

Can LLM-based Financial Investing Strategies Outperform the Market in Long Run?

Weixian Waylon Li¹ Hyeonjun Kim² Mihai Cucuringu^{3,4} Tiejun Ma¹

¹School of Informatics, University of Edinburgh

²Global Finance Research Center, Sungkyunkwan University

³Department of Mathematics, University of California Los Angeles

⁴Department of Statistics, University of Oxford

{waylon.li, tiejun.ma}@ed.ac.uk

Abstract

Large Language Models (LLMs) have recently been leveraged for asset pricing tasks and stock trading applications, enabling AI agents to generate investment decisions from unstructured financial data. However, most evaluations of LLM timing-based investing strategies are conducted on narrow timeframes and limited stock universes, overstating effectiveness due to survivorship and data-snooping biases. We critically assess their generalizability and robustness by proposing FINSABER¹, a backtesting framework evaluating timing-based strategies across longer periods and a larger universe of symbols. Systematic backtests over two decades and 100+ symbols reveal that previously reported LLM advantages deteriorate significantly under broader cross-section and over a longer-term evaluation. Our market regime analysis further demonstrates that LLM strategies are overly conservative in bull markets, underperforming passive benchmarks, and overly aggressive in bear markets, incurring heavy losses. These findings highlight the need to develop LLM strategies that are able to prioritise trend detection and regime-aware risk controls over mere scaling of framework complexity.

1 Introduction

Large language models (LLMs) are increasingly used in financial decision-making, especially for generating investment actions such as Buy, Hold, or Sell (Ding et al., 2024a; Fatouros et al., 2025). These so-called LLM *timing-based investing strategies* leverage LLMs’ ability to interpret historical and real-time data to autonomously trade. From sentiment-driven trading (Zhang et al., 2024a) to sophisticated multi-agent systems (Yu et al., 2023; Zhang et al., 2024b), a growing body of work has explored the potential of LLMs as autonomous financial agents.

¹Data and code available at <https://github.com/waylonli/FINSABER/>.

Backtesting is the standard method for assessing investment strategies, simulating them on historical data to evaluate profitability and robustness (Chan, 2021). However, current LLM investing research suffers from fragmented, underdeveloped evaluation practices. Most studies assess performance over short periods, on few stock symbols, and often omit code release, limiting reproducibility. As summarised in Table 1, several recent methods evaluate over under a year, with fewer than ten stocks, and benchmark only against naïve baselines like Buy-and-Hold. Such short horizons and narrow stock universes lead to three well-documented sources of bias: **survivorship bias** (Garcia and Gould, 1993), where delisted or failed stocks are omitted; **look-ahead bias** (Chan, 2021), where future information inadvertently influences past decisions; and **data-snooping bias** (Bailey et al., 2015), where strategy performance is inflated through repeated testing on the same data. These biases can result in misleading performance assessments and undermine the validity of claimed improvements over traditional methods. This raises a central question: **Can LLM-based investing strategies survive longer and broader robustness evaluations?**

Method	Type (Period, Symbols)	Code
MarketSenseAI	Sentiment Driven (1y3m, 100)	✗
TradingGPT	Multi Agents (N/A, N/A)	✗
FinMem	Multi Agents (6m, 5)	✓
FinAgent	Multi Agents (6m, 6)	✓
FinRobot	Multi Agents (N/A, N/A)	✓
TradExpert	Multi Agents (1y, 30)	✗
FinCon	Multi Agents (8m, 8)	✗
TradingAgents	Multi Agents (3m, 3)	✗
MarketSenseAI 2.0	Multi Agents (2y, 100)	✗

Table 1: Summary of current LLM-based investing strategies. Evaluation period and symbol coverage are combined for compactness.

While recent efforts such as Wang et al. (2025) and Hu et al. (2025) have begun to address backtesting of deep learning (DL)-based strategies and benchmarking of LLM-based time-series forecast-

ing, a standard framework for rigorously evaluating LLM investing strategies remains absent. Moreover, these works do not effectively tackle the mitigation of biases in quantitative trading. To fill this gap, we introduce **FINSABER**, a comprehensive framework for benchmarking LLM timing-based investing strategies that supports **longer backtesting periods**, a **broadener and more diverse symbol universe**, and explicit bias mitigation. Specifically, our main contributions are:

1. We propose FINSABER, the first comprehensive evaluation framework for LLM-based investing strategies that supports 20 years of multi-source data, including unstructured inputs such as news and filings, expands symbol coverage via unbiased selection, and mitigates survivorship, look-ahead, and data-snooping biases.
2. We empirically reassess prior claims and show that LLM advantages reported in recent studies often vanish under broader and longer evaluations, indicating that many conclusions are driven by selective or fragile setups.
3. We conduct regime-specific analysis and reveal that LLM strategies underperform in bull markets due to excessive conservatism and suffer disproportionate losses in bear markets due to inadequate risk control.
4. We offer guidance for future LLM strategy design, arguing that regime-awareness and adaptive risk management are more critical than increasing architectural complexity.

Altogether, our work provides empirical guidance for LLM-based investment research, advocating for the development of strategies that are able to adjust to dynamically-changing market conditions.

2 Related Works

Recent work using LLMs as investors directly employ LLMs to make investing decisions (Ding et al., 2024a). The most common approach leverages LLMs’ sentiment analysis capabilities, using either general-purpose LLMs (e.g., GPT, LLaMA (Touvron et al., 2023), Qwen (Bai et al., 2023)) or fine-tuned financial variants like FinGPT (Yang et al., 2023) to generate sentiment scores for trading decisions (Lopez-Lira and Tang, 2023; Wu, 2024; Kirtac and Germano, 2024; Zhang et al., 2024a). However, these approaches stop short of forming complete trading strategies, which require not only

directional forecasts, but also realistic liquidity sizing for mitigating impact, development of execution rules for trade timing and risk management, and incorporation of trading costs.

More advanced approaches move beyond sentiment scores by summarising and reasoning over multi-source financial text. For example, Fatouros et al. (2024) introduce a memory module that stores summarised financial data, retrieved during trading to guide decisions. Similarly, LLMFactor (Wang et al., 2024) learns to extract profitable factors from historical news aligned with price movements and applies them to future market forecasts.

A growing body of work incorporates LLM-based agents (Guo et al., 2024), where either one specialised agent or multiple collaborative agents are employed to perform financial analysis or predictions. Notable examples include FinMem (Yu et al., 2023), FinAgent (Zhang et al., 2024b), FinRobot (Yang et al., 2024), TradExpert (Ding et al., 2024b), FinCon (Yu et al., 2024), TradingAgents (Xiao et al., 2024) and MarketSenseAI 2.0 (Fatouros et al., 2025). Some models also incorporate reinforcement learning (RL) for iterative self-improvement (Ding et al., 2023; Koa et al., 2024).

3 Definitions of Investing Strategies

Timing-Based Strategies Timing-based strategies generate daily Buy (+1), Sell (−1), or Hold (0) signals based on market data such as prices, technical indicators, or model outputs. The objective is to capture short-term price movements through systematic trading rules.

Selection-Based Strategies Selection-based strategies identify subsets of assets expected to outperform based on ranking signals. Assets are selected periodically using top- k or thresholding. These strategies focus on cross-sectional alpha rather than timing individual trades.

4 Why Broader and Longer Evaluation Matters

Robust evaluation of financial strategies demands carefully designed backtests. Unlike typical machine learning tasks with large, clean datasets, financial data is noisy, nonstationary, and limited in scope. As a result, backtests are especially prone to three major sources of bias: **survivorship bias**, **look-ahead bias**, and **data-snooping bias**, each of which can inflate perceived performance and lead to misleading conclusions (Chan, 2021).

Survivorship Bias. This occurs when backtests include only currently active stocks while ignoring delisted or bankrupt assets. Such omissions systematically overstate returns and understate risk (Joubert et al., 2024). A common cause is using today’s S&P 500 constituents as the historical investment universe. This practice introduces what Garcia and Gould (1993) call “preinclusion bias”, also a form of look-ahead bias where future index membership influences past decisions. The impact is well-documented: Grinblatt and Titman (1989) and Elton et al. (1996) estimate annual return distortions between 0.1% and 0.9%, and Brown et al. (1992) show that even small distortions can misrepresent performance persistence.

Look-ahead Bias. Look-ahead bias arises when a strategy uses information that would not have been known at the time of decision-making (Chan, 2021). This includes selecting features, parameters, or symbols based on full-period outcomes, thereby introducing future knowledge into the backtest.

Data-snooping Bias. Also known as multiple testing bias, this occurs when repeated experimentation on the same dataset leads to overfitting. In finance, where sample sizes are small and the signal-to-noise ratio is very low, this bias is particularly problematic. Bailey et al. (2015) showed that evaluating strategies on overlapping data inflates false positive rates, and that standard hold-out validation techniques often fail to guard against this issue.

The Case for Broader and Longer Evaluation. Addressing these biases requires evaluating strategies across longer periods and broader asset universes. For daily trading, at least three years of data is generally recommended, while weekly and monthly strategies benefit from 10 to 20 years or more (Bailey et al., 2015). Gatev et al. (2006) tested pairs trading on 40 years of daily data, but Do and Faff (2010) extended this to 48 years and found profitability declined, highlighting the need for long-term evaluation. Likewise, recent deep learning models in finance rely on multi-year datasets to ensure robustness (Feng et al., 2019; Wang et al., 2022; Xia et al., 2024).

Stock selection is another critical factor. Many LLM-based investing studies selectively use only a small number of well-known stocks such as TSLA and AMZN. These are both historical winners, which limits generalisability and embeds both survivorship and look-ahead bias into the evaluation.

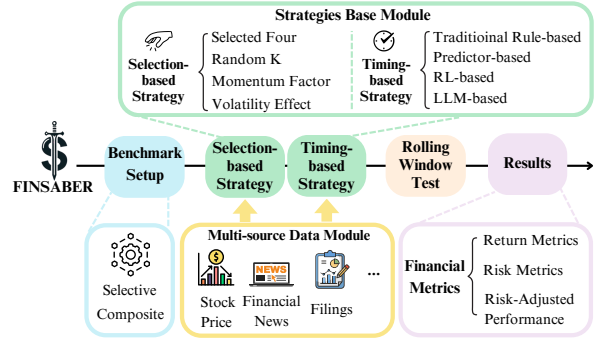


Figure 1: Overview of the FINSABER Backtest Framework. The central pipeline illustrates the backtesting process. The framework includes a Strategies Base Module (green), which covers both selection-based and timing-based strategies, and a Multi-source Data Module (yellow), integrating diverse financial data inputs.

Omitting delisted or underperforming stocks distorts performance metrics and presents an incomplete picture of real-world investing conditions.

Therefore, **backtests must address survivorship bias, look-ahead bias, and data-snooping bias explicitly**. Broader and longer evaluations, using historically accurate stock universes and spanning multiple market regimes, are essential for producing reliable, generalisable results that reflect real investing conditions.

5 FINSABER

As discussed in §4, existing evaluations of LLM-based investors suffer from survivorship bias, look-ahead bias, and data-snooping bias. These issues are largely due to limited evaluation periods and narrow stock selections. In this study, all subsequent findings and analyses are derived from our meticulously constructed backtesting framework, FINSABER², which systematically addresses biases and is specifically designed to meet the practical needs of LLM-based strategies, including the integration of unstructured, multi-source data.

FINSABER comprises three core modules: (1) a multi-source data module, (2) a modular strategy base, and (3) a bias-aware two-step backtesting pipeline. Figure 1 illustrates the framework.

Multi-source Data for LLM Benchmarking. LLM-based investing strategies utilise both structured and unstructured data such as historical stock prices, financial news, and company filings (10-K, 10-Q), spanning from 2000 to 2024. To prevent **look-ahead bias**, all data inputs are aligned with

²Financial INvesting Strategy Assessment with Bias mitigation, Expanded time, and Range of symbols

each backtest window using only information available prior to the start date. **Survivorship bias** is addressed by explicitly including delisted stocks, and open-source equivalents are provided for reproducibility (more detail in Appendix A).

Two-Step Pipeline for Bias Mitigation. FINSABER applies a two-step pipeline. First, *selection-based strategies* operate on regularly updated, historically accurate constituent lists, for example, the S&P 500 including delisted symbols, at each window. This further mitigates **survivorship bias** from the stock selection process, ensuring the evaluation is not restricted to a limited or selectively surviving set of stocks. Subsequently, *timing-based strategies* which covers rule-based, ML, RL, and LLM-driven approaches will be used to execute daily trading decisions. The modular strategy base is easily extensible for custom methods (see Appendix B). To mitigate **data-snooping bias**, rolling-window evaluations are performed over diverse and dynamically changing asset selections and extended time horizons. Window size and step are customisable, enabling realistic simulation across different market regimes. Together, this pipeline ensures broad symbol coverage and prevents overfitting to narrow datasets or short evaluation horizons.

Evaluation Metrics. FINSABER adopts three categories of evaluation metrics: *return*, *risk*, and *risk-adjusted performance*. Return metrics measure profitability and include Annualised Return (AR) and Cumulative Return (CR). Risk metrics quantify uncertainty and downside exposure, including Annualised Volatility (AV) and Maximum Drawdown (MDD). Risk-adjusted metrics assess capital efficiency and include the Sharpe Ratio (SPR) and Sortino Ratio (STR).

High returns alone do not imply strategy quality. Risk-adjusted metrics such as SPR and STR are more informative, especially in finance where capital efficiency and downside risk are critical (Chan, 2021). These metrics are standard in the literature (Cont, 2001; DeMiguel et al., 2007) and widely used in recent LLM-based investing benchmarks (Zhang et al., 2024b; Yu et al., 2024). Formal definitions and formulas are provided in Appendix C.

6 Experiments

Our experiments address methodological flaws in prior LLM-based investing evaluations identified in §4, specifically survivorship and data-snooping

biases from selective stock choices and short evaluation periods. We demonstrate how these practices inflate results and illustrate how FINSABER enables fairer assessments.

Specifically, our experiments include two parts: (1) **Pitfalls of selective evaluation:** Replicating previously reported results on select periods and symbols, then extending this evaluation period to demonstrate performance deterioration. (2) **Fair and robust comparisons:** Implementing systematic stock-selection methods to explicitly mitigate survivorship and data-snooping biases for fairer LLM assessments. We only consider go-long positions, aligning with current LLM strategies.

6.1 Pitfalls of Selective Evaluation

Revisiting Reported Claims. We begin by replicating earlier evaluation setups that demonstrated the effectiveness of LLM investing strategies on TSLA, NFLX, AMZN, and MSFT during the previously reported period (6 October 2022 to 10 April 2023). Additionally, we incorporate broader benchmarks, including traditional rule-based, ML, and DL methods. Previous studies omit key details such as exact risk-free rates and transaction costs. Thus, we explicitly set a historical average risk-free rate of 0.03 and use Moomoo’s³ standard US commission fee (\$0.0049/share, minimum \$0.99/order), comparable to HSBC and TradeUp⁴ platforms.

Table 2 summarises these results. Our analysis reveals that **LLM investors are not universally superior, even within their originally favoured setups**. Specifically, *FinMem* shows significant outperformance only for TSLA, whereas traditional benchmarks remain competitive or superior for other symbols. These observations caution against overly optimistic interpretations from selective evaluations. *FinAgent*, another LLM-based method, performs comparably to *FinMem* on NFLX and MSFT but generally does not offer consistent improvements across the set. Additionally, we find that **LLM-based strategies exhibit high annual volatility and substantial maximum drawdowns**, indicating a high-risk profile. This observation underscores the importance of explicit risk assessments in evaluating such strategies.

Additionally, our results highlight that **the choice of underlying language model dramatically impacts strategy performance**. The orig-

³https://www.moomoo.com/ca/support/topic10_122?lang=en-us

⁴<https://www.tradeup.com/pricing/detail>

Symbol	Type	Strategy	SPR \uparrow	CR \uparrow	MDD \uparrow	AV \downarrow
FinMem Selection (2022-10-06 to 2023-04-10)						
TSLA	Rule Based	Buy and Hold	-0.342	-20.483	-52.729	55.910
		SMA Cross	-0.293	-5.540	-18.517	38.602
		WMA Cross	0.215	3.741	-18.492	42.062
		ATR Band	-0.595	-19.142	-39.599	42.161
		Bollinger Bands	-0.769	-24.747	-44.655	45.366
		Turn of The Month	0.219	3.639	-11.642	31.042
	Predictor	ARIMA	0.601	15.007	-24.446	41.402
		XGBoost	0.331	6.213	-35.374	37.729
	RL	A2C	-	-	-	-
		DDPG	-	-	-	-
		PPO	-0.343	-20.425	-52.594	55.861
		SAC	-	-	-	-
		TD3	-	-	-	-
	LLM	FinMem (GPT-4o-mini)	0.927	19.940	-30.144	48.638
		FinMem (GPT-4o)	0.404	5.312	-36.351	54.434
		FinMem (reported)	2.679	61.776	-10.800	46.865
		FinAgent	-	-	-	-
NFLX	Rule Based	Buy and Hold	1.326	43.079	-20.184	41.523
		SMA Cross	-1.020	-8.285	-15.942	20.477
		WMA Cross	-0.803	-6.004	-14.290	19.826
		ATR Band	0.150	2.992	-12.231	19.314
		Bollinger Bands	-0.558	-4.996	-13.244	16.754
		Turn of The Month	0.559	8.383	-10.641	17.194
	Predictor	ARIMA	1.159	23.783	-15.043	25.749
		XGBoost	0.770	10.134	-11.246	14.928
	RL	A2C	-	-	-	-
		DDPG	-	-	-	-
		PPO	-	-	-	-
		SAC	-	-	-	-
		TD3	1.325	42.872	-20.121	41.448
	LLM	FinMem (GPT-4o-mini)	1.704	32.549	-13.018	34.766
		FinMem (GPT-4o)	0.896	16.244	-15.234	38.209
		FinMem (reported)	2.017	36.449	-15.850	36.434
		FinAgent	1.543	41.167	-20.417	51.030
AMZN	Rule Based	Buy and Hold	-0.460	-13.250	-31.546	35.624
		SMA Cross	-0.420	-4.433	-18.910	27.084
		WMA Cross	-0.563	-6.121	-21.030	26.831
		ATR Band	0.622	11.007	-15.842	23.272
		Bollinger Bands	-0.402	-7.105	-20.615	26.559
		Turn of The Month	-0.037	0.039	-14.892	20.722
	Predictor	ARIMA	-0.225	-4.752	-20.046	26.899
		XGBoost	1.955	42.468	-8.816	25.135
	RL	A2C	-0.462	-13.186	-31.397	35.496
		DDPG	-0.462	-13.186	-31.397	35.496
		PPO	-0.576	-9.485	-22.761	24.169
		SAC	-	-	-	-
		TD3	-0.462	-13.186	-31.397	35.496
	LLM	FinMem (GPT-4o-mini)	0.297	2.800	-2.744	10.247
		FinMem (GPT-4o)	-0.968	-20.091	-31.164	40.896
		FinMem (reported)	0.233	4.885	-22.929	42.658
		FinAgent	-1.108	-6.113	-9.317	13.257
MSFT	Rule Based	Buy and Hold	0.974	21.171	-14.192	28.327
		SMA Cross	1.515	18.289	-8.746	20.821
		WMA Cross	1.334	16.576	-8.883	21.503
		ATR Band	1.036	12.979	-7.709	15.005
		Bollinger Bands	2.115	31.619	-3.475	18.243
		Turn of The Month	-0.034	0.970	-11.955	15.097
	Predictor	ARIMA	2.245	44.777	-7.121	22.636
		XGBoost	0.895	12.678	-10.734	16.721
	RL	A2C	0.973	21.059	-14.120	28.117
		DDPG	-	-	-	-
		PPO	-	-	-	-
		SAC	-	-	-	-
		TD3	-	-	-	-
	LLM	FinMem (GPT-4o-mini)	-0.554	-7.104	-14.588	25.969
		FinMem (GPT-4o)	0.792	12.834	-13.555	33.884
		FinMem (reported)	1.440	23.261	-14.989	32.562
		FinAgent	1.252	21.438	-14.502	32.952
COIN	Rule Based	Buy and Hold	0.170	-12.729	-54.402	86.718
		SMA Cross	0.566	15.215	-33.219	90.853
		WMA Cross	0.517	12.655	-33.259	94.180
		ATR Band	0.933	25.169	-22.906	38.716
		Bollinger Bands	-0.795	-24.371	-40.733	43.904
		Turn of The Month	0.713	21.242	-21.522	49.357
	LLM	FinMem (GPT-4o-mini)	-1.674	-8.196	-8.196	11.582
		FinMem (GPT-4o)	-0.601	-38.368	-60.210	94.212
		FinMem (reported)	0.717	34.983	-35.753	89.752
		FinAgent	0.313	-10.598	-56.641	111.434

Table 2: Backtest performance of benchmark strategies over the previously reported period (2022-10-06 to 2023-04-10) where LLM investing strategies were shown to be effective. “-” metrics indicate no trading activities were triggered. Top in **red** and second-best in **blue**.

inal *FinMem* uses GPT-4; however, switching to GPT-4o or GPT-4o-mini significantly alters results.

Interestingly, larger models (GPT-4o) do not consistently outperform smaller ones (GPT-4o-mini), suggesting that model size alone is not predictive of superior performance and that model selection itself may introduce data-snooping biases. Further evidence in Appendix D reinforces this instability, where even a slight two-month extension of the evaluation period results in substantial variation for LLM-based strategies.

Extending the Evaluation Period. To further illustrate the limitations of short evaluation horizons, We extend the evaluation period (2004–2024) using the same four symbols (TSLA, NFLX, AMZN, MSFT) to assess LLM performance robustness over the long term.

Table 3 summarises these extended period results. Crucially, **extending the evaluation horizon significantly diminishes the perceived superiority of LLM investors**. Over two decades, traditional strategies like *Buy and Hold* consistently rank among the top performers across most symbols. TSLA is the only case where LLM investors (*FinMem*, *FinAgent*) clearly lead in AR, while for NFLX, AMZN, and MSFT, *Buy and Hold* or other strategies match or outperform them. This suggests previously reported LLM advantages are likely short-lived, potentially hand-picked, and highly sensitive to the evaluation period.

It is crucial to note that we cannot yet conclude that benchmark strategies cannot outperform the market. As mentioned, backtesting only on popular stocks may inadvertently introduce survivorship bias, as these stocks have gained popularity due to past success during prolonged bull markets. Thus, expanding the range of symbols is essential to ensure a more systematic and unbiased evaluation.

6.2 Fair Comparisons with the Composite Approach

To overcome the aforementioned biases, we introduce the **Composite** evaluation setup within FINS-ABER. This setup integrates systematic *selection-based strategies* to expand and diversify the stock universe, explicitly addressing survivorship and data-snooping biases. Specifically, we use three unbiased stock selection approaches (details in Appendix B.2): RANDOM FIVE, MOMENTUM FACTOR (Muller and and, 2010), and VOLATILITY EFFECT (Blitz and Vliet, 2007). Each selection strategy is executed at the start of every rolling window, serving as a rebalancing mechanism.

Symbol	Type	Strategy	SPR \uparrow	STR \uparrow	AR \uparrow	MDD \uparrow	AV \downarrow
Selected 4 (2004-01-01 to 2024-01-01)							
TSLA	Rule Based	Buy and Hold	0.631	0.915	37.767	-50.839	45.242
		SMA Cross	0.011	0.106	16.352	-42.975	29.144
		WMA Cross	0.218	0.375	19.743	-31.866	27.471
		ATR Band	0.021	0.062	-0.277	-38.516	27.262
		Bollinger Bands	0.193	0.294	4.279	-37.166	26.267
		Trend Following	0.139	0.624	17.342	-13.508	14.750
		Turn of The Month	0.207	0.352	7.868	-27.904	23.595
	Predictor	ARIMA	0.681	1.003	24.138	-30.450	27.612
		XGBoost	0.142	0.370	10.877	-22.901	19.537
		A2C	0.172	0.249	3.875	-27.367	22.890
	RL	DDPG	0.446	0.653	26.265	-21.626	21.892
		PPO	0.469	0.663	28.189	-46.810	40.156
		SAC	0.119	0.190	6.654	-11.042	9.902
		TD3	0.417	0.604	23.336	-33.725	30.233
	LLM	FinMem	0.641	1.069	42.153	-34.234	35.030
		FinAgent	0.206	0.649	38.591	-36.930	38.302
NFLX	Rule Based	Buy and Hold	0.622	0.952	23.919	-48.119	41.704
		SMA Cross	-0.126	0.015	1.219	-32.856	22.373
		WMA Cross	-0.091	-0.008	1.586	-34.378	23.275
		ATR Band	0.118	0.273	1.018	-37.646	24.373
		Bollinger Bands	0.075	0.381	0.284	-34.002	23.090
		Trend Following	-0.248	-0.094	3.433	-15.066	14.528
		Turn of The Month	0.288	0.489	7.122	-21.646	17.168
	Predictor	ARIMA	0.659	1.035	19.022	-27.567	25.514
		XGBoost	0.202	0.355	4.957	-21.301	17.302
		A2C	0.171	0.243	4.359	-20.960	16.129
	RL	DDPG	0.211	0.340	7.370	-27.922	22.317
		PPO	0.541	0.814	19.279	-39.615	33.630
		SAC	0.186	0.285	8.397	-9.545	9.216
		TD3	0.291	0.431	10.900	-21.451	19.304
	LLM	FinMem	0.293	0.622	12.566	-27.721	26.876
		FinAgent	-0.419	0.621	22.543	-20.466	22.838
AMZN	Rule Based	Buy and Hold	0.551	0.829	15.997	-36.842	30.859
		SMA Cross	-0.265	-0.247	0.582	-24.203	17.142
		WMA Cross	-0.144	-0.100	2.281	-22.867	16.329
		ATR Band	0.453	1.014	5.639	-19.922	15.166
		Bollinger Bands	0.020	0.126	0.900	-23.757	15.764
		Trend Following	-0.041	0.438	7.153	-10.265	12.342
		Turn of The Month	-0.033	-0.013	1.479	-20.472	15.726
	Predictor	ARIMA	0.339	0.504	7.523	-20.612	19.115
		XGBoost	-0.587	-0.366	1.200	-13.659	11.106
		A2C	0.165	0.247	3.925	-14.841	11.654
	RL	DDPG	0.252	0.391	7.059	-10.153	9.306
		PPO	0.505	0.767	13.831	-29.128	24.392
		SAC	0.179	0.257	4.438	-14.093	11.665
		TD3	0.382	0.597	11.738	-21.942	19.149
	LLM	FinMem	0.188	0.340	5.695	-28.296	24.786
		FinAgent	0.364	0.663	12.699	-25.516	25.390
MSFT	Rule Based	Buy and Hold	0.461	0.620	11.237	-25.457	21.787
		SMA Cross	-0.172	-0.167	1.948	-16.610	12.051
		WMA Cross	-0.226	-0.233	0.827	-16.708	12.083
		ATR Band	0.316	0.636	5.719	-11.893	10.881
		Bollinger Bands	-0.053	-0.028	1.578	-16.102	11.933
		Trend Following	-0.366	-0.205	1.966	-8.846	7.714
		Turn of The Month	-0.264	-0.343	-0.180	-14.309	10.438
	Predictor	ARIMA	0.304	0.466	8.207	-15.227	13.819
		XGBoost	0.171	0.322	5.890	-10.523	10.335
		A2C	0.279	0.380	7.478	-13.447	11.933
	RL	DDPG	0.300	0.403	7.838	-14.007	12.351
		PPO	0.344	0.463	8.589	-16.697	14.410
		SAC	0.216	0.288	5.329	-14.866	11.835
		TD3	0.050	0.070	1.405	-9.491	6.648
	LLM	FinMem	0.203	0.293	4.567	-19.270	17.891
		FinAgent	0.285	0.432	11.123	-18.596	18.863

Table 3: Backtest performance of benchmark strategies for previously reported LLM-selected symbols over an extended period (2004-01-01 or earliest available to 2024-01-01). Top in **red** and second-best in **blue**.

To mitigate survivorship bias, we use historical constituent lists, specifically S&P 500 for US market, at each evaluation period’s start and explicitly include delisted symbols. To address data-snooping bias, we evaluate a large and diversified symbol universe: 91, 84, and 63 total distinct symbols for RANDOM FIVE, MOMENTUM-based, and VOLATILITY-based selection, respectively. These counts reflect all unique symbols encountered across rolling windows, where stocks are reselected in each window, preventing cherry-picking and short-horizon bias.

Selection Strategy	Type	Timing Strategy	SPR \uparrow	STR \uparrow	AR \uparrow	MDD \uparrow	AV \downarrow
Random 5 S&P 500 (91 symbols)	Rule Based	Buy and Hold	0.315	0.456	6.694	-35.130	27.410
		SMA Cross	-0.298	-0.290	0.446	-22.292	15.774
		WMA Cross	-0.299	-0.305	0.232	-22.754	15.528
		ATR Band	0.232	0.425	5.119	-21.535	16.113
		Bollinger Bands	0.129	0.288	3.521	-22.487	16.290
		Trend Following	-0.389	-0.198	2.525	-8.587	8.223
		Turn of The Month	0.015	0.072	2.870	-18.582	13.542
	Predictor	ARIMA	0.255	0.434	6.928	-21.691	17.504
		XGBoost	-0.055	0.028	3.089	-17.160	13.075
	RL	A2C	0.086	0.122	1.902	-9.220	6.887
		DDPG	0.089	0.125	0.844	-15.185	11.624
		PPO	0.179	0.256	3.282	-18.395	13.783
		SAC	0.097	0.142	1.389	-16.058	12.375
		TD3	0.173	0.248	3.682	-14.471	11.565
	LLM	FinMem	-0.253	0.114	-0.094	-24.243	21.214
		FinAgent	0.094	0.323	4.477	-28.059	26.387
Momentum Factor (84 symbols)	Rule Based	Buy and Hold	0.384	0.694	9.916	-32.596	37.421
		SMA Cross	-0.251	0.008	2.109	-19.438	20.050
		WMA Cross	-0.169	0.051	3.674	-18.651	20.330
		ATR Band	0.197	0.595	4.314	-19.407	20.038
		Bollinger Bands	0.114	0.702	1.881	-19.451	21.555
		Trend Following	0.119	0.531	6.380	-15.726	18.696
		Turn of The Month	0.056	0.662	3.197	-18.108	18.055
	Predictor	ARIMA	0.542	1.043	13.257	-18.277	22.892
		XGBoost	0.094	1.525	6.131	-12.754	17.238
	RL	A2C	0.105	0.171	2.488	-14.452	14.815
		DDPG	0.209	0.366	5.973	-14.255	16.499
		PPO	0.185	0.308	1.939	-23.177	25.527
		SAC	0.195	0.321	5.591	-12.235	16.144
		TD3	0.186	0.293	3.464	-14.593	14.953
	LLM	FinMem	0.025	0.170	3.649	-23.335	28.078
		FinAgent	0.104	0.534	13.950	-20.675	30.635
Volatility Effect (63 symbols)	Rule Based	Buy and Hold	0.703	1.291	7.898	-14.146	14.720
		SMA Cross	-0.568	-0.544	0.781	-9.296	8.665
		WMA Cross	-0.665	-0.348	1.908	-8.481	8.573
		ATR Band	-0.026	0.120	2.798	-8.032	7.951
		Bollinger Bands	-0.077	0.029	2.503	-7.618	7.774
		Trend Following	0.230	0.619	5.503	-8.115	9.297
		Turn of The Month	-0.156	-0.095	2.881	-6.889	7.233
	Predictor	ARIMA	0.325	0.838	4.898	-9.111	9.807
		XGBoost	-0.108	-0.055	2.775	-6.676	7.077
	RL	A2C	0.421	0.795	4.620	-4.428	5.149
		DDPG	0.373	0.725	3.768	-6.681	6.693
		PPO	0.514	0.972	5.805	-8.757	9.461
		SAC	0.402	0.810	3.527	-4.821	5.030
		TD3	0.269	0.394	4.610	-5.442	5.992
	LLM	FinMem	-0.228	0.483	4.061	-10.860	11.641
		FinAgent	0.241	0.527	4.954	-10.268	11.502

Table 4: Backtest performance under the **Composite setup**, using three different selection strategies across historical S&P 500 constituents (2004–2024), including delisted symbols. Top in **red** and second-best in **blue**.

Table 4 summarises these comprehensive evaluations. Results obtained through this unbiased and systematic approach **further validate our previous findings from the selected-four evaluation**. Specifically, both the RANDOM FIVE and MOMENTUM-based selections reinforce the conclusion that the previously claimed superiority of LLM investors is largely driven by selective evaluation setups. For instance, in the RANDOM FIVE setup, *Buy and Hold*, *ATR Band* and *ARIMA* outperform *FinMem* and *FinAgent* in terms of risk-adjusted metrics. Similarly, *ARIMA* and simple rule-based strategies often perform better than LLM-based methods under the MOMENTUM-based selection. In the VOLATILITY-based selection, traditional methods dominate even more clearly: *Buy and Hold* achieves the highest Sharpe (0.703), stability (1.291), and AR (7.898%), while *PPO* and *ARIMA* again show strong all-round performance. LLM-based methods lag behind, with *FinAgent* offering moderate returns but lower Sharpe (0.241) and larger drawdowns. Notably, our reported LLM per-

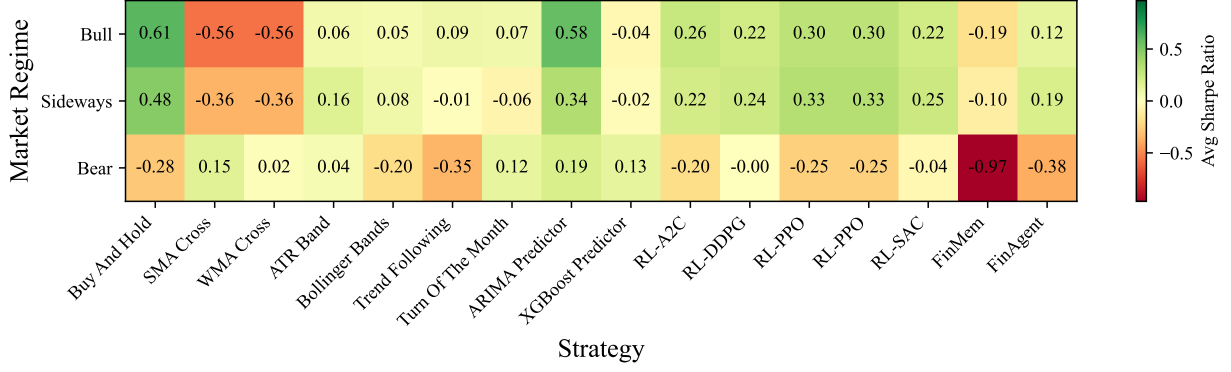


Figure 2: Average Sharpe ratio by regime for all benchmarking strategies. **Green = strong, red = weak.**

formances do not adjust for potential data leakage: given the use of pretrained models like GPT-4o, the LLMs may have seen parts of the data during training—yet they still fail to outperform traditional strategies under fair evaluation, casting further doubt on their real-world advantage.

Nevertheless, it is important to acknowledge **LLM-based strategies still show potential regarding absolute annual returns.** For instance, *FinAgent* achieves the highest AR (13.950%) in the MOMENTUM-based selection setup. However, the relatively weaker performance observed in SPR (0.104) and MMD metrics suggests a clear need for improved risk management within LLM-driven approaches before they can be reliably adopted in practice.

Moreover, by comparing *Buy and Hold* with different selection strategies, we clearly identify the relative effectiveness of each selection strategy: VOLATILITY selection (Sharpe 0.703) outperforms MOMENTUM (0.384), which in turn surpasses RANDOM FIVE (0.315). RL-based methods exhibit the clearest alignment with selection quality. Strategies like *PPO*, *SAC*, and *TD3* systematically achieve their best performance under the VOLATILITY selection and degrade under the other two. This suggests **RL methods are more sensitive to the quality of the stock candidates.** Among LLM strategies, *FinAgent* exhibits a greater dependency on selection quality than *FinMem*.

Overall, these results not only confirm our earlier insights but also underscore the critical importance of unbiased, systematic stock-selection methodologies for accurately assessing the true capabilities and limitations of LLM-based investing strategies.

7 Market Regime Analysis

Another key question in evaluating LLM-based investing strategies is whether they adapt appro-

priately across varying market conditions. Financial markets exhibit time-varying predictability and uncertainty across different economic, financial, and political regimes (Kim et al., 2011). Some strategies may exploit these variations, while others may struggle to adapt. Distinct market environments—bull, bear, and sideways—present unique challenges and opportunities: bull markets reward aggressive positioning and high exposure, bear markets require effective risk management, and sideways markets test a strategy’s ability to navigate uncertainty in the absence of clear trends. By decomposing performance across these regimes, it is possible to determine whether strategies are overly conservative and miss opportunities during bullish periods, or excessively aggressive and incur significant losses during downturns. Understanding these regime-specific behaviours is essential for interpreting the strengths and weaknesses of LLM-based investing strategies and for guiding their future development (Hui and Chan, 2018).

We label each calendar year based on the annual return of the S&P 500: $R_y = \frac{P_T - P_0}{P_0}$, where P_0 and P_T are the adjusted closing prices on the first and last trading days of year y . A year is classified as **bull** if $R_y \geq +20\%$, **bear** if $R_y \leq -20\%$, and **sideways** otherwise. The $\pm 20\%$ threshold follows standard industry convention (Zweig, 2019).

To analyse regime-specific performance, we employ our composite setup using the three selection strategies outlined in §6.2. For each timing strategy, we retrieve the SPR within each 1-year window from Table 4. These are then averaged per $\{\text{strategy}, \text{regime}\}$ pair to produce stable performance indicators across market conditions. Figure 2 illustrates the results, with **green** indicating strong SPR and **red** signifying the opposite.

Traditional rule-based and predictor-based methods still set the standard. *ATR Band*, *Turn of the*

Month and *ARIMA* deliver positive Sharpe in every regime, while *Buy and Hold*, our passive yardstick, posts 0.61 in bulls, 0.48 in sideways markets and only -0.28 in bears. No active strategy surpasses this passive SPR in the bull regime, suggesting that many strategies, including the LLM ones, may struggle to fully capitalise on strong up-trends.

RL algorithms sit in the middle. *A2C* and *DDPG* pick up part of the upside and limit losses; *PPO* and *SAC* swing with volatility and underperform *ARIMA* once conditions turn.

The LLM strategies perform poorly. *FinAgent* records Sharpe 0.12 in bulls and -0.38 in bears; *FinMem* slides from -0.19 to -0.97. Both are too cautious when risk is rewarded and too aggressive when it is penalised. *FinAgent* is the better of the two, halving the bear-market shortfall relative to *Buy and Hold* and keeping a small positive Sharpe in neutral conditions, but it still trails every rule-based or predictor benchmark.

These results suggest two directions for future LLM investors. A first direction concerns improving trend-detection so the strategy can at least match passive equity beta in up-markets. A second direction relates to embedding explicit regime-aware risk controls that scale exposure down as volatility or draw-down risk rises. Balancing risk-taking and risk management, aggression and defence, rather than increasing model size, appears the key to closing the gap with traditional methods.

8 Findings and Takeaways

The evaluations in §6 and §7 reveals structural limitations in current LLM-based investing strategies. These findings expose methodological flaws in recent studies and highlight the need for more rigorous and domain-aware evaluation.

First, the fragility of LLM-derived alpha under realistic conditions aligns with the Efficient Market Hypothesis (EMH) (Fama, 1970): asset prices reflect all available information, making it difficult to consistently achieve above-average returns through information-based trading. Despite access to large textual corpora, current LLMs fail to extract signals that persist beyond short windows, suggesting that observed gains may stem from narrow evaluation setups rather than genuine alpha.

Second, scaling laws observed in NLP and other domains do not translate effectively to financial applications. While larger models often yield better performance in NLP tasks (Kaplan et al., 2020),

our results indicate that increasing model size does not reliably improve investing performance. For instance, *FinMem* with GPT-4o does not outperform its smaller variant. This suggests that financial markets impose intrinsic limits on extractable signals, regardless of model capacity (Harvey and Liu, 2013). Moreover, simple models like *ARIMA* or rule-based systems often outperform LLMs on risk-adjusted metrics. This indicates that complexity without financial logic such as trend detection, volatility targeting, or drawdown control, adds little value (and, 2001; Kabiri et al., 2023).

Third, LLMs poorly adapt to market regimes. They tend to be overly conservative in bull markets and excessively aggressive in downturns, contradicting the Adaptive Markets Hypothesis (AMH), which stresses dynamic risk management across shifting conditions (Lo, 2004).

Finally, biased evaluation designs, such as short evaluation periods and selectively chosen symbols, systematically exaggerate results. In finance, this is not merely a technical flaw but a practical risk, as misleading models can drive poor allocation decisions (Sullivan et al., 1999). FINSABER addresses this by enforcing rigorous, bias-aware benchmarking, setting a baseline for credible evaluation in the financial LLM field.

9 Conclusion

We reassess the robustness of LLM *timing-based investing strategies* using FINSABER, a comprehensive framework that mitigates backtesting biases and extends both the evaluation horizon and symbol universe. Our results show that the perceived superiority of LLM-based methods deteriorates under more robust and broader long-term testing. Regime analysis further reveals that current strategies miss upside in bull markets and incur heavy losses in bear markets due to poor risk control.

We identify two priorities for future LLM-based investors: (1) enhancing uptrend detection to match or exceed passive exposure, and (2) including regime-aware risk controls to dynamically adjust aggression. Addressing these dimensions rather than increasing framework complexity is the key to building practical, reliable strategies.

Finally, our cost analysis (Appendix F) shows that large-scale LLM backtesting is financially intensive. Future work should pursue cost-efficient model designs and incorporate API costs into performance evaluation.

Limitations

There are several limitations to our current study. First, we did not individually tune the traditional rule-based strategies for each rolling evaluation window. Typically, applying domain-specific market insights to optimise parameters can significantly enhance the performance of these methods. However, we argue that our current configuration remains valid and effectively demonstrates the competitive disadvantage faced by LLM strategies. Indeed, tuning the parameters of traditional rule-based strategies would likely elevate their performance further, reinforcing rather than undermining our main conclusions.

Second, our evaluation has not fully eliminated look-ahead bias. Pre-trained LLMs, due to their inherent training corpus, may inadvertently contain stock-related information from historical periods overlapping our test sets. Despite this potential data leakage, the observed underperformance of LLM strategies strengthens our critical assessment. Explicitly addressing this look-ahead concern through controlled model training or careful exclusion of financial data from training corpora will be an important avenue for future research.

Third, to ensure experiment reproducibility, we restricted our analysis to publicly available data, excluding proprietary sources such as private newsfeeds, earning transcripts, or expert analyses. Nonetheless, the FINSABER framework was deliberately designed to be modular and extensible, allowing researchers with access to private data to easily integrate additional information sources. Our primary goal remains providing a rigorous, long-term evaluation pipeline that minimises selective reporting. Researchers lacking proprietary data can fully replicate our results using openly accessible resources.

References

- R. Cont and. 2001. [Empirical properties of asset returns: stylized facts and statistical issues](#). *Quantitative Finance*, 1(2):223–236.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jinguang Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *Preprint*, arXiv:2309.16609.
- David H. Bailey, Jonathan Michael Borwein, Marcos M. López de Prado, and Qiji Jim Zhu. 2015. [The probability of backtest overfitting](#). *ERN: Econometric Modeling in Financial Economics (Topic)*.
- David Blitz and Pim Vliet. 2007. [The volatility effect: Lower risk without lower return](#). *The Journal of Portfolio Management*, 34.
- J. Bollinger. 2002. [Bollinger on Bollinger Bands](#). McGraw-Hill Education.
- George Edward Pelham Box and Gwilym Jenkins. 1990. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., USA.
- Stephen J Brown, William Goetzmann, Roger G Ibbotson, and Stephen A Ross. 1992. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4):553–580.
- E.P. Chan. 2021. [Quantitative Trading: How to Build Your Own Algorithmic Trading Business](#). Wiley Trading. Wiley.
- Tianqi Chen and Carlos Guestrin. 2016. [Xgboost: A scalable tree boosting system](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.
- Rama Cont. 2001. [Empirical properties of asset returns: stylized facts and statistical issues](#). *Quantitative Finance*, 1(2):223–236.
- Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. 2007. [Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?](#) *The Review of Financial Studies*, 22(5):1915–1953.
- Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024a. [Large language model agent in financial trading: A survey](#). *Preprint*, arXiv:2408.06361.
- Qianggang Ding, Haochen Shi, and Bang Liu. 2024b. [Tradexpert: Revolutionizing trading with mixture of expert llms](#). *Preprint*, arXiv:2411.00782.
- Yujie Ding, Shuai Jia, Tianyi Ma, Bingcheng Mao, Xiuze Zhou, Liuliu Li, and Dongming Han. 2023. [Integrating Stock Features and Global Information via Large Language Models for Enhanced Stock Return Prediction](#). Papers 2310.05627, arXiv.org.
- Binh Do and Robert Faff. 2010. [Does simple pairs trading still work?](#) *Financial Analysts Journal*, 66(4):83–95.

- Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. [Fnspid: A comprehensive financial news dataset in time series](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4918–4927, New York, NY, USA. Association for Computing Machinery.
- Edwin J Elton, Martin J Gruber, and Christopher R Blake. 1996. Survivor bias and mutual fund performance. *The review of financial studies*, 9(4):1097–1120.
- Eugene F Fama. 1970. [Efficient capital markets](#). *Journal of Finance*, 25(2):383–417.
- George Fatouros, Kostas Metaxas, John Soldatos, and Manos Karathanassis. 2025. [Marketsenseai 2.0: Enhancing stock analysis through llm agents](#). *Preprint*, arXiv:2502.00415.
- Georgios Fatouros, Konstantinos Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2024. [Can large language models beat wall street? unveiling the potential of ai in stock selection](#). *Preprint*, arXiv:2401.03737.
- Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. [Temporal relational ranking for stock prediction](#). *ACM Trans. Inf. Syst.*, 37(2).
- CB Garcia and FJ Gould. 1993. Survivorship bias. *Journal of Portfolio Management*, 19(3):52.
- Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. 2006. [Pairs Trading: Performance of a Relative-Value Arbitrage Rule](#). *The Review of Financial Studies*, 19(3):797–827.
- Mark Grinblatt and Sheridan Titman. 1989. Mutual fund performance: An analysis of quarterly portfolio holdings. *Journal of business*, pages 393–416.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. [Large language model based multi-agents: A survey of progress and challenges](#). *Preprint*, arXiv:2402.01680.
- Campbell Harvey and Yan Liu. 2013. [Backtesting](#). *SSRN Electronic Journal*, 42.
- Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu tao Xia, Dawei Cheng, and Changjun Jiang. 2025. [Fintsb: A comprehensive and practical benchmark for financial time series forecasting](#). *Preprint*, arXiv:2502.18834.
- Eddie Hui and Ka Kwan Kevin Chan. 2018. [Optimal trading strategy during bull and bear markets for hong kong-listed stocks](#). *International Journal of Strategic Property Management*, 22:381–402.
- Jacques Joubert, Dragan Sestovic, Illya Barziy, Walter Distaso, and Marcos Lopez de Prado. 2024. The three types of backtests. *Available at SSRN*.
- Moulay Slimane Kabiri, Cherif El Msiyah, and Otheman Nouisser. 2023. [Risk budgeting: A tactical asset allocation approach for retirement reserve funds in morocco](#). *Journal of Financial Risk Management*, 12(02):203–223.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- Jae H Kim, Abul Shamsuddin, and Kian-Ping Lim. 2011. [Stock return predictability and the adaptive markets hypothesis: Evidence from century-long us data](#). *Journal of Empirical Finance*, 18(5):868–879.
- Kemal Kirtac and Guido Germano. 2024. [Sentiment trading with large language models](#). *Finance Research Letters*, 62:105227.
- Kelvin J.L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. [Learning to generate explainable stock predictions using self-reflective large language models](#). In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4304–4315, New York, NY, USA. Association for Computing Machinery.
- Xiao-Yang Liu, Ziyi Xia, Hongyang Yang, Jiechao Gao, Daochen Zha, Ming Zhu, Christina Dan Wang, Zhao-ran Wang, and Jian Guo. 2024. Dynamic datasets and market environments for financial reinforcement learning. *Machine Learning - Springer Nature*.
- Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. 2021. FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. *ACM International Conference on AI in Finance (ICAIF)*.
- Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. 2022. [Finrl: deep reinforcement learning framework to automate trading in quantitative finance](#). In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, New York, NY, USA. Association for Computing Machinery.
- Andrew Lo. 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 30(5):15–29.
- Alejandro Lopez-Lira and Yuehua Tang. 2023. [Can chatgpt forecast stock price movements? return predictability and large language models](#). *Preprint*, arXiv:2304.07619.
- John J. McConnell and Wei Xu. 2008. [Equity Returns at the Turn of the Month](#). *Financial Analysts Journal*, 64(2):49–64.
- C Muller and M Ward and. 2010. [Momentum effects in country equity indices](#). *Studies in Economics and Econometrics*, 34(1):111–127.

- Ryan Sullivan, Allan Timmermann, and Halbert White. 1999. [Data-snooping, technical trading rule performance, and the bootstrap](#). *The Journal of Finance*, 54(5):1647–1691.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Heyuan Wang, Tengjiao Wang, Shun Li, Jiayi Zheng, Shijie Guan, and Wei Chen. 2022. [Adaptive long-short pattern transformer for stock investment selection](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3970–3977. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. [Llmfactor: Extracting profitable factors through prompts for explainable stock movement prediction](#). *Preprint*, arXiv:2406.10811.
- Saizhuo Wang, Hao Kong, Jiadong Guo, Fengrui Hua, Yiyang Qi, Wanyun Zhou, Jiahao Zheng, Xinyu Wang, Lionel M. Ni, and Jian Guo. 2025. [Quantbench: Benchmarking ai methods for quantitative investment](#). *Preprint*, arXiv:2504.18600.
- Cole Wilcox, Eric Crittenden, and Blackstar Funds. 2005. [Does trend following work on stocks](#). In *The Technical Analyst*, volume 14, pages 1–19.
- Ruoxu Wu. 2024. [Portfolio performance based on llm news scores and related economical analysis](#). *SSRN Electronic Journal*.
- Hongjie Xia, Huijie Ao, Long Li, Yu Liu, Sen Liu, Guangnan Ye, and Hongfeng Chai. 2024. [Cisthpan: Pre-trained attention network for stock selection with channel-independent spatio-temporal hypergraph](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):9187–9195.
- Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024. [Tradingagents: Multi-agents llm financial trading framework](#). *arXiv preprint arXiv:2412.20138*.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. [Fingpt: Open-source financial large language models](#). *FinLLM Symposium at IJCAI 2023*.
- Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and Christina Dan Wang. 2024. [Finrobot: An open-source ai agent platform for financial applications using large language models](#). *Preprint*, arXiv:2405.14767.
- Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. 2023. [Finmem: A performance-enhanced llm trading agent with layered memory and character design](#). *Preprint*, arXiv:2311.13743.
- Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, GUO-JUN XIONG, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. [Fincon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Haohan Zhang, Fengrui Hua, Chengjin Xu, Hao Kong, Ruiting Zuo, and Jian Guo. 2024a. [Unveiling the potential of sentiment: Can large language models predict chinese stock price movements?](#) *Preprint*, arXiv:2306.14222.
- Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. 2024b. [A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist](#). *Preprint*, arXiv:2402.18485.
- Jason Zweig. 2019. [Where did this ‘bull market’ come from, anyway?](#) *The Wall Street Journal*.

A Data Collection

Our multi-source data comprises daily stock prices, daily financial news, and 10-Q and 10-K filings.

Daily Stock Prices. We collect daily price data for over 7,000 U.S. equities spanning from 2000 to 2024. In addition to actively traded stocks, our dataset includes delisted symbols that were historically part of the S&P 500 index, based on the archived constituent list⁵. This inclusion enhances the historical completeness of our dataset and mitigates survivorship bias within the context of index-based evaluations.

Financial News. The financial news dataset, initially compiled by [Dong et al. \(2024\)](#), comprises 15.7 million records pertaining to 4,775 S&P 500 companies, spanning the years 1999 to 2023. We have organised the news by aligning it with the respective companies and indexing it by date.

10K & 10Q Filings. We collect 10-K and 10-Q filings for companies included in the Russell 3000 index, sourced from the US Securities and Exchange Commission (SEC) EDGAR database. These filings are publicly available and accessed

⁵<https://github.com/fja05680/sp500>

via the SEC-API⁶, which allows programmatic retrieval and parsing. We preprocess the HTML documents and segment them into standardized sections, such as Risk Factors, MD&A, and Financial Statements, to support fine-grained analysis. Each filing is indexed by company identifier and filing date to enable alignment with other datasets.

Extensibility. All datasets used in this framework can be seamlessly substituted with proprietary or higher-resolution alternatives if available. Researchers may incorporate paid datasets such as premium financial news (e.g., Alpaca Markets⁷, Refinitiv⁸), earnings call transcripts, analyst research reports, or other modalities including video or audio. Integration is supported through the implementation of a custom dataset class, allowing modular and flexible replacement of any data stream within the pipeline.

B FinSABER Strategies Base

B.1 Timing-based Strategies

Open-Source LLM investors. This category includes *FinMem* (Yu et al., 2023) and *FinRobot* (Yang et al., 2024). We acknowledge other works, such as *FinAgent* (Zhang et al., 2024b), *FinCon* (Yu et al., 2024) but they are not (yet) open-source, which prevents us from generating backtesting results.

Traditional Rule-Based (Indicator-Based) Strategies. We implement and cover several well-known traditional rule-based (indicator-based) investing strategies, such as *Buy and Hold*, *Simple Moving Average Crossover*, *Weighted Moving Average Crossover*, *ATR Band*, *Bollinger Bands* (Bollinger, 2002), *Trend Following* (Wilcox et al., 2005), and *Turn of the Month* (McConnell and Xu, 2008). These strategies typically rely on one or multiple technical indicators or domain-based rules to generate timely buy/sell signals, aiming to exploit identifiable market patterns or anomalies.

It is noteworthy that **traditional strategies are often overlooked**, with many existing works focusing solely on *Buy and Hold*. However, other established strategies listed above have also endured over time and demonstrated their effectiveness.

ML/DL Forecaster-Based Strategies In contrast to fixed rules or indicator-based triggers, these

strategies rely on data-driven models—often statistical or neural network forecasters—to predict future price movements. Specifically, they buy or hold if an uptrend is indicated and sell (or go short) otherwise. This can be viewed as a relatively naive application of ML/DL forecasters, but it is widely used as a benchmark method for such models. Although one could consider the forecast output as a type of “indicator”, the reliance on predictive algorithms capable of uncovering complex patterns sets these methods apart from purely rule-based approaches. We include the well-known ARIMA (Box and Jenkins, 1990) and XGBoost (Chen and Guestrin, 2016) in this category and also cover forecasters based on LLMs, differentiating them from the LLM investors.

RL-Based Strategies Additionally, we incorporate RL-based strategies to evaluate their performance in investing environments. Specifically, we implement Advantage Actor-Critic (A2C), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), Twin Delayed Deep Deterministic Policy Gradient (TD3), and Soft Actor-Critic (SAC), leveraging the FinRL framework (Liu et al., 2022, 2024). These RL algorithms are widely used in financial markets due to their ability to learn optimal investing policies from historical data and adapt to evolving market conditions. A2C and PPO, as policy-gradient methods, optimise investing decisions while ensuring stability, whereas DDPG, TD3, and SAC extend this capability to continuous action spaces, rendering them well-suited for asset allocation and portfolio optimisation tasks.

B.2 Selection-based Strategies

(1) RANDOM FIVE, which involves randomly selecting five stocks from historical S&P 500 constituents available at each rolling evaluation period’s start date, (2) MOMENTUM FACTOR Selection, which selects stocks based solely on past returns at each period’s start date, following a momentum factor-based strategy (Muller and and, 2010), and (3) VOLATILITY EFFECT Selection, which selects stocks with the lowest historical return volatility, capturing the empirically observed inverse relationship between volatility and risk-adjusted performance (Blitz and Vliet, 2007).

⁶<https://sec-api.io/>

⁷<https://alpaca.markets/>

⁸<https://www.lseg.com/en>

C Evaluation Metrics

We group evaluation metrics into three categories, each targeting a distinct aspect of strategy performance:

C.1 Return Metrics

Annualised Return (AR). Measures the geometric average yearly return: $R_{\text{annual}} = (1 + C)^{\frac{252}{T}} - 1$ where $C = \prod_{t=1}^T (1 + R_t) - 1$.

Cumulative Return (CR). Total portfolio return over the test period: $C = \prod_{t=1}^T (1 + S_t \cdot R_t) - 1$, where S_t is the position and R_t is the market return.

C.2 Risk Metrics

Annualised Volatility (AV). Volatility of daily returns, scaled annually: $\sigma_{\text{annual}} = \sigma_{\text{daily}} \times \sqrt{252}$.

Maximum Drawdown (MDD). Largest drop from peak portfolio value: $\text{MDD} = -\max_{t \in [1, T]} \left(\frac{\text{Peak}_t - V_t}{\text{Peak}_t} \right)$.

C.3 Risk-adjusted Performance Metrics

Sharpe Ratio (SPR). Return per unit of total volatility: $\text{SPR} = \frac{\overline{R_t} - \frac{r_f}{252}}{\sigma_{\text{daily}}} \times \sqrt{252}$.

Sortino Ratio (STR). Return per unit of downside deviation: $\text{STR} = \frac{\overline{R_t} - \frac{r_f}{252}}{\sigma_{\text{down}}} \times \sqrt{252}$.

Where $\overline{R_t}$ is the average daily return, r_f is the annual risk-free rate, and σ_{down} is the standard deviation of negative returns.

D Extra Results on Selective Symbols

Tables 2 and 5 further substantiate our findings by highlighting the performance instability of *FinMem* and *FinAgent* when extending evaluation periods even marginally. Specifically, extending the evaluation by just two months beyond the originally reported periods (Yu et al., 2023) results in notable inconsistencies in critical performance metrics. It should be noted that the results for the LLM strategies are retrieved from Yu et al. (2024), while the traditional rule-based results presented are based on our implementations.

For instance, *FinMem* exhibited a drastic change in cumulative returns for MSFT from a reported 23.261% down to -22.036%, and a reduction in Sharpe ratios from 1.440 to -1.247. Similarly, for NFLX, the Sharpe ratio for *FinMem* shifted dramatically from a reported 2.017 to -0.478. These examples underscore the sensitivity of LLM-based

investing strategies to minor shifts in market conditions and reinforce our argument about the necessity of comprehensive and temporally robust evaluations to accurately assess the reliability and generalisability of these models.

E Technical Details

FINSABER Implementation. The backtesting framework and traditional rule-based strategies in FINSABER are implemented using BackTrader⁹ and Papers With Backtest¹⁰. Reinforcement learning-based methods are implemented using FinRL (Liu et al., 2021). FINSABER supports two operational modes: “LLM” mode and “BT” mode. The “LLM” mode is tailored for strategies that leverage multi-modal inputs, including financial news and regulatory filings. In contrast, the “BT” mode is built directly on BackTrader, offering robust support for traditional rule-based strategies while maintaining a familiar interface to facilitate easy migration from standard BackTrader workflows.

Experiment Rolling Windows. We apply a rolling-window evaluation setup to ensure temporal robustness and reduce data-snooping bias. For the **Selected 4** evaluation, we use a 2-year rolling window with a 1-year step, and allow strategies to use up to 3 years of prior data for training. For the **Composite** setup, we adopt a more frequent rebalancing scheme with a 1-year rolling window and a 1-year step, allowing up to 2 years of prior data. This adjustment reflects the observation that rebalancing every two years may be too infrequent to capture changing market dynamics. All experiments span the benchmark period from 2004 to 2024.

Parameters of Strategies. Table 6 summarises the key hyperparameters used for each benchmark strategy in our experiments. These settings are largely drawn from standard defaults commonly used in the public implementations. For traditional rule-based strategies, optimal parameter selection often requires domain expertise or practitioner experience. Our goal is not to optimise each strategy’s absolute performance, but to provide a fair and consistent baseline under a unified evaluation framework. We encourage future researchers to explore parameter optimisation techniques (e.g., grid

⁹<https://www.backtrader.com/>

¹⁰<https://paperswithbacktest.com/>

Type	Strategy	TSLA			AMZN			NIO			MSFT		
		SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD
FinCon Selection (2022-10-05 to 2023-06-10)													
Rule Based	Buy And Hold	0.247	2.056	-54.508	0.150	2.193	-32.177	-0.858	-51.569	-53.563	1.071	32.629	-14.452
	SMA Cross	-0.151	-3.973	-23.173	0.599	13.731	-18.910	0.810	22.047	-17.976	1.641	32.057	-8.746
	WMA Cross	1.104	32.058	-18.492	0.513	11.765	-21.030	-0.771	-9.412	-18.732	1.526	30.344	-8.883
	ATR Band	-0.554	-22.136	-39.599	0.494	11.007	-15.842	0.681	24.684	-21.229	0.827	12.979	-7.709
	Bollinger Bands	-0.249	-12.756	-44.655	-0.381	-7.105	-20.615	0.940	25.476	-16.623	1.759	31.619	-3.475
	Turn of The Month	0.928	27.850	-11.642	0.123	3.487	-14.892	0.874	31.344	-17.995	0.407	7.744	-11.955
LLM	FinGPT	0.044	1.549	-42.400	-1.810	-29.811	-29.671	-0.121	-4.959	-37.344	1.315	21.535	-16.503
	FinMem	1.552	34.624	-15.674	-0.773	-18.011	-36.825	-1.180	-48.437	-64.144	-1.247	-22.036	-29.435
	FinAgent	0.271	11.960	-55.734	-1.493	-24.588	-33.074	0.051	0.933	-19.181	-1.247	-27.534	-39.544
	FinCon	1.972	82.871	-29.727	0.904	24.848	-25.889	0.335	17.461	-40.647	1.538	31.625	-15.010
Type	Strategy	AAPL			GOOG			NFLX			COIN		
		SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD	SPR	CR	MDD
FinCon Selection (2022-10-05 to 2023-06-10)													
Rule Based	Buy And Hold	0.906	24.558	-19.508	0.683	20.884	-20.278	1.594	77.367	-20.421	0.024	-23.761	-54.402
	SMA Cross	1.423	21.054	-6.030	0.382	8.497	-17.035	-0.855	-8.393	-18.545	0.232	1.286	-35.559
	WMA Cross	1.648	25.257	-6.114	0.635	13.659	-14.985	-1.009	-9.479	-18.531	0.087	-7.461	-40.883
	ATR Band	0.241	4.522	-5.159	0.067	2.616	-13.522	0.522	10.739	-12.231	0.777	25.169	-22.906
	Bollinger Bands	-	-	-	0.365	7.526	-13.522	-0.182	-0.710	-13.244	-0.705	-24.371	-40.733
	Turn of The Month	0.098	3.337	-12.498	0.343	7.188	-13.519	0.987	18.942	-10.641	-0.020	-8.999	-33.895
LLM	FinGPT	1.161	20.321	-16.759	0.011	0.242	-26.984	0.472	11.925	-20.201	-1.807	-99.553	-74.967
	FinMem	0.994	12.397	-11.268	0.018	0.311	-21.503	-0.478	-10.306	-27.692	0.017	0.811	-50.390
	FinAgent	1.041	20.757	-19.896	-1.024	-7.440	-10.360	1.960	61.303	-20.926	-0.106	-5.971	-56.882
	FinCon	1.597	27.352	-15.266	1.052	25.077	-17.530	2.370	69.239	-20.792	0.825	57.045	-42.679

Table 5: Backtest performance of traditional rule-based (indicator-based) strategies and *FinCon* over the selective period (2022-10-05 to 2023-06-10), as presented in Yu et al. (2024), evaluated using four metrics: cumulative return (CR), Sharpe ratio (SPR), annual volatility (AV), and maximum drawdown (MDD). The best metrics are highlighted in red, while the second best are marked in blue. “-” metrics across the board indicate no trade signals were triggered.

Strategies	Parameters
SMA Cross	short_window=10, long_window=20
WMA Cross	short_window=10, long_window=20
ATR Band	atr_period=14, multiplier=1.5
Bollinger Band	period=20, devfactor=2.0
Trend Following	atr_period=10, period=20
Turn of the Month	before_end_of_month_days=5, after_start_of_month_business_days=3
ARIMA	order=(5,1,0)
XGBoost	num_boost_round=10, n_estimators=1000
RL	total_timesteps=50000
FinMem	model=gpt-4o-mini, top_k=3, embedding_model=text-embedding-ada-002, chunk_size=5000
FinAgent	model=gpt-4o-mini, trader_preference=aggressive_trader, top_k=5, previous_action_look_back_days=14

Table 6: Default parameter settings for benchmark strategies.

search, Bayesian tuning) if desired.

F LLM Strategies Cost Analysis

To better understand the practical deployment of LLM-based investing strategies, we monitor the API costs associated with running backtests on the **Composite** experiment with VOLATILITY EFFECT selection as a representative example. The cost for backtesting *FinAgent* was \$198.24, while *FinMem* incurred a significantly lower cost of \$31.79 using GPT-4o mini. This reflects the higher prompt complexity and more frequent calls involved in *FinAgent*’s multi-agent decision-making process.

Extrapolating from these numbers, we estimate

that completing all **Composite** experiments required approximately \$700 in LLM API costs. The **Selected 4** setup likely incurred even greater cost, given its larger rolling window size and the increased volume of financial news associated with these selectively popular symbols.

FinAgent was roughly 6 times more expensive than *FinMem* in our tests. Importantly, these figures only account for LLM generation costs (i.e., chat/completions endpoints), and do not include the cost of generating embeddings (e.g., via text-embedding-ada-002), which would further increase the total budget.

This observation raises a practical considera-

tion for future research: when evaluating LLM-driven strategies, computational cost should be factored into the financial metrics, particularly for real-world deployment scenarios. Incorporating API usage cost into risk-adjusted performance metrics (e.g., Sharpe or Sortino) could provide a more holistic picture of strategy efficiency.

Recommendation. For researchers with limited budget, we recommend adopting open-source LLMs (e.g., LLaMA, Qwen, Mistral) for benchmarking and prototyping. These models can be deployed locally or via cost-effective cloud infrastructure, significantly reducing evaluation costs while enabling reproducible experimentation.