Master's Thesis

Evaluating LLM's Understanding of Portfolio Theory: Risk Profiles and Optimal Portfolios

Hanyong Cho

Department of Management of Technology

Graduate School of Management of Technology

Korea University

August 2025

Evaluating LLM's Understanding of Portfolio Theory: Risk Profiles and Optimal Portfolios

By

Hanyong Cho

---

under the supervision of Professor Jang Ho Kim

A thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science.

Department of Management of Technology

Graduate School of Management of Technology

Korea University

June 2025

The thesis of Hanyong Cho has been approved
by the thesis committee in partial fulfillment of
the requirements for the degree of
Master of Science.

June 2025

_____ Committee Chair: Jang Ho Kim
_____ Committee Member: Joonyup Eun
_____ Committee Member: Junghun Kim

Evaluating LLM's Understanding of Portfolio Theory: Risk Profiles and Optimal Portfolios

by Hanyong Cho

Department of Management of Technology

under the supervision of Professor Jang Ho Kim

ABSTRACT

This study designs a benchmark dataset based on portfolio theory to evaluate the investment decision-making capabilities of major large language models (LLMs) such as GPT-4, Gemini 1.5 Pro, and Llama 3.1-70B. The benchmark consists of multiple-choice questions derived from mathematically well-defined objective functions, allowing for the comparison of LLM selection accuracy across various investment goals and constraints. Through this framework, we quantitatively assess how rationally these LLMs can make investment decisions grounded in portfolio theory. Experimental results show that GPT achieves the highest accuracy on risk-based objective functions and maintains robust performance even under complex constraints. In contrast, Gemini demonstrates strength in return-focused tasks but exhibits significant performance degradation as the problem structure becomes more complex. Llama consistently shows the lowest accuracy across all scenarios. Furthermore, analysis of the inherent investment risk

preferences of each model revealed clear differences between models and varying levels of sensitivity to the assigned hypothetical investor profiles.

Keywords: Large language models, Benchmark dataset, Portfolio theory, Asset allocation, Risk profile

LLM의 포트폴리오 이론 이해 평가: 위험 성향과 최적 포트폴리오

조한용

기술경영학과

지도교수: 김장호

초록

본 연구는 GPT-4, Gemini 1.5 Pro, Llama 3.1-70B와 같은 주요 대형 언어모델(LLM)의 투자 의사결정 능력을 평가하기 위해, 포트폴리오 이론에 기반한 벤치마크 데이터셋을 설계하였다. 본 벤치마크는 수학적으로 정답이 존재하는 목적 함수 기반의 객관식 문제들로 구성되며, 다양한 투자 목적과 제약조건 하에서 LLM의 선택 정확도를 비교하였다. 본 연구는 GPT-4, Gemini 1.5 pro, Llama 3.1-70B와 같은 주요 대형 언어모델(LLM)이 포트폴리오 이론에 기반한 투자 의사결정 상황에서 얼마나 합리적인 판단을 내릴 수 있는지를 정량적으로 평가하였다. 실험

결과, GPT는 위험 기반 목적 함수에서 가장 높은 정확도를 보이며 복잡한 제약조건에서도 안정적인 성능을 유지하는 반면, Gemini는 수익률 중심 문제에 강점을 보였으나 구조가 복잡해질수록 성능이 저하되었고, Llama는 전반적으로 낮은 정확도를 나타냈다. 또한 각 LLM이 내재적으로 지닌 투자 위험 성향을 분석한 결과, 모델 간 명확한 차이가 존재하며, 가상의 투자자 특성에 대한 민감도 또한 상이한 것으로 나타났다.

중심어: 대형 언어 모델, 벤치마크 데이터셋, 포트폴리오 이론, 자산배분, 위험 성향

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

With the rapid advancement of artificial intelligence technology, Large Language Models (LLMs) have garnered significant attention as tools to assist or replace human decision-making across various industries. According to Zhao et al. (2023), LLMs demonstrate strengths in understanding and solving complex problems based on their natural language processing capabilities, and their potential for application in fields requiring highly specialized knowledge, such as finance, healthcare, and law, is highly evaluated. In this context, efforts to objectively evaluate the level of judgment capabilities and applicability of LLMs in real-world work environments are actively underway. This chapter examines the background and research objectives of such studies.

1.1 Research Background

As LLMs rapidly advances, it is being utilized in various fields. Accordingly, research to quantitatively evaluate not only the capabilities of LLM but also its practical applicability and limitations (Chang et al., 2024; Brown, 2020). In such evaluations, standardized benchmarks serve as a standard for comparing models and analyzing performance, and are a key tool for determining the level at which LLMs can be utilized. Existing representative benchmarks include those focusing on general language understanding (Wang et al., 2018; Wang et al., 2019), common-sense reasoning (Talmor et al., 2018; Zellers et al., 2019), and academic knowledge (Hendrycks et al., 2020), contributing to measuring the general performance of LLMs across various task types.

Meanwhile, the need for domain-specific benchmarks to evaluate understanding and reasoning in specific domains has also emerged, with recent efforts focusing on evaluating LLM performance in specific domains such as finance (Chen et al., 2021; Xie et al., 2024), healthcare (Kim et al., 2024), and law (Chalkidis et al., 2021). In the financial sector, LLMs are being used in

various tasks such as investment strategy formulation (Li et al., 2025), risk assessment (Golec and AlabdulJalil, 2025), news analysis (Dolphin et al., 2024), and financial fraud detection (Kadam, 2024), with their potential continuing to expand. Reflecting this, various finance-specific benchmarks have been proposed; however, most of them focus on natural language processing-centric tasks (e.g., document summarization, information extraction, question-answering). That is, they remain at the level of evaluating language processing capabilities, such as extracting specific information from financial documents (Hamad et al., 2024) or answering questions (Choi et al., 2025), and do not sufficiently reflect the quantitative analysis, time-series prediction, and decision-making-based reasoning required in actual financial tasks.

To address these limitations, this study proposes a new benchmark framework to quantitatively evaluate how reasonably and consistently LLM can make judgments in real-world financial decision-making. The following sections describe the objectives of this study and the two-stage experimental design to achieve them.

1.2 Research Objectives

This study aims to design a benchmark framework based on portfolio theory to quantitatively evaluate the investment decision-making ability of LLM. To achieve this, two stages of research are conducted.

1.2.1 Risk Profiles

Portfolio theory is one of the core theories of modern finance, explaining the process by which investors make choices between expected returns and risk. In this context, an investor's risk preference serves as the most fundamental concept. Against this backdrop, accurately diagnosing and quantifying an investor's risk preference is crucial in the actual stages of portfolio construction. Currently, standardized investor preference assessment questionnaires are

widely used in the financial industry and academia to measure investor preferences. These questionnaires calculate scores based on various factors such as investment period, experience, behavioral characteristics, age, and asset size, and these scores are used to recommend appropriate portfolio compositions.

Meanwhile, with the recent introduction of LLMs into the financial advisory field, it has become important to understand the inherent investment tendencies of LLMs. In particular, evaluating whether LLMs can reveal specific investment tendencies without prior information and how well they reflect explicitly inputted investor characteristics to make reasonable portfolio decisions is a key factor in assessing the potential impact of LLMs on future financial advisory services.

To this, this study analyzes the inherent risk tendencies of LLMs and experimentally verifies how their responses change when risk tendencies are explicitly assigned. Using state-of-the-art LLMs such as GPT-4o, Gemini 1.5 Pro, and LlaMA 3.1-70B, experiments were conducted based on a widely used investor risk preference diagnostic questionnaire. Additionally, tests were performed under conditions where investor preferences were pre-assigned (e.g., "You are a risk-averse investor"), analyzing how sensitively each model reflects such information.

1.2.2 Optimal Portfolios

In the second study, based on these prior studies, we design a portfolio theory-based benchmark framework to quantitatively evaluate the actual investment decision-making ability of LLMs.

Portfolio theory (Markowitz, 1952) is a theory that explains how investors should combine assets given expected returns and risk levels, and it is one of the core foundations of modern finance. Investors construct portfolios by

considering the correlations between various assets and the returns and volatility of individual assets, and they can achieve asset allocation results that align with specific objectives (e.g., minimize volatility, or maximize return). This process is mathematically defined by an objective function and constraints, meaning that there is a definitive answer and evaluation criteria can be clearly established based on various investment objectives. This makes portfolio theory essential knowledge for evaluating the investment decision-making capabilities of LLMs. In particular, portfolio theory can mathematically explain the trade-off between returns and risks that arise in actual investment decisions, making it advantageous for quantitatively analyzing how well LLMs understand and apply such judgment criteria.

Therefore, this study aims to construct various optimization problems based on portfolio theory and evaluate how LLM makes decisions in actual investment decision-making situations. Accordingly, this study calculated theoretical optimal portfolios for each investment objective and used them as answers, then evaluated the accuracy of LLM's responses. The correct portfolios are calculated based on traditional portfolio theory, using volatility, return, Sharpe ratio, maximum drawdown (MDD), and conditional value-at-risk (CVaR) as metrics in the objective functions, and problem types are structured according to each investment objective. LLM evaluation is structured as multiple-choice questions, with each question presenting a portfolio selection task for a single investment objective. Each question includes one correct answer (optimal portfolio) and three incorrect answers, and the LLM must select the most appropriate portfolio.

CHAPTER 2. LITERATURE REVIEW

2.1 Portfolio optimization

Portfolio optimization is a core concept in modern financial theory that involves establishing asset allocation strategies to achieve investor objectives under given constraints(Kim et al., 2021; Kim et al., 2024). The foundation of this theory is the modern portfolio theory (MPT) proposed by Markowitz (1952), which states that investors should construct portfolios based on expected return, variance (or standard deviation), and covariance. This provides a mathematical explanation of the risk-return tradeoff and introduces the concept of the efficient frontier.

Generally, the mean-variance model calculates the optimal portfolio that aligns with the investor's objectives through the optimization problem given by (1):

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \mathrm{w}^{\mathrm{T}}\Sigma\mathrm{w} \\ \text{subject to} \quad & \mathrm{w}^{\mathrm{T}}\mu \geq \mu_0 \\ & \mathrm{w}^{\mathrm{T}}\mathbf{1} = 1 \\ & w_i \geq 0 \end{aligned} \tag{1}$$

Here, $\mathrm{w} \in \mathbb{R}^n$ denotes the portfolio weight vector, $\Sigma \in \mathbb{R}^{nxn}$ is the covariance matrix, $\mu \in \mathbb{R}^n$ represents the expected returns of the assets, and $\mathbf{1} \in \mathbb{R}^n$ is a vector consisting of all ones. The scalar μ_0 indicates the minimum required return of the portfolio.

Furthermore, portfolio optimization can set various forms of objective functions according to the investor's goals.

2.1.1 Minimize Volatility

The volatility minimization strategy focuses on minimizing the overall risk (dispersion) of the portfolio and is suitable for investors with a strong risk aversion. This strategy derives the optimal diversified portfolio by considering the covariance structure between assets without any constraints on expected returns. Mathematically, it is expressed as Equation 1. The objective function $w^T \Sigma w$ represents the variance of the portfolio, and by minimizing it, investors can achieve efficient diversification. The solution derived from this strategy is called the Global Minimum Variance Portfolio (GMV), which is the portfolio located at the far left of the Efficient Frontier. The GMV represents the combination with the lowest risk among all portfolios that can be constructed with the given assets, and it serves as a benchmark for the most conservative investment strategy regardless of market conditions.

In practice, GMV serves as the starting point for risk-based asset management strategies (Choueifaty and Coignard, 2008) and is useful for establishing long-term investor strategies, stable asset management for pension funds, and constructing defensive portfolios in extreme market conditions. Additionally, GMV can be applied in practice by imposing additional constraints (e.g., lower or upper bounds), making GMV-based strategies a widely adopted approach that combines theoretical validity with practical applicability.

2.1.2 Maximize Return

The expected return maximization strategy is one of the most basic forms of portfolio optimization, aiming to maximize future returns by adjusting asset allocations. This method seeks to maximize the overall expected return of the portfolio by allocating a higher proportion to assets with higher expected returns. The most significant characteristic of this strategy is that it does not

consider risk. In other words, risk indicators such as covariance or volatility are not included in the objective function, and investors are only required to select portfolios with high expected returns. Therefore, this strategy is suitable for investors with a very high risk tolerance or those seeking high returns in the short term.

It can be expressed as the following optimization problem given by (2):

$$\begin{aligned} \underset{\mathrm{w}}{\text{maximize}} \quad & \mathrm{w}^{\mathrm{T}}\mu \\ \text{subject to} \quad & \mathrm{w}^{\mathrm{T}}\mathbf{1} = 1 \\ & w_i \geq 0 \end{aligned} \tag{2}$$

Here, w is the asset-specific investment weight vector, μ is the expected return vector of the asset, and $\mathrm{w}^{\mathrm{T}}\mathbf{1} = 1$ means that the total investment weight of the asset is 100%.

This model has the advantage of being simple to interpret and calculate, but it ignores risk factors, so it may suggest portfolios with high volatility to investors. We note that solving the above equation always results in an optimal portfolio allocation where 100% is invested in the asset with the highest expected return. Therefore, in actual asset management, this model is often used with additional risk constraints (e.g., volatility caps, asset weight constraints).

2.1.3 Maximize Sharpe ratio

Sharpe ratio is an indicator that measures the risk-adjusted return efficiency of a portfolio, quantifying how much excess return can be obtained per unit of risk (standard deviation). It was proposed by Sharpe (1966) based on Markowitz's mean-variance model and has since become a key indicator for investment performance evaluation and asset allocation strategies.

The Sharpe ratio is defined as follows given by (3):

$$Sharpe\ Ratio = \frac{\mathrm{w}^{\mathrm{T}}\mu - r_f}{\sqrt{\mathrm{w}^{\mathrm{T}}\Sigma\mathrm{w}}} \tag{3}$$

Here, r_f represents the risk-free rate of return, and $\sqrt{\mathrm{w}^{\mathrm{T}}\Sigma\mathrm{w}}$ represents the standard deviation of the portfolio. At this point, $\mathrm{w}^{\mathrm{T}}\mu - r_f$ represents the excess return of the portfolio, which is not simply the expected return, but rather a benchmark that indicates how much higher the return is compared to investing in risk-free assets. Maximizing Sharpe ratio reflects a strategic approach that seeks to achieve as much excess return as possible under a given level of risk, or to achieve the same return with the least amount of risk.

Sharpe ratio maximization is defined as the following optimization problem given by (4):

$$\begin{aligned} \underset{\mathrm{w}}{\text{maximize}} \quad & \frac{\mathrm{w}^{\mathrm{T}}\mu - r_f}{\sqrt{\mathrm{w}^{\mathrm{T}}\Sigma\mathrm{w}}} \\ \text{subject to} \quad & \mathrm{w}^{\mathrm{T}}\mathbf{1} = 1 \\ & w_i \geq 0 \end{aligned} \tag{4}$$

This problem is more complex to optimize than typical linear or quadratic programming problems because it has a nonlinear fractional objective function. Numerically, since both the numerator and denominator are functions of portfolio weight, indirect methods such as linearization or log transformation are often used. For example, fixing the denominator and optimizing only the numerator, or applying the Lagrangian method to transform the problem are representative approaches.

Sharpe ratio-based strategies are theoretically considered an excellent choice, but they also have some limitations in practice. First, the covariance matrix and expected return vector are estimates that incorporate future uncertainty, so the absolute value of the Sharpe ratio itself may be sensitive to prediction errors. Second, when the denominator becomes extremely small, Sharpe ratio may

artificially increase excessively, potentially leading to a distorted portfolio composition that overly underestimates risk. Third, since it does not reflect asymmetric risk or tail risk, portfolios with the same Sharpe ratio may have very different actual loss risks.

Nevertheless, Sharpe ratio is the most intuitive performance metric that considers both return and risk simultaneously, and it is widely used as a standard measure of risk-adjusted return.

2.1.4 Minimize Maximum Drawdown

MDD is an indicator that shows how much the value of a portfolio has fallen from its peak to its lowest point during a specific period, measuring the largest loss that an investor can experience. It is not an indicator of average risk or volatility, but rather reflects the degree of exposure to the most unfavorable loss scenario, making it a very important decision-making criterion for conservative asset managers and investors with a high risk aversion.

MDD is defined as follows given by (5):

$$\text{Maximum drawdown(MDD)} = \max_{t\in[0,\mathrm{T}]}\left(\frac{\max_{s\in[0,t]}V(s) - V(t)}{\max_{s\in[0,t]}V(s)}\right) \qquad (5)$$

Here, V(t) is the value of the portfolio at time *t*, and $\max_{s\in[0,T]}\mathrm{V(s)}$ is the highest portfolio value up to time *t*. Ultimately, MDD measures how much assets have fallen from their past peak at a specific point in time *t*, and the largest value during the entire period becomes the maximum drop.

Minimization of MDD is defined as the following optimization problem given by (6):

$$\begin{aligned} \underset{\mathrm{w}}{\text{minimize}} \quad & \mathrm{MDD(w)} \\ \text{subject to} \quad & \mathrm{w}^{\mathrm{T}}\mathbf{1} = 1 \end{aligned} \qquad (6)$$

$$w_i \geq 0$$

In this formula, w is the asset weight vector, and the objective function MDD(w) is the maximum drawdown value derived from the past period under the given asset weights.

2.1.5 Minimize Conditional Value-at-Risk (CVaR)

Value-at-Risk (VaR) is a representative indicator for measuring financial risk, representing the maximum loss that a portfolio is expected not to exceed during a given period at a given confidence level. For example, if the VaR is $1,000 at a 95% confidence level, it means that there is a 95% probability that the loss will not exceed $1,000 during that period.

VaR is defined as follows given by (7):

$$\mathrm{VaR}_{\alpha\%}(X) = inf\{x; F_x(x) \geq \alpha\%\} \tag{7}$$

VaR has the advantage of being simple and intuitive to calculate, but it has the disadvantage of evaluating risk based on a single threshold, which does not take into account extreme losses. To overcome this limitation, Rockafellar and Uryasev (2000) proposed the concept of conditional value-at-risk (CVaR), also known as expected shortfall, and demonstrated that the problem of minimizing CVaR can be efficiently solved as a convex optimization problem. CVaR is defined as the average loss value of the loss interval exceeding VaR and is used to measure the average of the worst losses that an investor may experience in extreme risk situations.

CVaR is defined as follows given by (8):

$$\mathrm{CVaR}_{\alpha\%}(\mathrm{X}) = \mathbb{E}[X|X > \mathrm{VaR}_{\alpha\%}(X)] \tag{8}$$

CVaR is widely used as a more conservative risk indicator that includes low-probability, high-loss events such as market shocks and black swans by

calculating the average value in the loss interval exceeding VaR.

CVaR minimization is defined as follows: given by (9)

$$\begin{aligned} \underset{\mathrm{w}}{\text{minimize}} \quad & \mathrm{CVaR}_{\alpha\%}(\mathrm{w}) \\ \text{subject to} \quad & \mathrm{w}^{\mathrm{T}}\mathbf{1} = 1 \\ & w_i \geq 0 \end{aligned} \tag{9}$$

CVaR-based strategies are more sensitive to tail risk and more conservative than VaR, making them suitable for institutional investors with a high risk aversion, such as pension funds, insurance companies, and hedge funds. They are particularly useful for minimizing extreme losses and strengthening portfolio resilience, and their application is increasing in areas such as ESG investing and SRI strategies. However, limitations exist, such as sensitivity to the level of confidence α and estimation errors due to the lack of tail data. Despite these limitations, CVaR has established itself as a risk measurement metric that simultaneously reflects investors' psychological loss aversion and the uncertainty of financial markets.

These theories are evolving to become more sophisticated by reflecting various real-world constraints. Asset-specific minimum and maximum investment weights, transaction costs, liquidity constraints, and number of securities constraints (cardinality constraints) must be considered in the actual portfolio construction process, making the optimization problem computationally more complex.

Furthermore, various techniques are being proposed to reflect the uncertainty of the investment environment, changes in the market over time, and individual investor objectives. Notable examples include robust optimization, machine learning-based asset allocation, and dynamic programming, which are gaining attention as flexible and realistic portfolio optimization strategies that go beyond traditional mean-variance theory(Kim et al., 2024; Lee et al., 2024).

Recently, these advanced portfolio optimization techniques have been actively utilized in robo-advisors and algorithmic trading systems, functioning as core elements of automated asset management tools. In this context, verifying how accurately LLMs can understand and apply portfolio theory has emerged as an important research topic.

2.2 Benchmark Dataset

With the rapid advancement of LLMs, the importance of benchmark datasets for objectively evaluating model performance has been increasingly emphasized. In particular, LLMs have demonstrated natural language-based reasoning and problem-solving capabilities across various domains, and domain-specific benchmarks for quantitatively evaluating these capabilities are actively being developed. Hendrycks (2020), Pal et al. (2022), Fei et al. (2023), and Chen et al. (2021) have each used their respective benchmarks to assess how well LLMs understand real-world domain-specific knowledge in fields such as science, medicine, law, and finance.

2.2.1 General Benchmark Dataset

Representative general-purpose evaluation metrics include MMLU (Massive Multitask Language Understanding) and HellaSwag.

MMLU (Hendrycks, 2020) is a benchmark designed to measure the broad domain knowledge and reasoning abilities of LLMs. It consists of approximately 15,000 multiple-choice questions across 57 sub-subjects. These subjects include mathematics, physics, chemistry, computer science, history, economics, medicine, law, and other fields of humanities, social sciences, and natural sciences, with problem difficulty ranging from high school to graduate school levels. Each question is in an objective format with a clear correct answer, comprehensively evaluating whether an LLM can perform a series of processes based on specialized knowledge, including context understanding,

information retrieval, concept inference, and logical selection. Notably, MMLU places a high emphasis on questions requiring higher-order reasoning rather than mere memorization, making it a key benchmark for comparing the performance of recent models such as GPT-4, Claude, and Gemini.

On the other hand, HellaSwag (Zellers et al., 2019) is a natural language inference (NLI) dataset designed to evaluate common-sense reasoning capabilities. It measures whether an LLM can select the most natural and logical next sentence within the flow of everyday context. It consists of over 70,000 questions, each comprising a short prompt and four completion options. All options are grammatically correct and semantically possible sentences, but only one is considered the most appropriate description in context. HellaSwag was created by mixing sentences written by humans and sentences generated by GPT-2-based language models, designed to test reasoning abilities that require logical connections, understanding of causality, and social context awareness, rather than simple statistical probability predictions.

While such general-purpose benchmarks are effective for evaluating the overall language understanding and reasoning abilities of LLMs, they primarily focus on the relationships between natural language sentences or the associations between concepts, and thus have limitations in directly measuring mathematical model-based judgment capabilities.

2.2.2 Financial Benchmark Dataset

Recently, various benchmark datasets have been developed to evaluate the applicability of LLM in the financial domain, with FinQA, ConvFinQA, and FinBEN being widely used. These benchmarks are designed to measure the domain-specific understanding of models from various perspectives, including financial document-based question answering, mathematical calculation inference, and natural language interpretation capabilities.

FinQA (Chen et al., 2021) is a benchmark that evaluates the ability to respond to natural language questions through mathematical reasoning and calculations based on semi-structured financial statements and report data. Each problem consists of a table containing financial data, a natural language question, an intermediate reasoning path, and a final numerical answer. It requires formula-based inference, ratio calculations, and conditional operations, focusing on evaluating whether the model can derive the correct answer based on a logical mathematical reasoning framework.

ConvFinQA (Chen et al., 2022) is a benchmark that extends the structure of FinQA to reflect multi-turn dialogue scenarios (conversational setting), simulating the question-answering process in real financial consultation environments. The model must consider the context of the user's previous questions and responses to perform appropriate reasoning and calculations for the current question. This process requires complex capabilities such as the ability to maintain conversational flow through contextual memory, step-by-step numerical reasoning, and appropriate invocation of calculation modules, which are considered core elements in the design of practical robo-advisor systems or financial chatbots.

FinBEN (Xie et al., 2024) is a large-scale financial domain benchmark constructed based on financial news, company reports, and market analysis data collected from Bloomberg, Yahoo Finance, SEC filings, and other sources. It includes a total of 10 sub-tasks, including financial question answering (FinQA), summarization, sentiment analysis, information retrieval, and document classification. Notably, FinBEN is centered on unstructured text used in real-world financial markets, focusing on evaluating how well LLMs can accurately understand domain-specific vocabulary, context, and structure, and perform linguistic reasoning based on this understanding. Recently, various state-of-the-art LLMs such as GPT-4, Claude, and Gemini have been evaluated

and compared using this benchmark, establishing it as the de facto standard benchmark for assessing financial language understanding.

These financial benchmarks excel at evaluating various capabilities such as domain-specific information interpretation, document response, and numerical reasoning. However, most are centered on natural language-based question-answering or classification/summarization tasks, and thus have limitations in directly evaluating quantitative financial models, particularly mathematically structured optimization problems like portfolio theory.

CHAPTER 3. METHODOLOGY: RISK PROFILES

This chapter analyzes the response patterns of LLM regarding investors' risk preferences and evaluates whether the provided pre-profile information is reflected in the responses.[1] To this end, we quantitatively compared and analyzed LLM's investment preferences based on a widely used questionnaire for diagnosing investor preferences.

Although there is no universally accepted method for assessing investors' risk preferences, this study adopts a questionnaire-based approach that is widely recognized both in academic literature and industry practice. Prior research, including Tertilt and Scholz (2018), highlights the practicality and consistency of this method in identifying investor profiles. Because such questionnaires typically cover similar themes—such as age, investment horizon, experience, and behavioral tendencies—we chose a representative questionnaire for evaluating the risk preferences of large language models (LLMs). The questionnaire comprises seven multiple-choice items: two items (Q1–Q2) measure investment horizon and are used to calculate the "time horizon score," while the remaining five items (Q3–Q7) evaluate self-reported risk tolerance, producing a "risk tolerance score."[2] A summary of all items is presented in Table 1.

The LLMs evaluated in this study include GPT-4o, Gemini 1.5 Pro, and LLaMA 3.1-70B. To assess each model, we provided a standardized default prompt, followed by sequentially inputting each question. Responses were

---

[1] Tables in this section are originally presented in Cho and Kim (2024).

[2] We reference the investor profile questionnaire from Charles Schwab, which was retrieved in early 2025 from https://www.schwab.com/resource/investment-questionnaire.

recorded after each item, and the full survey was repeated 100 times for every model, with each repetition occurring in a new conversation session.

Default prompt:	Suppose you are an individual investor. Please answer the following questions one by one. Each question is multiple choice. Select one answer for each question.

To determine whether LLMs incorporate investor-specific traits when provided, we conducted supplementary experiments. These tests were designed to assess the models' sensitivity to assigned characteristics. For instance, by altering the initial prompt to state, "You are a risk-averse investor," we examined whether the model's responses aligned with a risk-averse profile. We also tested variations in characteristics such as age, wealth level, and investment experience.

The specific prompt configurations for each condition are outlined in Table 2. Each variation was tested 30 times. For comparative analysis between LLMs and across profile conditions, we utilized both mean and median scores. To assess statistical significance, ANOVA and Kruskal-Wallis tests were applied. Given that the LLM responses were categorical with minimal variance, and Levene's test rejected the assumption of homogeneity of variance, we followed Hecke (2012) in including Kruskal-Wallis results alongside ANOVA, as it offers greater reliability when the assumptions of normality and equal variances do not hold.

Table 1 Summary of investor profile questionnaire

		Question summary	Answer choices
Time horizon score	Q1	Investment period (years)	'Less than 3 years' to '11 years or more'
	Q2	Spending period after withdrawal (years)	'Less than 2 years' to '11 years or more'
Risk tolerance score	Q3	Investment knowledge	'None', 'Limited', 'Good', 'Extensive'
	Q4	Risk taking willingness	'Lower than average', 'Average', 'Above average'
	Q5	Investments currently own or owned	'Cash', 'Bond', 'Stocks', 'International funds'
	Q6	Reaction to investing loss	'Sell all', 'Sell some', 'Do nothing', 'Buy more'
	Q7	Acceptable investment outcomes	'Low risk, low return' to 'High risk, high return'

Table 2 Profile values and corresponding prompts

Category	Values	Prompt (first line)
risk_appetite	{'risk-averse', 'risk-neutral', 'risk-seeking'}	Suppose you are a [*risk_appetite*] individual.
age	{20, 30, 40, 50}	Suppose you are an individual investor in your [*age*]s.
wealth	{'below average wealth', 'average wealth', 'above average wealth'}	Suppose you are an individual with [*wealth*].
investing_ experience	{'no investing experience', 'some investing experience', 'professional investing experience'}	Suppose you are an individual with [*investing_experience*].

CHAPTER 4. METHODOLOGY: OPTIMAL PORTFOLIOS

In this chapter, we aim to quantitatively evaluate whether LLM can make reasonable decisions in investment decision-making situations based on portfolio theory.

To achieve this, we designed a benchmark framework that generates multiple-choice problems considering various investment objectives and constraints, and analyzes whether the LLM can select the optimal portfolio from the provided options. This benchmark has the advantage of being able to calculate the correct answer based on the mathematical structure of portfolio theory, thereby clearly distinguishing between correct and incorrect answers for each problem. Additionally, the ability to generate an infinite number of problem variations by combining input values—specifically, asset classes, investment objectives, and investment periods—is one of the core design features of this framework. Furthermore, the prompts presented to the LLM in this framework are composed of questions and options. The question describes a specific investment objective and market conditions, while the choices represent different asset allocations within a portfolio. An example is as follows:

Question	You are a portfolio manager. Your task is to select the optimal portfolio based on portfolio theory. Make your decision using asset returns during the specified investment period and the information below: Objective: Lowest Volatility Assets: BND, GSG, VTI Date: 2020-01-01 to 2024-12-31

Choices	(1) BND: 0.417, GSG: 0.031, VTI: 0.551 (2) BND: 0.026, GSG: 0.205, VTI: 0.769 (3) BND: 0.209, GSG: 0.686, VTI: 0.105 (4) BND: 0.892, GSG: 0.016, VTI: 0.093

In this study, based on portfolio theory, we constructed portfolio optimization problems corresponding to five investment objectives. Each objective is shown in Table 3.

The optimization problem defined for this purpose calculates the solution based on the expected return vector, the covariance matrix between assets, and the asset weight vector. The asset weights are constrained such that their total sum is 1 and none of the asset weights are negative. Additionally, in some problems, minimum or maximum bounds (lower/upper bounds) are added to individual asset weights to reflect constraints that may arise in real-world investment environments.

Each problem requires the LLM to select the optimal asset allocation from four portfolios given the conditions of investment objective, asset list, and investment period. The correct answer is the portfolio obtained through the optimization process described above that aligns with the investment objective, while the remaining three are considered incorrect answers. In particular, incorrect options directly influence the difficulty and evaluation reliability of the problem, so systematic generation is essential rather than simply generating them randomly. In fact, the difficulty of the problem can vary significantly depending on how the incorrect options are constructed. Therefore, to more accurately evaluate the LLM's understanding of portfolio theory, we applied an incorrect option generation method based on the following principles.

Table 4 describes the four methods used based on the mathematical characteristics of the portfolio and the differences in performance based on the

objective function. The first is the "distance-based" method, which selects incorrect answers based on the Euclidean distance between the answer portfolio and the correct answer portfolio falling within a certain range. This method presents choices based on structural differences in composition and has the advantage of excluding portfolios that are too similar or too different from the correct answer, thereby adjusting the level of difficulty appropriately. When the asset weight vector of the correct answer portfolio is w^* and the asset weight vector of the i-th incorrect answer portfolio is w_i, the Euclidean distance d_i between them is defined as follows given by (10) and (11):

$$d_i = \| w^* - w_i \| = \sqrt{\sum_{j=1}^{n} \left(w_j^* - w_{ij}\right)^2} \tag{10}$$

$$\theta_{min} \leq d_i \leq \theta_{max} \tag{11}$$

Furthermore, incorrect answer choices are only accepted if they satisfy equation 11.

Here, θ_{min} and θ_{max} are the lower and upper limits allowed in distance-based selection, and n denotes the total number of assets. This approach structurically maintains the differences between options while adjusting the similarity to the correct answer to a certain level, making it useful for evaluating the sophisticated judgment capabilities of LLM.

The second is the "threshold-based" approach, which selects portfolios that differ from the correct portfolio and the objective function value (e.g., volatility or return) by a certain level or more as incorrect, thereby encouraging performance-based judgment. In this approach, a portfolio is classified as incorrect if the relative difference between the performance metrics of the

correct portfolio and the candidate portfolio falls within a predefined acceptable range $[\delta_{min}, \delta_{max}]$.

At this point, if each candidate portfolio satisfies the following conditions, it is considered an incorrect answer choice given by (12):

$$\delta_{min} \leq \frac{|S^* - S_i|}{\max(|S^*|, \varepsilon)} \leq \delta_{max} \tag{12}$$

Here, S^* is the objective function value of the correct portfolio (e.g., volatility or return), S_i is the objective function value of the i-th candidate portfolio, and ε is a small value used to avoid dividing by zero; in this study, 10^{-6} was used. This mathematically defines that the incorrect portfolio is sufficiently distinguishable from the correct portfolio in terms of performance, enabling an evaluation of whether the LLM can recognize performance-based quantitative differences.

The third is the "quantile-based" method, which generates m random portfolios, classifies them into quantiles based on the objective function values, and uses the inefficient portfolios corresponding to the lower quantiles as incorrect answers.

Fourth, the "dual criteria" approach generates incorrect answer choices by simultaneously considering the difference in weight (similarity in composition) and performance difference (performance based on the objective function) between the correct answer portfolio and the incorrect answer choices. Unlike existing methods that only consider numerical deviations, the dual criteria method is designed to evaluate whether LLM can comprehensively interpret the multidimensional characteristics of a portfolio by providing alternatives that differentiate themselves from the correct answer in both the structural composition of the portfolio and investment performance. In this method, only portfolios where both Euclidean distance and relative performance difference

fall within the predefined threshold intervals $[\theta_{min},\theta_{max}]$ and $[\delta_{min},\delta_{max}]$ are selected as incorrect answer candidates. Subsequently, to ensure diversity among the candidates, the distance between candidates is also considered. Additionally, to prevent excessive similarity among incorrect answer options, a diversity condition is applied between selections. This is to ensure that the composition of the options is sufficiently different from one another, enabling the LLM to perform meaningful comparisons. Specifically, the newly selected candidate portfolio w_i is included as the final incorrect option only if its Euclidean distance from the already selected incorrect portfolios w_j is greater than a certain threshold τ. This condition is expressed as follows given by (13):

$$\forall j < i, \| w^* - w_i \| \geq \tau \tag{13}$$

Here, w_i is the i-th candidate portfolio newly considered, and w_j is the j-th incorrect portfolio previously selected. τ is the minimum distance between portfolios and is typically set at 0.4 to 0.6. This method is effective in preventing overlap between options and evaluating whether LLM can recognize structural diversity rather than simple numerical similarity.

In addition, to see if LLM can recognize changes in problem difficulty and respond appropriately, we added optimization constraints to the option composition process.

First, the “no constraint” approach is used to construct a portfolio freely without any restrictions on asset weightings and is used as the most basic benchmark.

Second, the ‘Asset Constraint’ method controls the complexity of portfolio composition by fixing the number of assets included, thereby facilitating structural comparisons. In this study, we used k (3, 5, 7) assets to further subdivide the difficulty levels. Generally, as the number of assets increases, the number of possible combinations grows exponentially, leading to higher

selection complexity and making it more challenging for LLM to derive the correct answer. Additionally, we applied weight constraints differentially based on the number of assets to reflect the realism of each portfolio.

Specifically, all portfolios were subject to the same total constraint $w^{\mathrm{T}}\mathbf{1} = 1$, and different lower bounds and upper bounds were applied to the weight w_i of each asset. The constraints according to the number of assets are as follows given by (14), (15), and (16):

- 3-assets

$$\text{Subject to} \quad \begin{aligned} & w^{\mathrm{T}}\mathbf{1} = 1 \\ & 0.1 \leq w_i \leq 0.6 \end{aligned} \tag{14}$$

- 5-assets

$$\text{Subject to} \quad \begin{aligned} & w^{\mathrm{T}}\mathbf{1} = 1 \\ & 0.05 \leq w_i \leq 0.4 \end{aligned} \tag{15}$$

- 7-assets

$$\text{Subject to} \quad \begin{aligned} & w^{\mathrm{T}}\mathbf{1} = 1 \\ & 0.05 \leq w_i \leq 0.3 \end{aligned} \tag{16}$$

By combining asset number and asset ratio constraints in this way, we systematically established experimental conditions for adjusting problem difficulty and evaluating LLM decision-making capabilities.

Third, the “lower bound” method prevents excessively low allocations to specific assets by assigning a minimum weight to all assets. The constraints applied are defined as follows given by (17):

$$\text{Subject to} \quad \begin{aligned} & w^{\mathrm{T}}\mathbf{1} = 1 \\ & \iota \leq w_i \end{aligned} \tag{17}$$

Here, ι denotes the minimum weight limit applicable to all assets i, and w_i

denotes the weight allocated to asset ι.

As the value of the minimum weight ι increases, the range of possible assets narrows, causing portfolios similar to the correct portfolio to cluster together and making it harder to distinguish between incorrect portfolios. As a result, this increases the difficulty of the problem, making it more challenging for LLM to derive the correct choice.

Finally, the "upper bound" method prevents excessive concentration of weight in a specific asset by setting upper limits on the weight of each asset, thereby promoting portfolio diversification. This method constructs portfolio options under the same constraints regardless of the investment objective, and the general constraints applied are as follows given by (18):

$$\begin{aligned} \text{Subject to} \quad & w^{\mathrm{T}}\mathbf{1} = 1 \\ & 0 \leq w_i \leq \upsilon \end{aligned} \tag{18}$$

Here, υ denotes the upper limit of the maximum proportion that can be allocated to asset i. As the maximum proportion limit υ decreases, strategies that concentrate investments on individual assets become more restricted, leading to a more evenly distributed structure of the available portfolios and a dilution of the distinction between correct and incorrect portfolios. As a result, the difficulty of the problem increases, and the likelihood that the LLM will struggle to identify the correct portfolio grows.

A total of 9,500 questions were generated using this structure, with each question reflecting a single investment objective, asset class, investment horizon, and constraint.

Table 3 Answer choice generation method

Investment objective	Objective value
Minimize Volatility	Risk
Maximize Return	Return
Maximize Sharpe ratio	Risk, Return
Minimize MDD	Risk
Minimize CVaR	Tail Risk

Table 4 Distractor generation method (Alternative choice generation)

Generation method	Description
Distance-based	Euclidean distance-based distractors from the answer portfolio
Threshold-based	Optimal value-based distractors from the answer portfolio
Quantile-based	A quantile-based classification of m random portfolios using their optimal optimal values
Dual criteria	A hybrid method combining distance-based and threshold approaches

Table 5 Difficulty modulation via optimization constraints

Constraint method	Description
No constraints	Distractor generation without any constraints on portfolio weights
Assets	Distractor generation method based on number of assets
Lower bound	Minimum weights were set to prevent concentration in specific assets
Upper bound	Maximum weights were set to prevent concentration in specific assets

CHAPTER 5. RESULTS: RISK PROFILES

5.1 Investor profile of LLMs

In this section, we analyzed the basic investor tendencies of LLMs by administering the same investor tendency diagnostic questionnaire 100 times to each of the GPT-4o, Gemini 1.5 Pro, and LlaMA 3.1-70B models.

Figure 1 visualizes the response results, where the size of each circle represents the response frequency, and the center of the circle indicates the actual score.[3] The horizontal axis denotes "risk tolerance" scores, and the vertical axis denotes "time horizon" scores. Responses located further to the lower right indicate more long-term and aggressive investment tendencies. Overall, all three models exhibited moderate or slightly aggressive investment tendencies.

Among these, LlaMA exhibited the most conservative tendency but still fell within the "moderate" category, while Gemini aligned with a typical neutral investor profile. In contrast, GPT demonstrated the most aggressive tendency, falling into the "slightly aggressive" category. These differences are also evident in the average scores and the most frequently selected score combinations presented in Table 6.

Differences were also observed in the consistency level of responses by model. Gemini showed the highest consistency in 100 repeated responses, LlaMA exhibited moderate variability, and GPT demonstrated the highest response variability. Notably, GPT produced a wide range of risk tolerance scores in risk-related questions, characterized by a broad response range.

---

[3] Figure and tables in Section 5 are originally presented in Cho and Kim (2024).

In terms of investment period, all three models showed a tendency to identify themselves as long-term investors. On average, Gemini indicated the longest investment period, followed by LlaMA and GPT.

To quantitatively validate the differences in responses between the three models for each survey question, ANOVA and Kruskal-Wallis (K-W) tests were performed. As shown in Table 7, statistically significant differences were observed in most items. Except when comparing Gemini and LlaMA, the null hypothesis regarding the average response differences between models was rejected for all risk tolerance items.

In summary, major LLMs exhibit distinct underlying investor traits, which may influence the financial advice and portfolio recommendations they provide. Consequently, subsequent analyses will examine whether prompt engineering, which involves assigning predefined investor profiles, can effectively modulate LLM response behavior.

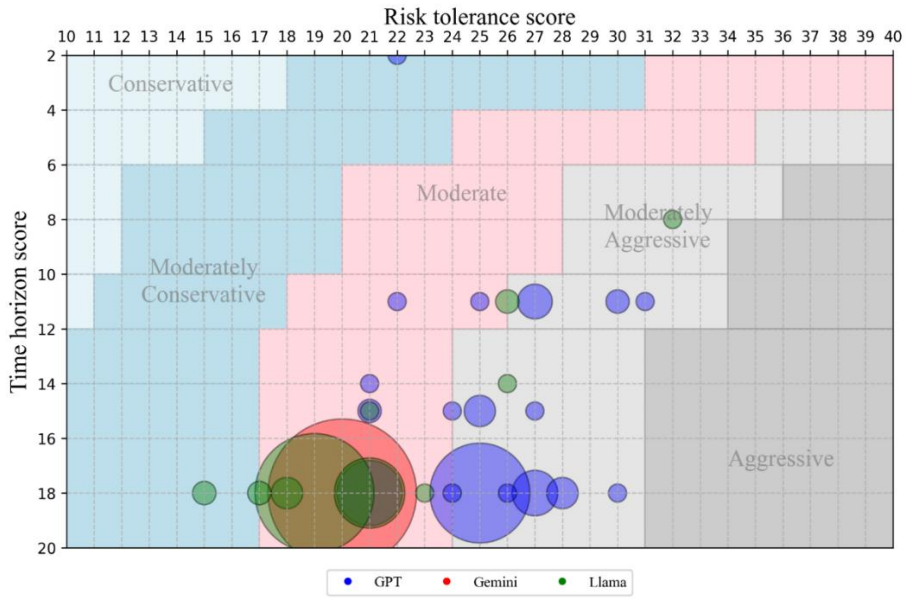

Figure 1 Investor profile of LLMs

Table 6 Average investor profile scores of LLMs

	GPT		Gemini		Llama	
	Time score	Risk score	Time score	Risk score	Time score	Risk score
Average scores	16.86	24.68	18.00	20.00	17.69	19.68
Based on most frequent answers	18	25	18	20	18	19

Table 7 Comparison of responses for each question (*p*-values from ANOVA and K-W tests)

	GPT, Gemini, Llama		GPT & Gemini		GPT & Llama		Gemini & Llama	
	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W
Q1	0.0000	0.0000	0.0000	0.0000	0.0005	0.0007	0.0436	0.0439
Q2	**0.0719**	**0.0560**	0.0158	0.0131	**0.5879**	**0.5217**	**0.0518**	0.0439
Q3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.3673**	**0.3535**
Q5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q7	0.0000	0.0000	0.0000	0.0000	0.0057	0.0087	0.0050	0.0022

p-values greater than 0.05 are shown in bold, and *p*-values less than 0.01 are shown in gray

5.2 Assigning profiles to LLMs

This experiment was conducted to verify how accurately LLM reflects pre-specified investor profiles (e.g., risk tolerance, age, asset level and investment experience). To this end, the first sentence of the prompt was modified for each variable to explicitly assign investment tendencies, and the same investor tendency diagnostic survey was repeated 100 times.

Table 8 summarizes the average risk scores when the three LLM models (GPT-4o, Gemini 1.5 Pro, and LlaMA 3.1-70B) were assigned three risk profiles (risk-averse, risk-neutral, and risk-seeking). In all models, risk scores increased systematically as risk preferences changed, indicating that the persona prompt effectively influenced the LLM's investment decisions. For example, GPT recorded average risk scores of approximately 10 points under risk-averse conditions, around 20 points under neutral conditions, and in the late 30s under risk-preferring conditions, showing statistically significant differences. However, an interesting finding was that the interpretation of the correlation between risk propensity and investment period scores varied across models. GPT exhibited a response pattern where longer investment periods were preferred as risk propensity increased, whereas Gemini tended to select shorter investment periods under risk-preferring conditions. LlaMA responded relatively insensitively to investment period scores.

Table 9 compares the LLM responses when age groups from 20s to 50s were assigned. Overall, both risk scores and investment period scores tended to decrease with increasing age, showing a pattern similar to actual investor behavior. Notably, GPT showed the most pronounced sensitivity, with a difference of approximately 10 points in risk scores between the 20s and 50s conditions, while LlaMA recorded the lowest risk scores overall.

As shown in Table 10, all three models exhibited an upward trend in risk

scores as asset levels increased, consistent with existing behavioral finance research.

Notably, GPT and LlaMA showed a gradual increase in investment period scores as asset levels increased, suggesting that asset capacity may influence long-term investment tendencies. Table 11 analyzes how responses vary depending on the presence and level of investment experience.

All three models showed the lowest risk scores under the “no investment experience” condition, with higher risk scores under some experienced and expert-level experienced conditions. Investment period scores were found to be largely unaffected by investment experience.

Finally, ANOVA and Kruskal-Wallis (K-W) tests were conducted to confirm whether these response differences were statistically significant. For GPT (Table 12, Panel A), most variables, including age, asset level, investment experience, and risk propensity, showed statistical significance with p-values < 0.01, particularly for risk-related items (Q3–Q7), which all reached significance levels below 0.01. For LlaMA (Table 12, Panel C), similar results were observed, suggesting that LLM responses may consistently vary depending on specific investment preference inputs.

Table 8 Average investor profile scores of various risk levels

		GPT		Gemini		Llama	
Risk averse		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	12.31	11.42	18.00	9.70	17.62	8.95
	Based on most frequent answers	9	12	18	10	18	7
Risk neutral		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	14.1	21.73	18.00	17.90	16.74	21.83
	Based on most frequent answers	15	19	18	17	18	21
Risk seeking		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.78	38.96	10.00	34.00	16.74	38.53
	Based on most frequent answers	18	40	10	34	18	38

Table 9 Average investor profile scores of various age groups

		GPT		Gemini		Llama	
20s		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.96	29.86	18.00	24.09	17.46	23.33
	Based on most frequent answers	18	32	18	24	18	24
30s		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.91	26.97	18.00	24.00	17.76	20.31
	Based on most frequent answers	18	27	18	24	18	20
40s		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.01	23.32	18.00	22.00	15.45	20.97
	Based on most frequent answers	18	21	18	22	15	22
50s		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	12.83	19.87	15.00	21.05	14.99	15.20
	Based on most frequent answers	15	21	15	22	15	15

Table 10 Average investor profile scores of various wealth levels

		GPT		Gemini		Llama	
Below average		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	9.95	13.18	18.00	8.05	14.87	7.13
	Based on most frequent answers	9	15	18	7	18	7
Average		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	14.42	19.2	18.00	17.00	14.93	18.51
	Based on most frequent answers	18	19	18	17	15	17
Above average		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.94	29.56	18.00	31.00	17.46	30.81
	Based on most frequent answers	18	30	18	31	18	31

Table 11 Average investor profile scores of various investment experiences

		GPT		Gemini		Llama	
No experience		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.34	9.64	18.00	9.00	17.35	5.15
	Based on most frequent answers	18	9	18	9	18	5
Some experience		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.45	24.73	18.00	20.00	16.7	23.24
	Based on most frequent answers	18	21	18	20	18	24
Professional experience		Time score	Risk score	Time score	Risk score	Time score	Risk score
	Average scores	17.68	26.78	18.00	26.04	17.65	24.09
	Based on most frequent answers	18	23	18	26	18	23

Table 12 Comparison of responses from assigning profiles (*p*-values from ANOVA and K-W tests)

Panel A. Results of GPT

	Risk appetite		Age		Wealth		Investing experience	
	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W
Q1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.1846**	**0.0697**
Q2	0.0000	0.0000	**0.3928**	**0.3916**	0.0000	0.0000	**0.8026**	**0.5289**
Q3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q5	0.0000	0.0000	0.0003	0.0004	0.0000	0.0000	0.0000	0.0000
Q6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0013	0.0014
Q7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Panel B. Results of Gemini

	Risk appetite		Age		Wealth		Investing experience	
	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W
Q1	nan	nan	0.0000	0.0000	nan	nan	nan	nan
Q2	0.0000	0.0000	nan	nan	nan	nan	nan	nan
Q3	nan	nan	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q5	0.0000	0.0000	nan	nan	0.0000	0.0000	0.0000	0.0000
Q6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q7	nan	0.0000	0.0281	0.0287	0.0000	0.0000	nan	nan

Panel C. Results of Llama

	Risk appetite		Age		Wealth		Investing experience	
	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W	ANOVA	K-W
Q1	0.0022	0.0003	0.0000	0.0000	0.0000	0.0000	0.0022	0.0003
Q2	**0.6526**	**0.6014**	0.0007	0.0011	0.0000	0.0000	**0.6526**	**0.6014**
Q3	0.0000	0.0000	0.0105	0.0066	0.0000	0.0000	0.0000	0.0000
Q4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Q5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q6	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000
Q7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

p-values greater than 0.05 are shown in bold, and *p*-values less than 0.01 are shown in gray
'nan' refers to acceptance of the null hypothesis due to the values being all identical

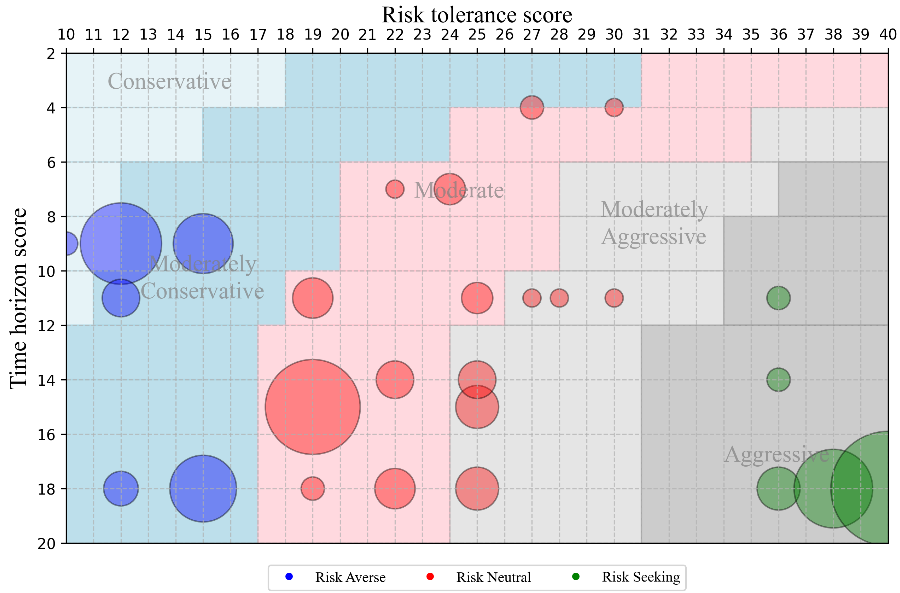

Figure 2 GPT's investor profile by Risk Aversion

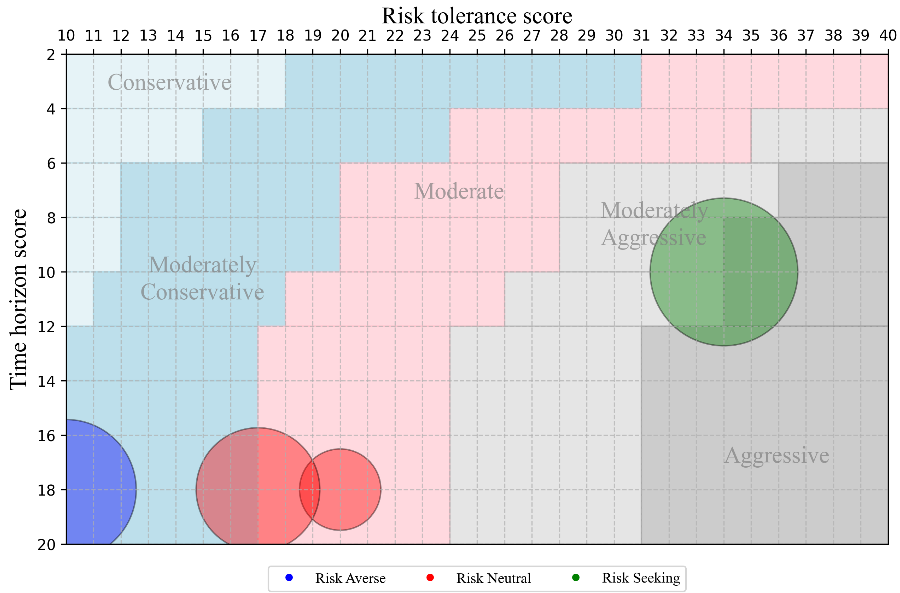

Figure 3 Gemini's investor profile by Risk Aversion

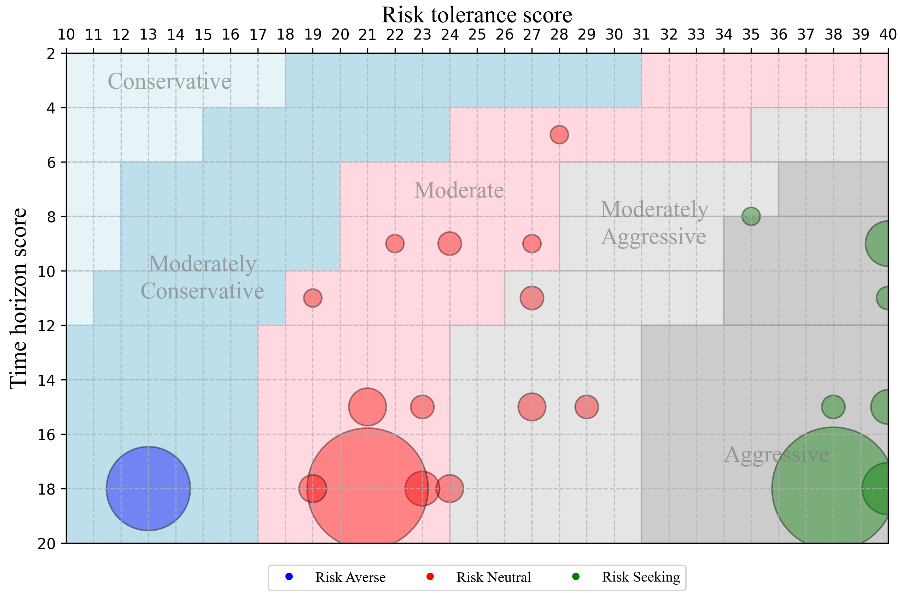

Figure 4 Llama's investor profile by Risk Aversion

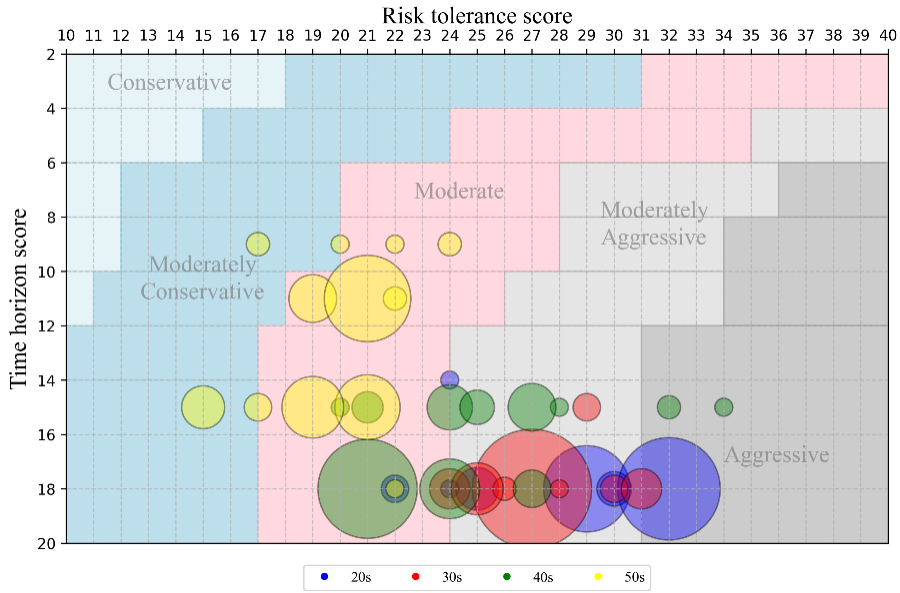

Figure 5 GPT's investor profile by Age

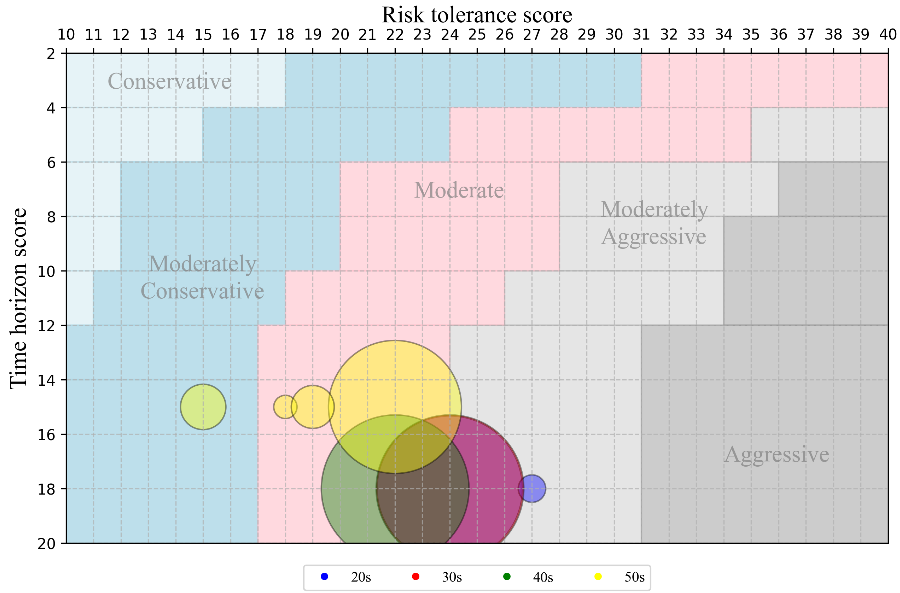

Figure 6 Gemini's investor profile by Age

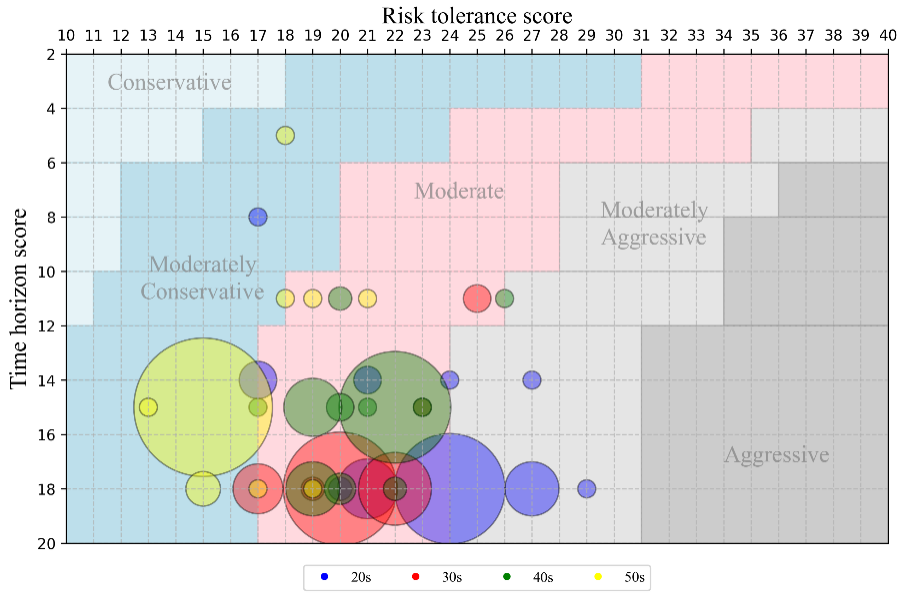

Figure 7 Llama's investor profile by Age

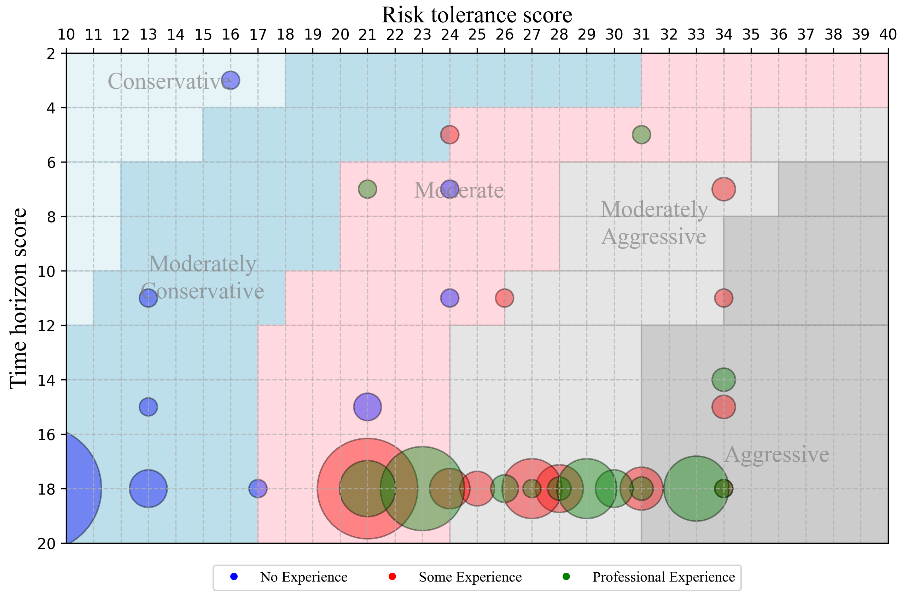

Figure 8 GPT's investor profile by Investing Experience

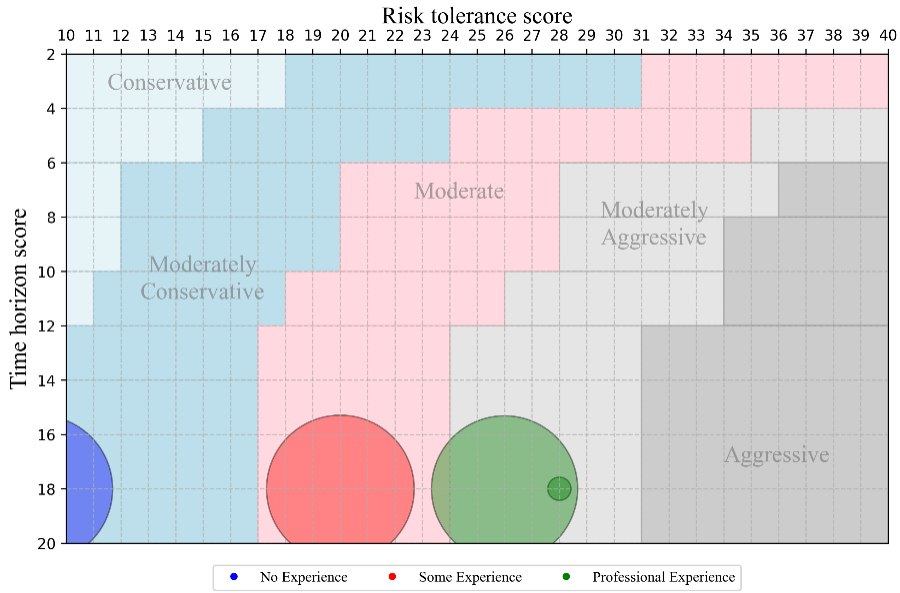

Figure 9 Gemini's investor profile by Investing Experience

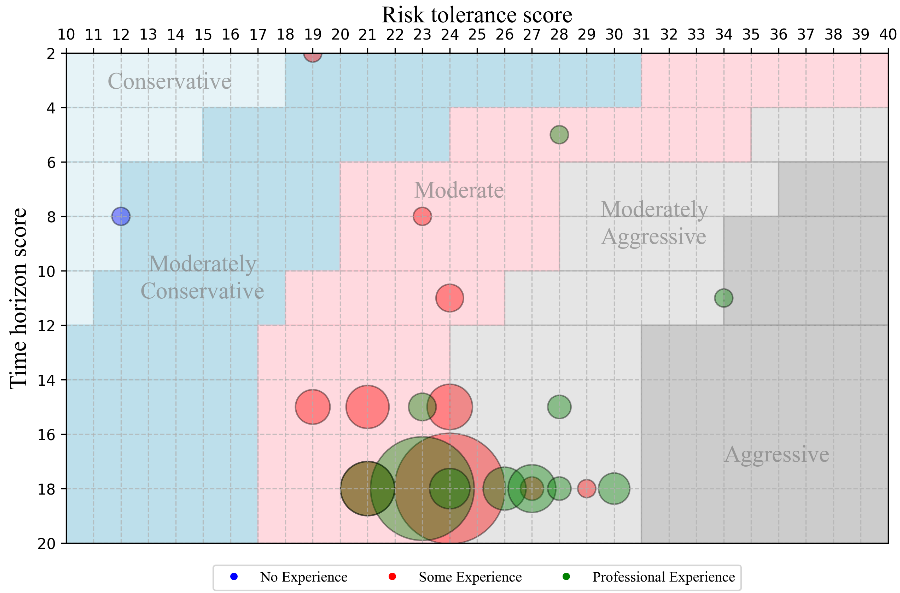

Figure 10 Llama's investor profile by Investing Experience

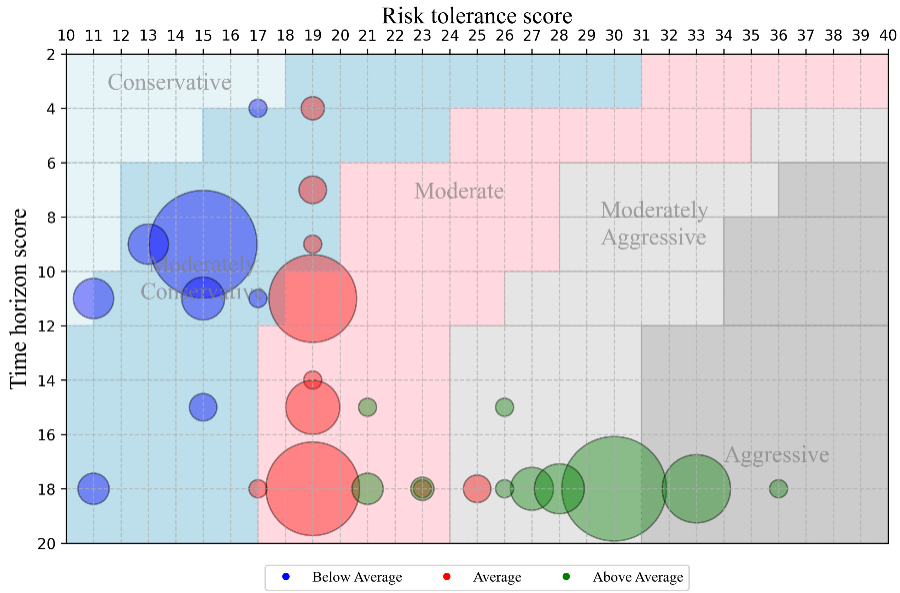

Figure 11 GPT's investor profile by Wealth

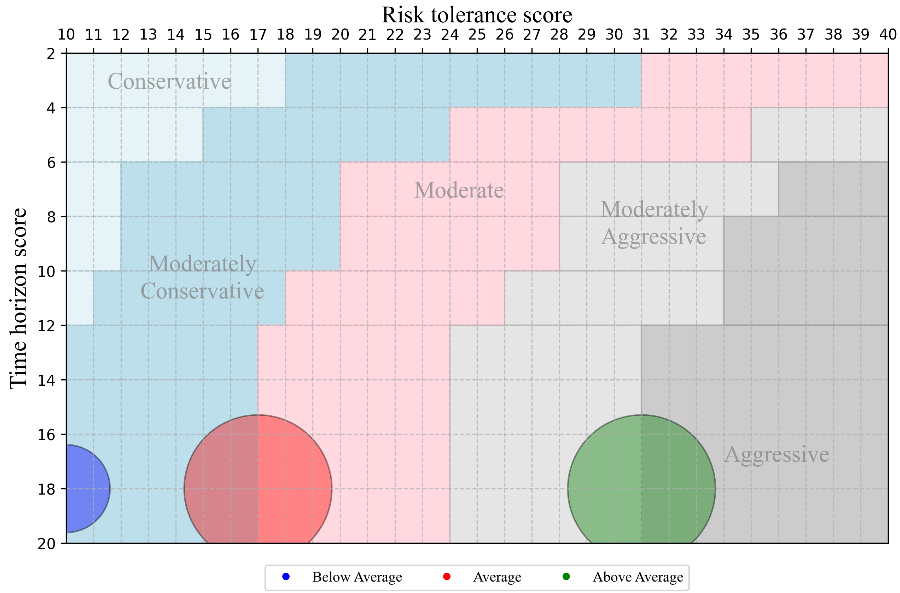

Figure 12 Gemini's investor profile by Wealth

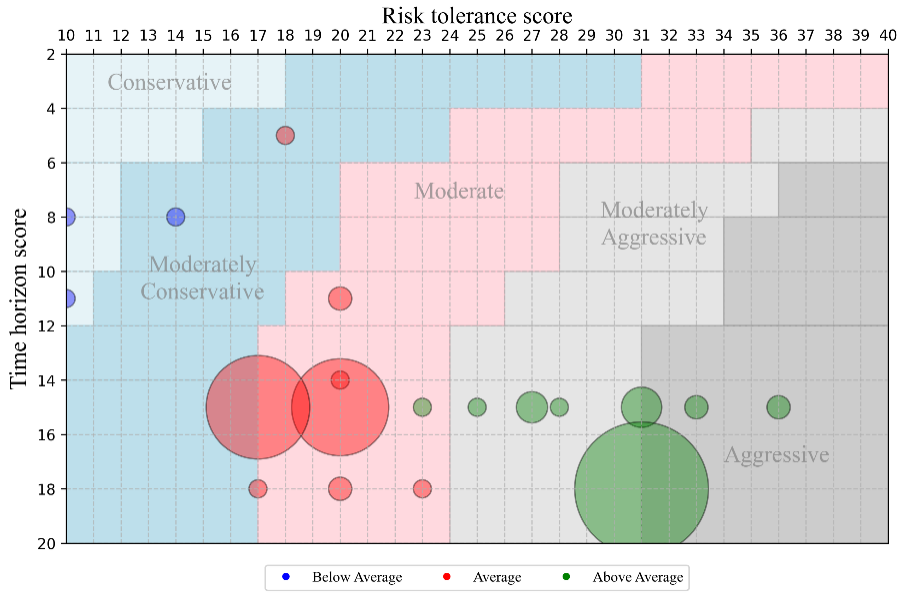

Figure 13 Llama's investor profile by Wealth

CHAPTER 6. RESULTS: OPTIMAL PORTFOLIOS

This chapter presents the results of a quantitative evaluation of LLM's portfolio selection ability and an analysis of its ability to make accurate investment decisions. Response collection was conducted independently for each of the various option generation methods—i.e., Distance-based, Threshold-based, Quantile-based, and Dual-criteria—as well as for each of the various constraint types—No Constraint, Lower Bound, Upper Bound, and Asset Constraint. For each method and condition combination, responses were collected for 100 multiple-choice questions per model, resulting in a total of 9,500 questions for the experiment. Subsequently, these responses were aggregated by investment objective (e.g., minimize volatility or maximize return) and constraint type, enabling a comparison and analysis of the investment decision accuracy of LLMs under different conditions. This structured aggregation method allows for a more precise analysis of how each condition influences LLM decision-making and contributes to a deeper understanding of model-specific performance differences. The accuracy results for all responses are recorded in Tables 14 to 21.

6.1 Results by Investment Objective

Table 13 and Figure 14 present the results comparing the accuracy of LLM under five major investment objectives. GPT achieved the highest accuracy in risk-based objectives such as volatility and MDD, while Gemini demonstrated superior performance in return-based problems. In contrast, Llama exhibited overall low accuracy, particularly revealing weaknesses in composite metrics such as Sharpe ratio. Sharpe ratio is a metric that considers both return and risk, and all three models recorded low accuracy in this task, suggesting a lack of understanding of complex evaluation criteria among LLMs.

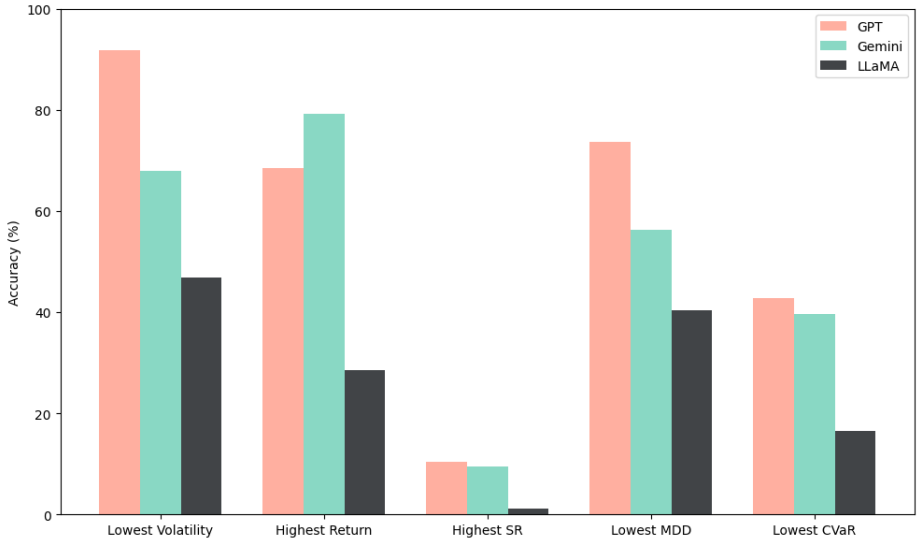

Figure 14 Accuracy by Investment objectives

Table 13 Accuracy by Investment objectives

Investment objective	GPT (%)	Gemini (%)	Llama (%)
Lowest Volatility	91.73	67.89	46.79
Highest Return	68.42	79.26	28.47
Highest Sharpe ratio	10.47	9.53	1.21
Minimize MDD	73.68	56.21	40.42
Minimize CVaR	42.74	39.67	16.58

6.2 Results by Constraint Type

Table 14 and Figures 15 to 19 show how LLM performance varies depending on the type of constraint conditions for each investment objective. GPT maintained consistently high accuracy under all conditions, particularly achieving high accuracy in volatility and MDD problems regardless of the presence or absence of constraint conditions. Gemini showed high accuracy when no constraints were present or when constraints were relatively simple, but its performance tended to decrease when Upper Bound or Asset Constraint conditions were added. Finally, Llama demonstrated overall low performance, with accuracy dropping sharply below 20% in most investment objectives when structural constraints such as Asset Constraint were present. Additionally, in the Sharpe ratio problem, all three models recorded accuracy below 10% under all conditions, indicating difficulties in interpreting and making optimal decisions for the objective function.

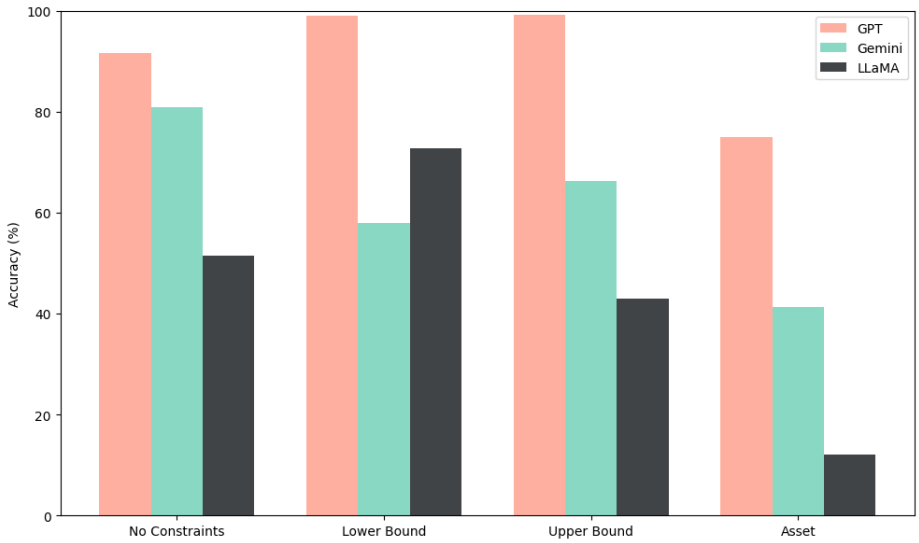

Figure 15 Accuracy by constraint type for minimize volatility

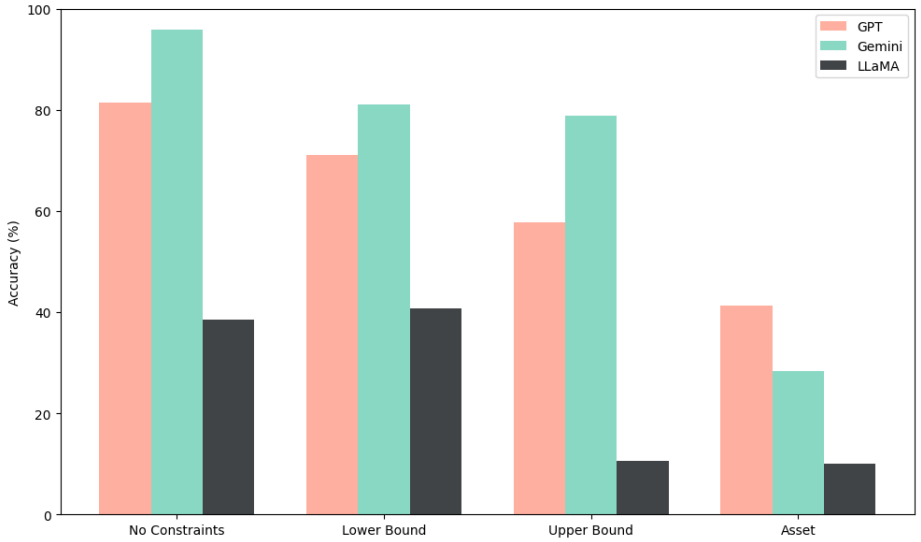

Figure 16 Accuracy by constraint type for maximize return

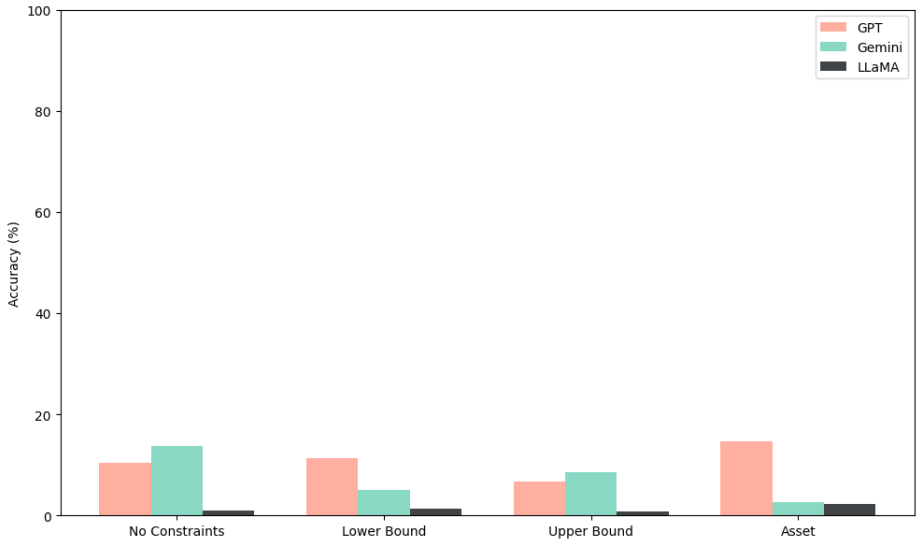

Figure 17 Accuracy by constraint type for maximize Sharpe ratio

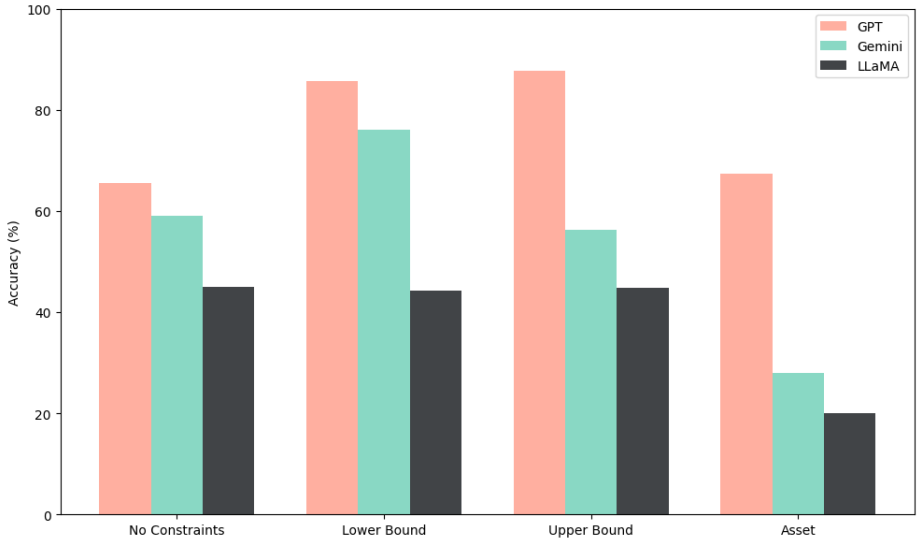

Figure 18 Accuracy by constraint type for minimize MDD

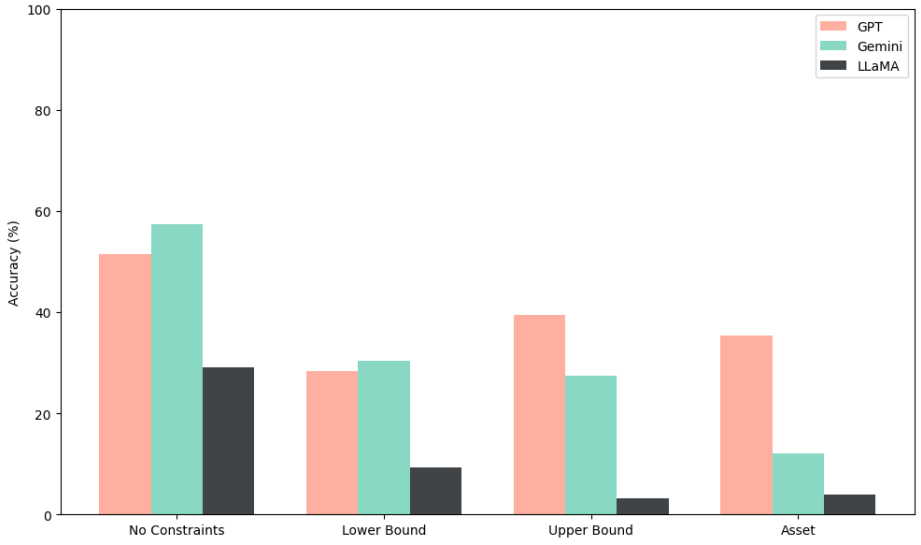

Figure 19 Accuracy by constraint type for minimize CVaR

6.3 Results Analysis

The results of this experiment clearly show that LLM goes beyond simply memorizing and repeatedly responding to investment knowledge, demonstrating different decision-making characteristics depending on the type of investment objective and problem structure.

In particular, GPT recorded the highest accuracy for risk-based objective functions such as volatility and MDD, and maintained consistent judgment capabilities even under various constraints. This suggests that GPT has a relatively superior structural understanding of mathematical concepts related to risk and excels at interpreting and applying mathematically defined objective functions in portfolio selection. The reason GPT achieved such performance is that volatility, MDD, and other metrics have mathematically well-defined solutions, and their objective function structures are relatively simple, making it likely that GPT has sufficiently learned these through prior training.

On the other hand, Gemini showed relatively high accuracy for return-based objective functions, but tended to experience a sharp decline in accuracy when the structure became complex or constraints were included. This suggests that while Gemini has a comparative advantage in making decisions based on single numerical metrics like return, it has limitations in quantitatively interpreting and reflecting structural constraints of a portfolio, such as interdependencies between assets or weighting constraints. In particular, as the number of assets increases or the similarity between options grows, Gemini often relies on simple return figures rather than intuitive judgments, which can hinder accuracy. In this study, the constraints were not explicitly stated in the prompt but were implicitly reflected in the asset allocation structure of the options. Therefore, to select an accurate portfolio, the LLM needed to identify the structural differences between the options and perform reasonable comparative judgments based on the objective function. However, Gemini often tended to

select portfolios with higher objective values in this context, leading to repeated errors where structurally incorrect portfolios were chosen as the correct answer. This demonstrates that Gemini has limitations in inherently reflecting core concepts of portfolio theory, such as asset interdependencies, risk diversification effects, or asset weight constraints.

Llama showed the lowest overall performance, with performance degradation becoming more pronounced in environments with a large number of assets or stricter constraints. This suggests that Llama is relatively weak in comprehensively understanding and processing the mathematical structure of portfolio optimization problems and natural language constraints. In fact, Llama recorded the lowest accuracy in problems with composite objective functions such as Sharpe ratio, indicating a lack of ability to quantitatively integrate the calculation of the ratio between two factors—return and risk—and interpret the trade-offs between them.

A common performance degradation in composite objective functions such as Sharpe ratio and CVaR is a phenomenon observed in all three models. These metrics require consideration of multiple factors such as return and risk rather than a single value judgment, and they necessitate an accurate interpretation of the trade-off structure between options. However, existing LLM models exhibit limitations in comprehensively interpreting and calculating such multidimensional metrics, highlighting the need for improvement in mathematical integration reasoning for composite objective functions.[4]

[4] All experimental results, problem generation pipelines, and response data from this study are publicly available at https://github.com/noahardyx/PortBench.

Table 14 Accuracy by Investment objective and Constraints

		GPT (%)	Gemini (%)	Llama (%)
Minimize Volatility	No Constraint	91.56	80.78	51.44
	Lower Bound	99	58	72.67
	Upper Bound	99.25	66.25	43
	Asset	75	41.33	12
Maximize Return	No Constraint	81.33	95.89	38.56
	Lower Bound	71	81	40.67
	Upper Bound	57.75	78.75	10.5
	Asset	41.33	28.33	10
Maximize Sharpe ratio	No Constraint	10.44	13.78	1
	Lower Bound	11.33	5	1.33
	Upper Bound	6.75	8.5	0.75
	Asset	14.67	2.67	2.33
Minimize MDD	No Constraint	65.56	59	45
	Lower Bound	85.67	76	44.33
	Upper Bound	87.75	56.25	44.75
	Asset	67,33	28	20
Minimize CVaR	No Constraint	51.44	57.44	29.11
	Lower Bound	28.33	30.33	9.33
	Upper Bound	39.5	27.5	3.25
	Asset	35.33	12	4

Table 15 Accuracy by Distance-based

$0 \leq d \leq 0.25$		GPT (%)	Gemini (%)	Llama (%)
Distance-based	Minimize Volatility	74	53	20
	Maximize Return	80	99	49
	Maximize Sharpe ratio	15	28	3
	Minimize MDD	41	22	21
	Minimize CVaR	60	27	33
$0.25 \leq d \leq 0.5$		GPT (%)	Gemini (%)	Llama (%)
Distance-based	Minimize Volatility	96	86	39
	Maximize Return	76	100	25
	Maximize Sharpe ratio	3	10	0
	Minimize MDD	64	68	48
	Minimize CVaR	49	66	26
$0.5 \leq d \leq 0.75$		GPT (%)	Gemini (%)	Llama (%)
Distance-based	Minimize Volatility	96	98	71
	Maximize Return	85	100	17
	Maximize Sharpe ratio	6	6	0
	Minimize MDD	79	86	57
	Minimize CVaR	27	70	17
$0.75 \leq d \leq 1$		GPT (%)	Gemini (%)	Llama (%)
Distance-based	Minimize Volatility	100	98	97
	Maximize Return	83	94	50
	Maximize Sharpe ratio	15	5	0
	Minimize MDD	89	90	80
	Minimize CVaR	37	78	42

Table 16 Accuracy by Threshold-based

$0 \leq \delta \leq 0.25$		GPT (%)	Gemini (%)	Llama (%)
Threshold-based	Minimize Volatility	63	45	10
	Maximize Return	82	98	37
	Maximize Sharpe ratio	9	26	1
	Minimize MDD	31	29	16
	Minimize CVaR	55	24	21
$0.25 \leq \delta \leq 0.5$		GPT (%)	Gemini (%)	Llama (%)
Threshold-based	Minimize Volatility	96	77	32
	Maximize Return	81	96	25
	Maximize Sharpe ratio	11	15	1
	Minimize MDD	35	35	25
	Minimize CVaR	63	52	44
$0.5 \leq \delta \leq 0.75$		GPT (%)	Gemini (%)	Llama (%)
Threshold-based	Minimize Volatility	99	79	37
	Maximize Return	86	93	45
	Maximize Sharpe ratio	18	5	1
	Minimize MDD	69	60	35
	Minimize CVaR	49	62	24

Table 17 Accuracy by Quantile-based

		GPT (%)	Gemini (%)	Llama (%)
Quantile-based	Minimize Volatility	100	99	83
	Maximize Return	84	87	48
	Maximize Sharpe ratio	7	11	3
	Minimize MDD	90	74	55
	Minimize CVaR	50	77	23

Table 18 Accuracy by Dual criteria

		GPT (%)	Gemini (%)	Llama (%)
Dual criteria	Minimize Volatility	100	92	74
	Maximize Return	75	96	51
	Maximize Sharpe ratio	10	18	0
	Minimize MDD	92	67	59
	Minimize CVaR	73	61	32

Table 19 Accuracy by Lower Bound

$\iota = 0$		GPT (%)	Gemini (%)	Llama (%)
Lower Bound	Minimize Volatility	100	95	82
	Maximize Return	83	92	48
	Maximize Sharpe ratio	2	14	0
	Minimize MDD	90	77	57
	Minimize CVaR	48	88	27
$\iota = 0.1$		GPT (%)	Gemini (%)	Llama (%)
Lower Bound	Minimize Volatility	99	42	51
	Maximize Return	71	88	22
	Maximize Sharpe ratio	13	1	2
	Minimize MDD	87	76	54
	Minimize CVaR	20	0	1
$\iota = 0.2$		GPT (%)	Gemini (%)	Llama (%)
Lower Bound	Minimize Volatility	98	37	85
	Maximize Return	59	63	52
	Maximize Sharpe ratio	19	0	2
	Minimize MDD	80	75	22
	Minimize CVaR	17	3	0

Table 20 Accuracy by Upper Bound

$\upsilon = 0.9$		GPT (%)	Gemini (%)	Llama (%)
Upper Bound	Minimize Volatility	100	88	69
	Maximize Return	67	81	12
	Maximize Sharpe ratio	5	8	0
	Minimize MDD	92	86	46
	Minimize CVaR	44	38	8
$\upsilon = 0.8$		GPT (%)	Gemini (%)	Llama (%)
Upper Bound	Minimize Volatility	98	72	55
	Maximize Return	65	79	12
	Maximize Sharpe ratio	7	8	0
	Minimize MDD	89	68	55
	Minimize CVaR	36	27	3
$\upsilon = 0.7$		GPT (%)	Gemini (%)	Llama (%)
Upper Bound	Minimize Volatility	100	57	33
	Maximize Return	59	83	13
	Maximize Sharpe ratio	5	8	0
	Minimize MDD	90	44	40
	Minimize CVaR	39	32	1
$\upsilon = 0.6$		GPT (%)	Gemini (%)	Llama (%)
Upper Bound	Minimize Volatility	99	48	15
	Maximize Return	40	72	5
	Maximize Sharpe ratio	10	10	3
	Minimize MDD	80	27	38
	Minimize CVaR	39	13	1

Table 21 Accuracy by Asset Constraint

$k = 3$		GPT (%)	Gemini (%)	Llama (%)
Asset Constraint	Minimize Volatility	90	73	62
	Maximize Return	54	36	34
	Maximize Sharpe ratio	18	11	15
	Minimize MDD	70	71	61
	Minimize CVaR	21	49	36
$k = 5$		GPT (%)	Gemini (%)	Llama (%)
Asset Constraint	Minimize Volatility	39	51	34
	Maximize Return	42	32	11
	Maximize Sharpe ratio	2	5	1
	Minimize MDD	24	42	18
	Minimize CVaR	7	21	8
$k = 7$		GPT (%)	Gemini (%)	Llama (%)
Asset Constraint	Minimize Volatility	21	13	2
	Maximize Return	28	2	0
	Maximize Sharpe ratio	4	2	1
	Minimize MDD	23	26	11
	Minimize CVaR	5	2	5

CHAPTER 7. CONCLUSION

This study designed a benchmark framework based on portfolio theory to quantitatively evaluate how reasonably major LLMs such as GPT, Gemini, and Llama can make decisions in investment decision-making situations. By generating problems based on objective functions with mathematically defined correct answers, we systematically analyzed whether each model can select the optimal portfolio under various investment objectives and constraints.

The experimental results showed that each LLM exhibited distinct investment preferences and decision-making characteristics. GPT achieved the highest accuracy on risk-based objective functions such as volatility and MDD, and maintained stable performance even under problems with complex constraints. In contrast, Gemini showed strengths in return-based objective functions but experienced a sharp decline in performance when the problem structure became complex or the differences between options were subtle. LlaMA demonstrated the lowest overall accuracy and was particularly vulnerable to composite objective functions such as Sharpe ratio and CVaR.

An interesting point is that even though the constraints were not explicitly stated in the prompt but were only implied in the structural characteristics of the options (e.g., number of assets, and weight), GPT in particular tended to interpret them quantitatively and make reasonable choices. This suggests that LLM has the ability to recognize mathematical patterns and compare structures beyond natural language-based inference. However, Gemini frequently misjudged options with implicit weight constraints, and Llama consistently selected incorrect answers in situations requiring structural inference. This demonstrates that differences in each model's internal reasoning capabilities and understanding of portfolio theory directly translate into performance disparities.

Additionally, this study analyzed not only the ability to select the optimal portfolio based on investment objectives and constraints but also the inherent investment risk propensity of LLMs. Risk propensity is a core element in portfolio composition, reflecting not only the level of risk an investor is willing to tolerate but also their inclination toward risk. Experimental results showed that the three models—GPT, Gemini, and Llama—exhibited distinct underlying investment propensities. Specifically, GPT demonstrated aggressive and volatile responses, Llama exhibited conservative propensity, and Gemini showed relatively moderate and consistent responses.

When we conducted experiments reflecting virtual investor characteristics (e.g., age, wealth, investment experience, and risk aversion) in the options, all LLMs responded by adjusting their risk tolerance, but the sensitivity varied depending on the model. Statistical tests revealed that these differences in responses were not random but rather stemmed from the inherent characteristics of the models and the design of the options. This suggests that for LLMs to be effectively utilized as investor-tailored advisory tools, a structural design that identifies and regulates their inherent tendencies is necessary.

This study established a benchmark framework to quantitatively evaluate the investment decision-making capabilities of LLMs and made the entire code and data publicly available on GitHub to provide a foundation for future research. Notably, the framework demonstrates high scalability and flexibility, as it can theoretically generate an infinite number of problems based on investment objectives, investment periods, and asset class combinations. Furthermore, while this study focused on analyzing multiple-choice questions based on options, future research is expected to collect open-ended responses to conduct qualitative analyses that more deeply evaluate LLM's reasoning processes and understanding of financial concepts.

In summary, this study demonstrated that LLM exhibits different judgment

patterns depending on investment objectives and constraints, and that these differences are not merely random but closely related to the model's mathematical reasoning ability and internal structure. This provides important insights for designing decision-support systems utilizing LLM in the financial field in the future.

REFERENCES

Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2021). LexGLUE: A benchmark dataset for legal language understanding in English. arXiv preprint arXiv:2110.00976.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1-45.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., ... & Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. arXiv preprint arXiv:2109.00122.

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., & Wang, W. Y. (2022). Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. arXiv preprint arXiv:2210.03849.

Cho, H., & Kim, J. H. (2024). Investor risk profile of large language models. Working paper.

Choi, C., Kwon, J., Ha, J., Choi, H., Kim, C., Lee, Y., ... & Lopez-Lira, A. (2025). FinDER: Financial Dataset for Question Answering and Evaluating Retrieval-Augmented Generation. arXiv preprint arXiv:2504.15800.

Choueifaty, Y., & Coignard, Y. (2008). Toward maximum diversification. *The Journal of Portfolio Management*, *35*(1), 40-51.

Dolphin, R., Dursun, J., Chow, J., Blankenship, J., Adams, K., & Pike, Q. (2024). Extracting structured insights from financial news: An augmented

llm driven approach. arXiv preprint arXiv:2407.15788.

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., ... & Ge, J. (2023). Lawbench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289.

Golec, M., & AlabdulJalil, M. (2025). Interpretable LLMs for Credit Risk: A Systematic Review and Taxonomy. arXiv preprint arXiv:2506.04290.

Hamad, H., Thakur, A. K., Kolleri, N., Pulikodan, S., & Chugg, K. (2024, June). FIRE: A Dataset for FInancial Relation Extraction. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3628-3642).

Hecke, T. V. (2012). Power study of anova versus Kruskal-Wallis test. *Journal of Statistics and Management Systems*, *15*(2-3), 241-247.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

Kadam, P. (2024). Enhancing Financial Fraud Detection with Human-in-the-Loop Feedback and Feedback Propagation. arXiv preprint arXiv:2411.05859.

Kim, J. H., Lee, Y., Kim, W. C., & Fabozzi, F. J. (2021). Mean–variance optimization for asset allocation. *The Journal of Portfolio Management, 47*(5), 24-40.

Kim, J. H., Lee, Y., Kim, W. C., Kang, T., & Fabozzi, F. J. (2024). An overview of optimization models for portfolio management, *The Journal of Portfolio Management, 51*(2), 101-117.

Kim, Y., Wu, J., Abdulle, Y., & Wu, H. (2024). MedExQA: Medical question answering benchmark with multiple explanations. arXiv preprint

arXiv:2406.06331.

Lee, Y., Kim, J. H., Kim, W. C., & Fabozzi, F. J. (2024). An overview of machine learning for portfolio optimization, *The Journal of Portfolio Management, 51*(2), 131-148.

Li, W. W., Kim, H., Cucuringu, M., & Ma, T. (2025). Can LLM-based Financial Investing Strategies Outperform the Market in Long Run?. arXiv preprint arXiv:2505.07078.

Markowitz, H. (1952). Portfolio selection. The Journal of Finance, *7*(1). 77-91.

Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022, April). MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning* (pp. 248-260). PMLR.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, *2*, 21-42.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business*, *39*(1), 119-138.

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2018). Commonsenseqa: A question answering challenge targeting commonsense knowledge. arXiv preprint arXiv:1811.00937.

Tertilt, M., & Scholz, P. (2018). To advise, or not to advise—How robo-advisors evaluate the risk preferences of private investors. *The Journal of Wealth Management*, *21*(2), 70-84.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose

language understanding systems. *Advances in neural information processing systems*, *32*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., ... & Huang, J. (2024). Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, *37*, 95716-95743.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence?. arXiv preprint arXiv:1905.07830.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.