Finance Research Letters 78 (2025) 107126

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Can AI beat a naive portfolio? An experiment with anonymized data

Marcelo S. Perlin *, Cristian R. Foguesatto, Fernanda M. Müller, Marcelo B. Righi

Federal University of Rio Grande do Sul, Washington Luis 855, Porto Alegre RS, 90010-460, Rio Grande do Sul, Brazil

ARTICLE INFO

JEL classification:
G110
G170

Keywords:
LLM
Regenerative AI
Artificial intelligence
Gemini
Investments
ChatGPT
Large language models

ABSTRACT

Using anonymized data from the United States (U.S.) market, we evaluate the performance of Google's main LLM (Large Language Model) Gemini 1.5 Flash in making investment decisions. Unlike other studies, we query the LLM for different investment horizons (1 to 36 months) and types of financial information (financial data, price data, and a combination of both). Running a total of 30,000 simulations for 1,522 companies over 20 years of data, we find that Gemini does not consistently outperform a naive portfolio and the S&P 500 index in terms of returns and Sharpe ratios. Additionally, our findings indicate a decline in risk adjusted investment performance as the investment horizon extends.

1. Introduction

Large Language Models (LLMs) are potentially groundbreaking computational innovations in modern society. The ability of LLMs to process and analyze vast amounts of unstructured data, coupled with their ability to generate human-like text, has raised the question of whether they can be leveraged to make efficient decisions in real life. Particular to the case of financial problems, the use of LLMs can help in cases such as trading and portfolio management, managing risk, text mining, and applications of AI in financial advisory services (Brenner and Meyll, 2020; Cucchiara, 2023; Li et al., 2023; Romanko et al., 2023).

Due to its potential, analyzing LLMs' predictive performance is an upward trend in academic research in Finance. Ko and Lee (2024) examines the potential of Open AI's ChatGPT in asset selection and diversification. Using three years of data from 20 large stocks in the United States (U.S.) market, the authors find that ChatGPT's selections are significantly better in diversity than randomly selected assets. In a similar study, Pelster and Val (2024) investigates the performance of ChatGPT in predicting future earnings in a *live* setting, using only past data for analysis and sample from July to September 2023. The authors conclude with positive findings for the return of a strategy based on the ChatGPT's output. Similarly, Kim (2023) investigates ChatGPT's investment recommendations of asset classes under different economic conditions, concluding that ChatGPT can improve portfolio efficiency.

Evaluating the potential of LLM in analyzing textual data, Kirtac and Germano (2024) analyzed the performance of LLMs in the sentiment analysis and portfolio construction of 965,375 U.S. financial news articles from 2010 to 2023. The authors find that ChatGPT outperforms its competitors with accurate predictions of stock market movements after the news. In the same topic, Lopez-Lira and Tang (2023) also investigates the performance of an LLM model in analyzing stock news, finding a positive correlation between forecasted sentiment and the stock's returns. Going further, Li et al. (2024) examines the possibility of ChatGPT acting as a financial advisor for predicting a Chinese firm's performance. The authors conclude that ChatGPT can correct the optimistic

* Corresponding author.
E-mail address: marceloperlin@gmail.com (M.S. Perlin).

https://doi.org/10.1016/j.frl.2025.107126
Received 27 August 2024; Received in revised form 26 February 2025; Accepted 27 February 2025
Available online 11 March 2025

biases of human analysts. For a comprehensive overview of large language models available in the field of Finance, we recommend consulting (Lee et al., 2025).

Given this backdrop, previous literature shows the relevance of LLM in Finance. However, while these efforts have addressed the issue from various angles, determining the appropriate methodology for assessing the predictive performance of an LLM model remains a crucial topic of discussion. If the LLM model was trained on past data and can identify the stocks in the sample, it can also perceive their historical performance and produce deceitful predictions (Pelster and Val, 2024). One critique of the main literature, as highlighted in Ko and Lee (2024) and Pelster and Val (2024), is that it uses non-anonymized data and lacks extensive performance testing. Therefore, we innovate in this study by proposing a unique approach of utilizing a large set of real but anonymized financial data, ensuring the experiment's integrity while maintaining the confidentiality of sensitive information about past performance.

The main objective of this study is to ascertain whether Gemini, Google's LLM, can, in a *blind* experiment, pick stocks and consistently outperform benchmark investment strategies. Our setup circumvents one of the main issues in dealing with historical data in LLM research, ensuring the LLM cannot identify historical performance (Pelster and Val, 2024). The research design involves querying the LLM model with different investment horizons and types of financial information. It enables a comprehensive assessment of its capabilities in navigating complex investment scenarios, including the current yield rate at fixed-income markets.

We contribute to the literature in several significant ways. First, we introduce a novel blind experimental design, which ensures a more rigorous assessment of the forecasting capabilities of LLMs when applied to extensive financial datasets (Pelster and Val, 2024). This approach addresses a critical gap in prior research by mitigating potential biases from non-anonymized data. Second, we expand the understanding of LLM performance in investment contexts by providing robust evidence that, contrary to some prior studies (Kim, 2023; Ko and Lee, 2024), LLMs do not consistently outperform simple investment strategies. Specifically, our findings reveal that Gemini, the LLM model evaluated, fails to outperform either a naive equal-weight portfolio or the S&P 500 regarding total return or Sharpe ratios (return/risk). Furthermore, our study reinforces that LLMs require additional contextual information, such as news headlines, to enhance their forecasting accuracy (Lopez-Lira and Tang, 2023; Ma et al., 2024). Finally, in alignment with recommendations by Ko and Lee (2024), we extend the evaluation of AI-driven portfolio management to long-term scenarios by employing Gemini as an alternative to ChatGPT. This provides a broader perspective on the applicability and limitations of LLMs in diverse financial contexts.

2. Methods and data

In this study, we set up a computational experiment in which we provide Gemini 1.5 Flash, Google's main LLM, anonymized and randomized information about past stock prices and financial statements of U.S. companies.[1] We subsequently instruct Gemini to allocate a total investment of 10,000 USD across five distinct stocks, with the option to include investments at the prevailing interest rate yield during the period. Gemini's performance is evaluated based on the portfolio's composition and its projected future value.

Compared to previous studies, we innovate by creating three sets of pdfs: one containing only financial information, another including only past adjusted prices, and a third combining both. The rationale for using different formats is to simulate reports with varying content. By separating these types of information, we can identify which specific elements from the report the model relies on to generate its performance. The inclusion of past prices is supported by the principles of chart analysis, which aim to predict future price movements based on historical patterns.

Each iteration of the simulation follow these steps:

1. A random date between 2004-01-01 (YYYY-MM-DD) and 2023-12-31 is selected using uniform probabilities;
2. Based on the previous date, we randomly select 5 stocks[2] from a sample of companies that respect all the following rules:
    - Prices and a history of financial documents (income statement, balance sheet, and cash-flow statement) from the last 5 years, counting from the random date, must be available;
    - There must be no penny prices (lower than 1 USD) in the past 5 years of price history;
    - In the past 5 years, the stock is traded with a daily average of financial volume higher than 250,000 dollars.
3. With the 5 random stocks from previous step, we use Quarto to build a single pdf with **past** information for the 5 selected companies in the previous 5 years, with three versions of the file:

    only-financials : a pdf version with only financial information (major registries in the Income Statement, Balance Sheet, and Cashflow statement);

    only-prices : a version with only past adjusted prices of the stock, presented as a figure in the pdf;

    prices-and-financials : a version with financial information and past prices.

[1] Our choice of companies from the United States is due to several reasons, such as the quality and availability of data, the substantial market size and its significant share of the global market capitalization, the highly advanced state of the market, the liquidity of the market, and because it has been used in previous studies with objectives similar to ours (Ko and Lee, 2024).

[2] In our study, we considered portfolios with five stocks because larger portfolios significantly increases the number of tokens in each API calls, and greater computing time, which could have made this study more challenging to execute. Additionally, this sample size found in the portfolio management literature demonstrates good performance. See, for example, Day et al. (2024).

All the numeric data in the pdf, fundamentals and stock prices, are anonymized using a linear random factor[3] between 0.01 and 0.99. By multiplying all numerical values by a single random number, the origin of the data is masked, but all financial ratios remain intact. For example, net margins (profit/sales) are identical in both the anonymized and original datasets. While anonymized data is common in the literature; see, for example, Chang et al. (2024), we believe that we are first to use it is in a simulation study with LLMs. We also remove any identification of the time period.

4. Based on the pdf from the previous step, we construct a query (details on Section 2.1) where we ask Gemini to divide a total of 10.000 USD among the 5 anonymized companies for a given time horizon of investment. We also include the possibility of investing in a current interest rate at the period of the random date.
5. Based on previous choices, we construct Gemini's portfolio for the given future investment horizon and compare its performance to benchmarks (S&P 500 and a naive portfolio with equal weights).

For step 3, we gather **real** financial statements and daily stock prices of U.S. companies from 2004 to 2024, covering 20 years of data and several market cycles. All stock prices are adjusted for dividends, splits, and any other corporate event that can affect the wealth of the investor. After sampling the dates and stocks, we end up with financial information and adjusted prices of 1,522 companies[4] traded in the U.S. equity market.[5] It is also important to highlight that liquidity filters are applied in the simulation steps. Therefore, by focusing on stocks with high liquidity, we are using a sample of companies biased towards mid to large caps.

2.1. The query

We use Gemini version 1.5 Flash, which is the most cost-effective version at the moment of writing this article.[6] All queries used Gemini version gemini-1.5-flash-002, which is the latest stable available model for the 1.5-flash group at the time of writing this article (2024-12-10). Similar to Romanko et al. (2023) and Ko and Lee (2024), we used prompts to achieve the paper's main objective. Using the API, every query in Gemini consists of two parts: system instructions (SI) and prompt. The SI part of the query is used to give context to the LLM model, and steer its behavior towards particular use cases (see Google's help files for details). For this study, we use the following text with placeholders for variables:

System instructions *You are a financial analyst with expertise in analyzing the past performance of companies and picking winning stocks. In this task, you are analyzing past information about 5 publicly traded companies.*

Prompt *Today, you have 10,000 USD to invest for the next N_MONTHS months. You have 6 choices, 5 companies traded in the financial exchange, or invest in the risk-free rate named RISK-FREE-ASSET, which is currently yielding RF_YIELD of return per year. Return the response as json, with 6 elements, with the following structure in each element: "company", "investment"*

The placeholder variables in the query are set as follows:

N_MONTHS The investing horizon in months. We set h = [1, 6, 12, 36];

RF_YIELD The current future yield rate. We use the most recent yield rate of the 5-year U.S. Treasury yield (ticker FVX) available on a random date from the simulation.

Additionally, we attach a pdf file from step 03 of the simulation in every query. Fig. 1 shows an example of a two-page pdf file that includes anonymized financial statements and prices of Apple INC (AAPL). From the heading, one can see that the company's name is company_1-1, with the main information regarding its balance sheet and income statement from the previous 5 years. Notice that, in the pdf, all the numeric data is anonymized, and no dates are presented.

2.2. Simulations

For every combination of the type of pdf (only-financials, only-prices, prices-and-financials) and horizon (h = [1, 6, 12, 36], in months), we conduct 2,500 simulations. Our experiment covers 30,000 simulations and queries to Gemini. We aimed for a sufficient number of simulations to ensure the study's results were stable and unaffected by adding more simulations. However, computational costs posed a limitation, and 2,500 simulations per type of pdf were deemed reasonable.

We evaluate the performance of Gemini in each simulation by independently analyzing[7]:

[3] The range of the factor was an arbitrary choice, and based on our analyses, it did not impact the results.

[4] In our sample, we retained the financial companies, which represent 15% of the dataset. Future research may consider excluding financial companies or analyzing them separately.

[5] All the data was obtained from https://eodhd.com/, a high-quality and popular platform for downloading financial information for companies around the world. We are grateful for the sponsorship of https://eodhd.com/ for this and other studies by our research group.

[6] We would also like to express our gratitude to Google for funding this research through credits provided by Google for Education program. This support enabled us to significantly scale our computational experiment. On a side note, we contacted OpenAI, asking for credits for using ChatGPT in a research setting. We have had no response so far.

[7] The portfolio evaluation criteria were based on indicators commonly used in both the literature and industry. Furthermore, studies with objectives similar to ours, such as Romanko et al. (2023), have employed similar indicators.

Financial Report

Report of company Company_1-1

Financial Information of Company Company_1-1

Balance Sheet
Values are in Millions USD

	Year 5	Year 4	Year 3	Year 2	Year 1
Assets	$9,251.04	$9,493.72	$8,137.07	$7,347.69	$5,864.39
Liabilities	$6,540.75	$6,102.99	$4,893.00	$4,328.60	$3,042.80
Stock Holder Equity	$2,710.29	$3,390.73	$3,244.07	$3,019.09	$2,821.59

Income Statement
Values in Millions USD

	Year 5	Year 4	Year 3	Year 2	Year 1
Revenue	$6,718.25	$5,798.49	$5,454.61	$5,911.84	$4,623.81
Gross Profit	$2,576.03	$2,230.67	$2,131.44	$2,368.28	$1,784.24
EBITDA	$2,069.16	$1,808.62	$1,784.04	$2,086.52	$1,563.56
Operating Income	$1,793.37	$1,551.70	$1,582.64	$1,801.77	$1,328.07
Net Income	$1,505.84	$1,223.04	$1,155.66	$1,350.61	$999.41

Cashflow Statement
Values in Millions USD

	Year 5	Year 4	Year 3	Year 2	Year 1
Net Income	$1,505.84	$1,223.04	$1,155.66	$1,350.61	$999.41
Cashflow from Operations	$1,958.70	$1,608.72	$1,665.02	$2,055.63	$1,510.45
Cashflow from Investing	$406.39	-$1,174.86	-$1,162.99	-$1,423.46	-$571.14
Cashflow from Financing	-$2,222.83	-$438.79	-$518.12	-$448.13	-$949.80
Free CashFlow	$1,621.95	$1,285.07	$1,322.33	$1,765.04	$1,262.22

1

(a) Page 1

Historical Stock Prices

2

(b) Page 2

Fig. 1. Example of Anonymous Financial Report with Fundamental and Price Information.

Return of Gemini's portfolio. A vector of returns from the portfolio suggested by Gemini, for the h months ahead from the sampling date D_0:

$$R^{Gemini}_{j,D_0+t} = \sum_{i=1}^{N} W^{Gemini}_{i} R_{i,D_0+t}, \tag{1}$$

where $R^{Gemini}_{j,t}$ is the return of simulation j for time t, $W^{Gemini}_{j,i}$ is the weight of asset i at simulation j, given by Gemini, and $R_{i,t}$ is the adjusted return of asset i at time t.

Return of Naive portfolio. A vector of returns from the benchmark portfolio with equal weights, including the risk-free asset:

$$R^{N}_{j,D_0+t} = \sum_{i=1}^{N} \frac{1}{6} R_{i,D_0+t}. \tag{2}$$

Notice that each simulation includes 5 stocks and one interest rate. We also include the fixed-income investment in the portfolio, which justifies assigning a naive weight of $\frac{1}{6}$ rather than $\frac{1}{5}$. We use the 5-year U.S. Treasury yield from the simulation as the future risk-free return. This is comparable to buying a bullet bond at current market prices and keeping it for the given time horizon.

Return of S&P 500 index. A vector of returns from the S&P 500 index:

$$R^{S\P500}_{j,D_0+t} = \frac{P^{S\&P500}_{D_0+t}}{P^{S\&P500}_{D_0+t-1}} - 1, \tag{3}$$

where $P^{S\&P500}_{t}$ is the value of the index at time t.

Total Return of Portfolio. The total future return from the portfolio suggested by Gemini:

$$TotalRet_j = \prod_{t=1}^{h} \left(1 + R^{Gemini}_{j,D_0+t}\right) - 1. \tag{4}$$

Table 1
Descriptive statistics of raw financial data.

GIC sector	Number of companies	Median Ret. per year	Std. Dev. of Ret. per year	Median ROE	Median net margin
All sectors	1634	7.99%	9.05%	11.22%	7.48%
Communication services	79	3.63%	8.65%	10.59%	7.56%
Consumer discretionary	211	7.55%	9.58%	14.55%	5.16%
Consumer staples	75	8.58%	6.45%	16.68%	6.06%
Energy	106	6.30%	8.79%	10.62%	6.36%
Financials	242	6.29%	7.29%	10.90%	20.07%
Health care	196	6.43%	12.58%	7.96%	4.55%
Industrials	245	10.06%	7.44%	13.53%	5.88%
Information technology	213	9.61%	10.75%	10.39%	7.03%
Materials	114	8.41%	6.72%	11.72%	6.10%
Real estate	95	6.68%	6.61%	6.07%	15.51%
Utilities	58	9.24%	4.75%	9.34%	8.96%

Notes: This table shows descriptive statistics for 1634 companies for the US equity market between 2004 and 2024. All financial data was obtained from https://eodhd.com/.

Annualized Total Return of Portfolio. The annually comparable total future return from the portfolio suggested by Gemini:

$$AnnualRet_j = (1 + TotalRet_j)^{\frac{12}{h}} - 1, \tag{5}$$

where h is the trading horizon.

Sharpe ratio. Using the vector of future returns of the suggested portfolio, we calculate the Sharpe ratio as the excessive mean return divided by the standard deviation (σ). The risk free rate R_f is defined as the 5-year treasury bond yield.

$$S_j = \frac{E\left(R^{Gemini}_{i,t}\right) - R_f}{\sigma\left(R^{Gemini}_{i,t}\right)}. \tag{6}$$

Beta and Alpha. Using the vector of future returns of the suggested portfolio, we calculate the beta (systemic risk) and alpha (unconditional constant return) from the market model (7), estimated from a standard ordinary least squares (OLS) model.

$$R^{Gemini}_{j,D_0+t} = \alpha + \beta R^{S\&P500}_{j,D_0+t} + \epsilon_{j,D_0+t}, \tag{7}$$

where α is a measure of intrinsic returns, unrelated to market conditions, β represents the systematic risk, and ϵ represents the residual error of the ordinary least squares model.

We evaluate Gemini's performance by comparing it against the benchmarks: the **S&P 500 index** and a **naive portfolio**. We consider a naive portfolio due to the good results observed with this strategy (see DeMiguel et al., 2009). Besides calculating medians of return and risk, we check the proportion of cases where Gemini presented higher return, higher Sharpe ratio, and positive and significant Jensen's alpha for each pair of trading horizons and type of pdf. If, based on past data, Gemini is a consistent and over-performing investor, then it must be able to present a proportion of over-performance higher than 50% against the benchmarks.

3. Results

Table 1 presents the descriptive statistics of the data sampled in this simulation exercise. We have 1,522 companies, with a reasonable distribution of companies across the industries. Notice that the median annual returns are positive, indicating a positive period for the American equity market. Across all companies, we find a median return of 7.94% per year. We also find positive median ROE (*return on equity*) and median net margin (*profit by revenues*) across all sectors.

In Fig. 2, we present the distribution of investments for all simulations across types of pdfs and investment horizons (h). Across all time horizons and data types, the investment distributions are heavily skewed to the left, indicating that a large proportion of the simulations resulted in lower invested values. As the time horizon increases, there seems to be a slight shift in the distribution towards higher invested values, especially when prices are included in the input data. This suggests that including price information and having a longer time horizon can potentially lead to higher investment outcomes, although the overall skew towards lower values remains. Furthermore, the inclusion of both price and financial data generally produces a slightly wider spread of outcomes compared to using only financial data. This result is very interesting and shows how Gemini can respond differently according to the structure of the query, justifying our methodological choice of including the type of pdf and investment horizon in the analysis. This result aligns with the findings of Pelster and Val (2024), who investigated ChatGPT's capability in asset selection for portfolio diversification.

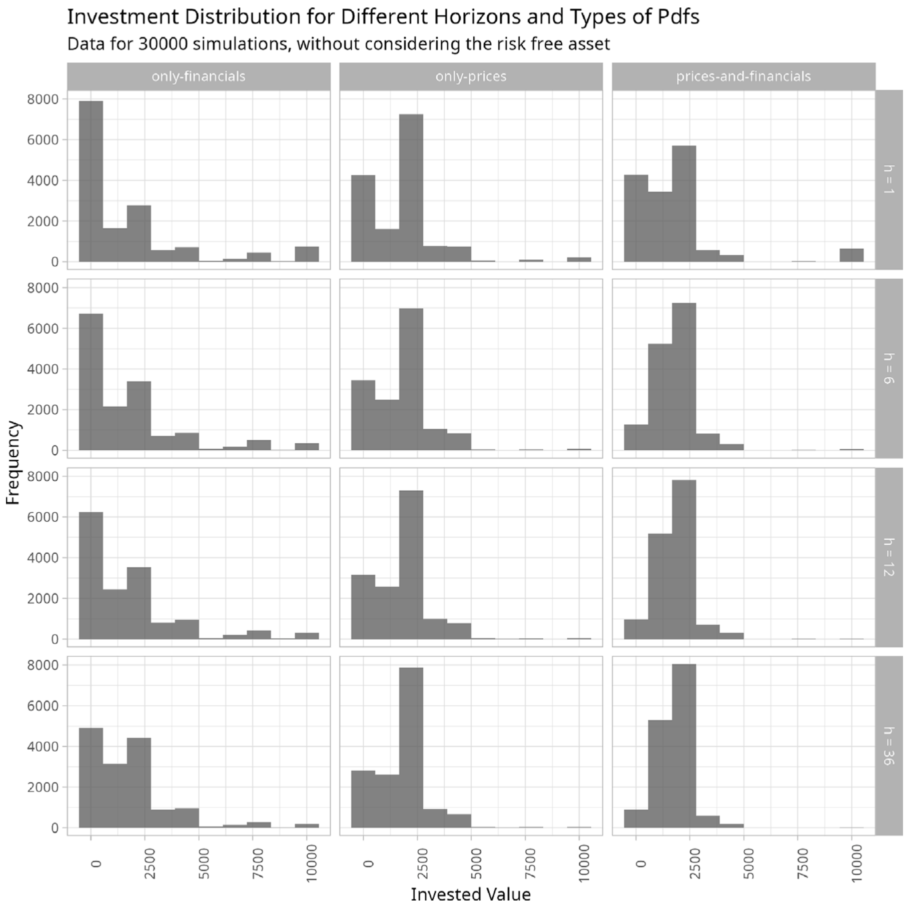

Fig. 2. Distribution of Investment for all simulations. For each combination of type of pdf and investment horizon (months), this figure shows the frequency of investment values.

Our next result is the performance of the LLM model in its investment decisions, as shown in Table 2. First, we compare the portfolio's risk, measured by its Beta (systematic risk), against the benchmarks. Overall, the median Beta values in the Gemini portfolio are close to one, presenting systemic risk close to the market index itself (Beta equals one). The odd case is for *only-financials*, where the betas are in the 0.75 range. Regarding Gemini portfolios, the *only-prices* portfolio presents the highest Beta across all horizons, suggesting greater exposure to systemic risk.

In columns 4 to 6 of Table 2, we have the annualized median returns for each pair of h and pdf type. Since we are comparing different investment horizons, the annualization brings all total returns to the same unit, the return equivalent per year. Please note that all returns in Table 3 are annualized and subject to outliers. An extreme monthly return for the index can result in outlier values for annualized returns, thereby inflating median figures. This explains the seemingly suspicious median annual return of 22.24% for the S&P 500 over a one-month window. When comparing Gemini against its benchmarks, we verify that the median return for Gemini is the lowest in most pairs of pdf types and horizons. We also find that, as the horizon increases, the median returns tend to decrease in all investments, including the benchmarks. Surprisingly, when comparing the median returns across different types of pdf, we find that *only-prices* presented a higher median for most trading horizons. The *only-prices* portfolios also have, when we consider h = 6, 12, and 36, superior median annualized returns than the S&P 500 index. For this same h, the *only-financials* portfolio tends to have the lowest annualized returns. This finding may suggest that excluding prices in portfolio construction may limit performance in terms of returns.

A notable observation from the Sharpe ratios (columns 7 to 9) is that despite lower returns, the Gemini portfolio achieves competitive risk-adjusted performance relative to the Naive portfolio. However, despite similar Sharpe ratios, the median annualized Sharpe ratios of the Naive portfolio is higher than the Gemini portfolio in almost all cases. The S&P 500 portfolio consistently maintains the highest Sharpe ratios, particularly in the short term, indicating a superior balance of risk and return in its diversified structure. These findings suggest that the Gemini portfolio's effectiveness is context-dependent: while it struggles to outperform traditional benchmarks on raw returns, its competitive risk-adjusted performance in certain conditions — especially when leveraging price data — indicates potential for refinement in the modeling approach. Additionally, the consistent underperformance of the *only-financials* portfolio highlights the critical role of price data in driving investment success.

Table 2
Results for investment performance of gemini and benchmarks.

Horizon (Months)	Median beta (Systemic risk)		Median annualized returns			Median annualized Sharpes		
	Gemini	Naive	Gemini	Naive	SP500	Gemini	Naive	SP500
Only-financials								
1	**0.72**	0.92	9.21%	19.53%	25.24%	**1.058**	1.254	1.833
6	**0.72**	0.91	9.29%	14.19%	14.09%	**0.872**	0.895	1.143
12	**0.73**	0.91	8.81%	13.31%	13.47%	0.795	0.791	0.942
36	**0.81**	0.91	9.67%	12.23%	11.37%	0.685	0.702	0.744
Only-prices								
1	1.02	0.92	**22.26%**	19.57%	25.26%	**1.254**	1.254	1.833
6	0.99	0.91	**14.25%**	14.19%	14.09%	0.859	0.895	1.143
12	1.01	0.91	**13.67%**	13.31%	13.47%	0.776	0.791	0.942
36	1.01	0.91	12.63%	12.23%	11.37%	0.677	0.702	0.744
Prices-and-financials								
1	0.97	0.92	**17.88%**	19.53%	25.24%	**1.067**	1.254	1.833
6	**0.94**	0.91	**13.85%**	14.19%	14.09%	**0.879**	0.895	1.143
12	0.96	0.91	13.27%	13.31%	13.47%	0.787	0.791	0.942
36	0.97	0.91	12.48%	12.23%	11.37%	0.688	0.702	0.744

Notes: This table shows descriptive statistics from comparing the investment performance of Gemini against two benchmarks : SP500 and a equal weighted portfolio. The results are based on 30,000 simulations, 2.500 for each pair of horizon and type of pdf. For each combination of trading horizon and type of pdf, we present the median beta (systemic risk) from the market model (Eq. (7)), median annualized returns (Eq. (5)) and median annualized Sharpe ratios (Eq. (6)). The five best values in each column, highest for return/Sharpe and lowest for beta, are highlighted in bold.

Table 3
Proportion results of Gemini simulation.

Horizon (Months)	Comparing against SP500		Comparing against naive portfolio		% of J. Alpha Pos. and Sign.
	% of returns above	% of Sharpe above	% of returns above	% of Sharpe above	
Only-financials					
1	46.28%	42.80%	46.28%	47.64%	**6.92%**
6	41.96%	41.52%	41.84%	46.96%	**6.04%**
12	41.48%	41.52%	40.96%	46.72%	**7.04%**
36	40.92%	42.08%	39.88%	45.68%	**9.32%**
Only-prices					
1	**53.14%**	43.50%	**55.06%**	32.13%	**4.00%**
6	**53.12%**	42.68%	**56.40%**	36.72%	**6.68%**
12	**54.28%**	41.84%	**57.72%**	33.80%	**5.92%**
36	**55.28%**	41.20%	**61.44%**	27.80%	**7.44%**
Prices-and-financials					
1	**50.00%**	41.08%	**50.32%**	42.04%	**4.76%**
6	**51.24%**	42.00%	**52.64%**	47.16%	**6.12%**
12	**53.16%**	41.96%	**54.12%**	42.64%	**5.84%**
36	**54.44%**	41.76%	**58.04%**	38.12%	**7.52%**

Notes: This table shows the results of comparing the investment performance of Gemini against two benchmarks : SP500 and a equal weighted portfolio. This results are based on 30.000 simulations, 2.500 for each case. For each combination of trading horizon and type of pdf, we present the proportion of portfolios with total returns/Sharpe ratios higher than the benchmark. The last column of the table shows the percentage of alpha (see Eq. (7)) with positive sign and statistical significance at 10%.

Table 3 describes the proportion of cases where the Gemini portfolio presented better results than its benchmarks regarding returns and Sharpe ratios. For returns, the worst performance was for *only-financials*, with proportions around 40%. When looking at the results for *only-prices* and *prices-and-financials*, we find proportions higher than 50%, reaching a maximum of 61.44% for $h = 36$ against the naive portfolio. In terms of returns, the best results are found for the *only-prices* portfolio, which surpasses the Naive portfolio in returns in all time horizons.

However, when looking at the trade-off between return and risk, measured by the Sharpe ratios, the results are worse, with proportions ranging from 27.8% to 47.64%. Additionally, it is worth noting that the *only-prices* portfolio, the one with highest proportion of excessive returns, struggles the most in terms of Sharpe ratios against the Naive portfolio in longer horizons ($h = 36$, with only 27.8%). This highlights the limited risk-adjusted performance of this approach.

Another interesting result from Table 3 is that the proportions of higher Sharpe ratios get lower as the trading horizon increases. This result suggests that, for higher time horizons, the Gemini portfolio performance becomes worse than its benchmarks. Looking at the last columns of Table 3, we find a low proportion of positive and significant Jensen's alpha, ranging from 4.76% to 9.32%. This means that the Gemini portfolio has only been able to present consistent returns above the market index in a small proportion of the simulations. Interestingly, the *prices-and-financials* portfolio shows the highest proportion of significant positive alpha at

$h = 36$ (9.32%), suggesting that combining price and financial data may slightly improve long-term performance. However, even this proportion remains relatively low, underscoring the difficulty of achieving statistically significant outperformance.

4. Conclusions and limitations

This study assesses the trading performance of Gemini, a prominent regenerative AI model developed by Google, through an anonymized experiment using real-world data. Our analysis benchmarks Gemini against two standards: a naive portfolio and the S&P 500 index. The findings reveal no significant empirical evidence that Gemini consistently outperforms these benchmarks in terms of return and risk. Specifically, Gemini's success rate, measured by the proportion of cases with better performance than the benchmarks, is approximately 52%, close to a unbiased coin flip, indicating no substantial advantage over naive investment strategies. Furthermore, the model's performance in Sharpe ratios is even less impressive, with most simulations yielding results below 50%. A notable trend is the decline in risk adjusted performance as the trading horizon extends, indicating potential limitations in Gemini's predictive capabilities over longer periods.

A key limitation of this study is its exclusive focus on Google's Gemini as the sole large language model (LLM) evaluated. Previous research has predominantly analyzed ChatGPT (Romanko et al., 2023; Ko and Lee, 2024; Pelster and Val, 2024; Ma et al., 2024), which employs distinct architectures, training data, and optimization strategies. These differences could lead to variations in investment decision-making processes and outcomes. While we suspect this study's core conclusions are generalizable to other LLMs, it is unclear how significant anonymization shapes our results. This ambiguity highlights a promising avenue for future research, particularly in evaluating the impact of data anonymization on model performance.

Future research should also examine the relative effectiveness of Gemini and other LLMs under dynamic trading strategies and market conditions (Kim, 2023). Comparative studies provide valuable insights into the performance of models like Gemini and ChatGPT, offering a more nuanced understanding of their suitability for investment decision-making. Furthermore, incorporating tail risk measures into future analyses could enhance our understanding of these models' resilience in extreme market scenarios. Such efforts would clarify the role of anonymization and contribute to a more comprehensive evaluation of LLMs' potential as tools for financial decision-making.

CRediT authorship contribution statement

Marcelo S. Perlin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Cristian R. Foguesatto:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Fernanda M. Müller:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Marcelo B. Righi:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Gemini and ChatGPT in order to improve language and readability of the final text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Data availability

The authors do not have permission to share data.

References

Brenner, L., Meyll, T., 2020. Robo-advisors: A substitute for human financial advice? J. Behav. Exp. Financ. 25, 100275.
Chang, V., Hahm, N., Xu, Q.A., Vijayakumar, P., Liu, L., 2024. Towards data and analytics driven B2B-banking for green finance: A cross-selling use case study. Technol. Forecast. Soc. Change 206, 123542.
Cucchiara, R., 2023. How AI is transforming scientific research and discovery. Nature URL: https://www.nature.com/articles/d43978-023-00095-8. (Accessed: 20 February 2025).
Day, M.-Y., Yang, C.-Y., Ni, Y., 2024. Portfolio dynamic trading strategies using deep reinforcement learning. Soft Comput. 28 (15), 8715–8730.
DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? Rev. Financ. Stud. 22 (5), 1915–1953.
Kim, J.H., 2023. What if ChatGPT were a quant asset manager. Financ. Res. Lett. 58, 104580.
Kirtac, K., Germano, G., 2024. Sentiment trading with large language models. Financ. Res. Lett. 62, 105227.
Ko, H., Lee, J., 2024. Can ChatGPT improve investment decisions? From a portfolio management perspective. Financ. Res. Lett. 64, 105433.
Lee, J., Stevens, N., Han, S.C., 2025. Large language models in finance (FinLLMs). Neural Comput. Appl. 1–15.
Li, X., Feng, H., Yang, H., Huang, J., 2024. Can ChatGPT reduce human financial analysts' optimistic biases? Econ. Political Stud. 12 (1), 20–33.
Li, Y., Wang, S., Ding, H., Chen, H., 2023. Large language models in finance: A survey. In: Proceedings of the Fourth ACM International Conference on AI in Finance. pp. 374–382.
Lopez-Lira, A., Tang, Y., 2023. Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619.
Ma, F., Lyu, Z., Li, H., 2024. Can ChatGPT predict Chinese equity premiums? Financ. Res. Lett. 105631.
Pelster, M., Val, J., 2024. Can ChatGPT assist in picking stocks? Financ. Res. Lett. 59, 104786.
Romanko, O., Narayan, A., Kwon, R.H., 2023. ChatGPT-based investment portfolio selection. In: Operations Research Forum, Vol. 4. Springer, p. 91.