. This is attributed to ChatGPT's reliance on knowledge-based inference rather than data-intensive approaches, empowering it to construct factor calculation algorithms without the obligatory financial data.

In the contemporary research milieu, there has been an unmistakable ascendancy of Large Language Models (LLMs) such as ChatGPT. These linguistic behemoths have demonstrated unparalleled competencies across multifarious tasks, prompting scholars to explore their applicability in financial domains. In a seminal contribution, Araci (2019) delineated FinBERT—a financial lexicon-oriented model inspired by the BERT architecture. It set new benchmarks in financial sentiment analysis. Jiang *et al.* (2022) harnessed LLMs in forecasting financial returns, adeptly gleaning insights from a corpus of financial news articles. Breaking traditional bounds, Wu *et al.* (2023) introduced BloombergGPT, tailored specifically for financial tasks, registering remarkable proficiency.

Transitioning to the realm of personal finance, Lakkaraju *et al.* (2023) delineated 13 cardinal queries by LLM, encapsulating a spectrum of banking dimensions, from accounts to credit instruments, and extending to temporal deposits. This exploration also encompassed multifaceted interactions, such as high-value acquisitions and financial advisory engagements. Albeit their rigorous efforts, the resultant insights were

somewhat muted. Yu *et al.* (2023b) highlighted the intricacies and impediments in leveraging machine learning for financial chronologies, shedding light on challenges such as data cross-referencing and historical data integration. Yet, their research underscored the latent potential of LLMs in decoding financial news timelines and asset pricing dynamics.

In a magisterial exposition, Xie *et al.* (2023) unveiled PIXIU—a confluence of FinMA (enhanced with LLaMA), a pioneering financial LLM, a curated fine-tuning dataset, and a multifarious evaluation benchmark. On a parallel note, Zhang *et al.* (2023a) presented a meticulous backtesting paradigm tailored to the nuances of the Chinese financial ecosystem. They evaluated a cohort of LLMs, including ChatGPT, for their aptitude in extracting sentiment indicators from Chinese news abstracts. The revelations ignited discourses on the underpinnings of LLM efficiency in sentiment analysis. Shah and Chava (2023) orchestrated a comparative study of ChatGPT against other generative LLMs, illuminating nuances in data annotation, performance deltas, and the broader implications of generative models in finance. Their observations accentuated that while ChatGPT exhibited resilience even without labeled data, models with fine-tuning typically demonstrated superior outcomes. Culminating the discourse, Son *et al.* (2023) delineated the vast potentialities of LLMs in financial applications, emphasizing the burgeoning prowess of astute financial inferencing, particularly when armed with refined instructional paradigms and voluminous data reservoirs.

A similar study is the work of Wang *et al.* (2023). They introduce a groundbreaking approach to alpha mining in quantitative investment research, emphasizing the integration of human-AI interaction and a novel prompt engineering algorithmic framework. This new paradigm, embodied in the Alpha-GPT system, leverages large language models to creatively and effectively translate the ideas of quant researchers into actionable trading signals, demonstrating significant advancements over traditional methods through a series of alpha mining experiments. Our research, while in the same realm of leveraging AI for alpha generation, diverges significantly in its approach and focus. Unlike Alpha-GPT, our methodology primarily relies on the autonomous capabilities of GPT-4, without initial human-generated ideas. We provide the AI with data and prompts, guiding it to generate novel and complex alpha factors independently. This method brings forth the AI's ability to creatively utilize vast amounts of information and identify patterns that might be overlooked by human researchers. Our work stands out in its attempt to push the boundaries of AI's role in alpha generation, exploring how it can not only assist but also lead the process of uncovering valuable market insights.

This study pioneers the fusion of auto-generating factors with the sophisticated GPT-4 large language model, culminating in the generation of return-maximizing factors specific to the domain of stock investment. The procedural elegance lies in its simplicity: we provide ChatGPT with concise information regarding our data structure, encompassing variables such as opening prices, closing prices, and trading volume. This input serves as a directive for ChatGPT to architect novel factors and concurrently generate the requisite computational code. The code generated by GPT can be directly executed

in Python to derive profitable investment strategies. Human experts are essentially limited to the task of copying the code from GPT to their local Python terminal, negating the need for any manual intervention.

The empirical results bear testimony to the exceptional caliber of factors contrived by ChatGPT. A salient observation is that over fifty percent of these factors engender alphas that remain inexplicable by both the three-factor and five-factor models. Remarkably, the apex factor yielded an annualized return rate of 66.16%. A considerable proportion of these factors manifest significant alpha magnitudes, eclipsing those formulated by human experts anchored in financial scholarship. Furthermore, they markedly supersede factors extrapolated through data mining techniques. This underscores the unparalleled efficacy and latent potential of ChatGPT in the realm of factor generation, paving the way for novel paradigms in stock investment.

Both static and dynamic multi-factor investment portfolios curated with these factors demonstrate robust returns and statistically noteworthy alphas. This attests to the general efficacy of GPT-4 generated factors, corroborating their capability to enhance investment returns.

In encapsulation, the present research marks a watershed moment in asset pricing academia. By adroitly intertwining GPT-4 with the art of factor generation, we have birthed factors of unparalleled quality. This augments the repertoire of tools available to investors and scholars, proffering an astute and efficient avenue for factor discernment and origination, thus expanding the horizon for stock investment pursuits.

The distinctiveness and innovative attributes of this investigation can be delineated as follows:

- (1) **Efficacy in Factor Generation:** The GPT model showcases outstanding prowess in factor generation, producing factors with impressive return metrics. Such remarkable performance positions the GPT as an invaluable tool, equipping academicians and financiers with advanced methodologies for factor identification and creation.
- (2) **The GPT model presents a groundbreaking paradigm shift in research methodologies.** While traditional asset pricing approaches demand extensive human intervention, often characterized by factor ideation, laborious iterations and validation cycles that span prolonged periods, the GPT model dramatically trims this timeline. Harnessing its autonomous capabilities, GPT can conceive and validate factors in a fraction of the time typically required by human experts. This rapidity not only revolutionizes the transition from factor ideation to practical implementation but also empowers scholars to focus more on advanced topics, such as refining trading algorithms, instead of being mired in the time-consuming intricacies of factor discovery.
- (3) **Paradigmatic Shift in Asset Pricing Research:** Our proposition to incorporate the GPT model into the tapestry of asset pricing scholarship heralds a radical departure from entrenched research archetypes. Predominant empirical asset pricing frameworks are anchored in a symbiotic fusion of human sagacity and financial erudition for factor postulation and its

subsequent corroboration. The GPT model's advent, with its capacity to emulate extensive manual research hours in a temporally compressed manner, suggests an impending tectonic shift. This indicates that through the automated feature generation and selection facilitated by GPT, one can expediently discern novel efficacious factors, thereby accelerating advancements in the domain of asset pricing research.

- (4) **Revealing a New Technical Approach to Factor Research:** The facility to engage in discourse with ChatGPT and swiftly tap into its expansive financial knowledge repository positions it as a game-changer. By doing so, researchers can expeditiously derive predictive pricing factors, virtually obviating temporal inefficiencies inherent to human-centric factor ideation. This modus operandi is poised to redefine the fabric of factor research, catalyzing paradigmatic evolutions in both scholarly and industrial milieus.

This methodology represents a marked departure from extant investigations that juxtapose sizable models with the domain of finance. Whereas antecedent scholarship was predominantly circumscribed to sentiment analysis, the present exposition accentuates the expansive financial acumen inherent to the GPT, spotlighting its superlative capacity for factor origination. The salient advantages encompass:

- (1) **Streamlined Execution with Optimal Efficiency:** This method allows for swift deployment, reducing computational delays and saving significant time for modeling engineers. While there may be concerns about the GPT's interface response latency and potential operational costs, this approach minimizes the need for excessive data input, keeping input concise with no more than 20 000 tokens.
- (2) **Expedited Deployment and Enhanced Efficacy:** This Strategy engenders rapid deployment, mitigating computational latencies and economizing on modeling engineers' temporal commitments. Albeit the interface response latency of GPT might be perceived as protracted and incurring operational expenses, the algorithm circumvents inundating the model with copious data, restricting interactions to dialogues not exceeding 20 000 lexemes. This enterprise primarily enjoins researchers to architect the backtesting and visualization subroutines, with the factor computational logic being directly bequeathed by the GPT.
- (3) **Elevated Economic Significance:** Factors engendered pursuant to this methodology possess pronounced economic salience. Each factor is delineated with a lucid mathematical expression, complemented by an executable Python script conferred by the GPT. Furthermore, GPT provides a detailed description of the economic logic behind each factor, thereby streamlining the journey for scholars to extract economic insights.

To encapsulate, the integration of GPT in the realms of factor generation and asset pricing scholarship not only culminates in premium factor returns but also orchestrates a dramatic abatement in human latencies. This augments research efficacy, propelling asset pricing paradigms towards

heightened efficiency and sagacity. The resultant innovations enhance the robustness and acumen of investment adjudications and the refinement of investment strategies.

2. Process of GPT’s factor generation

In this study, we position ChatGPT as a surrogate financial savant, tasking it with the origination of unprecedented features (or factors) poised to elucidate prospective stock returns. To bolster the external validity of our findings while tempering computational overheads, ChatGPT remains privy solely to the schema of data storage, remaining oblivious to the substantive data therein. To obviate potential ambiguities inherent in textual delineations or mathematical notations proffered by ChatGPT, we solicit the direct extrapolation of the computational algorithm pertaining to the factors from ChatGPT. Subsequent to this, scholars are enjoined to transplant this computational algorithm onto their local computational environments for the actual factor derivation.

We refrain from commissioning ChatGPT for the actual computation of factor magnitudes. Contrarily, our mandate for ChatGPT is to elucidate the computational modality germane to the factor. Broadly, the factor magnitude vector, encompassing all stocks at temporal juncture  $t$ , is formulated as a function contingent on market idiosyncrasies spanning temporal juncture  $t$  and its antecedents:  $z_{it} = f(X_t, X_{t-1}, \dots)$ . Here,  $X_t$  epitomizes the cross-sectional data at the temporal juncture  $t$ . ChatGPT, with an acumen of the dataframe storage of  $(X_t, X_{t-1}, \dots)$ , discerns its columnar designations and the semantics inherent to each column’s contents. Anticipations are vested in ChatGPT to proffer a potentially lucrative function  $f$  drawn from its expansive knowledge reservoir, ensuring the innovativeness of  $f$ . To further insulate against the peril of ChatGPT reverting to factors already chronicled in the extant literature, we impose a stipulation for the manifestation of a multifaceted factor. As a further safeguard, a meticulous manual scrutiny of the extrapolated factor is undertaken to affirm its novelty.

Segueing into the particulars of our methodology, we shepherd GPT in an incremental fashion to extrapolate the computational regimen for a nascent factor. This procedural schema is graphically delineated in the form of a flowchart, as captured in figure 1.

Primarily, invoking prompts within ChatGPT emerges as an indispensable procedure. To commission factor generation from GPT-4, the model is instructed to conceptualize itself as an investment fund manager tasked with the origination of novel factors. The available data comprises stock names, date, opening and closing prices, as well as the highest and lowest prices, and trading volume. The prompt we used was:

‘You are a fund manager, and I need you to brainstorm new factors for me. You have panel data for many stocks stored as a dataframe named “data”. I require you to provide me with a Python program that can generate a new column as a factor. Your factor should be original and innovative. Think creatively. The data columns include stockname, date, open, high, low, close, volume, and amount. You don’t need to name the factor. Ensuring originality is more important’.

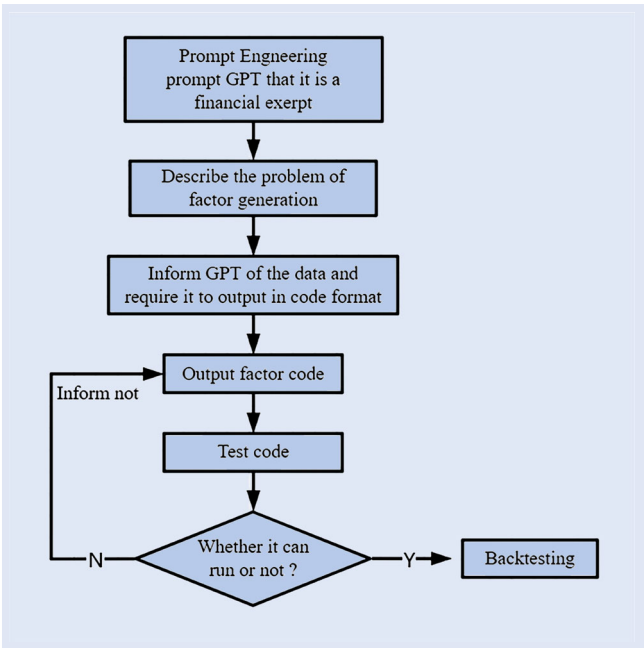


Figure 1. Guiding ChatGPT in factor generation. This figure illustrates how we directed GPT to produce useful factors and implemented them through code.

In our test, upon receiving this instruction, GPT proposed a factor called ‘Normalized Price Oscillation’, which was not highly original. Therefore, we further prompted, ‘please create a more complex and original factor’.

Following this directive, GPT-4 elucidated a multifaceted factor and detailed its computational mechanism:

- (1) Volatility-Momentum Oscillator (VMO): We first calculate the 5-day momentum, which is the difference between the current closing price and the closing price 5 days ago. We then divide this by the 5-day standard deviation of the close prices to normalize for volatility.
- (2) Volume Delta to Volatility (VDV): For each day, calculate the change in volume from the previous day (Volume Delta). We then divide this by the 5-day standard deviation of the volume to normalize for volatility in the volume.
- (3) Factor: Multiply the VMO and VDV together. The resulting factor is a complex measure that incorporates price momentum, price volatility, changes in volume, and volatility of volume.

Post this elucidation, ChatGPT proffered the correlative computational algorithm.

For the inception of a successive factor, a mere prompt stating ‘please create another one’ suffices, with GPT persistently delivering new factors.

It warrants mention that all factors emanating from GPT are predicated solely on the primordial data columns: stock name, date, open, high, low, close, and volume. Nonetheless, GPT’s code may occasionally betray inconsistencies. Certain instances witness the generation of a composite factor, where a component may adopt a misaligned data structure. Specifically, there might be a computation of a value for each stock, rather than a daily value. In sporadic cases, GPT’s code might



be riddled with errors, either aborting execution or engendering an unending loop. Operationalizing this, factors returning sparse data points are jettisoned. For factors marred by coding discrepancies, feedback is dispatched, highlighting the error, often accompanied by the aberrant code. This serves to galvanize GPT into rectifying and reissuing the correct code. For code instances culminating in infinite values owing to zero denominators, an infinitesimal adjustment is incorporated (e.g. 0.001 for price-aligned denominators or 10 for volume-correlated ones).

If experimenters are not satisfied with the complexity of the factors generated, they can guide GPT to create more complex factors by further prompting with ‘please create a more complex factor’. It should be noted that we cannot guarantee that GPT’s factor generation process is the same each time; the above is just an example from our experiments. In our subsequent robustness tests, we continued to follow the same prompting method in newly opened windows. In multiple experimental results, GPT consistently generated a relatively simple factor on the first attempt, and a more complex, original factor after the enhanced prompts. Experimenters can use different prompts according to their needs.

### 3. Factors and data

Leveraging the computational prowess of GPT-4, we synthesized a cohort of factors, subsequently isolating 35 that flawlessly integrated within the coding framework. Conscious of the ‘forward-looking’ bias and the temporal constraints of GPT’s training data, which is delimited at September 2021, our backtesting of main result employed data from October 2021 to December 2022, sourced from the CRSP’s U.S. equity market dataset. In order to further assess the effects of the factors, we conducted a backtesting analysis over the period from 2003 to 2022.

## 4. Empirical results

### 4.1. Single factor portfolio

Our empirical methodology comprised a daily rebalanced long-short factor return analysis from October 2021 through December 2022. In the purview of a long-only modus operandi, equity investments were channeled towards stocks manifesting factor values in the upper 50th percentile. Conversely, for the long-short paradigm, long allocations were directed towards the top 50% of stocks by factor value, juxtaposed with short positions in the lower 50%. In the main results, we use the signs of long-short portfolios from January 2000 to September 2021 to determine the trading direction of the factor, and use data after October 2021 for out-of-sample testing. To be specific, a performance assessment of each factor was executed spanning 2000 to September 2021. Factors demonstrating positive performance retained their original values during the factor return analysis, while negatively performing factors underwent value inversion prior to the analysis after October 2021. The out-of-sample performance

Table 1. Performance of factors: long-short portfolio.

Factor	Annualized Return	Sharpe Ratio	Max Drawdown
factor1	0.078983	0.78658	0.067298
factor2	0.13694	1.3725	0.041848
factor3	0.16409	1.6346	0.060855
factor4	0.41127	3.2263	0.06438
factor5	0.091482	0.96611	0.059092
factor6	−0.15742	−1.8293	0.23157
factor7	0.066023	0.61637	0.067433
factor8	0.14235	1.5274	0.037539
factor9	0.04314	0.32961	0.075877
factor10	0.028006	0.17803	0.11063
factor11	0.22168	1.3942	0.085803
factor12	0.052999	0.39502	0.10849
factor13	0.25392	2.5068	0.044595
factor14	0.10636	0.77598	0.070162
factor15	−0.071784	−1.0284	0.16745
factor16	0.05146	0.46792	0.082357
factor17	0.16762	1.6377	0.047413
factor18	−0.12006	−1.4221	0.18121
factor19	0.19857	1.801	0.066468
factor20	0.16044	1.0433	0.053437
factor21	0.17613	1.0999	0.11092
factor22	0.17991	1.8579	0.040127
factor23	0.11402	0.67006	0.12132
factor24	0.55347	4.4913	0.087487
factor25	0.19581	1.9152	0.039251
factor26	−0.0036311	−0.16865	0.09649
factor27	0.35218	2.819	0.065002
factor28	0.6616	4.4491	0.13316
factor29	0.035456	0.19077	0.10019
factor30	0.4428	3.9695	0.034246
factor31	0.37547	2.8379	0.088788
factor32	0.10318	0.92899	0.067779
factor33	−0.071541	−0.54518	0.22742
factor34	0.036353	0.18361	0.139
factor35	0.43463	3.9651	0.055648

Notes: The table displays the returns of factors generated by GPT when implemented in long-short portfolios. A long-short portfolio means going long on stocks ranked in the top 50% based on their factor values and shorting those in the bottom 50%. Our factor return test period is from October 2021 to the end of 2022. We use the positive and negative performance of long-short total returns for each factor from 2000 to September 2021 as the reference. If a factor’s return is negative from 2000 to September 2021, we take the opposite of the factor values for the following factor return testing; otherwise, we conduct the testing using the original factor values.

is comprehensively elucidated in table 1. Under the long-short schema, merely 5 factors registered negative returns, while the lion’s share of 30 factors championed positive returns. Factor28 emerged as the epitome of performance, realizing an astoundingly robust annualized return of 66.16%, closely trailed by factor24, boasting 55.347%. Factors 4, 30, and 35 also underscored their potency with returns transcending 40%. Cumulatively, the high-performing echelon of factors is characterized by 9 entities exceeding 20% in annualized returns, while 21 surpassed the 10% threshold.

Examining this through the lens of the Sharpe ratio, a predominantly direct correlation between return magnitude and Sharpe ratios is evident, albeit with outliers like factor11. The pinnacle of Sharpe ratios is represented by factor24,

registering 4.4913, narrowly eclipsing factor28. Notably, this signifies factor24's marginally diminished volatility, a claim substantiated by its maximum drawdown rates. Factors exhibiting returns beyond the 40% benchmark consistently flaunted Sharpe ratios beyond 3.

In terms of drawdown metrics, a substantial 24 factors were characterized by drawdown rates beneath 10%. Of these, only 6 factors navigating the positive return space surpassed the 10% drawdown threshold. It's pivotal to highlight the peak drawdown rate, which, for the high performing cadre, was capped at 13.9%. Factor30 was the paragon of resilience, with a minuscule drawdown rate of 3.4246%, thereby bolstering its Sharpe credentials. Several other factors also manifested commendable drawdown resistance, cementing their robustness.

The long-only strategy is dissected in table 2, unveiling the annual returns, Sharpe ratios, and peak drawdowns corresponding to each factor. A striking observation is the ubiquitously positive returns across all factors. Factor6 manifests the least potent performance, with an annual return of just 8.7699%, ironically contriving negative returns under a long-short modus operandi. All other contenders transcend the 10% annualized return mark. Factor28 ascends as the performance titan, charting an annualized return of 53.899%. Nipping at its heels are factor4 and factor24, showcasing returns of 44.842% and 44.324%, respectively. A high-performing league emerges with seven factors surmounting 30% in annual returns. This elite cadre, in addition to the aforementioned, envelops factors 27, 30, 31, and 35. Notably, factors 30 and 35 vault beyond the 40% echelon. Zooming out, 24 factors within the long-only architecture eclipse the 20% annualized return threshold.

Navigating through the terrain of risk-adjusted returns, Factor28 remains unchallenged with its Sharpe ratio north of 1.9. Trailblazing behind are factors 30, 24, and 4, each brandishing sharpe ratios soaring beyond 1.6. Barring factors that languish in the negative under the long-short framework, every other factor musters a Sharpe ratio above 0.7, with a laudable 18 breaching the 1.0 landmark.

Juxtaposing the long-only and long-short strategies reveals a predilection of the former towards elevated drawdowns. Factor28, while leading in returns and Sharpe metrics, also bears the brunt of the highest drawdown, peaking at 27.278% among its positively returning peers in the long-short realm. In stark contrast, factor20 emerges as the paragon of resilience, restricting its drawdown to a modest 17.993%.

A subsequent layer of nuance is introduced by evaluating alpha returns, showcased in table 3. The analytical spectrum bifurcates into two temporal windows: an extended span from 2003 to 2022, using 2000–2002 returns as a directional reference, and a narrower window mirroring our earlier tests, anchored to 2000–September 2021 reference returns, spanning October 2021 through December 2022.

The tableau reflects compelling alpha generation for the September 2021–December 2022 stretch. Nine factors – 4, 13, 21, 24, 27, 28, 30, 31, and 35 – indomitably stand out, their alphas achieving statistical significance at the stringent 1% threshold. Bar factor13 and factor21, the cohort, under the long-only prism, breaches the 30% return barrier. We also discern alphas of factors 19, 22, and 25 attaining significance

Table 2. Performance of factors: long-only portfolio.

Factor	Annualized Return	Sharpe Ratio	Max Drawdown
factor1	0.20479	0.89204	0.20155
factor2	0.23443	1.0099	0.20684
factor3	0.24808	1.0758	0.18583
factor4	0.44842	1.6328	0.26959
factor5	0.21132	0.91828	0.20943
factor6	0.087699	0.37461	0.23565
factor7	0.20013	0.85673	0.22782
factor8	0.24238	1.0312	0.2245
factor9	0.18744	0.80099	0.23814
factor10	0.17976	0.76965	0.25036
factor11	0.2772	1.0771	0.23132
factor12	0.19183	0.80766	0.26495
factor13	0.29238	1.278	0.18683
factor14	0.21933	0.89947	0.22586
factor15	0.12955	0.56631	0.22303
factor16	0.19352	0.83108	0.24896
factor17	0.24964	1.1161	0.24227
factor18	0.10531	0.41937	0.25588
factor19	0.2671	1.0763	0.23943
factor20	0.24549	1.4548	0.17993
factor21	0.25458	0.88686	0.26787
factor22	0.25659	1.1862	0.22937
factor23	0.22314	0.83267	0.26979
factor24	0.44324	1.8314	0.24462
factor25	0.26593	1.1539	0.19595
factor26	0.16408	0.73379	0.25523
factor27	0.34354	1.3997	0.22706
factor28	0.53899	1.9339	0.27278
factor29	0.18552	0.77443	0.25873
factor30	0.38943	1.8877	0.184
factor31	0.35378	1.3401	0.24271
factor32	0.21809	1.0455	0.23394
factor33	0.13028	0.4832	0.29923
factor34	0.18621	0.76978	0.23569
factor35	0.38338	1.5258	0.23638

This table shows the returns of factors created by GPT in long-only portfolios. For this strategy, long-only involves choosing stocks from the top 50% based on their factor values. Our factor return testing period spans from October 2021 to the end of 2022. We use the positive or negative sign of the long-short total returns of each factor from January 2000 to September 2021 as a reference. If a factor has a negative returns during this period, we conduct the factor return testing by taking the opposite of the factor values. Conversely, if a factor has a positive return during this period, we conduct the testing using the original factor values.

at the 5% level, and a spate of factors – 2, 3, 6, 8, 11, 17, and 22 – securing their place at the 10% significance echelon. Factor28 is particularly stellar, parading an alpha of 0.74355, emblematic of its prowess in consistently outmaneuvering market returns. Post value inversion, its long-short strategy return of 66.16% vastly outpaces its peers, consonant with its high alpha. Factor24's alpha prowess is also palpable at 0.61935, harmonizing with its formidable 55.347% return under the long-short schema. Factors 4, 30, 31, and 35 also engrave their alpha supremacy, each registering above 0.4, reflecting their long-short strategy returns that comfortably exceed 37%.

Notably, among the top five factors in terms of alpha values—factor28, 24, 35, 31, and 30—their annualized returns under daily rebalancing in the long-short strategy consistently

Table 3. Alphas of factors.

Factor	short term	long term	Factor	short term	long term
factor1	0.081366 (1.05)	0.037995** (2.38)	factor19	0.22016** (2.50)	0.078177*** (4.37)
factor2	0.14555* (1.75)	0.048151*** (3.08)	factor20	0.064678 (0.86)	0.15804*** (9.43)
factor3	0.16033* (1.89)	0.074474*** (4.59)	factor21	0.24133*** (2.96)	−0.0021094 (−0.13)
factor4	0.41524*** (3.80)	0.11757*** (4.63)	factor22	0.13958* (1.77)	−0.0058011 (−0.38)
factor5	0.10443 (1.39)	0.055657*** (3.60)	factor23	0.13558 (1.08)	0.23501*** (9.00)
factor6	−0.15809* (−1.92)	0.0085058 (0.56)	factor24	0.61935*** (6.51)	0.35665*** (17.97)
factor7	0.052289 (0.65)	0.026176* (1.68)	factor25	0.19768** (2.26)	0.038568** (2.18)
factor8	0.13276* (1.74)	0.050085*** (3.27)	factor26	−0.050981 (−0.65)	0.025383 (1.52)
factor9	0.030934 (0.35)	0.068305*** (4.02)	factor27	0.36163*** (3.40)	0.20942*** (10.04)
factor10	0.019998 (0.24)	0.041312** (2.43)	factor28	0.74355*** (6.80)	0.39464*** (17.96)
factor11	0.22122* (1.71)	0.23327*** (8.84)	factor29	0.025872 (0.23)	0.22233*** (10.11)
factor12	0.05269 (0.56)	0.068138*** (3.47)	factor30	0.42559*** (4.83)	0.20825*** (11.87)
factor13	0.27217*** (3.13)	0.060105*** (3.50)	factor31	0.43864*** (4.52)	0.14078*** (7.56)
factor14	0.11712 (1.09)	0.20045*** (9.36)	factor32	0.037085 (0.47)	−0.012 (−0.75)
factor15	−0.070954 (−0.98)	−0.0065781 (−0.44)	factor33	−0.062677 (−0.48)	0.17787*** (6.97)
factor16	0.065698 (0.85)	−0.043872*** (−2.65)	factor34	−0.0051649 (−0.04)	0.11937*** (4.88)
factor17	0.16279* (1.89)	0.053537*** (3.22)	factor35	0.48099*** (5.62)	0.17422*** (11.15)
factor18	−0.11913 (−1.45)	0.019631 (1.31)			

Notes: This table depicts the alpha of each GPT-generated factor within long-short portfolios. We present the portfolio alphas for two distinct time periods. 'Short term' refers to the alphas of long-short factor portfolios from October 2021 to the end of 2022, with signs determined by the positive or negative sign of each factor's return from 2000 to September 2021. 'long term' represents the alphas of long-short factor portfolios from 2003 to the end of 2022, with signs determined by the positive or negative sign of each factor's returns from 2000 to 2022. Statistical significance is denoted as: \*\*\* for 1% significance, \*\* for 5% significance, and \* for 10% significance.

exceed those in the long-only strategy, further highlighting the strong predictive capabilities of these factors. Among the factors with statistical significance at the 10% level, factor8 holds the lowest alpha value, yet it still reaches 0.13636. In absolute terms, only 6 factors exhibit three-factor and five-factor alphas below 0.5.

The alphas obtained from testing spanning from 2003 to 2022 are even more prominent. Notably, only 5 factors exhibit negative alphas, while a total of 23 factors display positive alphas that are statistically significant at the 1% level. Additionally, 3 factors exhibit alphas that are statistically significant at the 5% level, and 1 factor demonstrates significance at the 10% level. Factor28 maintains its position with the highest alpha at 0.39464, followed by factor24 at 0.35665. Factor11, 14, 23, 27, 29, and 30 also exhibit alphas exceeding 0.2. This demonstrates the enduring predictive power of the factors generated by GPT, as evidenced by their ability to consistently deliver significant market outperformance over a 20-year backtesting period.

#### 4.2. Multi-factor portfolio investment

Among the quintessential factors delineated by their alpha magnitudes—namely factor28, 24, 35, 31, and 30—their annualized returns under a daily rebalancing regime in the long-short portfolio consistently outstrip those accrued under the long-only framework. This accentuates the potent prognosticative prowess of these specified factors. Delving into factors with a statistical significance demarcated at the 10% threshold, factor8 registers the most diminutive alpha magnitude, albeit clocking in at 0.13636. In absolute metrics, a mere sextet of factors manifest three-factor and five-factor alphas that fall below the 0.5 demarcation.

The alphas extrapolated from the temporal analysis encompassing 2003 to 2022 are even more conspicuous. A paucity of factors, totaling five, exhibit negative alphas. In stark contrast, a compendium of 23 factors unveil positive alphas achieving statistical significance at the rigorous 1% standard. Further dissection reveals 3 factors with alphas significant at the 5% criterion, while a singular factor exhibits significance at the



10% marker. Factor28 perpetuates its preeminence with an alpha cresting at 0.39464, with factor24 in proximate pursuit at 0.35665. Additional constituents, namely factor11, 14, 23, 27, 29, and 30, demonstrate alphas transcending the 0.2 threshold. This underscores the sustained predictive efficacy of the factors engineered by GPT, as corroborated by their consistent capability to outpace the market over two decades of backtesting.

Beyond the rudiments of aggregation, an elementary dynamic factor aggregation strategies was also probed. The only deviation between dynamic and static aggregations hinges on the historical return dictating a factor's polarity. This elementary algorithm culminated in a model return of 0.37593.

For a comprehensive assessment of the efficacy of factors molded by GPT-4, we embarked on backtesting multi-factor investments anchored to historical datasets. On a diurnal basis, we normalized factor magnitudes based on antecedent data and forged both long-only and long-short portfolios grounded in the aggregation of these normalized values. The normalization leveraged was the Min-Max normalization method, explicated as:

$$\frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Initially, a static multi-factor portfolio was postulated. Explicitly, the polarity of each factor—either positive or negative—was deduced from their long-short returns spanning 2000 to September 2021. Subsequent aggregation returned  $average_{t,stock}$ . The term *sign* epitomizes the vector that ascertains the positive or negative delineation of each factor.  $sign_i$  resonates with the polarity of the returns for the  $i$ th factor from 2000 to September 2021, oscillating between values of 1 or  $-1$ . The computation is elucidated as:

$$average_{t,stock} = \frac{1}{n} \sum_{i=1}^n factor_{t,stock}^i \times sign_i \quad (2)$$

Where  $factor_{t,stock}$  symbolizes the factor magnitude of the stock on day  $t$ , and  $factor_{t,stock}^i$  designates the  $i$ th factor. Given the ensemble of 35 factors,  $n$  is consequently set at 35.

Following the articulated procedure for multi-factor portfolio configuration and its subsequent daily recalibration, the return dynamics are encapsulated in table 4.

With respect to the long-only portfolio, it registered an annual return of 45.161%, complemented by a Sharpe ratio of 1.6295. When juxtaposed against single-factor portfolios, the cumulative return of the multi-factor ensemble may marginally lag behind the apex performer, factor 28. However, it eclipses the returns of all other individual factors. The maximum drawdown metric, recorded at 25.439%, aligns closely with the majority of the factors when assessed within the paradigm of long-only strategies.

Transitioning to the long-short landscape, the annualized return underwent a decrement to 39.728% relative to its long-only counterpart. Yet, the Sharpe ratio soared to an impressive 2.4854, with the maximum drawdown receding to a mere 7.1795%. Such metrics insinuate that despite the attenuation in absolute returns relative to the long-only spectrum,

Table 4. Multi-factor portfolio returns.

Static Multi-Factor Portfolio Returns			
Strategy	Annualized return	Sharpe ratio	Max drawdown
long-only	0.45161	1.6295	0.25439
long-short	0.39728	2.4854	0.071795
Dynamic Multi-Factor Portfolio Returns - Referenced by long-only Returns			
Strategy	Return	Sharpe ratio	Max drawdown
long-only	0.34044	1.3822	0.22944
long-short	0.37593	3.0931	0.061103
Dynamic Multi-Factor Portfolio Returns - Referenced by Long-Short Returns			
Strategy	Return	Sharpe ratio	Max drawdown
long-only	0.30262	1.2644	0.29177
long-short	0.30036	2.4478	0.12983

Notes: The table demonstrates the returns from static and dynamic multi-factor portfolios, both for long-only and long-short combinations, using factors generated by GPT. For the static method, we determine the positive or negative sign of each factor based on their returns from 2000 to September 2021, then calculate the average of these signs to create a new factor for testing. For the dynamic approach, the sign of the factor values is adjusted daily based on the cumulative positive or negative returns of either the long-only or long-short strategies associated with each factor.

the affiliated risk was substantially curtailed. Long-only portfolios inherently exhibit heightened susceptibility to market vicissitudes, paving the way for potential substantial capital erosions or appreciations during specific temporal windows. Conversely, long-short portfolios adeptly neutralize the capriciousness engendered by market fluctuations.

Our investigative trajectory further embraced an uncomplicated modus operandi for dynamically modulating the polarity of factor magnitudes. Explicitly, subsequent to the daily normalization of factor magnitudes predicated on their historical performance employing the Min-Max algorithm, the polarity (affirmative or negative) of every factor was anchored to the cumulative historical return up until the specified day.

To elucidate, assuming  $return_t$  embodies the return vector for every discrete factor on day  $t$ , and  $factor_{t,stock}$  epitomizes the Min-Max standardized factor magnitude of a specific stock on day  $t$ , we initially ascribe a polarity to the factor magnitude predicated on its return, culminating in the recalibrated factor magnitude,  $newfactor_{t,stock}$ :

$$newfactor_{t,stock} = factor_{t,stock} \times sign(return_t) \quad (3)$$

Subsequent to this, the diurnal multi-factor mean for the stock is deduced as:

$$average_{t,stock} = \frac{1}{n} \sum_{i=1}^n newfactor_{t,stock}^i, \quad (4)$$

with  $newfactor_{t,stock}^i$  representing the  $i$ th factor magnitude. This mean,  $average_{t,stock}$ , is then harnessed as the novel factor for portfolio orchestration.

In our empirical evaluations, we deployed both the individual factor returns stemming from daily recalibrated long-only and long-short portfolios as benchmarks for polarity determination. The resultant metrics are also encapsulated in table 4.

A discerning observation in a dynamically modulated setting, benchmarked against the long-only returns of individual factors, unveils that the annualized return for the long-only Strategy reached 0.34044. The long-short Strategy, in contrast, touched a commendable 0.37593. Notably, both these values markedly transcend those observed in the static configuration.

The analytical insights gleaned from the long-short returns unveil that the circumvention of certain market nuances endows the GPT-4 generated factor with a capacity to consistently realize stable and appreciable returns. The dynamism in factor value adjustment, which astutely accounts for sporadic market inversions, does dual service—mitigating the risks associated with reversals while capitalizing on potential profit avenues. Such dynamics further underscore the potency and utility of the GPT-derived factor in portfolio management.

Drawing a parallel between employing the daily rebalanced long-short portfolio return and the long-only returns as the yardstick for discerning the polarity of return outcomes, the dynamism in returns under the former seems less accentuated. The plausible reason for this attenuation could be that long-short returns, to a degree, fall short in echoing the undulating market currents. This results in a tardy capture of the factor's inversion signals, rendering the factor value adjustments less agile and efficient. Yet, the annualized return for the portfolios—both long-only and long-short—stands at 0.30262 and 0.30036 respectively. Paired with Sharpe ratios of 1.2644 and 2.4478, these figures denote tangible advancements over their static counterparts.

Delving deeper, the alphas computed for both static and dynamic scenarios are encapsulated in table 5. With all alphas consistently surpassing the 0.3 mark, the alphas pertaining to static investment and dynamic investment (with long-only data as a reference) both exhibit 1% statistical significance. Contrarily, the alpha for dynamic investment, referenced against long-short data, manifests a 5% statistical significance—potentially skewed by some preliminary setbacks. In a three-factor regression context, the zenith of alpha is seen for dynamic investments using daily long-only data, registering at 0.38067, while the static counterpart is close on its heels at 0.37762. Conversely, in a five-factor regression landscape, the alpha pinnacle is associated with static investments, recording 0.37034, while the dynamic counterpart referencing daily long-only data posts 0.36583—a decrement roughly amounting to 0.015 vis-à-vis the three-factor backdrop.

The numerically robust and statistically significant alphas further embellish the overarching credibility of the GPT-4 derived factor. Such empirical evidence attests to the proposition that investment frameworks pivoted on GPT-4 generated factors are poised to accrue elevated returns.

Table 5. Alphas of multi-factor strategies.

Factor	3factor-days	5factor-days
static	0.41809*** (3.01)	0.40985*** (2.97)
dynamic <sub>long-only base</sub>	0.38067*** (2.97)	0.36583*** (2.86)
dynamic <sub>longshort base</sub>	0.32770** (2.40)	0.32256** (2.38)

Notes: The table presents the alpha of static and dynamic multi-factor portfolios using factors generated by GPT in a long-short combination. 'long-only base' refers to using the factor's long-only returns as a positive or negative reference, and 'long-short base' means referencing the factor's long-short returns. \*\*\* denotes 1% statistical significance, \*\* denotes 5% statistical significance, and \* indicates 10% statistical significance.

For both the static and dynamically calibrated multi-factor portfolios, an additional layer of analysis was undertaken. Here, portfolios were architecturally designed by ranking factor magnitudes in descending order and then uniformly allocating weights to the top decile. The associated annualized return, the Sharpe ratio, and the maximum drawdown metrics are shown in table 6.

In these tabulations, the return differential between the quintile with the apex factor magnitude and its antithetical counterpart with the nadir factor magnitude is showcased. This essentially reflects the strategy of adopting a long-only stance on the upper decile stocks as per factor magnitudes while simultaneously shorting the bottom decile. Remarkably, under the static paradigm, this long-short portfolio configuration records an astounding annualized return of 0.88097. Transitioning to the dynamic framework, referenced against long-only returns, the return metric dips to 0.41481. However, a dynamic strategy based on long-short returns registers an enhanced annualized return of 0.69870. Furthermore, the Sharpe ratio for the quintile housing stocks with the zenith factor magnitude consistently overshadows its counterpart with the nadir factor magnitude by a margin exceeding a unit. In the static construct, the long-short portfolio's Sharpe ratio, clocking 2.4622, testifies to its stellar performance credentials.

Overall, equities characterized by elevated factor values tend to manifest superior returns and Sharpe ratios relative to their counterparts with diminished factor values. Notably, the annualized return does not exhibit a strict inversely proportional relationship with diminishing factor values, with group 2 serving as a prominent outlier. The quintile labeled as 'High' consistently boasts the pinnacle Sharpe ratio, registering 2.0391 under the dynamic investment modus operandi predicated on long-short returns. This translates to an annualized return of 0.73425, marking an approximate surplus of 0.163 over group 9, which posts an annualized return of 0.57093. A singular deviation from this trend is observed in the dynamic investment strategy referenced against long-only returns; here, the 'High' quintile's annualized return stands at 0.50977, trailing both group 9's 0.52126 and group 8's 0.613. Nevertheless, this figure substantially eclipses the returns of other quintiles. In the static paradigm, the annual return of

Table 6. Multi-factor portfolio returns - 10 bins.

Static Multi-Factor Portfolio Returns - 10 Bins			
Group	Annualized return	Sharpe ratio	Max drawdown
Low	0.90252	2.2864	0.30424
2	0.74673	2.0861	0.22777
3	0.20703	0.65923	0.28248
4	0.1914	0.58776	0.23074
5	0.21317	0.72800	0.26801
6	0.18109	0.53500	0.34179
7	-0.10458	-0.48012	0.31533
8	-0.09856	-0.46569	0.34026
9	0.27494	0.90091	0.21767
High	0.021541	0.042082	0.34141
H-L	0.88097	2.4622	0.1335
Dynamic Multi-Factor Portfolio Returns - Referenced by long-only Returns - 10 Bins			
Group	Annualized return	Sharpe ratio	Max drawdown
Low	0.094964	0.33951	0.39546
2	0.20934	0.68301	0.19487
3	-0.036588	-0.19603	0.3092
4	0.067688	0.19416	0.32577
5	0.011829	0.0011659	0.30439
6	0.34726	1.1524	0.21421
7	0.20533	0.63581	0.25718
8	0.61300	1.7134	0.22975
9	0.52126	1.5392	0.26295
High	0.50977	1.5949	0.27727
H-L	0.41481	1.4839	0.20772
Dynamic Multi-Factor Portfolio Returns - Referenced by Long-Short Returns - 10 Bins			
Group	Annualized return	Sharpe ratio	Max drawdown
Low	0.035556	0.090087	0.37699
2	0.23402	0.77437	0.17019
3	0.017148	0.021575	0.20962
4	-0.017581	-0.11558	0.23718
5	0.22608	0.6264	0.30014
6	0.22253	0.7258	0.27082
7	0.18375	0.56628	0.3168
8	0.34612	1.1166	0.30783
9	0.57093	1.6698	0.29316
High	0.73425	2.0391	0.29997
H-L	0.69870	2.0195	0.14967

Notes: The table displays the annualized returns, Sharpe ratio, and maximum drawdowns derived from both static and dynamic multi-factor portfolios based on factors generated by GPT. All stocks are categorized into ten groups based on the computed mean factor values, with a long-only strategy applied to each group. Specifically, the static approach first adjust the signs of factors according to the their long-short returns from 2000 to September 2022, while the dynamic method adjusts the sign of the factor values daily based on the cumulative returns' positive or negative direction of each factor for either long-only or long-short positions.

the 'High' quintile is gauged at 0.66492, markedly surpassing other cohorts, with group 9 recording 0.57613 and the remainder falling below the 0.4 threshold.

The empirical outcomes lend credence to the proposition that factors engineered by GPT are adept at returning

remarkable returns. These findings insinuate that such factors harbor the capacity for versatile integration into a diverse array of investment portfolio structures. It can be rationally posited that these economically cogent and efficacious factors, sculpted by GPT, are poised to consistently proffer optimized returns in pragmatic investment landscapes.

## 5. How GPT constructs the aforementioned factors

### 5.1. Principles and fundamental logic of GPT's factor generation

In seeking to decipher the methodology through which GPT operationalizes the generation of these factors, our inquisitiveness led us to request GPT to decompose its factor generation process. It articulated its factor formulation protocol in the subsequent manner:

- (1) Understanding the Requirements and Data: First, it's crucial to know the data we are working with (e.g. OHLCV data) and the requirements, such as the type of stocks, trading frequency, investment horizon, etc.
- (2) Exploring Existing Indicators: There's a rich library of existing technical and fundamental indicators in finance. Many factors are based on these well-known indicators, such as Moving Averages, Bollinger Bands, MACD, etc.
- (3) Combining Indicators: By combining different indicators, we can create complex factors that encapsulate different aspects of the market, such as trend, momentum, volume, and volatility.
- (4) Applying Financial Theories: Sometimes, the creation of factors is based on specific financial theories or investment philosophies. For example, some factors may focus on identifying value stocks, growth stocks, or market inefficiencies.
- (5) Experimenting with Mathematical Transformations: Applying mathematical or statistical transformations to price, volume, or other data can uncover hidden patterns or normalize the data in a useful way.
- (6) Grouping and Rolling Operations: Utilizing grouping by assets and applying rolling computations allows us to calculate factors in a way that respects the time-series nature of the data.
- (7) Iterative Testing and Refinement: Creating a factor often requires several iterations of testing, refinement, and validation. The effectiveness of a factor can be assessed through back-testing, out-of-sample testing, and statistical analysis.
- (8) Incorporating Domain Knowledge: Sometimes, unique insights from a particular industry, market conditions, or company specifics can inspire the creation of a novel factor.
- (9) Ensuring Practicality: Factors need to be computable from available data, robust to noise, and applicable in real-world trading scenarios.

The process through which GPT conjures these remarkably prognosticative factors remains an intricate conundrum



in financial economics. Nonetheless, endeavoring to delineate the operational intricacies underpinning the intricate outputs emanating from such an expansive computational model poses substantial challenges. ChatGPT orchestrates its computations via a conspicuously nonlinear neural network, and given that the GPT-4.0 iteration, which is the focal point of our study, is fortified with in excess of a hundred billion parameters, endeavors at interpretative dissection are inherently intricate. Providentially, the interactive capability of ChatGPT accords us the privilege to solicit its elucidation on the factor construction methodology. Through our exhaustive dialogues, we managed to crystallize three pivotal tenets.

*Applying Underlying Principles:* GPT uses the underlying mathematical and statistical properties of the indicators to analyze how they might combine meaningfully. For example, if one indicator measures trend and another measures volatility, combining them might provide insight into the momentum of a trending market.

*Considering Market Context:* Based on historical data and industry practices, GPT can make educated guesses on how certain indicators might behave together in different market conditions.

*Applying General Analytical Techniques:* GPT can apply standard mathematical and statistical techniques, such as normalization, weighting, and statistical correlations, to combine indicators in a logically coherent way.

Based on its self-interpretation, GPT benefits immensely from a vast Market Context and knowledge about the underlying semantics and logic of the factors. It can rationally combine foundational factors based on its learning of Underlying Principles. This combination adheres to the application of underlying principles and general Analytical Techniques while being inspired by its vast Market Context.

This insight is simultaneously intuitive and revelatory. It underscores ChatGPT's prowess as a formidable financial analyst, furnished with an extensive repository of fiscal acumen. Its capability to autonomously synthesize and amalgamate this vast expanse of financial information without fatigue, and at an inconsequential expense, foregrounds the transformative implications of expansive linguistic models for both the financial sector and scholarly endeavors.

ChatGPT elucidates: 'In the examples I provided, I was combining well-known technical indicators in different ways, sometimes including mathematical transformations and statistical measures, to create new factors. These combinations are often guided by intuition about what might be an effective signal (e.g. combining trend and volume information), but they would typically be followed by extensive empirical testing to confirm their effectiveness in practice'.

To encapsulate, through the adroit exploitation of established technical indicators and a profound application of financial intricacies, integrating mathematical operations—like computing standard deviations and averages—and by evaluating market prices, oscillations, trends, and inversion cues, GPT meticulously engineers market-optimized factors with the potential to procure substantial returns.

## 5.2. Basic factors and strategies referenced by GPT

To deeply probe the nexus between the factors manifested by GPT and canonical financial literature as well as pragmatic fiscal experiences, we embarked on an analytical journey scrutinizing the rudimentary operations and factors that GPT leverages. Predominantly, its foundational operations draw inspiration from the numpy and pandas libraries, embracing rudimentary functions such as groupby, rolling, std, kurt, while also extending its arm into the realm of sophisticated functionalities like Exponential Moving Average. As GPT elucidates, its elemental factors are meticulously curated from Financial Literature and Known Strategies.

Each factor GPT has unveiled is inherently a composite entity. Elucidated within table 7 are the elemental factors GPT leverages to synthesize these composite structures in the 35 factors we used, accompanied by their comprehensive explanations. Each factor emblematically encapsulates treatments of various stock metrics such as open price, close price, highs, lows, and volumes—these treatments range from calculating standard deviations, momentum to rate of change. Within these factors, one can discern traditional market technical indicators, epitomized by the likes of MACD and daily averages. Some components, however, elude direct correspondence to prevalent factors within extant databases, emerging more as GPT's ingenious improvisations. For instance, the factor12 element, christened as Trading Range Expansion (TRE), gauges the differential between the current trading band and its counterpart from two days hence. Furthermore, the Momentum and Intraday Range Interaction (MIRI) employed in Factor 5 is not a widely recognized financial metric or technical term, and it is not readily searchable on the internet. To calculate MIRI, one must first compute the 5-day and 10-day momentum of closing prices. Subsequently, the average intraday price range (defined as the difference between the highest and lowest prices) over the past 5 days and 10 days is determined. MIRI is defined as the product of the 5-day momentum and the 5-day average range divided by the product of the 10-day momentum and the 10-day average range. Table 8 presents the corresponding constituent factors used in each of the 35 factors.

Contrasting these factors with ones forged purely from data mining or stochastic genesis, it's evident that GPT's creations are underpinned by robust economic logic and are more semantically transparent. They transcend mere data dependency, thus potentiating them as instruments of potentially lucrative investment returns.

At a macroscopic lens, GPT's sophisticated factors gravitate towards capturing market volatilities, price trajectories, and momentum. A significant majority of these composite factors bear at least one foundational element which acutely addresses price gyrations or trade volume oscillations. GPT's meticulousness extends to recognizing reversal cues, frequently employing indicators such as the Commodity Channel Index (CCI), price reversal (PR), and the Parabolic Stop and Reverse (PSAR)—all hallmarks of reversal prognosticators. It often marries the Volume Acceleration (VA) and Directional Movement Index (DMI) to discern price trends, augmenting the predictive prowess of the factor. Out of the

Table 7. Description of constituent factors used in the 35 complex factors we used created by GPT.

Factor	Description
Volatility-Momentum Oscillator (VMO)	First calculate the 5-day momentum, which is the difference between the current closing price and the closing price 5 days ago. Then divide this by the 5-day standard deviation of the close prices to normalize for volatility.
Volume Delta to Volatility (VDV)	For each day, calculate the change in volume from the previous day (Volume Delta). We then divide this by the 5-day standard deviation of the volume to normalize for volatility in the volume.
Volatility and momentum interaction (VMI)	First, compute the 7-day and 14-day momentum of the closing prices. Then, divide the 7-day momentum by the 14-day standard deviation of the close prices, and the 14-day momentum by the 7-day standard deviation. Multiply these two resulting factors together.
Volume and time decay (VTD)	Compute the difference between today's volume and the volume from 3 days ago. Then, apply an exponential decay to this difference, where the decay factor decreases as we go back in time.
Relative Price Oscillator (RPO)	Calculate the 10-day and 20-day simple moving averages of the closing price. Then, subtract the 10-day from the 20-day and divide the result by the 20-day average to get the oscillator.
Volume Change Rate (VCR)	This is calculated as the rate of change of a 5-day simple moving average of the volume.
Discrete Fourier Transform (DFT)	Compute the magnitude of the DFT of the close price over a 30-day window.
Mean Reverting Scaled Oscillator (MRSO)	This factor takes into consideration the mean-reverting characteristic of financial data. Create an oscillator using the 7-day and 21-day moving averages and scale it using the standard deviation over the past 21 days.
Volume Impact (VI)	Calculate the volume change compared to the average volume over the past 10 days. Then, apply a sigmoid function to this ratio to keep the impact within a reasonable range.
Autocorrelation Measure (AM)(2)	This measure will capture the autocorrelation of the closing prices over the past 5 days.
Momentum and Intraday Range Interaction (MIRI)	Calculate the 5-day and 10-day momentum of the closing prices. Then, compute the intraday price range (high - low) and average it over the past 5 and 10 days. The MIRI is the 5-day momentum times the 5-day average range divided by (the 10-day momentum times the 10-day average range).
Volume and Temporal Decay (VTD)	Calculate the difference between today's volume and the volume from 3 days ago. Then, apply a linear decay factor to this difference, which reduces the weight of older volume data.
Weighted Close Price (WCP)	Calculate this as (close * volume) / volume. It gives a sense of how much the closing price is 'supported' by the volume.
Volatility Impact (VI)	We calculate the standard deviation of the past 10 days' closing prices as a measure of volatility. Then, we apply a logarithm to this measure to limit its impact.
Price-Volume Trend (PVT)	measures the cumulative volume-adjusted price change. It is calculated as the cumulative sum of the product of the price change rate and volume change rate.
Normalized Intraday Volatility (NIV)	This is the standard deviation of the intraday returns (close-to-close) over the past 10 days, normalized by the 20-day moving average of the closing prices.
Trading Range Divergence (TRD)	This is the divergence between the current intraday range (high - low) and its 5-day average, normalized by the 5-day average.
Volume Weighted Change (VWC)	Multiply the percentage change in price with the percentage change in volume.
Range Rate of Change (RRC)	Calculate this as the percentage change in the daily range (high - low).
Intraday Volatility Scaling (IVS)	Calculate this as the daily range (high - low) divided by the standard deviation of the daily range over the past 14 days.
Volume Acceleration (VA)	Calculate the percentage change in volume compared to 3 days ago.
Historical Price Movement (HPM)	Calculate the percentage change in the current closing price compared to the closing price 10 days ago.
Momentum Score (MS)	Calculate the current close price minus the close price 5 days ago.
Volume Spike (VS)	Calculate the percentage change in volume over the past 3 days.
Trading Range Fluctuation (TRF)	Compute the ratio of the current day's trading range (high - low) to the average trading range over the past 10 days.
Trend Score (TS)	Calculate the difference between the close price and its 10-day moving average. This score helps us detect the recent price trend.
Volume Dynamics (VD)	Calculate the ratio of the current volume to the 5-day moving average volume. A significant increase or decrease in this ratio indicates possible changes in trading interest.
Volatility Index (VI)	Calculate the standard deviation of the daily returns (close-to-close percentage change) over the past 20 days. Higher values indicate greater volatility.
Price-Volume Trend (PVT)	This measures the cumulative volume-adjusted price change. It is calculated as the cumulative sum of the product of price change rate and volume change rate.
Normalized Intraday Volatility (NIV)	This is the standard deviation of the intraday returns (close-to-close) over the past 10 days, normalized by the 20-day moving average of the closing prices.
Trading Range Divergence (TRD)	This is the divergence between the current intraday range (high - low) and its 5-day average, normalized by the 5-day average.
Volume Weighted Change (VWC)	Multiply the percentage change in price with the percentage change in volume.
Range Rate of Change (RRC)	Calculate this as the percentage change in the daily range (high - low).

(Continued).



Table 7. Continued.

Factor	Description
Intraday Volatility Scaling (IVS)	Calculate this as the daily range (high - low) divided by the standard deviation of the daily range over the past 14 days.
Volume Acceleration (VA)	Calculate the percentage change in volume compared to 3 days ago.
Historical Price Movement (HPM)	Calculate the percentage change in the current closing price compared to the closing price 10 days ago.
Momentum Score (MS)	Calculate the current close price minus the close price 5 days ago.
Volume Spike (VS)	Calculate the percentage change in volume over the past 3 days.
Trading Range Fluctuation (TRF)	Compute the ratio of the current day's trading range (high - low) to the average trading range over the past 10 days.
Trend Score (TS)	Calculate the difference between the close price and its 10-day moving average. This score helps us detect the recent price trend.
Volume Dynamics (VD)	Calculate the ratio of the current volume to the 5-day moving average volume. A significant increase or decrease in this ratio indicates possible changes in trading interest.
Volatility Index (VI)	Calculate the standard deviation of the daily returns (close-to-close percentage change) over the past 20 days. Higher values indicate greater volatility.
Price Acceleration (PA)	This is the difference between the current price change rate (current close price to previous close price) and the price change rate from 2 days ago (close price from 2 days ago to 3 days ago). A higher PA value might indicate an acceleration in price movement.
Trading Range Expansion (TRE)	This is the difference between the current trading range (high - low) and the trading range from 2 days ago. An increase in this value might signal an expansion in the trading range.
Price Reversal (PR)	Compute the difference between the current close price and the highest close price in the past 7 days. A negative value might indicate a potential price reversal.
Volume Rate of Change (VRC)(2)	Calculate the percentage change in volume over the past 5 days. This metric can signal changes in trading activity.
Intraday Price Volatility (IPV)	Compute the ratio of the intraday price range (high - low) to the closing price. A higher value can indicate higher intraday price volatility.
Twisted Price Momentum (TPM)	Calculate the difference between the current close price and the median close price over the past 10 days. This can help us identify unusual price changes.
Volume Variability (VV)	Compute the coefficient of variation (standard deviation divided by mean) of the volume over the past 5 days. A higher value might indicate significant changes in trading volume.
Intraday Price Efficiency (IPE)	Calculate the absolute value of the open price minus the close price divided by the day's high minus the low. A lower value indicates a more efficient intraday market.
Relative Price Change (RPC)	Compute the percentage change in closing price from the previous day divided by the percentage change in opening price from the previous day. This gives an idea about how the price is moving relative to the opening of the day.
Volume Momentum (VM)	This is the current volume divided by the maximum volume over the past 10 days. A higher value can indicate momentum in trading volume.
Daily Range Expansion (DRE)	Calculate the percentage change in the daily range (High - Low) from the previous day. This factor captures if the intraday price range is expanding or contracting.
Average True Range Rate of Change (ATRRC)	Compute the average true range (ATR) for the last 14 days and find the percentage change from the previous day. ATR is a measure of market volatility.
Standardized Closing Price (SCP)	This is the closing price standardized by the mean and standard deviation over the last 30 days. It shows how much the closing price deviates from the recent average.
Weighted Close Price (WCP)	This is the average of high, low, and double the close price for the day, giving more weight to the closing price.
Rate of Change in Volume (RCV)	This is the percentage change in volume compared to the previous day. This indicates a shift in trading activity.
Normalized Intraday Range (NIR)	This is the ratio of the difference between the day's high and low prices to the closing price, normalized by the maximum range observed over the past 10 days.
Cumulative Return (CR)	This is the running total of the daily return. It shows the overall profit or loss over a period of time.
Volume Change Momentum (VCM)	This is the difference between today's volume and the average volume over the past 5 days. This provides an indication of any sudden interest in the stock.
Daily Price Oscillation (DPO)	This is the difference between the highest and lowest prices divided by the closing price. This gives a sense of the relative volatility of the stock price during the day.
Volatility (VOL)	Compute the standard deviation of daily returns over the last 14 days. It provides a measure of risk associated with price changes.
Normalized Price Deviation (NPD)	This is the deviation of the closing price from its 30-day moving average, normalized by the 30-day standard deviation of the price. It shows how much the current price is deviating from the recent average.
Volume Adjusted Momentum (VAM)	This is the product of the 5-day momentum of the closing price and the 5-day average volume. It indicates the direction and intensity of the stock's movement.
Daily Range Percent (DRP)	This is the ratio of the difference between the day's high and low prices to the closing price. It indicates the day's volatility relative to the closing price.
Deviation from Simple Moving Average (DSMA)	This is the difference between the closing price and its 10-day simple moving average. It shows how much the current price deviates from the recent trend.
Exponential Moving Average Ratio (EMAR)	This is the ratio of the 5-day exponential moving average to the 20-day exponential moving average. It helps in identifying the short-term trend relative to the longer-term trend.

(Continued).

Table 7. Continued.

Factor	Description
Standard Deviation of Volume (SDV)(2)	Compute the standard deviation of volume over the last 14 days. It provides a measure of volatility in trading volume.
Price Range Breakout (PRB)	This is the closing price minus the maximum high price of the last 10 days. It shows if the stock is breaking out of its recent range.
Relative Strength Index (RSI)(3)	This is a momentum oscillator that measures the speed and change of price movements. It ranges from 0 to 100 and is often used to identify overbought and oversold conditions.
On Balance Volume (OBV) Ratio	This is the ratio of the difference between today's OBV and the previous day's OBV to the total OBV. It measures cumulative buying and selling pressure by adding the volume on up days and subtracting the volume on down days.
Price Change Acceleration (PCA)	This is the second derivative of the price, which measures how quickly the price change is accelerating or decelerating.
Rate of Change (ROC)(5)	This is a momentum-based technical indicator that measures the percentage change in price between the current price and the price a certain number of periods ago.
Accumulation/Distribution Line (ADL)	This is a volume-based indicator designed to measure the cumulative flow of money into and out of a security. It can help to identify buying or selling pressure.
Commodity Channel Index (CCI)(3)	This is a technical indicator that measures the difference between the current price and the historical average price.
Bollinger Band Width (BBW)	This is a technical analysis indicator derived from the standard Bollinger Bands. BBW is calculated by subtracting the lower band from the upper band and then dividing the result by the middle band.
Volume-price Trend (VPT)	This is a cumulative indicator that integrates volume and price data to depict the direction of a trend.
Detrended Price Oscillator (DPO)	This is an indicator in technical analysis that seeks to eliminate the long-term trends in prices by using a displaced moving average so it does not react to the most current price action.
Parabolic Stop and Reverse (PSAR)(2)	This is a method proposed by J. Welles Wilder Jr. to find potential reversals in the market price direction.
Williams %R(2)	This is a momentum indicator that measures overbought/oversold levels. It moves between 0 and -100, and it is typically used to identify buying or selling signals.
Triple Exponential Moving Average (TEMA)	It is a technical indicator that applies an increased weight to recent data points, making it more reactive to price changes.
Average True Range (ATR)(2)	This is a volatility indicator introduced by Welles Wilder that measures the degree of price volatility.
Moving Average Convergence Divergence (MACD)(3)	This is a trend-following momentum indicator that shows the relationship between two moving averages of a security's price.
On Balance Volume (OBV)(2)	It is a technical trading momentum indicator that uses volume flow to predict changes in stock price.
Stochastic Oscillator	This is a momentum indicator comparing a particular closing price of a security to a range of its prices over a certain period of time.
Directional Movement Index (DMI)	This is an indicator developed by J. Welles Wilder for identifying when a definable trend is present in an instrument. That is, the DMI tells whether an instrument is trending or not.
Chaikin Money Flow (CMF)	This is an oscillator that measures the amount of Money Flow Volume over a specific period. Money Flow Volume forms the basis for the Accumulation Distribution Line. Instead of a cumulative total of Money Flow Volume, Chaikin Money Flow simply sums Money Flow Volume for a specific look-back period, typically 20 or 21 days.
Ultimate Oscillator	This is a technical indicator invented by Larry Williams that uses the weighted average of three different time periods to reduce the volatility and false transaction signals that are associated with many other indicators that mainly rely on a single time period.
Average Directional Movement Index Rating (ADXR)	It measures the strength of the trend, irrespective of its direction. The trend strength is indicated by rising ADXR levels, while non-trending periods are characterized by falling ADXR.
Balance of Power (BoP)	It is a price-based indicator that estimates the balance of market power. Positive BoP values signal buyers' dominance, while negative BoP values indicate sellers are in control.
Aroon Oscillator(2)	This indicator is used to measure the strength of a trend by comparing the time it takes for the price to reach the highest and lowest points over a given time period.
Money Flow Index (MFI)	This is a momentum indicator that measures the flow of money into and out of a security over a specified period of time. It is related to the Relative Strength Index (RSI) but incorporates volume, whereas the RSI only considers price.
Chande Momentum Oscillator (CMO)	Developed by Tushar Chande, this is a technical momentum indicator similar to other momentum indicators such as the Relative Strength Index (RSI). However, the CMO aims to capture the inherent volatility of a security by oscillating between -100 and +100, with 0 as the baseline.
Bollinger Bands(2)	A technical analysis tool defined by a set of lines plotted two standard deviations (positively and negatively) away from a simple moving average (SMA) of the security's price.
Chaikin A/D Oscillator	It measures the momentum of the Accumulation Distribution Line using the MACD formula. This makes it an indicator of an indicator. The Chaikin A/D Oscillator fluctuates above/below the zero line.

Notes: The table showcases all 81 factors used in the 35 factors by GPT along-only with their explanations. Some factors share the same name but differ in calculation methods. The number in parentheses after the name of some factors indicates the total number of times the factor has been used; those without numbers have been used only once.

Table 8. The constituent factors used in each factor.

Factor	Constituent factor1	Constituent factor2	Constituent factor3	Constituent factor4
factor1	VMO	VDV		
factor2	VMI	VTD		
factor3	RPO	VCR	DFT	
factor4	ROC	ADL	CCI	
factor5	MIRI	VTD		
factor6	WCP	VI	AC	
factor7	PVT	NIV	TRD	
factor8	VWC	RRC	IVS	
factor9	RPO	VA	HPM	
factor10	MS	VS	TRF	
factor11	TS	VD	VI	
factor12	PA	TRE	VD	
factor13	PR	VRC	IPV	
factor14	TPM	VV	IPE	
factor15	RPC	VM	DRE	
factor16	PVT	ATRRC	SCP	
factor17	WCP	RCV	NIR	
factor18	CR	VCM	DPO	
factor19	VOL	VA	NPD	
factor20	VAM	DRP	DSMA	
factor21	EMAR	SDV	PRB	
factor22	RSI	OBV_ratio	PCA	
factor23	BBW	VPT	DPO	
factor24	PSAR	Williams %R	TEMA	
factor25	ATR	VROC	MACD	
factor26	OBV	ROC	Stochastic Oscillator	
factor27	RSI	CCI	ADX	
factor28	Bollinger Bands	CMF	Ultimate_Oscillator	
factor29	EMA	CCI	VPT	MFI
factor30	Williams %R	MACD	Aroon Oscillator	
factor31	ROC	MFI	CMO	
factor32	OBV	CCI	ROC	
factor33	Bollinger Bands	ATR	Momentum	
factor34	RSI	MACD	STD	
factor35	Aroon Oscillator	Chaikin A/D Oscillator	ROC	

Notes: The table presents the corresponding constituent factors used in each of the 35 factors. For detailed explanations of each constituent factor, refer to table 7.

rich tapestry of 31 factors proffered by GPT, a staggering 81 unique foundational components were identified. However, factoring in the nuances of code which rendered certain factors inoperable, this number burgeons to 105. This profound diversity underscores that, while GPT amalgamates a plethora of base factors, it seldom retreads the same pathway. Such a kaleidoscopic array is a testament to GPT's expansive financial lexicon and its proclivity for novel factorial synthesis.

### 5.3. GPT's economic analysis of representative factors

To more profoundly delineate the distinctive attributes inherent to the factors synthesized by ChatGPT, we shall dissect selected factors that exemplify potent representational capacity and unique attributes. We prompted GPT to deconstruct these factors through an economic lens. As per the insights drawn from GPT's elucidation, it unmistakably manifests a holistic comprehension of financial axioms pertinent to the devised factors, accompanied by nuanced rationales, notwithstanding the potential contention surrounding such interpretations.

Factor 4, an amalgamated entity, is engendered through the convolution of the Rate of Change (ROC), Accumulation/Distribution Line (ADL), and Commodity Channel Index (CCI). Concretely, ROC stands as a momentum-centric technical metric quantifying the percentage alteration juxtaposing the prevailing price to the antecedent price of a delineated period. It computes the rate of change grounded on a decacyclic timeline of closing price. In the economic paradigm, the rate of change serves as an instrument to gauge the tempo and magnitude of price oscillations, unmasking ephemeral market trends. Brisk price perturbations might resonate with market supply-demand dynamism or emergent contingencies. An augmented rate of change might hint at tempestuous market sentiments, potentially triggered by the dissemination of pivotal intel or unexpected market perturbations.

The Accumulation/Distribution Line (ADL) functions as a volume-anchored metric with an intent to quantify the aggregated monetary ingress and egress concerning a security. Its mathematical representation for a stock on the  $n$ th day is articulated as:

$$\sum_{t=1}^n \frac{(close_t - low_t) - (high_t - close_t)}{high_t - low_t} * volume_t \quad (5)$$

Table 9. Performance of factors from new conversation windows: long-short portfolio.

Factor	Annualized Return	Sharpe Ratio	Max Drawdown	Factor	Annualized Return	Sharpe Ratio	Max Drawdown
factor1_1	0.29779	2.8481	0.043772	factor6_1	0.048668	0.43499	0.045504
factor1_2	0.1818	1.7725	0.042725	factor6_2	−0.01459	−0.32053	0.14621
factor1_3	0.33967	3.3058	0.032042	factor6_3	0.24988	2.6	0.031042
factor1_4	0.13043	1.3207	0.052484	factor6_4	0.52389	4.6602	0.084969
factor1_5	−0.00078	−0.10601	0.16356	factor6_5	−0.10633	−1.3019	0.1406
factor2_1	−0.02414	−0.39243	0.082185	factor7_1	0.24595	1.8089	0.13214
factor2_2	0.12571	1.0673	0.057712	factor7_2	0.1997	1.3616	0.076892
factor2_3	0.1543	1.446	0.10238	factor7_3	0.25966	2.2649	0.078359
factor2_4	0.044074	0.34419	0.082259	factor7_4	0.4707	4.4871	0.071766
factor2_5	0.25978	2.6973	0.040711	factor7_5	0.75267	6.8203	0.026867
factor3_1	−0.08908	−1.257	0.14636	factor8_1	0.11401	1.0991	0.068365
factor3_2	0.49753	3.2999	0.075001	factor8_2	0.70539	7.7075	0.023888
factor3_3	0.18823	1.9664	0.041809	factor8_3	0.011082	−0.00271	0.091569
factor3_4	−0.06229	−0.70738	0.13477	factor8_4	0.42277	2.6802	0.1004
factor3_5	0.13753	1.3859	0.067603	factor8_5	0.45331	4.9563	0.024933
factor4_1	0.040718	0.26566	0.13141	factor9_1	0.32757	2.362	0.055154
factor4_2	0.46443	4.0802	0.070925	factor9_2	0.30588	3.1431	0.032327
factor4_3	0.086933	0.7477	0.10353	factor9_3	0.18357	1.595	0.047165
factor4_4	0.15554	1.3019	0.078935	factor9_4	0.43787	2.0408	0.23554
factor4_5	0.070693	0.5691	0.10335	factor9_5	0.50044	3.1423	0.10319
factor5_1	0.23756	1.9662	0.05434	factor10_1	−0.03648	−0.47319	0.11595
factor5_2	−0.01917	−0.27999	0.23972	factor10_2	0.028341	0.24124	0.042435
factor5_3	0.051866	0.3081	0.10707	factor10_3	0.13849	1.2093	0.073359
factor5_4	0.13751	1.2635	0.05946	factor10_4	0.081028	0.72757	0.063207
factor5_5	0.4323	5.0995	0.020236	factor10_5	0.24476	2.3778	0.052769

Notes: The table displays the returns of factors generated by GPT from new conversation windows when implemented in long-short portfolios. A long-short portfolio means going long on stocks ranked in the top 50% based on their factor values and shorting those in the bottom 50%. Our factor return test period is from October 2021 to the end of 2022. We use the positive and negative performance of long-short total returns for each factor from 2000 to September 2021 as the reference. If a factor's return is negative from 2000 to September 2021, we take the opposite of the factor values for the following factor return testing; otherwise, we conduct the testing using the original factor values.

It can assist in identifying the pressures to buy or sell, thereby revealing the direction of capital flows within the market. From an economic perspective, the movement of funds reflects the sentiments of market participants regarding securities. A superlative ADL delineation insinuates capital influx, potentially echoing amplified acquisition momentum, whilst an inferior ADL signifies capital egress, potentially resonating with intensified divestment momentum. Such capital maneuvers encapsulate market stakeholders' prognostications and assurance.

The Commodity Channel Index (CCI) quantitatively scrutinizes the discrepancy between the extant price and its chronological median. Within economic discourses, the CCI acts as an apparatus to ascertain whether a valuation deviates from its protracted median, pinpointing potential overbought or oversold market states. Its mathematical formulation is:

$$TP = \frac{high + low + close}{3} \quad (6)$$

$$CCI = \frac{TP - mean(TP)}{0.015 \times std(TP)} \quad (7)$$

CCI measures the extent of the current price's departure from its historical midpoint. Divergences amid the price and its chronological median might be attributed to supply-demand stasis and market appraisals. An elevated CCI metric might insinuate a notable divergence from the protracted average, potentially garnering market scrutiny and potentially heralding a market recalibration. When amalgamating these tripartite indicators, this composite metric offers perspicacious

revelations regarding the velocity and amplitude of price metamorphoses, the direction and vigor of capital streams, and the divergence of valuations from their historical benchmarks. On a day-trading echelon, should a stock exhibit an elevated value for this composite element, it might insinuate a robust positive rate of change, capital accretion, and an elevated CCI. This could signify an overbought market milieu. Retaining such equities under these circumstances might portend amplified risk quotients. In contrast, when the factor metric is minimized, signaling a pronounced negative rate of change, capital drainage, and a diminutive or adverse CCI, the market might be navigating an oversold stratum. Harnessing such equities during this phase might herald a market reversion, culminating in substantial returns. This delineation proffers a rationale for the pronounced returns concomitant with investments grounded in the inverse values of Factor 4.

Factor 13 is delineated as an amalgamation of price reversal, volume rate of change, and intraday price volatility metrics. The Price Reversal (PR) metric quantifies the disparity between the extant closing price and the zenith closing price over the preceding 7-day interval. A negative PR typically signals an imminent price reversal. The Volume Rate of Change (VRC), derived from a 5-day cyclical assessment, stands as a testament to the dynamism in trading volume, acting as a mirror to the market participants' trading predilections and liquidity fluxes. A pronounced VRC might intimate heightened market enthusiasm, potentially triggered by salient events or pivotal information, presaging potential shifts in

Table 10. Performance of factors from new conversation windows: long-only portfolio.

Factor	Annualized Return	Sharpe Ratio	Max Drawdown	Factor	Annualized Return	Sharpe Ratio	Max Drawdown
factor1_1	0.31015	1.3022	0.2177	factor6_1	0.18437	0.80607	0.21272
factor1_2	0.28204	1.1423	0.22545	factor6_2	0.15277	0.63568	0.24839
factor1_3	0.33052	1.3699	0.17863	factor6_3	0.28499	1.2683	0.21177
factor1_4	0.22862	0.95642	0.2385	factor6_4	0.42197	1.7347	0.25901
factor1_5	0.16492	0.65866	0.27313	factor6_5	0.10682	0.44558	0.28168
factor2_1	0.14804	0.66577	0.25921	factor7_1	0.28601	1.1235	0.28444
factor2_2	0.2284	0.97901	0.22665	factor7_2	0.26297	1.0952	0.19687
factor2_3	0.23717	1.0195	0.22422	factor7_3	0.30001	1.2168	0.23357
factor2_4	0.18207	0.77044	0.23718	factor7_4	0.39842	1.8968	0.22211
factor2_5	0.28995	1.287	0.21859	factor7_5	0.54079	2.582	0.15739
factor3_1	0.11551	0.4893	0.25893	factor8_1	0.22688	1.0112	0.22202
factor3_2	0.40882	1.5767	0.20574	factor8_2	0.51746	2.3207	0.17146
factor3_3	0.25414	1.1149	0.22069	factor8_3	0.17008	0.70377	0.26147
factor3_4	0.1289	0.48477	0.27931	factor8_4	0.37437	1.3904	0.23618
factor3_5	0.22875	1.0052	0.21741	factor8_5	0.38949	1.7679	0.18869
factor4_1	0.19076	0.76306	0.25726	factor9_1	0.3269	1.7324	0.18395
factor4_2	0.39832	1.5865	0.22237	factor9_2	0.31513	1.2948	0.21159
factor4_3	0.2138	0.919	0.23457	factor9_3	0.25469	1.0862	0.20194
factor4_4	0.24074	1.1069	0.19561	factor9_4	0.38341	1.2068	0.34055
factor4_5	0.19771	0.82763	0.2274	factor9_5	0.41027	1.5077	0.23934
factor5_1	0.28287	1.4281	0.19415	factor10_1	0.14446	0.62852	0.23167
factor5_2	0.15448	0.63897	0.28638	factor10_2	−0.13944	−0.65557	0.33027
factor5_3	0.18878	1.0193	0.19488	factor10_3	0.23231	0.99016	0.23343
factor5_4	0.22879	1.1041	0.17382	factor10_4	0.23227	0.95148	0.23157
factor5_5	0.38098	1.653	0.20958	factor10_5	0.30278	1.3064	0.18791

Notes: The table displays the returns of factors generated by GPT from new conversation windows when implemented in long-short portfolios. A long-short portfolio means going long on stocks ranked in the top 50% based on their factor values and shorting those in the bottom 50%. Our factor return test period is from October 2021 to the end of 2022. We use the positive and negative performance of long-short total returns for each factor from 2000 to September 2021 as the reference. If a factor's return is negative from 2000 to September 2021, we take the opposite of the factor values for the following factor return testing; otherwise, we conduct the testing using the original factor values.

price orientations. Intraday Price Volatility (IPV) calibrates the ratio between the intraday price amplitude (highest to lowest price differential) and the culminating price. A soaring IPV underscores pronounced intraday price gyrations, the genesis of which can be attributed to factors like market liquidity, trading momentum, or significant news influx. A surging IPV might encapsulate frenetic market sentiments or amplified trading fervor, opening up avenues for strategic trading maneuvers. Stocks characterized by diminished factor values might exhibit a negative PR, a pronounced and bullish VRC, and heightened IPV, hinting at a potential inflection from a bearish to a bullish trend. Thus, contrarian investment strategies tethered to the inverse of this factor's values might culminate in lucrative dividends.

Factor 21 is synthesized by weaving together the Exponential Moving Average Ratio (EMAR), Standard Deviation of Volume (SDV), and Price Range Breakout (PRB). The EMAR is articulated as a quotient of the 5-day and the 20-day exponential moving averages, serving as a barometer to assess short-term trends relative to long-term counterparts. An augmented EMAR typically signifies a robust recent price trend, whereas its diminished counterpart leans towards a prevailing long-term trend. The SDV, encapsulating the volatility of trading volume over a 14-day continuum, offers insights into trading volume dynamism. Economically construed, volume oscillations are emblematic of market stakeholders' trading zeal and engagement. An escalated SDV might be emblematic of pronounced shifts in market sentiment. The PRB metric,

illustrating the chasm between the closing price and the highest price over a 10-day cycle, acts as an indicator of price range ruptures. Such breakouts could be precursors to nascent market trends or pivotal price oscillations. Investing contrarily to the factor's orientation might proffer substantial returns, especially when accompanied by a declining PRB.

Factor 24 amalgamates the Parabolic Stop and Reverse (PSAR), Williams %R, and the Triple Exponential Moving Average (TEMA) metrics. The PSAR furnishes potential market inflection points, its placement beneath ascending prices and above descending ones. Elevated PSAR values could either be symptomatic of a robust price trend or herald impending reversals. Williams %R, a momentum-centric metric, demarcates overbought and oversold terrains, with its zenith values potentially presaging price recalibrations. The TEMA, with its predisposition to accentuate recent data, showcases heightened reactivity to price perturbations. Augmented TEMA values might resonate with robust price momentum, heralding lucrative trading prospects. A stock characterized by a soaring combined factor value might be nestled in a bearish market with burgeoning reversal potential, thereby signposting opportunistic buy signals.

Factor 27, synthesized by integrating the Relative Strength Index (RSI), Commodity Channel Index (CCI), and Directional Movement Index (DMI), provides a nuanced lens into market momentum and trends. Both the RSI and DMI proffer insights into market momentum and its trend, equipping investors with a strategic vantage point to decipher market



Table 11. Alphas of factors from new conversation windows.

Factor	short term	long term	Factor	short term	long term
factor1_1	0.30686*** (3.46)	0.11893*** (6.55)	factor6_1	0.045902 (0.59)	0.01045 (0.64)
factor1_2	0.19768** (2.29)	−0.0082992 (−0.50)	factor6_2	−0.030415 (−0.42)	−0.028965** (−2.06)
factor1_3	0.34586*** (3.97)	0.09427*** (5.53)	factor6_3	0.23369*** (2.81)	0.029834* (1.87)
factor1_4	0.13506* (1.69)	−0.012405 (−0.80)	factor6_4	0.5625*** (6.24)	0.31816*** (17.83)
factor1_5	0.014094 (0.14)	0.21141*** (10.52)	factor6_5	−0.10909 (−1.33)	0.0048027 (0.32)
factor2_1	−0.064727 (−0.83)	0.047638*** (3.08)	factor7_1	0.26119** (2.32)	−0.11278*** (−4.98)
factor2_2	0.12582 (1.31)	0.093728*** (5.02)	factor7_2	0.206* (1.65)	0.065805*** (2.85)
factor2_3	0.14975* (1.68)	0.12953*** (7.21)	factor7_3	0.2745*** (2.94)	0.080923*** (4.47)
factor2_4	0.025465 (0.30)	0.057016*** (3.42)	factor7_4	0.44248*** (5.40)	0.22039*** (13.13)
factor2_5	0.23564*** (2.89)	0.10228*** (6.51)	factor7_5	0.71688*** (8.37)	0.35131*** (20.75)
factor3_1	−0.091937 (−1.28)	−0.026149* (−1.82)	factor8_1	0.082612 (1.00)	0.081104*** (4.83)
factor3_2	0.50905*** (4.07)	0.19103*** (7.78)	factor8_2	0.68804*** (8.66)	0.23637*** (14.34)
factor3_3	0.17321** (2.12)	0.020946 (1.32)	factor8_3	−0.0068203 (−0.08)	0.054251*** (3.17)
factor3_4	−0.022603 (−0.27)	−0.090449*** (−5.71)	factor8_4	0.4521*** (3.64)	0.19747*** (7.82)
factor3_5	0.11738 (1.44)	0.01466 (0.84)	factor8_5	0.43586*** (5.54)	0.13789*** (8.63)
factor4_1	0.066104 (0.70)	0.03518* (1.79)	factor9_1	0.26632*** (3.00)	0.22155*** (12.34)
factor4_2	0.52915*** (5.99)	0.2634*** (12.42)	factor9_2	0.31841*** (4.00)	0.085818*** (5.64)
factor4_3	0.12785 (1.47)	0.060263*** (3.40)	factor9_3	0.17323* (1.77)	0.056103*** (2.87)
factor4_4	0.12949 (1.32)	0.098115*** (5.03)	factor9_4	0.61977*** (6.39)	−0.30918*** (−16.05)
factor4_5	0.048083 (0.51)	0.076921*** (4.01)	factor9_5	0.53904*** (4.41)	0.22865*** (9.00)
factor5_1	0.19799** (2.31)	0.090148*** (5.00)	factor10_1	−0.084776 (−0.94)	−0.01939 (−1.15)
factor5_2	−0.013874 (−0.14)	−0.127*** (−6.66)	factor10_2	0.074225 (1.49)	−0.0081852 (−0.93)
factor5_3	−0.015687 (−0.18)	0.0084252 (0.46)	factor10_3	0.11111 (1.17)	0.058197*** (3.24)
factor5_4	0.088154 (1.07)	−0.0051941 (−0.31)	factor10_4	0.050492 (0.59)	0.0090953 (0.55)
factor5_5	0.41609*** (5.59)	−0.15059*** (−8.96)	factor10_5	0.22147** (2.51)	0.076147*** (4.07)

Notes: This table depicts the alpha of each GPT-generated factor from new conversation windows within long-short portfolios. We present the portfolio alphas for two distinct time periods. ‘Short term’ refers to the alphas of long-short factor portfolios from October 2021 to the end of 2022, with signs determined by the positive or negative sign of each factor’s return from 2000 to September 2021. ‘long term’ represents the alphas of long-short factor portfolios from 2003 to the end of 2022, with signs determined by the positive or negative sign of each factor’s returns from 2000 to 2022. Statistical significance is denoted as: \*\*\* for 1% significance, \*\* for 5% significance, and \* for 10% significance.

oscillations. By amalgamating these metrics with the CCI, which offers a meticulous appraisal of price deviations in conjunction with the territory of overbought or oversold conditions, a more refined prognostication about potential price aberrations from their historical mean and imminent reversals becomes attainable. Elevated composite factor values, typified by surging RSI, CCI, and DMI indicators, often portend

an overbought market scenario, teetering on the brink of a substantive decline. Such conditions render stocks ripe for strategic short-selling endeavors.

Factor 35 harmonizes the intricacies of the Aroon Oscillator, Chaikin A/D Oscillator, and the Rate of Change (ROC) to offer a holistic understanding of market dynamics. The Aroon Oscillator, a trend-discerning metric, quantifies the

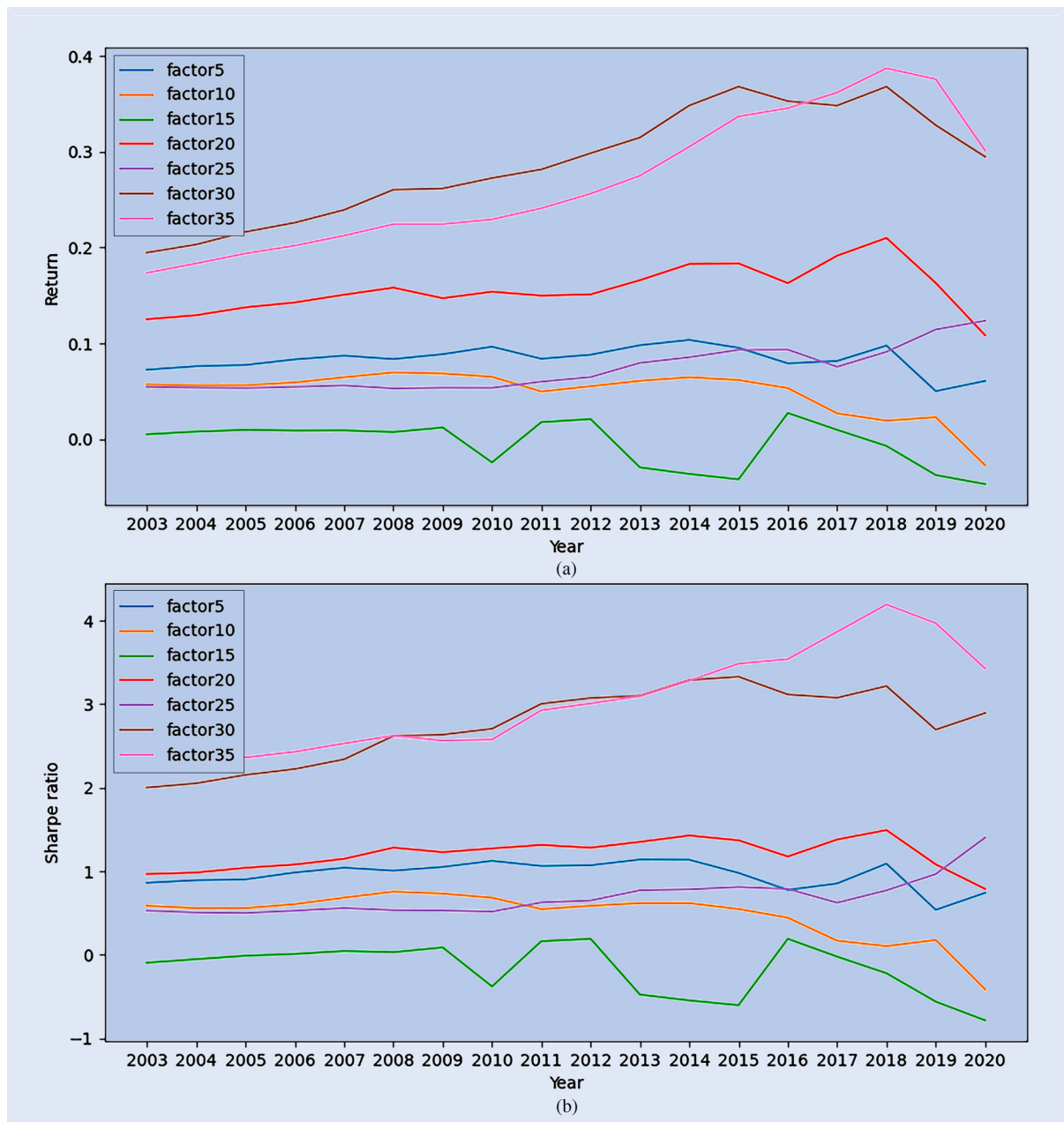


Figure 2. The change of annualized return and sharpe ratio across different test sets for seven factors. (a) Annualized Return across Different Test Sets and (b) Sharpe Ratio across Different Test Sets.

The following two charts illustrate the annualized returns and Sharpe ratios of seven factors—factor5, factor10, factor15, factor20, factor25, factor30, and factor35—across different test sets. The years on the horizontal axis represent the final year used to determine the sign of the factors. For instance, the point corresponding to the year 2010 indicates that the factor's sign is determined based on returns from 2000 to 2010, and then the factor is tested from 2011 to 2022.

vigor and potential persistence of prevailing trends by juxtaposing the Aron Up and Aron Down indicators. Concurrently, the Chaikin A/D Oscillator gauges the momentum embedded within the Accumulation/Distribution Line (ADL), serving as a barometer for market participants' sentiment and convictions regarding a particular security. The ROC, firmly ensconced in the momentum indicator echelon, calibrates the percentage differential between the contemporary prices and their historical counterparts from a predetermined temporal horizon. This confluence of indicators facilitates a panoramic evaluation of price fluctuations against the backdrop of overarching market forces. In scenarios where the composite factor value is ascendant, market

perceptions of the stock gravitate towards optimism, bolstered by formidable buying and selling impetuses. Such conditions often hint at a propitious milieu for initiating long positions on securities exemplifying elevated Factor 35 values.

When constructing new factors, ChatGPT does not rely on data mining but deduces based on its knowledge base. ChatGPT is explicitly aware of the mathematical meaning and financial logic behind each factor it writes. In light of the foregoing analysis, it becomes unequivocally manifest that ChatGPT possesses an astute grasp of intricate financial forecasting paradigms and is adept at crafting factors rooted in this nuanced learning.

Table 12. The number of factors with statistically significant alphas in different time period.

Time Period	1%	5%	10%	Sum
2004–2022	26	4	0	30
2005–2022	25	4	1	30
2006–2022	24	4	2	30
2007–2022	24	3	2	29
2008–2022	24	3	2	29
2009–2022	25	1	1	27
2010–2022	24	2	1	27
2011–2022	23	1	3	27
2012–2022	21	2	4	27
2013–2022	20	4	3	27
2014–2022	20	5	3	28
2015–2022	20	4	5	29
2016–2022	20	5	2	27
2017–2022	18	5	3	26
2018–2022	18	4	1	23
2019–2022	19	4	0	23
2020–2022	16	5	1	22
2021–2022	11	4	6	21

Notes: This table depicts the number of factors with statistically significant alphas within long-short portfolios in different time period. 1%, 5%, 10% represents 1% statistically significant, 5% statistically significant, and 10% statistically significant, respectively.

## 6. Robustness test

### 6.1. Average effectiveness of factors generated by GPT

Considering that ChatGPT is well known to have a ‘temperature’, implying that each run is uniquely different from another, we add 10 new conversation windows. In each window, we prompt GPT-4 to generate 5 factors that are correct in code and devoid of future information, resulting in a total of 50 factors. This is done to assess the average effectiveness of specific prompts on ChatGPT’s answers. For these additional factors, we continue to test them using the two methods previously mentioned. The first method is a short-term test, where the positivity or negativity of the factors is determined based on the returns from 2000 to September 2021, followed by empirical testing from October 2021 to December 2022. The second method is a long-term test, determining the sign of the factors based on returns from 2000 to 2002, and then conducting long-short empirical tests from 2003 to 2022. Tables 9 and 10 present the results of the long-short and long-only combinations in the short-term test, as well as the alphas for both short-term long-short and long-term long-short strategies. Our results show that, in the short-term test, 42 out of 50 factors have positive long-short returns, and 49 factors have positive long-only returns. Among these, under the long-short combination, the 5th factor of the 7th window (factor7\_5) achieves an annualized return of 0.75267, and the annualized return of factor8\_2 is also high at 0.70539. Additionally, 2 other factors have annualized returns greater than 0.5, specifically factor6\_4 and factor9\_5. Eleven factors, which constitute 20% of the total factors, have annualized returns exceeding 0.4. From the perspective of the Sharpe ratio, these factors also perform exceptionally well, with 33 factors in the short-term long-short test having a Sharpe ratio greater than 1, 11 of which were above 3, and 7 above 4.

Table 13. Performance of factors generated by GPT given simple prompts.

Long-short Portfolio			
Factor	Annualized Return	Sharpe Ratio	Max Drawdown
simple1	0.2805	1.7973	0.080651
simple2	0.15747	1.701	0.037159
simple3	0.45713	2.1984	0.21354
simple4	0.50055	2.4568	0.21737
simple5	0.024609	0.14688	0.077793
simple6	−0.00314	−0.17381	0.092351
simple7	0.18039	1.4917	0.069628
simple8	0.057701	0.35275	0.11564
simple9	0.46166	2.2219	0.23085
simple10	−0.40033	−4.1556	0.52052
simple11	−0.02878	−0.40406	0.11163
simple12	0.57432	3.6281	0.15564
simple13	0.57974	4.9122	0.11527
simple14	−0.23011	−2.9653	0.33284
simple15	0.40528	3.2272	0.070571
Long-only Portfolio			
Factor	Annualized Return	Sharpe Ratio	Max Drawdown
simple1	0.30028	1.2139	0.20116
simple2	0.23882	1.0372	0.23466
simple3	0.39012	1.242	0.3229
simple4	0.41032	1.3203	0.32386
simple5	0.17234	0.7475	0.23721
simple6	0.16182	0.67133	0.21028
simple7	0.25357	1.0255	0.23885
simple8	0.18894	0.77988	0.23881
simple9	0.39387	1.2567	0.32754
simple10	−0.04014	−0.28117	0.21849
simple11	0.14894	0.61045	0.26264
simple12	0.44722	1.6082	0.28405
simple13	0.45801	1.8224	0.25204
simple14	0.045009	0.16146	0.24707
simple15	0.36602	1.4572	0.2234

Notes: The table displays the returns of factors generated by GPT given simple prompts when implemented in long-short and long-only portfolios. A long-short portfolio means going long on stocks ranked in the top 50% based on their factor values and shorting those in the bottom 50%, while long-only involves choosing stocks from the top 50% based on their factor values. Our factor return test period is from October 2021 to the end of 2022. We use the positive and negative performance of long-short total returns for each factor from 2000 to September 2021 as the reference. If a factor’s return is negative from 2000 to September 2021, we take the opposite of the factor values for the following factor return testing; otherwise, we conduct the testing using the original factor values.

Factor8\_2 achieves the highest Sharpe ratio at 7.7075, and factor7\_5 has a Sharpe ratio of 6.8203. In the long-only scenario, the annualized returns of factor7\_5 and factor8\_2 are not as outstanding as in the long-short situation but are still above 0.5, at 0.54079 and 0.51746, respectively. A total of 17 factors have long-only returns higher than 0.3, and 31 factors have a Sharpe ratio greater than 1, indicating these factors could achieve high returns with relatively lower risk. From the perspective of maximum drawdown, the maximum drawdown rate of all factors in the long-short combination

Table 14. Alphas of factors by simple prompts.

Factor	short term	long term
simple1	0.3025** (2.29)	0.10617*** (4.35)
simple2	0.13797* (1.77)	0.030254** (2.04)
simple3	0.63129*** (6.65)	0.30509*** (16.23)
simple4	0.66238*** (6.97)	0.32776*** (17.46)
simple5	0.012792 (0.16)	0.012125 (0.76)
simple6	−0.0091832 (−0.13)	−0.033989** (−2.34)
simple7	0.20765** (2.10)	0.18742*** (8.90)
simple8	0.017534 (0.15)	0.10741*** (4.50)
simple9	0.62663*** (6.35)	0.30994*** (15.57)
simple10	−0.41264*** (−5.49)	−0.0060259 (−0.37)
simple11	−0.018094 (−0.20)	0.096959*** (5.24)
simple12	0.71126*** (7.02)	0.33301*** (14.53)
simple13	0.63526*** (7.13)	0.26665*** (15.48)
simple14	−0.25693*** (−3.56)	−0.033039** (−2.29)
simple15	0.43206*** (4.16)	0.27026*** (12.40)

Notes: This table depicts the alpha of each GPT-generated factor given simple prompts within long-short portfolios. We present the portfolio alphas for two distinct time periods. 'Short term' refers to the alphas of long-short factor portfolios from October 2021 to the end of 2022, with signs determined by the positive or negative sign of each factor's return from 2000 to September 2021. 'long term' represents the alphas of long-short factor portfolios from 2003 to the end of 2022, with signs determined by the positive or negative sign of each factor's returns from 2000 to 2022. Statistical significance is denoted as: \*\*\* for 1% significance, \*\* for 5% significance, and \* for 10% significance.

was less than 30%, with 34 factors having a maximum drawdown rate below 10%, and only that of two factors exceeding 20%. Factor5\_5 has the lowest maximum drawdown rate at 0.020236. In the long-only combination, 48 factors still have maximum drawdown rates below 30%, with 11 factors below 20%, indicating that these factors also had lower risk.

On the other hand, table 11 presents the alpha of short-term long-short and long-term long-short strategies, more clearly demonstrating the average effectiveness of the factors generated by GPT. In the short-term long-short strategy, the alphas of 27 factors are statistically significant, with 18 at 1% significance, 5 at 5% significance, and 4 at 10% significance. In the long-term test, we find 41 factors with positive returns, and 30 factors with positive alphas showing 1% statistical significance, and 2 factors with positive alphas at 10% significance. Additionally, 5 factors have negative alphas with 1%

Table 15. Performance of factors: long-short portfolio-specific types of factors.

Factor	Annualized Return	Sharpe Ratio	Max Drawdown
trend1	−0.12072	−1.2348	0.24177
trend2	−0.11726	−0.9782	0.38296
trend3	0.18169	1.7589	0.10188
trend4	0.097449	0.81791	0.14395
trend5	0.1364	1.1905	0.066119
trend6	0.078684	0.52671	0.092304
trend7	0.25631	2.7539	0.032792
trend8	0.11174	1.1322	0.084229
trend9	0.18485	1.8205	0.072463
trend10	0.087957	0.76738	0.083814
math1	0.2562	2.0276	0.087094
math2	0.11963	1.3673	0.030666
math3	−0.07804	−0.07238	0.62564
math4	0.49764	2.288	0.23452
math5	0.32481	2.7825	0.10636
math6	−0.195	−1.7623	0.28642
math7	0.10339	0.66518	0.088862
math8	−0.3161	−2.7006	0.59926
math9	−0.12348	−0.99288	0.21697
math10	−0.04532	−0.41246	0.16664
fourier1	−0.02581	−0.46462	0.13262
fourier2	0.19029	1.8817	0.10725
fourier3	0.18077	2.0631	0.033807
fourier4	0.21894	2.0381	0.090883
fourier5	0.044965	0.31704	0.17188
fourier6	0.4789	2.2456	0.22468
fourier7	−0.0069	−0.19738	0.1038
fourier8	−0.00614	−0.20581	0.12888
fourier9	0.26259	2.467	0.067168
fourier10	0.63804	5.3902	0.070187
behavior1	0.070887	0.54861	0.1026
behavior2	0.1251	1.0948	0.054247
behavior3	0.11321	1.0209	0.053649
behavior4	0.042011	0.2813	0.068155
behavior5	0.14026	1.5041	0.058069
behavior6	0.27752	2.9663	0.029298
behavior7	0.072398	0.59205	0.059803
behavior8	0.17655	1.6657	0.082268
behavior9	0.066735	0.55872	0.08282
behavior10	−0.04138	−0.50072	0.11371

Notes: The table displays the returns of specific types of factors generated by GPT when implemented in long-short portfolios. We specified the characteristics of the generated factors, including four categories: those related to trends, mathematically complex factors, factors associated with Fourier analysis, and factors correlated with behavioral finance. A long-short portfolio means going long on stocks ranked in the top 50% based on their factor values and shorting those in the bottom 50%. Our factor return test period is from October 2021 to the end of 2022. We use the positive and negative performance of long-short total returns for each factor from 2000 to September 2021 as the reference. If a factor's return is negative from 2000 to September 2021, we take the opposite of the factor values for the following factor return testing; otherwise, we conduct the testing using the original factor values.

statistical significance, and another 2 factors have negative alphas with 5% and 10% statistical significance respectively, suggesting that these 7 factors could achieve significant positive returns if their signs were reversed. These results further illustrate the strong ability of the original factors generated by GPT to achieve excess returns.

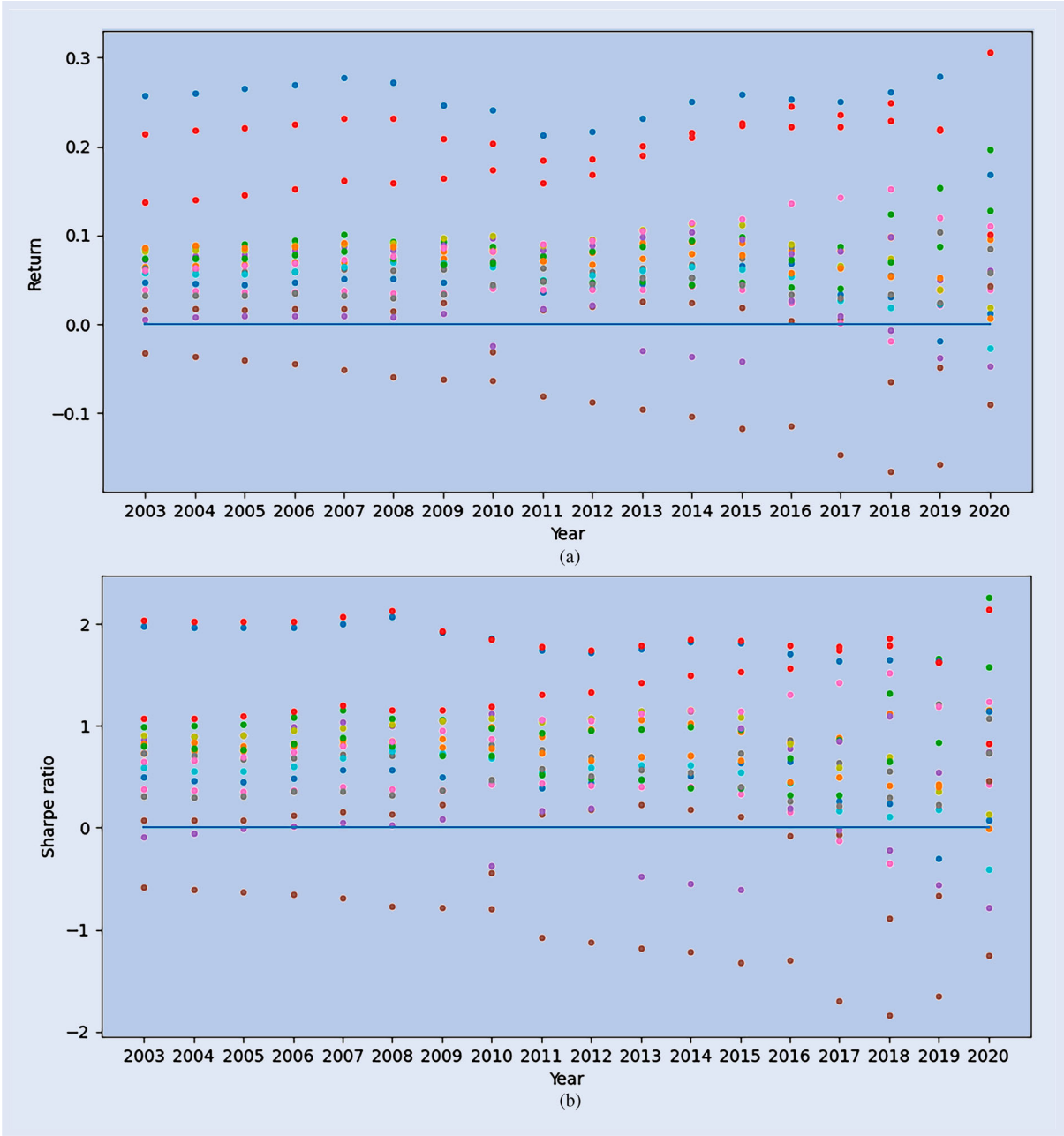


Figure 3. The distribution of annualized return and sharpe ratio across different test sets for all 35 factors. (a) Annualized Return across Different Test Sets and (b) Sharpe Ratio across Different Test Sets.

The following two charts illustrate the distribution of annualized returns and Sharpe ratios of all 35 factors across different test sets. The years on the horizontal axis represent the final year used to determine the sign of the factors. For instance, the point corresponding to the year 2010 indicates that the factor’s sign is determined based on returns from 2000 to 2010, and then the factor is tested from 2011 to 2022.

6.2. Testing on different time periods

Are the factors generated by GPT only effective in specific periods in the U.S. stock market? To further explore the robustness of the factors generated by GPT, we test them on different datasets. Specifically, we adopt a rolling method to investigate whether the change of time period causes significant fluctuations in factor returns and significant changes in alpha. In this method, under the framework of long-term testing, we add one year of data each time as the basis for determining the sign of the factors, correspondingly reducing one year of testing factors, until the test set was the data from 2021 to 2022. Consequently, we obtain 18 sets of test results.

Our findings indicate that, regardless of different periods or markets, the factors generated by GPT were quite robust.

Table 12 provides the number of statistically significant factors under different time period tests. It is evident that from the test set of 2004–2022 to 2016–2022, the number of factors with statistically significant alphas is between 26 and 30, and the number of factors with 1% statistical significance is always over 20. As the test set becomes smaller, market uncertainty causes a decline in the number of factors with statistically significant alphas, but the number of statistically significant factors in all time period test sets is still over 20. This indicates that these factors are generally robust and can generate significant alphas over a long period of testing.



Table 16. Performance of factors:long-only portfolio-specific types of factors.

Factor	Annualized Return	Sharpe Ratio	Max Drawdown
trend1	0.11508	0.4817	0.24602
trend2	0.11681	0.42395	0.32272
trend3	0.26626	1.1626	0.2012
trend4	0.23866	1.0731	0.19927
trend5	0.24494	1.0315	0.21736
trend6	0.21312	1.0143	0.21863
trend7	0.29833	1.3696	0.18585
trend8	0.22622	0.97766	0.1972
trend9	0.27461	1.2443	0.18483
trend10	0.22619	0.99115	0.21525
math1	0.28807	1.1385	0.2505
math2	0.21988	0.94558	0.23199
math3	-1.2941	-1.1947	1.6079
math4	0.41865	1.3132	0.33507
math5	0.32249	1.565	0.2291
math6	0.062838	0.26189	0.27799
math7	0.21971	0.87665	0.22962
math8	0.002027	-0.0396	0.36437
math9	0.10266	0.40305	0.27051
math10	0.14047	0.53355	0.26823
fourier1	0.1515	0.65299	0.22715
fourier2	0.25955	1.1429	0.20898
fourier3	0.25485	1.1427	0.22159
fourier4	0.27957	1.1969	0.23031
fourier5	0.18694	0.77318	0.3063
fourier6	0.40388	1.2774	0.32962
fourier7	0.16101	0.69123	0.24572
fourier8	0.16138	0.70519	0.26706
fourier9	0.29574	1.2119	0.24682
fourier10	0.48344	2.2186	0.20018
behavior1	0.19855	0.83614	0.24313
behavior2	0.22478	0.95984	0.21472
behavior3	0.2182	0.93855	0.20702
behavior4	0.18196	0.7644	0.22251
behavior5	0.23348	1.0156	0.22085
behavior6	0.29964	1.2566	0.20364
behavior7	0.1984	0.85578	0.20704
behavior8	0.25111	1.0698	0.22851
behavior9	0.19608	0.83742	0.23909
behavior10	0.14262	0.60012	0.23911

Notes: This table shows the returns of specific types of factors created by GPT in long-only portfolios. We specified the characteristics of the generated factors, including four categories: those related to trends, mathematically complex factors, factors associated with Fourier analysis, and factors correlated with behavioral finance. For this strategy, long-only involves choosing stocks from the top 50% based on their factor values. Our factor return testing period spans from October 2021 to the end of 2022. We use the positive or negative sign of the long-short total returns of each factor from January 2000 to September 2021 as a reference. If a factor has a negative returns during this period, we conduct the factor return testing by taking the opposite of the factor values. Conversely, if a factor has a positive return during this period, we conduct the testing using the original factor values.

Additionally, we also present specific factor return rates and Sharpe ratio changes. Figure 2 shows the changes in annualized return rates and Sharpe ratios of factor5, factor10, factor15, factor20, factor25, factor30, factor35 under different time period test sets. It can be seen that, as the test set decreases, the annualized return rates and Sharpe ratios of

the factors do not significantly decrease in most cases and even increase sometimes. Figure 3 shows the distribution of return rates and Sharpe ratios of all factors under different time period tests. It is clear that the change in test sets does not lead to a significant increase in outliers or negative points, and the overall distribution of returns and Sharpe ratios is relatively stable. Thus, we have reason to believe that the factors generated by GPT are still effective in different time period test sets.

## 7. The effects of prompt engineering

In the previous results, we prompt GPT to generate complex and original factors and enhance this process by using the prompt 'please create a more complex and original factor', leading GPT to produce a series of original factors combining multiple elements. This brings up the question: What changes would occur in the factors generated by GPT if we alter the prompt, and can it generate factors that meet specific characteristics?

### 7.1. Simplified prompts

For the previous results, we prompt GPT to generate factors that were complex and original, hoping for a brainstorming effect. In this section, our prompt merely require GPT to generate new factors and provide their codes, informing it of the data names we possessed. Under these conditions, GPT generate simpler factors, such as standard deviation of returns, the difference between the highest and lowest prices, and some slightly more complex, common market factors like the Relative Strength Index (RSI) Factor, Moving Average Convergence Divergence (MACD) Factor, Stochastic Oscillator Factor, and Accumulation/Distribution Line. These factors have been proven effective and applied in the market. Although the generated factors are not original creations of GPT, we test 15 such factors for profitability, with results for long-short and long-only in the short term presented in table 13. The results show that in the short-term test, 11 factors have positive long-short returns, and 14 have positive long-only returns. Interestingly, some of these factors achieve annualized returns of over 50%. Both simple12 and simple13 exceed an annualized return of 0.57, and Simple14 slightly surpasses 0.5, at 0.50055. Additionally, three other factors have long-short annualized returns exceeding 0.4. From a risk-adjusted perspective, simple13's Sharpe ratio even reaches 4.9122, with five other factors exceeding 2. Given that most of these factors are already recognized by the market, such profitability is understandable. Table 14 displays the five-factor alpha results for these factors, with 8 alphas statistically significant at 1% (6 positive, 2 negative), 2 at 5% (both positive), and 1 positive at 10% significance; 14 factors have positive long-only returns, and their corresponding alphas are positive and 1% statistically significant. In the long-term test, 12 factors have positive long-short returns, with 10 alphas statistically significant at 1% (all positive), and 3 at 5% significance (1 positive, 2 negative). These test

Table 17. Alphas of specific types of factors.

Factor	short term	long term	Factor	short term	long term
trend1	− 0.14542 (− 1.50)	− 0.012439 (− 0.67)	fourier1	− 0.056008 (− 0.78)	− 0.0063055 (− 0.44)
trend2	− 0.062261 (− 0.59)	0.13114*** (5.60)	fourier2	0.19735** (2.30)	0.056511*** (3.40)
trend3	0.14446* (1.67)	0.064558*** (3.73)	fourier3	0.15733** (2.15)	− 0.12326*** (− 8.19)
trend4	0.078312 (0.84)	− 0.0084658 (− 0.44)	fourier4	0.21633** (2.35)	− 0.078534*** (− 4.59)
trend5	0.13373 (1.41)	0.041968** (2.05)	fourier5	0.029685 (0.31)	− 0.049015*** (− 2.70)
trend6	0.030288 (0.29)	− 0.11011*** (− 5.24)	fourier6	0.65268*** (6.59)	0.31961*** (15.61)
trend7	0.23085*** (2.91)	0.03712** (2.21)	fourier7	0.0055289 (0.07)	− 0.0014433 (− 0.09)
trend8	0.11664 (1.46)	0.0011732 (0.07)	fourier8	− 0.023159 (− 0.30)	− 0.021279 (− 1.36)
trend9	0.1479* (1.74)	0.063074*** (3.77)	fourier9	0.28191*** (3.18)	− 0.10479*** (− 5.69)
trend10	0.052036 (0.58)	0.0032189 (0.16)	fourier10	0.6395*** (7.12)	0.37659*** (20.29)
math1	0.28684*** (2.80)	0.21717*** (10.20)	behavior1	0.05973 (0.61)	0.069842*** (3.53)
math2	0.12181* (1.70)	− 0.035397** (− 2.54)	behavior2	0.10246 (1.09)	0.091803*** (4.57)
math3	− 0.20779 (− 0.19)	0.0622 (0.06)	behavior3	0.12439 (1.38)	0.033337** (2.00)
math4	0.68216*** (6.86)	0.34928*** (17.36)	behavior4	0.044417 (0.45)	0.075359*** (3.78)
math5	0.29311*** (3.50)	0.077623*** (4.55)	behavior5	0.12114 (1.56)	0.073352*** (4.51)
math6	− 0.17489** (− 2.12)	− 0.061513*** (− 3.61)	behavior6	0.2863*** (3.69)	0.067574*** (4.37)
math7	0.093367 (0.76)	0.23712*** (9.74)	behavior7	0.0709 (0.76)	0.075556*** (4.02)
math8	− 0.27871*** (− 2.70)	− 0.16542*** (− 7.64)	behavior8	0.14911* (1.67)	0.077126*** (4.10)
math9	− 0.12475 (− 1.02)	0.20511*** (9.51)	behavior9	0.046666 (0.52)	0.052164*** (2.81)
math10	− 0.020575 (− 0.18)	0.24105*** (10.16)	behavior10	− 0.071364 (− 0.75)	0.063783*** (3.32)

Notes: This table depicts the alpha of each GPT-generated specific types of factor within long-short portfolios. We present the portfolio alphas for two distinct time periods. ‘Short term’ refers to the alphas of long-short factor portfolios from October 2021 to the end of 2022, with signs determined by the positive or negative sign of each factor’s return from 2000 to September 2021. ‘long term’ represents the alphas of long-short factor portfolios from 2003 to the end of 2022, with signs determined by the positive or negative sign of each factor’s returns from 2000 to 2022. Statistical significance is denoted as: \*\*\* for 1% significance, \*\* for 5% significance, and \* for 10% significance.

results suggest that simple prompts can lead GPT to generate common, simple factors, but since these are not original, they do not verify GPT’s ability to generate original factors. However, on the other hand, the factors provided by GPT are classic and effective, possibly offering useful, already discovered factors for researchers whose goal is not factor discovery. Some of these can even yield returns comparable to complex strategies, thereby reducing the time spent on factor collection and allowing more focus on technical and strategic research. However, these factors might also become ineffective due to their widespread use in market trading.

## 7.2. Generation of specific types of factors

So, can GPT generate factors of specific types that we designate? For this, we consider more complex prompts. We open

4 new windows, each adding a new directive to the original prompt, requesting that the generated factors meet one of the following four conditions: trend-based, mathematically complex, Fourier analysis-related, and related to behavioral finance. Based on this, we test 10 factors. Our results show that the prompts are effective for GPT-4, and the factors it generates indeed meet the aforementioned conditions. The respective long-short and long-only returns of these types of factors in the short term are presented in tables 15 and 16. It can be seen that all long-only returns are positive. Under the long-short strategy, the trend-based and behavioral finance categories generally have positive returns, but the returns for mathematically complex and Fourier analysis are less stable. The highest long-short return is from the tenth factor related to Fourier analysis, with an annualized return of 0.63804 and the highest Sharpe ratio of 5.3902. Among these 40 factors, only 5 have long-short returns over 0.3, and their long-only

returns also exceed 0.3. However, from the perspective of the Sharpe ratio, more than half, 21 factors, have a ratio over 1, with 11 exceeding 2. Even under the long-only combination, 19 factors have a Sharpe ratio over 1. Moreover, 20 factors have a maximum drawdown rate below 10%. This suggests that while these factors do not achieve very high returns overall, the risk of investment portfolios constructed from them is very low.

As illustrated in table 17, from the alpha perspective concerning long-short returns in the short-term assessment, over half of the factors within the mathematically complex and Fourier analysis-related categories demonstrate statistical significance, albeit with a mix of positive and negative values. In contrast, the trend-based and behavioral finance categories exhibit a smaller number of significant alphas, yet all are positively oriented. However, in long-term performance, all behavioral finance category factors have positive alphas that are 1% statistically significant. The other three categories also have more than half of their factors with significant alphas, but with both positive and negative values, indicating instability. Notably, the mathematically complex factors, in the long-term test, all have 1% significant alphas, but with 3 factors having negative alphas, meaning the signs of these factors were reversed. Overall, we believe GPT has better control over behavioral finance category factors, capable of generating more stable. However, for mathematically complex and Fourier analysis categories, although they can achieve high alphas, their stability is poorer, indicating that GPT cannot guarantee the effectiveness and stability of factors meeting specific mathematical requirements.

## 8. Conclusion

This study pioneers the integration of stock feature generation with the sophisticated capabilities of the GPT-4 language model, culminating in the construction of factors that exhibit profitability in the equity investment realm. Deviating markedly from extant literature on automated feature generation, the factors conceived by ChatGPT are anchored in knowledge-based reasoning, eschewing traditional data mining methodologies. Empirical evidence from our investigations corroborates the superior quality of GPT-derived factors, underscoring their robust predictive prowess, with certain factors realizing anomalously elevated returns. Consequently, GPT emerges as an indomitable instrument, proffering to both investors and scholars an enhanced and superior paradigm for factor discernment and genesis. Moreover, the adoption of GPT-centric factor generation markedly diminishes human temporal overhead, facilitating efficiency in scholarly endeavors.

Our scholarly exposition posits that GPT-derived factors amalgamate myriad indicators, each elucidating market price dynamics from multifaceted vantage points, thereby fostering a more encompassing market exegesis. These synthesized factors are emblematic of ChatGPT's profound acumen in the financial realm, delineating a cornucopia of insights and proffering a comprehensive foundation for investment strategies.

A salient revelation from our analysis is the heightened interpretability intrinsic to GPT-forged factors, juxtaposed against their counterparts generated through indiscriminate data mining. Delving into the factors conceived by GPT and probing its factorization mechanics, we discern that GPT adeptly harnesses the technical indicators within its purview, meticulously applying the intrinsic mathematical and statistical tenets of these indicators to craft factors imbued with profound economic ramifications.

To encapsulate, this scholarly endeavor proffers an avant-garde perspective and a potent instrument for factor discernment and genesis within the financial ambit, spotlighting the prodigious potentialities of GPT-infused factors for equity investments. We believe that this study will provide invaluable insights for market players and the academic community, sparking further interest and avenues for upcoming scholarly investigations.

## Acknowledgements

Special thanks to Hui Chen for his kindly advice. Thanks to Yuming Zeng for his contributions as a research assistant. Ke Tang acknowledges financial support from the Guoqiang Institute, Tsinghua University (2020QQG1018).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Yuhan Cheng  <http://orcid.org/0000-0001-5241-1344>

Ke Tang  <http://orcid.org/0000-0003-4049-030X>

## References

- Araci, D., Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019.
- Carhart, M.M., On persistence in mutual fund performance. *J. Finance.*, 1997, **52**, 57–82.
- Cui, C., Wang, W., Zhang, M., Chen, G., Luo, Z. and Ooi, B.C., Alphaevolve: A learning framework to discover novel alphas in quantitative investment. In *Proceedings of the 2021 International Conference on Management of Data*, pp. 2208–2216, 2021.
- Fama, E.F. and French, K.R., The cross-section of expected stock returns. *J. Finance.*, 1992, **47**, 427–65.
- Fama, E.F. and French, K.R., A five-factor asset pricing model. *J. FinancEcon.*, 2015, **116**, 1–22.
- Fang, J., Xia, S., Lin, J., Xia, Z., Liu, X. and Jiang, Y., Alpha discovery neural network based on prior knowledge. arXiv preprint arXiv:1912.11761, 2019.
- Fang, J., Lin, J., Xia, S., Xia, Z., Hu, S., Liu, X. and Jiang, Y., Neural network-based automatic factor construction. *Quant. Finance*, 2020, **20**, 2101–2114.

- Jiang, J., Kelly, B.T. and Xiu, D., Expected returns and large language models. Available at SSRN, 2022.
- Kanter, J.M. and Veeramachaneni, K., Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, 2015 (IEEE).
- Lakkaraju, K., Vuruma, S.K.R., Pallagani, V., Muppasani, B. and Srivastava, B., Can llms be good financial advisors?: An initial study in personal decision making for optimized outcomes. arXiv preprint arXiv:2307.07422, 2023.
- Lam, H.T., Thiebaut, J.-M., Sinn, M., Chen, B., Mai, T. and Alkan, O., One button machine for automating feature engineering in relational databases. arXiv preprint arXiv:1706.00327, 2017.
- Li, L., Wang, H., Zha, L., Huang, Q., Wu, S., Chen, G. and Zhao, J., Learning a data-driven policy network for pre-training automated feature engineering. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda, 2022.
- Shah, A. and Chava, S., Zero is not hero yet: Benchmarking zero-shot performance of llms for financial tasks. arXiv preprint arXiv:2305.16633, 2023.
- Sharpe, W.F., Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finance.*, 1964, **19**, 425–442.
- Son, G., Jung, H., Hahm, M., Na, K. and Jin, S., Beyond classification: Financial reasoning in state-of-the-art language models. arXiv preprint arXiv:2305.01505, 2023.
- Wang, S., Yuan, H., Zhou, L., Ni, L.M., Shum, H.-Y. and Guo, J., Alpha-gpt: Human-ai interactive alpha mining for quantitative investment. arXiv preprint arXiv:2308.00016, 2023.
- Wu, S., Irsoy, O., Lu, S., Dabrowski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D. and Mann, G., Bloomberggpt: A large language model for finance, 2023.
- Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A. and Huang, J., Pixiu: A large language model, instruction data and evaluation benchmark for finance. arXiv preprint arXiv:2306.05443, 2023.
- Yu, S., Xue, H., Ao, X., Pan, F., He, J., Tu, D. and He, Q., Generating synergistic formulaic alpha collections via reinforcement learning. arXiv preprint arXiv:2306.12964, 2023a.
- Yu, X., Chen, Z., Ling, Y., Dong, S., Liu, Z. and Lu, Y., Temporal data meets llm—explainable financial time series forecasting. arXiv preprint arXiv:2306.11025, 2023b.
- Zhang, C., Li, Y., Chen, X., Jin, Y., Tang, P. and Li, J., Doubleensemble: A new ensemble method based on sample reweighting and feature selection for financial data analysis. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 781–90, 2020a (IEEE).
- Zhang, T., Li, Y., Jin, Y. and Li, J., Autoalpha: An efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. arXiv preprint arXiv:2002.08245, 2020b.
- Zhang, H., Hua, F., Xu, C., Guo, J., Kong, H. and Zuo, R., Unveiling the potential of sentiment: Can large language models predict chinese stock price movements? arXiv preprint arXiv:2306.14222, 2023a.
- Zhang, T., Zhang, Z.A., Fan, Z., Luo, H., Liu, F., Liu, Q., Cao, W. and Jian, L., Openfe: Automated feature generation with expert-level performance. In *International Conference on Machine Learning*, pp. 41880–41901, 2023b (PMLR).