

Leveraging Large Language Models for Institutional Portfolio Management: Persona-Based Ensembles

Yoshia Abe

Graduate School of Information Science and Technology
The University of Tokyo
Tokyo, Japan
y-abe@isi.imi.i.u-tokyo.ac.jp (0009-0007-0885-8852)

Shuhei Matsuo

Graduate School of Information Science and Technology
The University of Tokyo
Tokyo, Japan
matsuo@isi.imi.i.u-tokyo.ac.jp

Ryoma Kondo

Graduate School of Information Science and Technology
The University of Tokyo
The Canon Institute for Global Studies
Tokyo, Japan
kondor@g.ecc.u-tokyo.ac.jp

Ryohei Hisano

Graduate School of Information Science and Technology
The University of Tokyo
The Canon Institute for Global Studies
Tokyo, Japan
hisano@g.ecc.u-tokyo.ac.jp

Abstract—Large language models (LLMs) have demonstrated promising performance in various financial applications, though their potential in complex investment strategies remains underexplored. To address this gap, we investigate how LLMs can predict price movements in stock and bond portfolios using economic indicators, enabling portfolio adjustments akin to those employed by institutional investors. Additionally, we explore the impact of incorporating different personas within LLMs, using an ensemble approach to leverage their diverse predictions. Our findings show that LLM-based strategies, especially when combined with the mode ensemble, outperform the buy-and-hold strategy in terms of Sharpe ratio during periods of rising consumer price index (CPI). However, traditional strategies are more effective during declining CPI trends or sharp market downturns. These results suggest that while LLMs can enhance portfolio management, they may require complementary strategies to optimize performance across varying market conditions.

Index Terms—Large language models, Finance, Prompt engineering, Persona, Ensemble method, Portfolio management

I. INTRODUCTION

Large language models (LLMs) exhibit a wide range of capabilities that extend beyond traditional natural language processing tasks. In the financial sector, LLMs are increasingly employed to enhance decision-making and improve operational efficiency. For example, BlackRock has explored innovative methods for classifying companies using LLMs [1]. Similarly, [2] used LLMs to extract structured environmental, social, and governance (ESG) data from sustainability reports to build a knowledge graph that facilitates deeper analysis of corporate sustainability practices. LLMs have also been employed to detect accounting fraud in the Management Discussion and Analysis sections of 10-K reports, surpassing existing benchmark models [3]. These examples and others [4] illustrate the expanding role of LLMs in finance, though

further exploration is required to understand their full potential in more complex investment strategies.

Narrowing the focus to investment-related applications, LLMs have shown promising results in various aspects of portfolio management. For instance, [5] demonstrated that assets selected by GPT models outperform randomly chosen assets in terms of diversification and average return. Although GPT excels in stock selection, optimization models perform better at portfolio allocation, prompting researchers to propose a strategy that combines the strengths of both [6]. Additionally, [7] showed that GPT models can create economically explainable factors based solely on their knowledge base, leading to the development of a new model based on these factors. However, despite these advances, most research remains focused on portfolio management at the individual stock level, with limited attention paid to institutional investors, who often manage portfolios at a more granular and complex level.

In finance, understanding investor attitudes is crucial, as decisions are shaped by beliefs, values, and preferences. Individual investors, for example, often make short-term decisions [8], whereas institutional investors typically adopt a long-term perspective and are less influenced by behavioral biases [9]. Additionally, research shows that gender differences influence risk perception and management among investment professionals, with women placing greater emphasis on risk reduction, particularly in extreme scenarios [10]. Modeling these diverse investment attitudes through LLMs could provide a powerful tool for personalizing financial strategies, especially for institutional investors who must account for various needs and behaviors when managing large portfolios.

Interestingly, just as individual investors vary in their preferences and performance, LLMs exhibit significant variation in their outputs depending on the specified persona [11], [12]. For example, [13] introduced DR-CoT prompting, in which LLMs use personas to mimic the diagnostic reasoning processes of

We would like to express our gratitude to The University of Tokyo Data Science School and The University of Tokyo Data Science Practicum (<https://dss.i.u-tokyo.ac.jp/>) for supporting this project.

medical professionals. Similarly, [14] employed personas (e.g., buyer, seller, and critic) in a gaming environment to evaluate whether LLMs can autonomously enhance their strategies through iterative interactions and mutual feedback. While these personas have proven effective in other fields, their application in finance, particularly in replicating investor attitudes or investment strategies, remains underexplored. Leveraging this capability could provide novel insights into portfolio management, especially in institutional settings, where nuanced or mixed decision-making is critical.

Building on this foundation, we explore the task of inputting economic indicator data into LLMs to predict the price movements of a portfolio consisting of stocks and government bonds, adjusting positions based on the obtained predictions in a manner similar to the portfolio management of institutional investors. Additionally, we investigate how the performance of LLMs varies depending on the specified persona, applying these differences in an ensemble approach to construct the final portfolio. Because the effectiveness of investment strategies can vary depending on the testing period, we compare LLM-based portfolio management strategies with baseline models in detail, analyzing the conditions under which LLMs are most effective. Furthermore, we qualitatively examine the reasoning behind the LLM’s predictions, enabling a deeper understanding of the decision-making process and the key information it focuses on.

We find that LLM-based predictions, particularly when combined with the ensemble approach, detect market declines well. Moreover, LLM-based investment strategies outperform the buy-and-hold strategy in terms of the Sharpe ratio during periods of rising consumer price index (CPI), whereas buy-and-hold performs better during a declining CPI. For other metrics, such as return, volatility, and maximum drawdown, different strategies tend to perform best depending on market conditions. Additionally, LLM-based strategies generally respond effectively to sharp market declines by reducing positions, though traditional strategies can offer better protection during rapid downturns.

The contributions of this study are as follows:

- 1) Prompts that enable LLMs to manage portfolios in line with institutional investor settings are designed.
- 2) Differences in the performance of portfolio strategies based on LLM personas are investigated and leveraged via ensemble methods.
- 3) Periods when LLM-based strategies excel are quantitatively analyzed and the LLM-generated rationales for each persona are qualitatively analyzed.
- 4) LLM-based strategies are shown outperform traditional methods in Sharpe ratio during rising CPI trends.

II. RELATED WORK

As mentioned above, LLMs have seen widespread application in finance in recent years. The related research can be broadly categorized into three areas: financial concept comprehension, academic applications, and investment decision-making [15]. Research on investment decision-making can be

further divided into studies focused on individual investors and those focused on institutional investors, with most current studies concentrating on the former. One of the few studies focusing on institutional investors is [7], which used GPT-4 to generate high-return equity investment factors, achieving an annualized return of up to 88% and a Sharpe ratio of 2.46, significantly outperforming traditional models. However, their approach is limited to stock prices and does not predict price movements in a portfolio consisting of both stocks and government bonds, nor does it adjust positions based on these predictions in a manner consistent with institutional portfolio management.

We also review prompt engineering, which is widely recognized as a crucial step in enhancing the capabilities of LLMs. The studies [12], [16] demonstrated that specifying a persona; providing concrete examples of investment strategies; and clearly defining the objective, output content, and format significantly improve response quality. In particular, studies on personas have shown that adjusting the explanation level based on the persona of the intended audience can be effective [17]. However, the focus was on financial concept comprehension, not comparing investment performance when the persona of the institutional investor was altered.

III. METHODS

A. Task Definition

We investigate the ability of LLMs to first predict price movements in a portfolio consisting of stocks and government bonds, and then adjust positions based on these predictions in a manner similar to institutional portfolio management. The portfolio is composed of 40% US equities and 60% US bonds. The LLM is tasked with predicting whether the portfolio value will rise or fall by more than 2% within the next 5 days based on data from the previous 10 days for the following seven numerical data indicators:

- A: 40% Stock, 60% Bond Portfolio (Return)
- B: US Stocks (Futures, Return)
- C: US 5-year Interest Rate
- D: US 30-year Interest Rate
- E: US Interest Rate Spread, 30–10 year
- F: Volatility Index (VIX)
- G: US Dollar Index

The output prediction is a three-class classification task, where “0,” “1,” and “2” indicate holding, falling, and rising, respectively. While various LLMs are available, we use GPT-4 (gpt-4-0613) [18], which is known for its strong performance across many tasks.

B. Prompt Design

The performance of LLMs varies significantly depending on prompt design, which has led to the development of various prompting methods [19], [20]. Table I lists common prompting methods. We use methods 1–9 to minimize the number of interactions with LLMs. The remaining methods require substantially more tokens and interactions with the LLMs, and hence we leave these for future work.

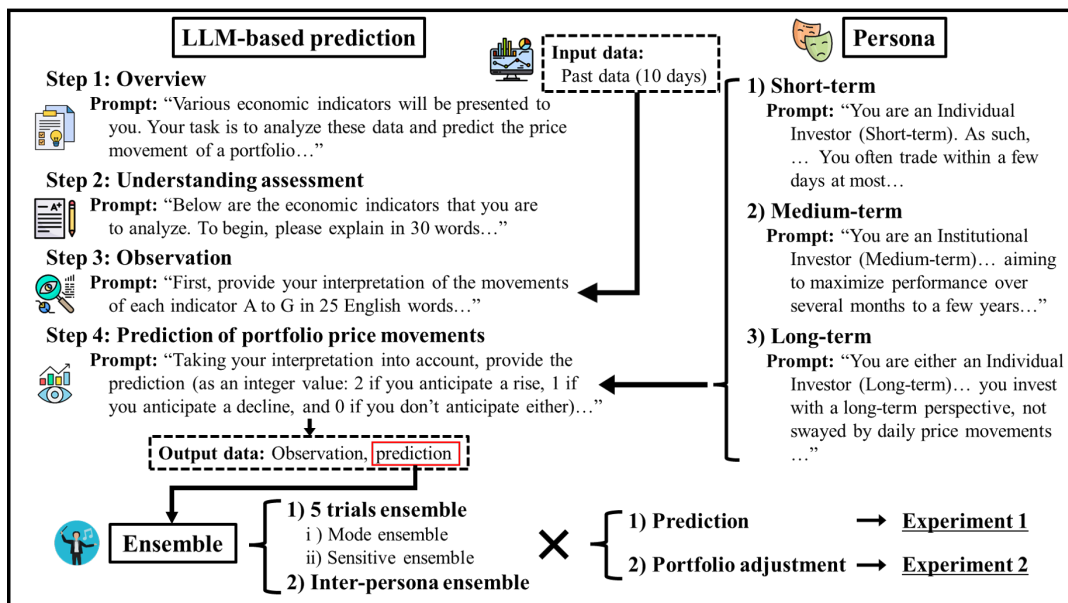


Fig. 1. Overview of our approach.

The following sequence of prompts was used in the experiment: First, an overview of the task and relevant considerations is provided as a system attribute prompt. Next, the LLM is queried to assess its understanding of various economic indicators. Then, numerical data from the past 10 days for various economic indicators are input, and the model is asked to interpret the trends. Finally, based on these observations and interpretations, the model is tasked with predicting the rise or fall of a portfolio. Our approach is illustrated in Fig. 1. The complete version of the prompt is available on our GitHub page¹.

IV. EXPERIMENTS

A. Overview of the Experimental Procedure

We use GPT-4 (gpt-4-0613), which was trained up to September 2021, in all of our experiments. To avoid data leakage and properly evaluate the predictive performance of the LLM, we focused on the period from October 2021 to January 2024, covering 593 weekdays. Moreover, we repeated the predictions five times to account for variability in the LLM outputs.

We conducted two experiments. Because LLM inference capabilities are heavily influenced by the persona specified in the prompt [12] and investment strategies often vary based on individual beliefs, values, and aims, we sought to leverage this heterogeneity. In the first experiment, we examined how different personas affect prediction accuracy. The performance was evaluated across multiple trials using metrics such as accuracy, precision, recall, and F1-score. Additionally, we assessed the performance of several ensemble methods to determine whether the predictions from different personas could be effectively integrated.

In the second experiment, we developed an investment strategy based on the LLM's price movement predictions and compared it with a baseline strategy that did not use an LLM. The evaluation period was divided into months, each characterized by economic indicator data (US CPI Total), and we determined the periods over which the LLM-based strategies performed better.

B. Experiment 1: Impact of Personas and Ensembles on Prediction Accuracy

We investigated how different personas affect prediction accuracy and evaluated performance based on the time span of their investments using the following three persona conditions:

- 1) Short term:** An individual investor who trades over a span of several days, with limited knowledge of investment and challenges in risk management.
- 2) Medium term:** An institutional investor trading over several months to a few years, with extensive knowledge of investment and robust risk management capabilities.
- 3) Long term:** Both individual and institutional investors who operate with a long-term perspective, spanning 20 to 30 years.

Additionally, we considered the following two ensemble methods to account for the variability in LLM predictions:

- 1) Mode:** The final prediction is the class with the most votes across five trials. In the case of a tie, the class with the smaller number is prioritized.
- 2) Sensitive:** If Class 2 (rise) or Class 1 (fall) appears in any of the five trials, that class is chosen as the final prediction. If both Class 1 and Class 2 are present, the class with the most votes is selected. In the event of a tie, Class 0 is chosen. We call this method "sensitive" because it is more likely to predict a rising or falling market than the mode method.

¹https://github.com/YoshiaAbe/llm_based_portfolio_management

TABLE I
PROMPTING METHODS. CHECK MARK ✓ INDICATES THAT THE METHOD WAS USED IN THE EXPERIMENTS IN THIS STUDY.

No.	Name and Explanation	Used
1	Use clear and concrete instructions [19], [20].	✓
2	Use delimiters such as ### [19], [20].	✓
3	Specify the output length, style, and similar factors in detail. [19], [20]	✓
4	Specify the output format, e.g., in JSON [19], [20].	✓
5	Use "Refrain from ..." instead of "Don't do ..." [19].	✓
6	Specify the persona that you want the LLM to behave as [12], [20].	✓
7	Ask the model to output its thoughts before its conclusion [20].	✓
8	Chain-of-Thought [21], Least-to-Most [22]: decompose the task into sub-tasks.	✓
9	Zero-shot Chain-of-Thought [23]: tell the LLM to "think step-by-step."	✓
10	Make the LLM evaluate whether it met the specified instruction [20].	
11	EchoPrompt [24]: make the LLM rephrase the question before answering.	
12	Few-shot Prompting [25]: show the LLM examples of correct answers.	
13	Contrastive Chain-of-Thought [26]: show the LLM examples of incorrect reasoning.	
14	Self-consistency [27]: the final answer is the majority choice among multiple outputs.	
15	Self-refine [28]: make the LLM provide feedback about itself and iteratively refine its output.	
16	Tree of Thoughts [29]: Manage and explore a chain of thought in a tree structure.	

C. Experiment 2: Adaptive Investment Strategies with LLM Predictions

In the second experiment, we evaluated the performance of investment strategies based on the predictive outputs from Experiment 1. Our strategy involved adjusting the position size of stocks and government bonds, ranging from 0.0 to 1.0, in increments of 0.2. The investment period spanned 593 weekdays, from October 2021 to January 2024, as in Experiment 1, with the position starting at 1.0 on the first day. For evaluation purposes, and to ensure sufficient data within each month, we focused on the 26 months between November 2021 and December 2023, excluding the initial and final months of the evaluation period used in Experiment 1.

We used the following actions for our investment strategy:

- 1) **Pattern 1:** Binary actions (increase or decrease position size);
- 2) **Pattern 2:** Ternary actions (increase, decrease, or maintain position size);
- 3) **Pattern 3:** Ternary actions (increase, decrease, or maintain position size), with adjustments.

In the first pattern, the position size decreases when the predicted class is 1 and increases when the predicted class is 0 or 2. Pattern 2 decreases the position size when the predicted class is 1, increases it when the predicted class is 2, and remains unchanged when the predicted class is 0. The final pattern, pattern 3, is similar to pattern 2 but differs in that if there is no change in position size for the past d_{flat} days, the position is gradually increased to 1.0 in steps of 0.2. In this study, d_{flat} was fixed at 5 days.

For comparison, we evaluated our strategy against the following baselines:

- 1) **Buy-and-hold:** This strategy maintains a constant position of 1.0 starting from the first day and continuing throughout the period.
- 2) **Continuous movement (CM):** This strategy tracks indicator A (40% stock, 60% bond portfolio return values) over the past d_{window} days. If the value rises continuously for $d_{\text{continuity}}$ days, the position is increased; if

it falls continuously, the position is decreased. In this experiment, d_{window} was fixed at 10 days and two values of $d_{\text{continuity}}$ were tested: 2 and 3 days (CM(D2) and CM(D3), respectively).

- 3) **Regression (RG):** This strategy also observes indicator A over the past d_{window} days and performs a linear regression with the days as the explanatory variable and indicator A values as the dependent variable. If the slope of the regression line is positive, the position is increased; if it is negative, the position is decreased. However, if the absolute value of the slope is below $s_{\text{threshold}}$, the position is not adjusted. In this experiment, d_{window} was fixed at 10 days and two values of $s_{\text{threshold}}$ were tested: 0.001 and 0.0005 (RG(S10) and RG(S5), respectively).

We use the following four metrics to evaluate the portfolio management strategies:

Return: The cumulative return R_{cumul} over period T is given by $R_{\text{cumul}} = \{\sum_i^n (1 + p_i r_i)\} - 1$, where p_i and r_i represent the position and return on the i -th day, respectively. The return metric is adjusted by dividing it by the average position size: $\frac{1}{\bar{p}} R_{\text{cumul}}$.

Volatility: The volatility V over period T is the standard deviation of $p_i r_i$, scaled by the square root of the number of trading days: $V = \sqrt{\frac{1}{n} \sum_i^n \{(p_i r_i) - \bar{p}_i r_i\}^2} \sqrt{n}$. The volatility metric is adjusted by dividing it by the average position size: $\frac{1}{\bar{p}} V$.

Maximum drawdown: The maximum drawdown D_{max} during period T is the largest decline in asset value from a previous peak. Let $R_{\text{cumul},i}$ represent the cumulative return up to day i . The drawdown D_i is calculated as $D_i = R_{\text{cumul}} - \max(R_{\text{cumul},1}, R_{\text{cumul},2}, \dots, R_{\text{cumul},i})$. The maximum drawdown is the lowest value of D_i over n days, multiplied by -1 : $D_{\text{max}} = -\min(D_1, D_2, \dots, D_n)$. The maximum drawdown metric is adjusted by dividing it by the average position size: $\frac{1}{\bar{p}} D_{\text{max}}$.

Sharpe ratio: The Sharpe ratio S for period T is calculated as $\frac{1}{\bar{V}} R_{\text{cumul}}$ and is used as the evaluation metric.

V. RESULTS

A. Results of Experiment 1: Impact of Personas and Ensembles on Prediction Accuracy

First, we compare the predictive accuracy of each persona in Fig. 2. The average accuracies across five trials for the short, medium, and long personas were 0.345, 0.333, and 0.352, respectively. When applying the mode ensemble within the same persona across trials, the accuracy improved to 0.361, 0.339, and 0.378, respectively. In contrast, the sensitive ensemble resulted in lower accuracy values of 0.314, 0.312, and 0.317, respectively. For all three personas, the mode ensemble consistently improved performance compared with the average values, whereas the sensitive ensemble caused performance to decline.

Second, we examine the performance of ensembles across different persona. For each persona, we generated prediction results using the mode or sensitive ensemble across the five trials. We then ensembled these predictions across the three personas, using either the mode or sensitive method, resulting in a total of four patterns (2x2). The results are presented in Table II. Among the four patterns, the highest accuracy of 0.366 was achieved by applying the mode ensemble across both the five trials and the three personas. The highest individual accuracy was obtained using the mode ensemble for the five trials of the long-term persona, which resulted in an accuracy of 0.378, as shown in the bottom panel of Fig. 2.

To further evaluate the predictive accuracy of the LLMs, we report the precision, recall, F1-score, correct counts, and predicted counts for each class in Table III. The results indicate that, while approximately half of the correct labels fall into class 0, the LLM tends to predict class 1 more frequently. Given that the chance level of accuracy when randomly selecting a class from 0, 1, or 2 with a uniform distribution is 0.333, it is clear that the LLM’s predictions outperform chance. Notably, the mode ensemble contributes to an improvement in overall accuracy.

TABLE II

ACCURACY OF INTER-PERSONA ENSEMBLES. ROWS AND COLUMNS INDICATE THE ENSEMBLE METHOD APPLIED ACROSS THE THREE PERSONAS AND WITHIN EACH PERSONA, RESPECTIVELY.

Inter-Persona Ensemble Method	Mode	Sensitive
Mode (across three personas)	0.366	0.324
Sensitive (across three personas)	0.319	0.307

TABLE III

PRECISION, RECALL, F1-SCORE, CORRECT COUNTS, AND PREDICTED COUNTS FOR EACH CLASS WHEN USING THE MODE ENSEMBLE FOR THE FIVE TRIALS OF THE LONG-TERM PERSONA.

Class	Precision	Recall	F1-score	Corr. Cnt.	Pred. Cnt.
0	0.551	0.327	0.411	281.0	167.0
1	0.380	0.570	0.456	172.0	258.0
2	0.202	0.243	0.221	140.0	168.0

For institutional investors, a model that predicts declines (class 1) with high precision and recall is particularly desirable

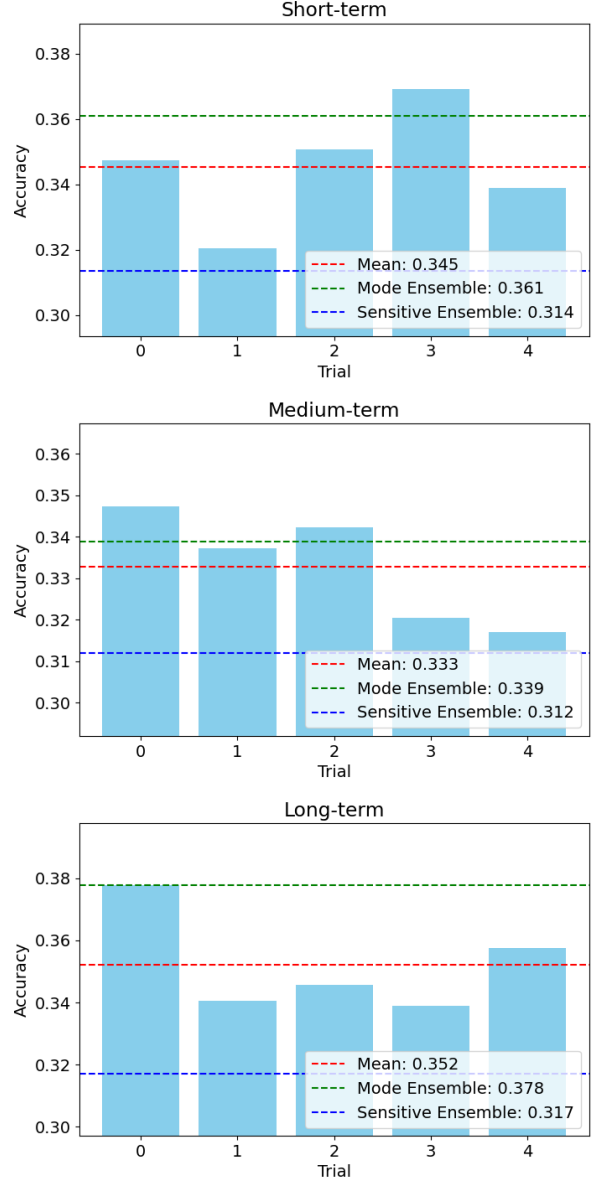


Fig. 2. Accuracy of five trials for each persona condition (short, medium, or long term)

to avoid significant losses. Therefore, we further investigated the F1-score for class 1, as shown in Fig. 3. The average F1-scores across five trials for the short-, medium-, and long-term personas were 0.449, 0.451, and 0.423, respectively. When applying the mode ensemble within the same persona across trials, the scores improved to 0.474, 0.479, and 0.456, respectively. The sensitive ensemble yielded scores of 0.469, 0.468, and 0.452, respectively. Unlike accuracy, the F1-scores of both the mode and sensitive ensembles were better than the average values.

The F1-score performance when using ensembles across different personas is presented in Table IV. The highest F1-

score (0.484) was achieved by applying the mode ensemble across the five trials, followed by a mode ensemble across the three personas. The precision, recall, F1-score, correct counts, and predicted counts for each class in this case are listed in Table V. Notably, while maintaining a precision of 0.378, the recall reached 0.674.

On the basis of these results, we conclude that the mode ensemble method, which uses majority voting for the final predictions, improves both accuracy and F1-score. Moreover, applying the mode ensemble across personas proves to be the most effective approach when focusing on the F1-score for predicting a declining market, which is critical for institutional investors.

TABLE IV

F1-SCORE (CLASS 1) COMPARISON OF INTER-PERSONA ENSEMBLE. ROWS AND COLUMNS INDICATE THE ENSEMBLE METHOD APPLIED ACROSS THE THREE PERSONAS AND WITHIN EACH PERSONA, RESPECTIVELY.

Inter-Persona Ensemble Method	Mode	Sensitive
Mode (across three personas)	0.484	0.466
Sensitive (across three personas)	0.477	0.461

TABLE V

PRECISION, RECALL, F1-SCORE, CORRECT COUNTS, AND PREDICTED COUNTS FOR EACH CLASS WHEN USING THE MODE ENSEMBLE FOR THE FIVE TRIALS, FOLLOWED BY THE MODE ENSEMBLE ACROSS THE THREE PERSONAS.

Class	Precision	Recall	F1-score	Corr. Cnt.	Pred. Cnt.
0	0.556	0.263	0.357	281.0	133.0
1	0.378	0.674	0.484	172.0	307.0
2	0.176	0.193	0.184	140.0	153.0

B. Results of Experiment 2: Adaptive Investment Strategies with LLM Predictions

We first examine the Sharpe ratio to evaluate overall performance, balancing profit and risk. Table VI presents the monthly Sharpe ratios for each strategy during the 26-month evaluation period. For each month, the best- and worst-performing strategies are highlighted in bold and underlined, respectively. Strategies that outperform the buy-and-hold strategy are indicated in red. The “nan” entries indicate periods where the Sharpe ratio could not be calculated due to zero volatility.

The LLM-based strategies (patterns 1, 2, and 3) used the prediction gained after the mode ensemble across the five trials, followed by the mode ensemble across the three personas. The red text reveals that these LLM-based strategies outperformed the buy-and-hold strategy in some periods but underperformed in others. Furthermore, examining the bold and underlined entries reveals that, for certain months, a strategy could be either the best or the worst performer among all strategies.

C. Results of Experiment 2: Comparison of Investment Strategies Across Different Time Periods

Evaluating which strategy performed best on a month-by-month basis, as described in the last subsection, does not fully

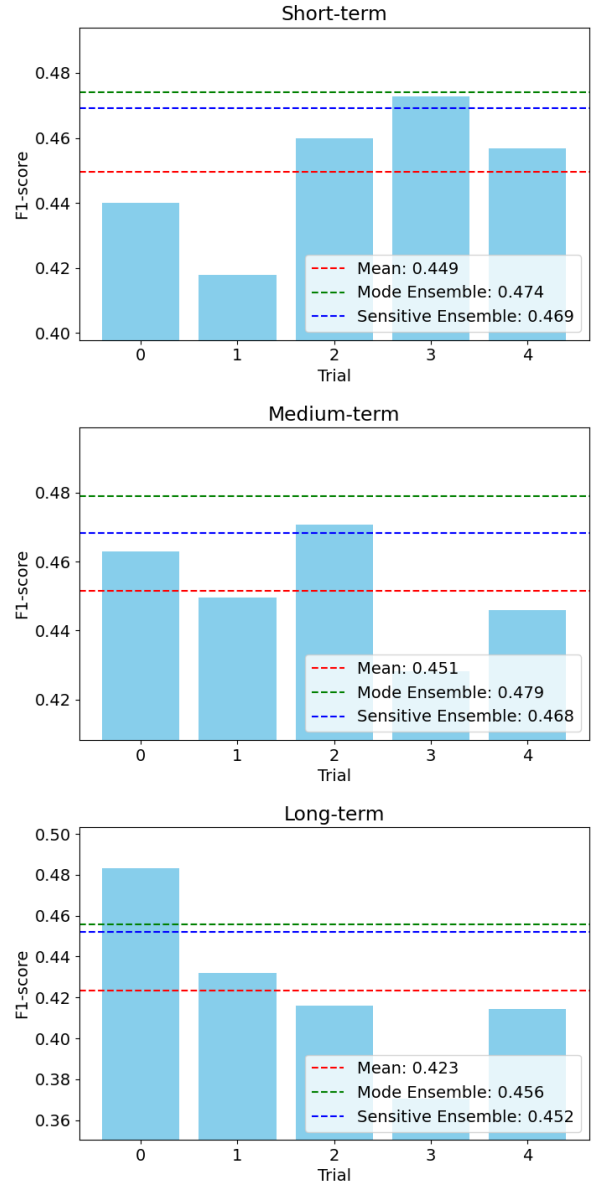


Fig. 3. F1-score (Class 1) of five trials for each persona condition (short term, medium term, and long term)

capture when the LLM-based portfolio management strategies are most effective. To address this, we conducted additional analysis, hypothesizing that the LLM-based strategies perform well during specific periods but not others.

Specifically, we used the US CPI Total indicator to divide the 26-month evaluation period into two types of periods, investigating the best-performing strategies in each. The periods were divided by calculating the 6-month moving average of the year-over-year change in the US CPI Total and determining whether it increased or decreased with respect to the previous month. As a result, the 26 months were categorized as follows: November 2021 to August 2022 and December 2023 were

TABLE VI

SHARPE RATIOS FOR EACH STRATEGY DURING THE EVALUATION PERIOD. FOR EACH MONTH, THE BEST-PERFORMING STRATEGY IS HIGHLIGHTED IN BOLD, WHEREAS THE WORST ONE IS UNDERLINED. STRATEGIES THAT OUTPERFORMED THE BUY-AND-HOLD STRATEGY ARE INDICATED IN RED.

Year-Month	Buy-and-hold	Pattern 1	Pattern 2	Pattern 3	CM(D2)	CM(D3)	RG(S10)	RG(S5)
2021-11	<u>0.097</u>	0.280	0.397	0.397	0.108	<u>0.097</u>	<u>0.097</u>	<u>0.172</u>
2021-12	1.003	0.559	0.424	<u>0.353</u>	0.574	0.736	1.003	0.709
2022-01	-1.421	-1.583	-1.547	-1.583	<u>-2.013</u>	-1.731	-1.421	-0.804
2022-02	-0.540	-1.092	<u>-1.489</u>	<u>-1.489</u>	-1.446	-0.690	-0.565	-0.860
2022-03	-0.196	-0.051	-0.306	-0.306	-0.731	-0.556	-0.614	-0.861
2022-04	-1.826	-1.432	-1.432	-1.432	-1.391	-1.167	<u>-1.943</u>	<u>-1.847</u>
2022-05	<u>0.136</u>	0.922	0.404	0.404	0.266	1.188	0.147	0.544
2022-06	-0.909	-1.521	-1.011	-1.011	-0.373	<u>-1.778</u>	-1.301	0.160
2022-07	2.116	2.034	1.993	1.993	<u>1.249</u>	1.557	2.131	1.797
2022-08	<u>-1.290</u>	-0.745	-0.833	-0.833	-0.324	-0.797	-1.091	-1.184
2022-09	-1.657	-0.846	-0.846	-0.846	-1.730	<u>-2.628</u>	-1.847	-1.269
2022-10	0.593	0.761	<u>-0.032</u>	<u>-0.032</u>	0.429	<u>-0.032</u>	0.320	0.074
2022-11	0.831	0.574	<u>-0.350</u>	-0.148	0.755	0.635	0.280	<u>-0.363</u>
2022-12	-1.236	-0.613	-0.975	-0.842	-1.190	-1.069	<u>-1.239</u>	-0.969
2023-01	1.613	1.424	1.312	1.360	0.880	<u>0.275</u>	1.601	1.436
2023-02	-1.164	-0.058	-0.415	-0.415	-0.216	-0.398	<u>-1.165</u>	-0.453
2023-03	1.539	1.020	<u>0.912</u>	<u>0.912</u>	1.552	1.507	1.584	1.373
2023-04	0.564	0.512	<u>0.155</u>	0.271	0.491	0.402	0.564	0.549
2023-05	-0.556	<u>-1.281</u>	-0.906	-0.711	0.110	nan	-0.556	-0.553
2023-06	0.802	<u>0.477</u>	<u>-0.243</u>	0.324	0.416	nan	0.802	0.568
2023-07	0.320	<u>-0.867</u>	-0.463	-0.663	-0.066	-0.053	0.320	-0.154
2023-08	-0.621	-0.012	-1.218	-1.218	<u>-1.970</u>	-1.173	-0.621	0.398
2023-09	-2.400	-2.214	-1.637	-1.637	<u>-2.521</u>	-2.430	-2.400	-1.589
2023-10	-0.935	-1.773	-0.376	-0.376	<u>-2.064</u>	-1.920	-0.895	-1.223
2023-11	2.187	<u>1.478</u>	<u>1.478</u>	<u>1.478</u>	1.757	1.520	1.832	1.499
2023-12	1.890	1.781	1.767	1.767	1.890	1.890	1.890	<u>0.886</u>

classified as upward trends (High), while September 2022 to November 2023 were classified as downward trends (Low).

We use the following two methods to compare the strategies:

- 1) **Best-mean:** The average value of the target performance metric was calculated during the High or Low period, and the strategy with the best average was selected.
- 2) **Win-ratio-buy-and-hold:** For each month in the High or Low period, the number of months in which a strategy outperformed the buy-and-hold strategy was counted, and the strategy with the highest win ratio was selected. The buy-and-hold strategy itself was excluded from this comparison.

Of the four evaluation metrics used, higher values are preferable for return and Sharpe ratio, while lower values are desirable for volatility and maximum drawdown.

Table VII presents the results. From the perspective of the Sharpe ratio, during High periods, the LLM-based strategy pattern 1 is the best strategy according to both the best-mean and win-ratio-buy-and-hold methods. Conversely, in Low periods, pattern 1 is the best in terms of win-ratio-buy-and-hold, but buy-and-hold performs best in terms of best-mean. This suggests that the buy-and-hold strategy may be more suitable during such periods. These results indicate that, in terms of the Sharpe ratio, LLM-based strategies can outperform basic strategies during certain periods, particularly when the CPI trend is upward (i.e., when the 6-month moving average of the year-over-year change in the US CPI Total is rising compared to the previous month).

When considering other evaluation metrics, the results are mixed. For example, in terms of return, the CM(D2) strategy performs best during High periods. For volatility, the RG(S10) strategy often performs best. Regarding maximum drawdown, during High periods, the RG(S5) strategy is the best according to both the best-mean and win-ratio-buy-and-hold methods.

TABLE VII

BEST STRATEGIES IN TERMS OF FOUR METRICS (RETURN, VOLATILITY, MAX DRAWDOWN AND SHARPE RATIO) WITH DIFFERENT CPI TRENDS.

Metric	CPI Trend	Best-mean	Win-ratio-buy-and-hold
Return	High	Buy-and-hold	CM(D2)
	Low	Buy-and-hold	Pattern 1
Volatility	High	Buy-and-hold	RG(S10)
	Low	RG(S10)	RG(S10)
Max drawdown	High	RG(S5)	RG(S5)
	Low	Buy-and-hold	Pattern 1
Sharpe ratio	High	Pattern 1	Pattern 1
	Low	Buy-and-hold	Pattern 1

To better clarify the differences among strategies, we report the position series for the buy-and-hold, pattern 1, pattern 2, pattern 3, CM(D2), RG(S5), and RG(S10) strategies in Fig. 4. Additionally, the portfolio value trends during the investment period (with the initial value set to 1 on the first day) are shown in Fig. 5. Because LLM predictions often classify movements as class 1 (decline), LLM-based strategies tend to reduce positions. Compared with the pattern 2 and pattern 3 strategies, the pattern 1 strategy increases the position even when class 0 is predicted, allowing the position to return to 1

more easily. This likely facilitates profits during upward trends. The differences between the pattern 2 and pattern 3 strategies were minimal.

Looking at the buy-and-hold strategy in Fig. 4, it appears that there were short-term sharp declines around June 2022, from August to October 2022, around January and March 2023, and again from September to October 2023. During the sharp declines in January and March 2023, not only the LLM-based strategies but also the CM(D2) and RG(S5) strategies reduced their positions to 0, effectively preventing a loss in value. However, during the significant declines from September to October 2022 and September to October 2023, only the LLM-based strategies (patterns 1, 2, and 3) reduced their portfolio positions to 0, thereby avoiding value depreciation. In contrast, during the sharp decline in June 2022, the LLM-based strategies were slow to respond, allowing the value to drop, whereas the CM(D2) and RG(S5) strategies managed to lower their positions to 0, partially mitigating the loss. From these observations, it can be concluded that LLM-based strategies can sometimes detect short-term sharp declines effectively, but there are also instances where basic strategies outperform them in responding to these declines.

Additionally, examining the buy-and-hold strategy from a macroscopic trend perspective, we observe a continuous downward trend during the High period from November 2021 to August 2022. Starting in September 2022, when the CPI Trend shifted to Low, the overall trend remained relatively flat despite some fluctuations (Fig. 5). Based on these results, we conclude that LLM-based strategies achieve a higher Sharpe ratio during periods of high CPI Trend, particularly when there is a macroscopic downward trend.

D. Qualitative Analysis of the Reasoning of the LLM

We analyzed the reasoning structure for each persona using the explanations GPT gave for the logic behind its predictions. For instance, when predicting a decline, GPT using a short-term persona stated, “Rising interest rates and VIX suggest market uncertainty, while a stronger dollar could pressure exports, potentially impacting the portfolio negatively.” For a rise, GPT using a long-term persona explained, “Despite market volatility, the overall growth in stock futures and steepening yield curve suggest potential economic growth, which could positively impact the portfolio.” We extracted the cause-and-effect relationships, ensuring GPT avoided hallucinations using the method in [30]. Minor phrasing variations were handled via phrase embedding and clustering. The prompt is provided in our repository.

Table VIII highlights the top 15 cause-and-effect relationships for each persona. Short-term predictions emphasized declines (class 1: 6.896; class 2: 0.853), driven by factors such as interest rate spreads, market volatility, “flattening yield curve,” and the value of the dollar. In the medium term, pessimism persisted (class 1: 6.725; class 2: 0.792), but there was a shift toward stability as the dollar’s impact diminished. Long-term predictions reflected growing optimism (class 1: 5.540; class 2: 1.473), with growth factors such as “potential

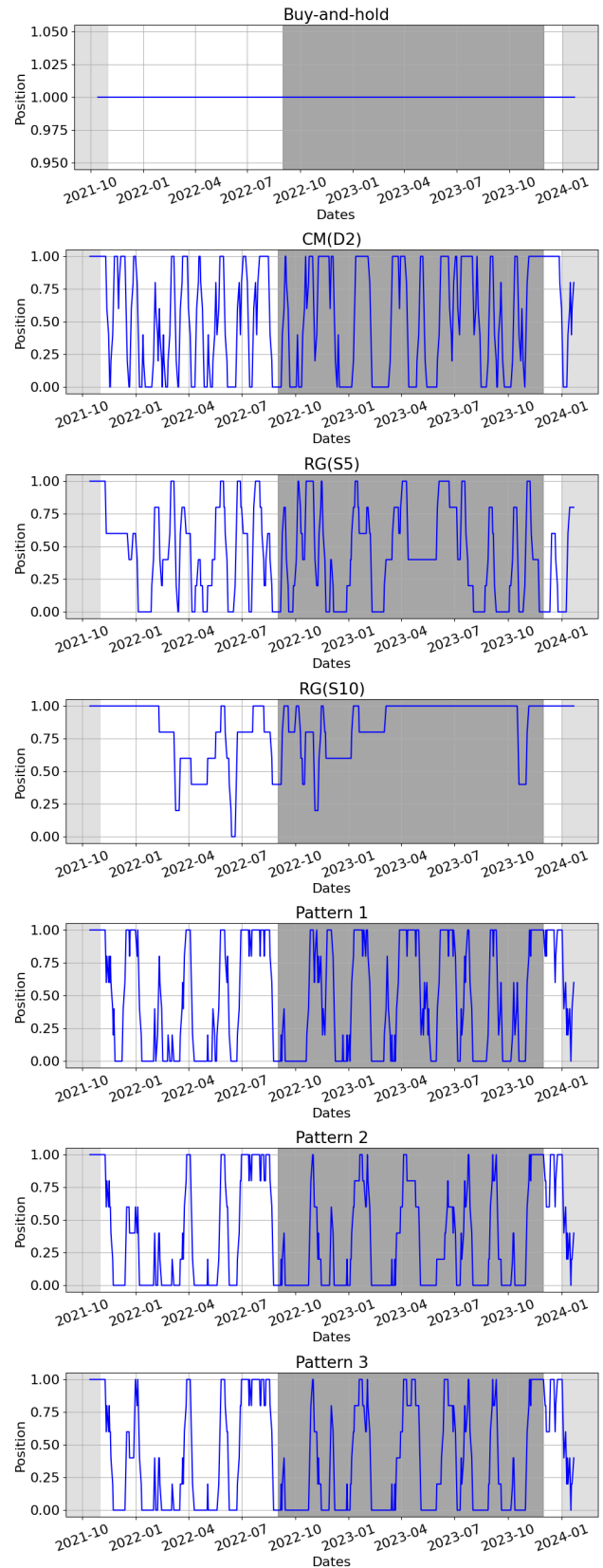


Fig. 4. Changes in the positions of the buy-and-hold, CM(D2), RG(S5), RG(S10), Pattern 1, Pattern 2, and Pattern 3 strategies. The background colors indicate the CPI trend: white for High, dark gray for Low, and light gray for periods outside the evaluation.

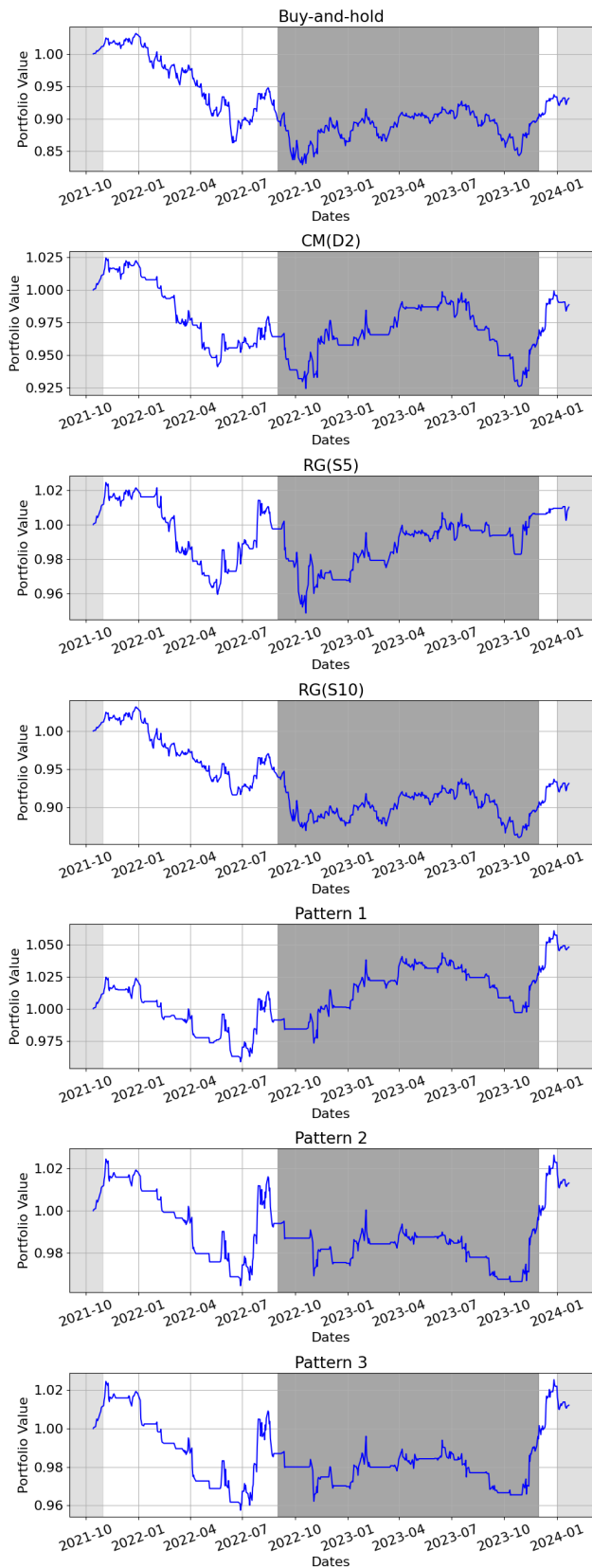


Fig. 5. Changes in the portfolio values of the buy-and-hold, CM(D2), RG(S5), RG(S10), pattern 1, pattern 2, and pattern 3 strategies. The background colors indicate the CPI trend: white for High, dark gray for Low, and light gray for periods outside the evaluation.

TABLE VIII
DIFFERENCES IN REASONING AMONG VARIOUS PERSONAS

Short-term persona		
Cause	Effect	Share
A decline is anticipated	Class 1 is predicted	4.898
Increasing interest rate spread	A decline is anticipated	2.169
The portfolio might face downward pressure	Class 1 is predicted	1.998
Increased market volatility	A decline is anticipated	1.771
Increasing interest rate spread	The portfolio might face downward pressure	1.300
Downward trend in the portfolio and stocks	A decline is anticipated	0.991
Flattening yield curve	A decline is anticipated	0.958
Potential for growth	Class 2 is predicted	0.853
General downward trend in the portfolio and futures	A decline is anticipated	0.747
Increased market volatility	The portfolio might face downward pressure	0.731
Increase in volatility	A decline is anticipated	0.666
Fluctuations in interest rates	A decline is anticipated	0.471
Stronger dollar	The portfolio might face downward pressure	0.439
Strengthening dollar	The portfolio might face downward pressure	0.431
Stronger dollar	A decline is anticipated	0.431
Medium term persona		
Cause	Effect	Share
A decline is anticipated	Class 1 is predicted	3.984
Increasing interest rate spread	A decline is anticipated	2.123
Increased market volatility	A decline is anticipated	1.957
The portfolio might face downward pressure	Class 1 is predicted	1.315
Flattening yield curve	A decline is anticipated	0.943
Increasing interest rate spread	The portfolio might face downward pressure	0.919
Downward trends in the portfolio and stocks	A decline is anticipated	0.840
potential for growth	2	0.792
A decline in the portfolio is anticipated	Class 1 is predicted	0.784
General downward trends in the portfolio and futures	A decline is anticipated	0.721
A decline in the portfolio's price is anticipated	Class 1 is predicted	0.642
Increased market volatility	The portfolio might face downward pressure	0.578
Increasing interest rate spread	A decline in the portfolio's price is anticipated	0.483
Increased market volatility	A decline in the portfolio is anticipated	0.452
General downward trend of the portfolio	A decline is anticipated	0.444
Long-term persona		
Cause	Effect	Share
A decline is anticipated	Class 1 is predicted	4.276
Increasing interest rate spread	A decline is anticipated	2.005
Increased market volatility	A decline is anticipated	1.860
Potential for growth	Class 2 predicted	1.103
Flattening yield curve	A decline is anticipated	1.087
General downward trends in the portfolio and futures	A decline is anticipated	0.894
The portfolio might face downward pressure	Class 1 is predicted	0.886
Downward trends in the portfolio and stocks	A decline is anticipated	0.644
Increasing interest rate spread	The portfolio might face downward pressure	0.612
Significant price movement is not anticipated	Class 0 is predicted	0.596
Increase in volatility	A decline is anticipated	0.499
Yield curve steepens	Potential for growth	0.491
A decline in the portfolio is anticipated	Class 1 is predicted	0.378
Potential growth in the portfolio	Class 2 is predicted	0.370
Fluctuations in interest rates	A decline is anticipated	0.362

for growth” (1.103) and “yield curve steepening” causing the “potential for growth” (0.491) to gain prominence. This suggests a shift from a decline-focused outlook to a more growth-oriented perspective, with the yield curve evolving from a source of pressure to an indicator of potential growth.

VI. CONCLUSION

In this paper, we showed that LLM-based predictions, particularly when combined with the mode ensemble, demonstrated strong performance in detecting market declines. LLM-based investment strategies outperformed the buy-and-hold strategy in terms of Sharpe ratio during periods of upward CPI trends, whereas the buy-and-hold strategy performed better during downward trends. Additionally, other strategies obtained better metrics such as return, volatility, and maximum drawdown, depending on market conditions.

One possible reason for the strong performance of our LLM-based strategies during the CPI uptrend is that baseline strategies, which consider only the past 10 days, rely solely on this limited data to adjust position sizes. This makes them highly sensitive to fluctuations within such a narrow window. In contrast, while our LLM strategy also examines only the last 10 days, it draws on its underlying knowledge to recognize that the recent price movements are part of a larger downward trend. This likely enabled the LLM strategy to make more informed adjustments, scaling down portfolio positions accordingly.

To illustrate, starting in December 2021, the U.S. began to acknowledge that rising inflation was not just a temporary consequence of COVID-19. This shift in understanding led to a series of interest rate hikes, especially rapid in the first year following December 2021. As a result, bond prices entered a range-bound phase, with CPI trends reflecting both upward and downward shifts. These patterns mirrored broader global trends observed in buy-and-hold strategy values, which tended to either decline or oscillate depending on the period.

ACKNOWLEDGMENT

This research was supported by JST SPRING GX project (Grant Number JPMJSP2108), the JST FOREST Program (Grant Number JPMJFR216Q) and The University of Tokyo Data Science School. We thank Kimberly Moravec, PhD, from Edanz for editing a draft of this manuscript.

REFERENCES

- [1] D. Vamvourellis, et al., “Company similarity using large language models,” arXiv preprint arXiv:2308.08031, 2023.
- [2] M. Trajanoska, R. Stojanov, and D. Trajanov, “Enhancing knowledge graph construction using large language models,” arXiv preprint arXiv:2305.04676, 2023.
- [3] I. Bhattacharya and A. Mickovic, “Accounting fraud detection using contextual language learning,” *Int. J. Account. Inf. Syst.*, vol. 53, p. 100682, 2024.
- [4] Y. Nie, Y. Kong, X. Dong, J. M. Mulvey, H. V. Poor, Q. Wen, and S. Zohren, “A survey of large language models for financial applications: Progress, prospects and challenges,” arXiv preprint arXiv:2406.11903, 2024.
- [5] H. Ko and J. Lee, “Can ChatGPT improve investment decisions? From a portfolio management perspective,” *Finance Research Letters*, vol. 64, p. 105433, 2024.
- [6] O. Romanko, A. Narayan, and R. H. Kwon, “ChatGPT-Based Investment Portfolio Selection,” *Operations Research Forum*, vol. 4, no. 4, p. 91, 2023.
- [7] Y. Cheng and K. Tang, “GPT’s idea of stock factors,” *Quantitative Finance*, vol. 0, no. 0, pp. 1–26, 2024.
- [8] B. Barber and T. Odean, “Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors,” *J. Finance*, vol. 55, no. 2, pp. 773–806, 2000.
- [9] A. Skiba and H. Skiba, “Institutional investors,” in *Financial Behavior: Players, Services, Products, and Markets*, H. K. Baker, G. Filbeck, and V. Ricciardi, Eds. New York: Oxford Academic, 2017.
- [10] R. A. Olsen and C. M. Cox, “The influence of gender on the perception and response to investment risk: The case of professional investors,” *J. Psychol. Financ. Mark.*, vol. 2, no. 1, pp. 29–36, 2001.
- [11] Y.M. Tseng et al., “Two tales of persona in LLMs: A survey of role-playing and personalization,” arXiv preprint arXiv:2406.01171, 2024.
- [12] B. Xu, A. Yang, J. Lin, Q. Wang, C. Zhou, Y. Zhang, and Z. Mao, “ExpertPrompting: Instructing Large Language Models to be Distinguished Experts,” arXiv:2305.14688, 2023.
- [13] C.-K. Wu, W.-L. Chen, and H.-H. Chen, “Large language models perform diagnostic reasoning,” arXiv preprint arXiv:2307.08922, 2023.
- [14] Y. Fu, H. Peng, T. Khot, and M. Lapata, “Improving language model negotiation with self-play and in-context learning from AI feedback,” arXiv preprint arXiv:2305.10142, 2023.
- [15] X. Li, H. Feng, H. Yang, and J. Huang, “Can ChatGPT reduce human financial analysts’ optimistic biases?,” *Economic and Political Studies*, vol. 12, no. 1, pp. 20–33, 2023.
- [16] Dhar, A. Datta and S. Das, “Analysis on Enhancing Financial Decision-making Through Prompt Engineering,” 2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), 2023, pp. 1–5.
- [17] T. Yue, D. Au, C.C. Au, and K. Lu, “Democratizing Financial Knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology,” *SSRN Electronic Journal*, 2023.
- [18] OpenAI, “GPT-4 Technical Report,” arXiv:2303.08774, 2023.
- [19] OpenAI, “Best practices for prompt engineering with the OpenAI API,” <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the>, Accessed on 2 May 2024.
- [20] OpenAI, “Prompt engineering,” <https://platform.openai.com/docs/guides/prompt-engineering>, Accessed on 2 May 2024.
- [21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv:2201.11903, 2023.
- [22] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, “Least-to-Most Prompting Enables Complex Reasoning in Large Language Models,” arXiv:2205.10625, 2023.
- [23] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models are Zero-Shot Reasoners,” arXiv:2205.11916, 2023.
- [24] R. R. Mekala, Y. Razeghi, and S. Singh, “EchoPrompt: Instructing the Model to Rephrase Queries for Improved In-context Learning,” arXiv:2309.10687, 2024.
- [25] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” arXiv:2005.14165, 2020.
- [26] Y. K. Chia, G. Chen, L. A. Tuan, S. Poria, and L. Bing, “Contrastive Chain-of-Thought Prompting,” arXiv preprint arXiv:2311.09277, 2023.
- [27] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-Consistency Improves Chain of Thought Reasoning in Language Models,” arXiv:2203.11171, 2023.
- [28] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhume, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, “Self-Refine: Iterative Refinement with Self-Feedback,” arXiv:2303.17651, 2023.
- [29] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, “Tree of Thoughts: Deliberate Problem Solving with Large Language Models,” arXiv:2305.10601, 2023.
- [30] R. Matsuoka, H. Matsumoto, T. Yoshida, T. Watanabe, R. Kondo, R. Hisano, “Hierarchical Narrative Analysis: Unraveling Perceptions of Generative AI,” arXiv preprint arXiv:2409.11032, 2024.