CVaR-based Portfolio with Investment Views Given by Large Language Models

Zhen Li, Liangyu Min*, *Member, IAENG*

***Abstract*—With the development of Large Language Models (LLMs), remarkable high quality output about professional financial consultation has already attracted the attention of the academic community, which stands to reason that some financial analysts relying to public market information may be rendered obsolete by these LLMs. However, some empirical studies show that the LLMs could not build and solve the mathematical problem accurately, especially when the specific investment goals are taken into account. Therefore, we aim to utilize the ability of analyzing huge public information of LLMs and the tools in the field of operations research to construct CVaR-based portfolios in this paper. According to the associated numerical experiments, the effectiveness and efficiency of the proposed method could be demonstrated, where Claude 3.5 Sonnet achieves the best comprehensive performance among the CVaR portfolios based on LLMs.**

***Index Terms*—Portfolio selection, Artificial Intelligence, Large Language Model, CVaR**

I. Introduction

THE advent of generative artificial intelligence has greatly influenced the research mode in lots of industries, also including the financial engineering. These LLMs with massive parameters are able to learn and analyze huge text datasets and efficiently generate detailed outputs for financial portfolio constructing or stock picking [1], [2]. However, model hallucination is a tricky problem for LLMs-based portfolio construction paradigm, thus managers may suffer from unreasonable financial suggestions given by LLMs, and form err-prone investment decisions. Although LLMs such as GPT-4 and Gemini have shown impressing advantage in information retrieval and logical reasoning to some extent, but they require a lot of human intervention when dealing with customized modelling and problem-solving tasks. In this paper, we aim to expand existing research and construct the portfolio model with specified risk and return objectives based on some investment views given by LLMs.

Cho [3] investigates the applicability of ChatGPT in empirical asset pricing, where multiple Twitter posts including macro and firm-specific news are used as prompts. Based on the generated buy and sell stock tickers, the author builds open-to-close intraday trading strategy and earn significant long-short returns. According to the performance analysis, ChatGPT is applicable to extract mispricing signals in text data.

Callanan et, al [4] examine the financial reasoning abilities of ChatGPT and GPT-4 in mock CFA exams (known as meticulous but practical assessment about financial expertise), where zero-shot, few-shot and chain-of-thought are the three prompting paradigms used in experiments. They also verify the hallucinations of LLMs in mathematical formula and calculations.

Romanko et, al [1] employ ChatGPT to build a universe of stocks from S&P 500 market index that worthwhile for investing. Based on that, they compare different portfolio optimization strategies performances on the stock universe given by ChatGPT. The provided empirical results show that by blending strengths of stock selection of LLMs and quantitative optimization techniques, some reliable financial decisions can be made.

Kong et, al [2] also leverage the advantages of LLMs and operation research tools on the field of financial engineering, where the Black-Litterman portfolio models based on stock universe given by ChatGPT and BARD are derived and constructed. According to the efficient frontiers given by the associated numerical experiments, the proposed Black-Litterman portfolios based on LLMs are superior to the classical mean-variance portfolio on the out-of-sample performance.

Niszczota & Abbas [5] assess the ability of GPT to serve as a financial robo-advisor via a financial literacy test. The test results illustrate that Davinci and GPT-3.5 score 66% and 65% on the financial literacy test, respectively. GPT-4 achieves significantly high score of 99%. Based on that, they point the advice-utilization from LLMs for researchers.

Wang et, al [6] propose a new alpha mining paradigm via GPT tools, where the associated prompt engineering algorithmic framework, Alpha-GPT, is also implemented. In this system, LLM works as a mediator between investors and alpha search, where the trading ideas could be translated into fitting expressions. Also, the provided numerical experimental results demonstrate that Alpha-GPT could dig out alphas as investors' ideas such as golden cross patterns, Bollinger bands, and price momentum.

Ko & Lee [7] investigate the potential of employing ChatGPT as the investment assistant for human experts. Also, the constructed portfolios based on the selections provided by ChatGPT outperform random portfolios in the associated numerical experiments, which elucidate the feasibility of building LLM-based investment strategies.

Ullah et, al [8] focus on the Pakistan market and demonstrate the positive and significant impact of ChatGPT in

Manuscript received Feb 10, 2025; revised May 31, 2025. This work is supported by the College Young Teacher Training Subsidy Plan of Shanghai (No. AG24-33811-3301), the Startup Fund for Advanced Talents of Putian University (2020008), Putian University 2024 Education and Teaching Research Program (JG202447), and the education reform project of Shanghai Business School (No. AG25-12213-3307-15).

Zhen Li is an associate professor in the School of Business, Putian University, Putian Fujian 351100, China (E-mail: 408053275@qq.com).

Liangyu Min is a lecturer in the Faculty of Business Information, Shanghai Business School, Shanghai 201400, China (Corresponding author to provide email: 21220001@sbs.edu.cn).

investment decision making process using the collect data from 388 individual investors. Moreover, they emphasize the moderating role of financial literacy in the relationship between ChatGPT usage dimensions and optimal results. And, the possible fraudulent information given by some generative AI tools should be seriously considered and regulated.

Oehler & Horn [9] compare the recommendations from academic literature and ChatGPT, and find that ChatGPT could provide better financial advice for one-time investments than robo-advisors. They propose that independent chatbots could be viewed as trustworthy financial information source, by which the human or robo-advisor recommendations could be double-checked. In the empirical research, the significant deviation among robo-advisors could be observed, which would pose negative effects on individual investors. Thus, a solid financial plan or informed decisions require the help of ChatGPT.

Existing studies have demonstrated the immense potential of LLMs in financial applications. Building on this foundation, we seek to apply selected LLMs to financial modelling and showcase the superiority of the LLM-based financial models. Conditional Value at Risk (CVaR) stands out as a valuable indicator in that it addresses limitations of the famous Value at Risk (VaR) metric by giving a more comprehensive view of risk, especially in some extreme scenarios. Moreover, CVaR is a coherent risk metric, suggesting it satisfies desirable properties such as subadditivity and monotonicity. Zinchenko & Asimit [10] consider the worst possible dependence scenario, where the risk preferences are modeled by the CVaR, coinciding with the lower-orthant stochastic ordering (partial information) of the underlying bivariate distributions for portfolios with two risks. And, their study contributes significantly to the field by establishing sharp lower and upper bounds for the CVaR-based risk levels.

Xu et, al [11] elucidate a large CVaR-based portfolio modelling method, where weight constraints on the standard CVaR-based framework are imposed. The proposed numerical experiments illustrate the efficacy of constructing sparse and stable portfolios from large assets, where no extreme positive and negative weights could be observed from the final portfolio.

Lotfi & Zenios [12] develop a robust mean-to-CVaR portfolio model, where the interval ambiguity in returns and covariance are considered. Their study suggests that covariance ambiguity could also induce bias even though enough explanations are provided. Normality assumption is used in their model derivation, that is, $\text{CVaR}_\alpha(r_p) = -\bar{r}_p + \kappa_{1-\alpha}\sigma_{r_p}$, where $\bar{r}_p$ and σ_{r_p} represent the mean and standard deviation of return rates, respectively. And, $\kappa_{1-\alpha}\phi(\Psi^{-1}(1-\alpha))$ with the normal density of ϕ and the cumulative distribution function of Ψ.

On basis of the existing researches, we construct the portfolio model with CVaR as the risk measure. LLMs are employed to deal with the potential conservatism of the investment with risk control as the main objective. Section II derives the CVaR portfolio model and Section III gives the operations and results of some LLMs. The portfolio framework and the associated algorithm are presented in Section IV. Section V illustrates the numerical experiments of the proposed framework, and Section VI concludes the study.

II. CVaR-based portfolio

CVaR is viewed as a valuable indicator in financial modelling due to it addresses limitations of the classical VaR metric and provides a more comprehensive point of portfolio risk, especially in some extreme scenarios.

Like VaR, CVaR also focuses on portfolio tail risk, measuring the expected loss in the worst-case scenarios beyond a specified confidence level. However, unlike VaR, which merely indicates the threshold loss at a given percentile, CVaR quantifies the average loss beyond this threshold, providing a more comprehensive view of extreme scenarios. It is wise for the portfolio manager to build risk-averse investment strategy employing CVaR as its metric. By emphasizing severe losses, CVaR is particularly well-suited for investors and institutions with a low tolerance for significant risks. For conservative managers, CVaR offers valuable insights into the depth of potential losses under extreme market conditions.

CVaR is a coherent risk measure, indicating it satisfies desirable properties such as subadditivity and monotonicity. Thus, CVaR is suitable for portfolio optimization as it produces smooth risk profile across asset returns. For N available risky assets, define $\boldsymbol{w}$ be the vector of portfolio weights, and $\sum_{i=1}^{N} w_i = 1$. X be a random variable representing portfolio returns, and α be the confidence level, for example, 95% or 99%, thus $1-\alpha$ denotes the tail probability.

For a portfolio loss $L(\boldsymbol{w}) = -\boldsymbol{w}^T X$, the VaR at level α is defined as follows:

$$P(L(\boldsymbol{w})) \geq \text{VaR}_\alpha(\boldsymbol{w}) = 1-\alpha$$

Based on the definition of VaR, the CVaR at level α refers to the expected loss beyond the corresponding VaR threshold, which can be calculated as follows:

$$\text{CVaR}_\alpha(\boldsymbol{w}) = \mathbb{E}[L(\boldsymbol{w})|L(\boldsymbol{w}) \geq \text{VaR}_\alpha(\boldsymbol{w})]$$

Equivalently, CVaR can be formulated as an optimization problem as follows:

$$\text{CVaR}_\alpha(\boldsymbol{w}) = \min_{\eta}\left(\eta + \frac{1}{1-\alpha}\mathbb{E}[L(\boldsymbol{w})-\eta]^+\right)$$

where $x^+ = \max(x, 0)$.

To construct a CVaR-based portfolio, the goal is to minimize CVaR at a specified confidence level α, subject to constraints on the portfolio weight $\boldsymbol{w}$, and the corresponding optimization problem with budget and non-shorting constraints is:

$$\min_{\boldsymbol{w},\eta}\left(\eta + \frac{1}{(1-\alpha)T}\sum_{t=1}^{T}\max(L_t(\boldsymbol{w})-\eta, 0)\right)$$
$$s.t. \quad \sum_{i=1}^{N} w_i = 1, w_i \geq 0$$

where T is the number of scenarios such as historical or simulated returns, and $L_t(\boldsymbol{w}) = -\boldsymbol{w}^T X_t$ is the loss of the portfolio in scenario t.

Further, the CVaR optimization can be formulated into the linear programming via auxiliary variable $u_t = \max(L_t(\boldsymbol{w}) - \eta, 0)$ as follows:

$$\min_{\boldsymbol{w},\eta,u_t} \left(\eta + \frac{1}{(1-\alpha)T} \sum_{t=1}^{T} u_t \right)$$
$$s.t. \quad \sum_{i=1}^{N} w_i = 1, w_i \geq 0$$

where η approximates the VaR at the α-level, u_t represent the excess loss over η in each scenario, and CVaR considers the average value.

III. Recommendations from LLMs

According to the conclusions from [1], LLM is effective in stock selection while not good at in assigning optimal portfolio weights. For example, one of the useful prompts could be "Assume you are a professional investment expert. Please use a range of accessible investing principles taken from some out-standing funds, construct a theoretical portfolio comprising of 15 stocks from the S&P 500 with the objective to outperform the S&P 500 index".

Table I shows the recommended results of OpenAI-o1 model, which is launched in September 2024, with enhanced reasoning capabilities. According to the feedback from vast users, OpenAI-o1 has superior performance compared to earlier LLMs, achieving higher scores on benchmarks such as American Invitational Mathematics Examination (AIME). Its recommendations cover several industries such as Technology, Consumer Discretionary, Healthcare, Financial, etc. Thus, a diversified portfolio could be constructed with the help of OpenAI-o1.

Table II gives the portfolio of Claude 3.5 Sonnet, which is considered as one of the state-of-the-art LLM developed by Anthropic, building upon the capabilities of its predecessor, Claude 3. As far as the Claude 3.5 Sonnet concerned, it is favored by lots of users due to impressive abilities of reasoning and understanding of complex instructions. The given portfolio involves 8 industries, where Technology and Healthcare are applauded by Claude 3.5 Sonnet.

Table III summarizes the investment strategy built by GPT 4 Omni, which is also called GPT-4o. It is a multilingual, multimodal generative pre-trained transformer developed by OpenAI and released in May 2024. The provided portfolio leverages a mix of industry-leading growth companies and defensive value investments to maximize long-term returns while limiting volatility.

Table IV presents the financial suggestions given by Llama 3.1 405B, which is released by Meta in July 2024. The 405B parameters version is one of the largest open-source models, designed for tasks requiring extensive computational resources. The portfolio allocation spreads on multiple industries, such as technology, healthcare, consumer discretionary, and energy.

To explore the potential of LLMs as investment experts, we employ several LLMs to construct CVaR-based portfolio models based on their recommended tickers. Existing research and industry reports highlight the remarkable performance of LLMs such as GPT-4, Llama, and Claude across diverse domains, including financial modelling. These models can be leveraged to support market analysis, risk management, and portfolio optimization. However, [1] has pointed that the optimized portfolio weight given by GPT models may lack reliability, whereas a two-phase approach incorporating human intervention proves more effective. With advancements in LLMs, we also compare the effectiveness of the zero-slot end-to-end paradigm with the proposed LLM-CVaR portfolio.

IV. Portfolio framework

Fig. 1 presents the flowchart of the proposed portfolio framework, where two types of investment strategies involved. The first one is classical two-stage portfolio modelling, and the CVaR portfolios based on the recommendations from LLMs are constructed. The second one applies the idea of end-to-end financial decision [13], where the portfolio weights are formed directly by prompts. Existing researches seldom evaluate the prompt-based portfolios from ChatGPT because they believe the given weights are merely demos without any optimization. However, things may be different using some upgraded LLMs as investment advisor.

Model hallucination remains a critical challenge in LLMs and generative systems, where outputs are generated that lack grounding in the provided data, context, or reality. Such outputs may include fabricated facts, inaccurate information, or logically inconsistent statements. To address this issue, the proposed framework incorporates a validation stage to verify the reliability of recommendations generated by LLMs.

A comparative analysis is conducted between the prompt-based portfolio and the CVaR-based portfolio, evaluating their performance in terms of risk and return using established financial metrics. This analysis also validates the effectiveness and efficiency of the prompt-based portfolio construction, offering valuable insights for to some individual investors.

V. Numerical experiments

In this section, we mainly discuss the results of numerical experiments designed to validate the performance of the proposed portfolios. The experimental data set could be retrieved from the website of Yahoo Finance. The experiments focus on evaluating some risk and return attributes of the involved portfolios, by which we can validate the effectiveness of the portfolios given by LLMs, especially from the perspective of tail risk.

A. 1/N portfolio for each LLM

We first construct $1/N$ portfolio for each recommendation from LLMs, and Table V gives the fundamental descriptive statistics for the equal-weighted portfolio returns. Also, we report the tail risk information via two indicators VaR@95 and CVaR@95. Note that GPT-4 and LLAMA, while providing the recommended portfolios, also give the weights of each constituent stock. Fig. 2 records the portfolio weights directly given by the two LLMs. It can be found that these portfolio weights are not optimized by

TABLE I
PORTFOLIO GIVEN BY OPENAI-O1.

Ticker	Sector	Rationale
AAPL	Technology	Market leader with a strong ecosystem, consistent revenue growth, and robust financials.
MSFT	Technology	Dominant position in software and cloud services, steady revenue streams, and innovative product development.
AMZN	Consumer Discretionary	Leader in e-commerce and cloud computing, with significant growth potential in multiple markets.
GOOG	Consumer Services	Strong presence in online advertising and search, with diversification into cloud services and emerging technologies.
NVDA	Technology	Pioneering in graphics processing units (GPU) and artificial intelligence (AI).
JNJ	Healthcare	Diversified healthcare company with stable earnings, strong dividend history, and a solid pharmaceutical pipeline.
V	Financial	Leading global payments technology company benefiting from the shift to digital payments.
HD	Consumer Discretionary	Largest home improvement retailer with consistent performance and growth in the housing market.
JPM	Financial	Strong banking institution with diversified services and solid track record of financial performance.
PG	Consumer Staples	Global consumer goods company with a wide product portfolio, strong brand recognition, and defensive qualities.
CAT	Industrial	Leading manufacturer of construction and mining equipment, positioned to benefit from infrastructure spending.
INTC	Technology	Major player in semiconductor manufacturing with potential for turnaround and value appreciation.
MRK	Healthcare	Strong pharmaceutical company with a promising pipeline and consistent dividend payouts.
DIS	Communication Services	Diversified entertainment giant with assets in streaming, theme parks, and media networks.
NEE	Utilities	Leader in renewable energy with growth prospects in clean energy initiatives.

TABLE II
PORTFOLIO GIVEN BY CLAUDE 3.5 SONNET.

Ticker	Sector	Rationale
MSFT	Technology	Cloud computing leadership, enterprise software dominance, strong recurring revenue.
AAPL	Technology	Brand power, ecosystem lock-in, service revenue growth.
AVGO	Technology	Semiconductor leadership, AI/ML exposure, strong dividend growth.
UNH	Healthcare	Market leader in healthcare, consistent earnings growth, strong competitive moat.
LLY	Healthcare	Strong drug pipeline, obesity/diabetes market potential, R&D effectiveness.
ABT	Healthcare	Diversified healthcare products, global presence, stable growth.
BRK.B	Financial	Diversified conglomerate, strong management, value investing approach.
SPGI	Financial	Rating agency moat, data services growth, high margins.
COST	Consumer	Membership model strength, consumer loyalty, operational efficiency.
MCD	Consumer	Brand strength, global presence, recession resistant.
CAT	Industrial	Infrastructure exposure, global reach, strong dealer network.
UNP	Industrial	Railroad monopoly, pricing power, operational efficiency.
GOOG	Communication	Search dominance, AI leadership, strong cash flow.
XOM	Energy	Scale advantages, integrated operations, dividend stability.
LIN	Material	Industrial gas leadership, pricing power, geographic diversification.

TABLE III
PORTFOLIO GIVEN BY GPT 4 OMNI.

Ticker	Sector	Rationale
AAPL	Technology	Strong ecosystem, expanding revenue streams, and global leadership.
MSFT	Technology	Leader in cloud computing, enterprise software, and AI durable competitive advantage.
NVDA	Technology	Global leader in GPUs and AI hardware, with strong demand due to AI-driven industry growth.
AVGO	Technology	Chip-making leader with exposure to 5G, AI, and data centers.
AMZN	Consumer Discretionary	E-commerce giant with strength in AWS cloud services.
TSLA	Consumer Discretionary	Leader in EVs and renewable energy, driving the sustainable energy revolution.
HD	Consumer Discretionary	Market leader in home improvement, benefiting from housing spending trends and resilient demand.
UNH	Healthcare	Diversified healthcare provider benefiting from long-term trends in aging demographics.
LLY	Healthcare	Biopharma giant with strong pipeline of obesity and diabetes treatments like Mounjaro.
JNJ	Healthcare	Stable dividend aristocrat with exposure to pharmaceuticals, medical services, and consumer healthcare.
JPM	Financial	Leading global bank with diversified revenue streams, strong management, and exceptional risk controls.
V	Financial	Global leader in payment processing and a key player in the shift to digital payments.
NEE	Energy & Utilities	Heavy focus on clean energy and renewable infrastructure.
XOM	Energy & Utilities	Integrated energy leader with strong free cash flow from oil and gas while investing in clean energy.
CAT	Industrial	Dominant player in construction and mining equipment benefiting from increased global infrastructure spending.

TABLE IV
PORTFOLIO GIVEN BY LLAMA 3.1 405B.

Ticker	Sector	Rationale
AAPL	Technology	Strong brand, innovative products, and growing services segment.
MSFT	Technology	Dominant player in cloud computing, AI and gaming.
NVDA	Technology	Leader in graphics processing units (GPUs) and AI computing.
AMZN	Technology	E-commerce giant with growing cloud services and advertising business.
UNH	Healthcare	Largest health insurer in the US, with a strong track record of growth.
JNJ	Healthcare	Diversified healthcare company with a strong pharmaceutical pipeline.
PFE	Healthcare	Leading pharmaceutical company with a strong pipeline of innovative medicines.
DIS	Consumer Discretionary	Media and entertainment giant with a strong brand and growing streaming services.
MCD	Consumer Discretionary	Fast-food chain with a strong broad and growing digital sales.
NKE	Consumer Discretionary	Leading athletic apparel brand with a strong track record of innovation.
3M	Industrial	Diversified industrial company with a strong track record of innovation.
CAT	Industrial	Leading heavy machinery manufacturer with a strong global presence.
XOM	Energy	Integrated energy company with a strong record of dividend payments.
NEE	Energy	Leading renewable energy company with a strong growth potential.

TABLE V
1/N PORTFOLIO FROM LLMS.

LLM	N	Mean	Std.	Skew.	Kurt.	CV	VaR@95	CVaR@95
OpenAI-o1	15	0.0751%	0.0137	-0.1812	12.42	18.23	0.0380	0.0561
Claude	15	0.0851%	0.0129	-0.4101	14.59	15.23	0.0358	0.0563
GPT-4	15	0.1054%	0.0147	-0.5882	11.51	13.96	0.0422	0.0626
LLAMA	14	0.0686%	0.0131	-0.2788	13.63	19.06	0.0378	0.0549

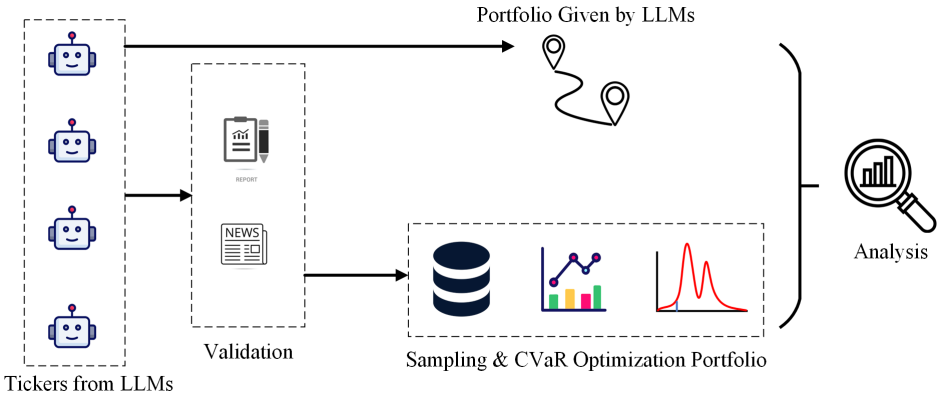

Fig. 1. Flowchart of the proposed portfolio framework.

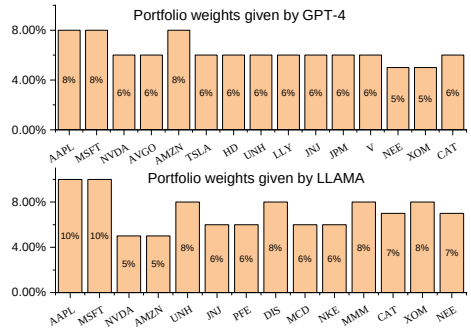

Fig. 2. Portfolio weights directly given by GPT-4 & LLAMA.

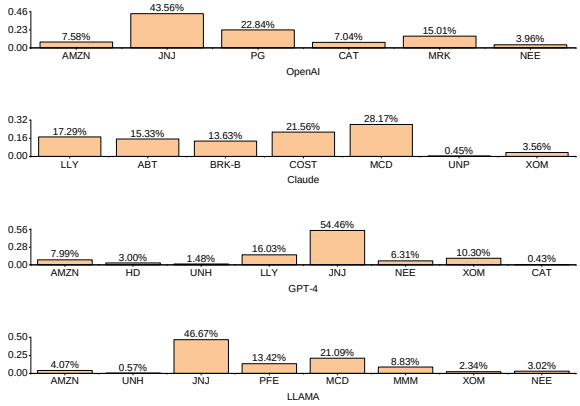

Fig. 3. CVaR portfolio weights of LLMs.

specified investment models, due to the resulted weights are almost the same as the equal weight scheme.

According to the performance of $1/N$ strategy [14] applied on LLMs, GPT-4 gives the portfolio with the highest mean return of 0.1054%, as well as the largest standard deviation of 0.0147. While Claude presents the most stable portfolio with the lowest volatility of 0.0129, but its return ranks the second highest position of 0.0851 among the 4 LLMs. All of the 4 portfolios show negative-skewed, indicating possible long left tail and high chance of large losses. The portfolios kurtosis values suggest the non-normality and high probability of extreme outliers, for example, large gain or losses. CV illustrates the relative risk by comparing the standard deviation to the mean return, that is, $\text{CV} = \frac{\sigma}{\mu}$. The portfolio recommended by GPT-4 has the lowest CV of 13.96, which is appealing to rational investors, whereas the portfolio given by LLAMA shows the highest CV of 19.06, requiring high risk tolerance for investors.

Tail risk can be reflected via VaR@95 and CVaR@95, with $\alpha = 0.95$. Claude's $1/N$ portfolio achieves the lowest VaR@95 of 0.0358 while the second highest CVaR@95 of 0.0563. LLAMA reaches the lowest CVaR@95 of 0.0549 and the second lowest VaR@95 of 0.0378. As far as the $1/N$ portfolio concerned, the tail risk is roughly consistent with the volatility level according to Table V.

B. CVaR-based portfolios

In CVaR-based modelling, we add the return constraint $R_t \geq 0$ for economic sense, due to negative portfolio return is meaningless to rational investors. Fig. 3 presents portfolio weight of each LLM recommendation, where the tickers with nearly zero weights are omitted. Note that GPT-4 and LLAMA consider the most possible constituent stocks, where 8 tickers are covered. To further measure the diversification of portfolio, the following two indicators are introduced, entropy and Herfindahl-Hirschman Index (HHI).

TABLE VI
WEIGHT ANALYSIS OF CVAR@95.

LLM	N	Entropy	HHI
OpenAI-o1	6	1.4941	0.2767
Claude	7	1.6933	0.1991
GPT-4	8	1.4258	0.3444
LLAMA	8	1.5211	0.2912

Entropy evaluates how evenly the portfolio's capital is distributed among different assets, which can be measured by the formula of $H = -\sum_{i=1}^{n} w_i \ln(w_i)$. Likewise, HHI also quantifies the diversification level of an investment, whose formula is $\text{HHI} = \sum_{i=1}^{n} w_i^2$. Table VI gives the weight analysis of the constructed CVaR portfolios, where Claude has the highest entropy of 1.6933 with the lowest HHI of 0.1991, illustrating highly diversified portfolio. However, GPT-4 shows the lowest entropy of 1.4258 with the highest HHI of 0.3444, indicating relatively high concentration.

Fig. 4 calculates the corresponding entropy and HHI using different α levels in CVaR optimization portfolios.

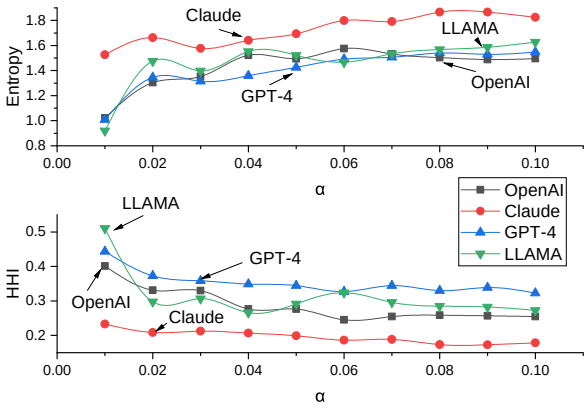

Fig. 4. Entropy & HHI of CVaR@95 using different α.

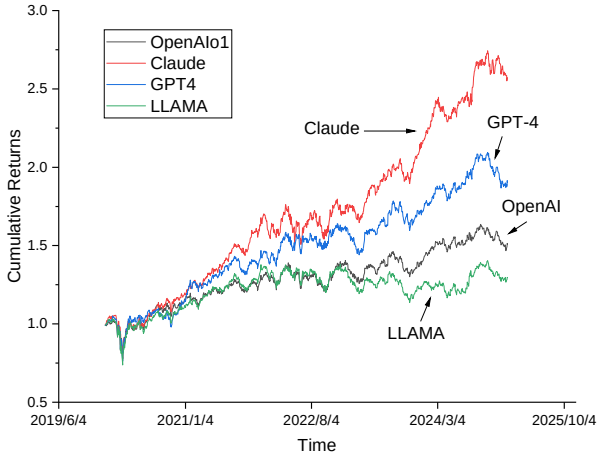

Fig. 5. Cumulative returns of CVaR@95 portfolios.

TABLE VII
PORTFOLIO PERFORMANCE OF CVAR@95.

LLM	APY	STD	SR	MDD	CR
OpenAI-o1	0.0962	0.1695	0.5673	0.2451	0.3924
Claude	0.2043	0.1818	1.1241	0.2597	0.7868
GPT-4	0.1440	0.1802	0.7991	0.2745	0.5244
LLAMA	0.0666	0.1737	0.3837	0.2852	0.2337

As far as the portfolio diversification concerned, Claude shows the clear superiority among the four advanced LLMs, where significantly high level of entropy and low HHI can be observed.

C. Performance analysis

Some risk-adjusted financial indicators such as Sharpe ratio (SR), and Calmar ratio (CR) could be used for comprehensively evaluating the constructed portfolios, whose formulas are as follows:

$$\text{SR} = \frac{\text{APY}}{\text{STD}}$$

where APY is the portfolio annual return, and STD is the portfolio annual standard deviation (volatility).

$$\text{CR} = \frac{\text{APY}}{\text{MDD}}$$

where MDD is the maximum drawdown, which is a key metric that measures the largest peak-to-through decline of an investment over a given period before new peak is achieved:

$$\text{MDD} = \frac{\max(\text{Peak Value} - \text{Trough Value})}{\max(\text{Peak Value})}$$

It could assess downside risk by quantifying how much a portfolio can lose from its highest point to the lowest point before recovering. Table VII presents the associated financial indicators for the CVaR@95 portfolio optimization, where Claude achieves the highest SR of 1.1241 and the highest CR of 0.7868, giving the best comprehensive portfolio performance.

OpenAI provides the most conservative portfolio, whose volatility is 0.1695 and MDD is 0.2451. Both of them are the lowest level among the four LLMs. For the conservative investors prioritizing capital preservation over high returns, OpenAI-o1 is a trustworthy LLM. However, LLAMA exhibits disappointing portfolio performance, neither its risk nor return indicator is the worst among the four LLMs.

Fig. 5 visualizes the cumulative returns of the LLM-based portfolios, where the CVaR@95 portfolio using tickers recommended by Claude dominates other CVaR@95 portfolios significantly. Table VIII summarizes the descriptive statistics of the proposed CVaR@95 portfolios. Comparing to the $1/N$ strategy, CVaR portfolio presents better ability on risk-control, where lower CVaR@95, VaR@95, and standard deviation can be reached via the optimization procedure. However, $1/N$ strategy shows higher daily return than CVaR portfolio, demonstrating its robustness.

Figs. 6 & 7 consider different combinations of (α, R_t) in modelling the CVaR optimization portfolio, and the corresponding SR values are calculated. In total, Claude gives the CVaR portfolios with relatively high level of SR values, where the robust performance can be observed. GPT-4 has the potential to obtain high SR values when the R_t is set around 0.0015. The poor performance of LLAMA's CVaR@95 portfolio can be forgave due to it can reach acceptable SR level using higher (α, R_t) combinations.

VI. CONCLUSIONS

CVaR optimization is a widely adopted approach for mitigating portfolio tail risk. Its effectiveness is evident when comparing the $1/N$ strategy with the CVaR@95 portfolio, where the latter achieves a lower risk level than the equal-weighted benchmark.

In this study, four prominent LLMs, OpenAI-o1, Claude, and GPT-4, are evaluated using identical prompts. OpenAI-o1, Claude, and GPT-4 each construct portfolios comprising 15 constituent stocks, meeting the investor's requirements. In contrast, LLAMA selects only 14 stocks, failing to satisfy the specified criteria. Numerical results show that the CVaR@95 optimization portfolio based on Claude is the most diversified strategy, also achieves the best risk and return indicators such SR and CR. Nonetheless, LLAMA does not provide wise financial suggestions since the CVaR@95 portfolio based on its recommendation presenting unsatisfactory performance.

TABLE VIII
CVaR PORTFOLIO FROM LLMS.

LLM	Mean	Std.	Skew.	Kurt.	CV	VaR@95	CVaR@95
OpenAI-o1	0.0381%	0.0106	0.4187	11.78	27.98	0.0275	0.0447
Claude	0.0811%	0.0114	-0.1529	14.65	14.12	0.0296	0.0485
GPT-4	0.0571%	0.0113	0.3408	10.21	19.86	0.0284	0.0455
LLAMA	0.0264%	0.0109	0.1859	13.79	41.37	0.0267	0.0459

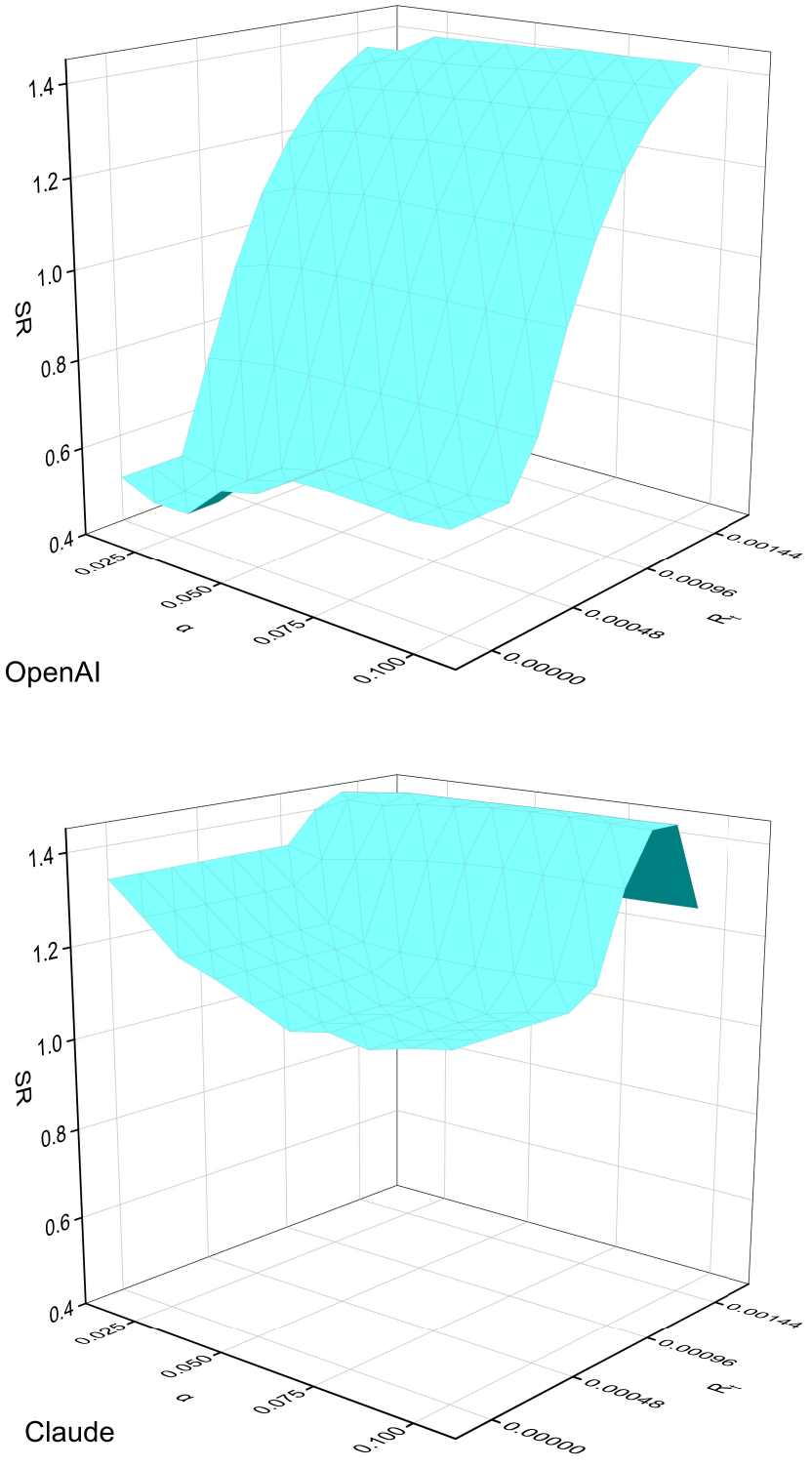

Fig. 6. LLM-CVaR portfolio Sharpe ratios with different combinations of (α, R_t).

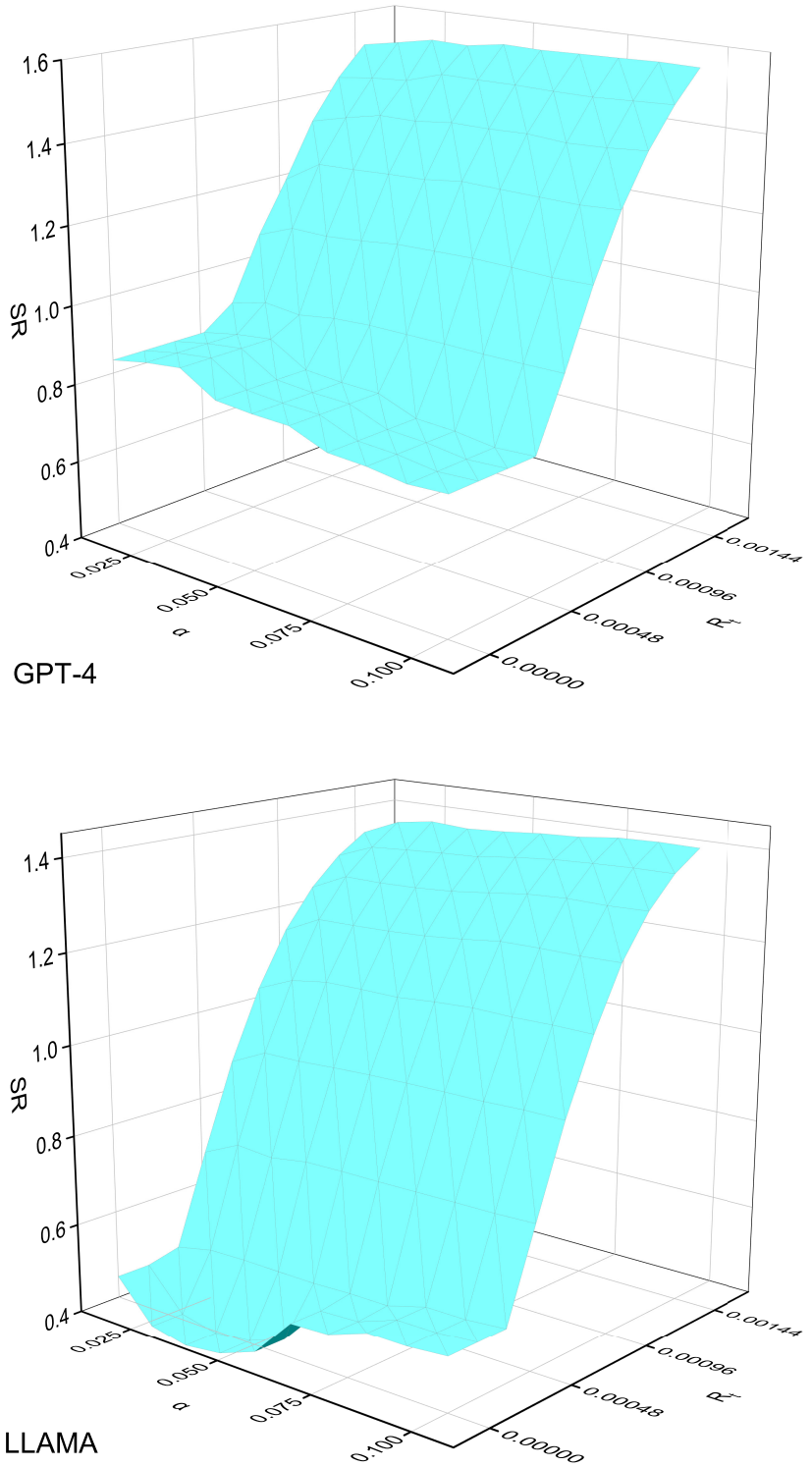

Fig. 7. LLM-CVaR portfolio Sharpe ratios with different combinations of (α, R_t).

Future research would investigate the CVaR portfolio with multiple investment objectives. Also, we will compare the performance with the solutions directly given by some advanced LLMs.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the associated professor from the School of Information Management and Engineering, Shanghai University of Finance & Economics, Jianjun Gao, for his professional guidance. The authors also thank the professor from the School of Information Management and Engineering, Shanghai University of Finance & Economics, Dongmei Han, for her kindly help.

REFERENCES

[1] O. Romanko, A. Narayan, and R. H. Kwon, "Chatgpt-based investment portfolio selection," in *Operations Research Forum*, vol. 4, no. 4. Springer, 2023, p. 91.

[2] X. Kong, L. Min, D. Lin, and Z. Li, "Black-litterman portfolio optimization with asset universe given by large language models." *IAENG International Journal of Computer Science*, vol. 51, no. 8, pp. 976–984, 2024.

[3] S. Cho, "Can chatgpt generate stock tickers to buy and sell for day trading?" *Available at SSRN 4759311*, 2024.

[4] E. Callanan, A. Mbakwe, A. Papadimitriou, Y. Pei, M. Sibue, X. Zhu, Z. Ma, X. Liu, and S. Shah, "Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams," *arXiv preprint arXiv:2310.08678*, 2023.

[5] P. Niszczota and S. Abbas, "Gpt has become financially literate: Insights from financial literacy tests of gpt and a preliminary test of how people use it as a source of advice," *Finance Research Letters*, vol. 58, p. 104333, 2023.

[6] S. Wang, H. Yuan, L. Zhou, L. M. Ni, H.-Y. Shum, and J. Guo, "Alpha-gpt: Human-ai interactive alpha mining for quantitative investment," *arXiv preprint arXiv:2308.00016*, 2023.

[7] H. Ko and J. Lee, "Can chatgpt improve investment decisions? from a portfolio management perspective," *Finance Research Letters*, vol. 64, p. 105433, 2024.

[8] R. Ullah, H. B. Ismail, M. T. I. Khan, and A. Zeb, "Nexus between chat gpt usage dimensions and investment decisions making in pakistan: Moderating role of financial literacy," *Technology in Society*, vol. 76, p. 102454, 2024.

[9] A. Oehler and M. Horn, "Does chatgpt provide better advice than robo-advisors?" *Finance Research Letters*, vol. 60, p. 104898, 2024.

[10] Y. Zinchenko and A. V. Asimit, "Modeling risk for cvar-based decisions in risk aggregation," *Journal of Risk and Financial Management*, vol. 16, no. 5, p. 266, 2023.

[11] Q. Xu, Y. Zhou, C. Jiang, K. Yu, and X. Niu, "A large cvar-based portfolio selection model with weight constraints," *Economic Modelling*, vol. 59, pp. 436–447, 2016.

[12] S. Lotfi and S. A. Zenios, "Robust mean-to-cvar optimization under ambiguity in distributions means and covariance," *Review of Managerial Science*, pp. 1–26, 2024.

[13] A. N. Elmachtoub and P. Grigas, "Smart "predict, then optimize"," *Management Science*, vol. 68, no. 1, pp. 9–26, 2022.

[14] Victor, DeMiguel, Lorenzo, Garlappi, Raman, and Uppal, "Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?" *Review of Financial Studies*, 2009.