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

Weixian Waylon Li
waylon.li@ed.ac.uk
University of Edinburgh
United Kingdom

Hyeonjun Kim
Sungkyunkwan University
South Korea

Mihai Cucuringu
University of California Los Angeles, University of Oxford
United States, United Kingdom

Tiejun Ma
tiejun.ma@ed.ac.uk
University of Edinburgh
United Kingdom

## 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 FINSABER[1], 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.

## CCS Concepts

**• General and reference → Evaluation**; **Empirical studies**; **• Computing methodologies → Intelligent agents**.

## Keywords

Automated trading, LLM investors, Backtest, Benchmark

## 1 Introduction

Large language models (LLMs) are increasingly used in financial decision-making, especially for generating investment actions such as `Buy`, `Hold`, or `Sell` [12, 19]. 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 [55] to sophisticated multi-agent systems [53, 56], a growing body of work has explored the potential of LLMs as autonomous financial agents.

Backtesting is the standard method for assessing investment strategies, simulating them on historical data to evaluate profitability and robustness [8]. 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** [22], where delisted or failed stocks are omitted; **look-ahead bias** [8], where future information inadvertently influences past decisions; and **data-snooping bias** [2], 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 | Eval Period | Eval Symbols | Code |
|---|---|---|---|
| MarketSenseAI | 1 year 3 months | 100 | ✗ |
| TradingGPT | N/A | N/A | ✗ |
| FinMem | 6 months | 5 | ✓ |
| FinAgent | 6 months | 6 | ✓ |
| FinRobot | N/A | N/A | ✓ |
| TradExpert | 1 year | 30 | ✗ |
| FinCon | 8 month | 8 | ✗ |
| TradingAgents | 3 months | 3 | ✗ |
| MarketSenseAI 2.0 | 2 years | 100 | ✗ |

**Table 1: Summary of current LLM-based investing strategies.**

While recent efforts such as Wang et al. [45] and Hu et al. [27] have addressed benchmarking for deep learning (DL)-based trading and LLM-based time-series forecasting, comprehensive evaluation of LLM-based investing strategies remains unaddressed. Separately, FinBen [50] provides a thorough FinLLM benchmark covering multiple tasks, including decision-making. However, as a broad FinLLM benchmark, FinBen's backtesting still relies on a limited, hand-picked symbol set, which contains the aforementioned biases and lacks a professional backtesting pipeline or systematic comparison with traditional strategies. To fill this gap, we introduce **FINSABER**, a comprehensive framework for benchmarking LLM timing-based investing strategies that supports **longer backtesting periods**, a **broader and more diverse symbol universe**, and **explicit bias mitigation**. Specifically, our main contributions are:

[1] Data and code available at https://github.com/waylonli/FINSABER.

(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 [12]. The most common approach leverages LLMs' sentiment analysis capabilities, using either general-purpose LLMs (e.g., GPT, LLaMA [42], Qwen [1]) or fine-tuned financial variants like FinGPT [51] to generate sentiment scores for trading decisions [32, 38, 47, 55]. 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. [20] introduce a memory module that stores summarised financial data, retrieved during trading to guide decisions. Similarly, LLMFactor [44] 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 [25], where either one specialised agent or multiple collaborative agents are employed to perform financial analysis or predictions. Notable examples include FinMem [53], FinAgent [56], FinRobot [52], TradExpert [13], FinCon [54], TradingAgents [49] and MarketSenseAI 2.0 [19]. Some models also incorporate reinforcement learning (RL) for iterative self-improvement [14, 33].

## 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 and technical indicators. 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.

## 4 Biases and Robustness Challenges in Backtesting LLM Investors

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 [8].

*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 [29]. A common cause is using today's S&P 500 constituents as the historical investment universe. This practice introduces what Garcia and Gould [22] 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 [24] and Elton et al. [17] estimate annual return distortions between 0.1% and 0.9%, and Brown et al. [6] 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 [8]. 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. [2] 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.

*Bias-Mitigation Requires 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 [2]. Gatev et al. [23] tested pairs trading on 40 years of daily data, but Do and Faff [15] 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 [21, 43, 48].

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. 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.

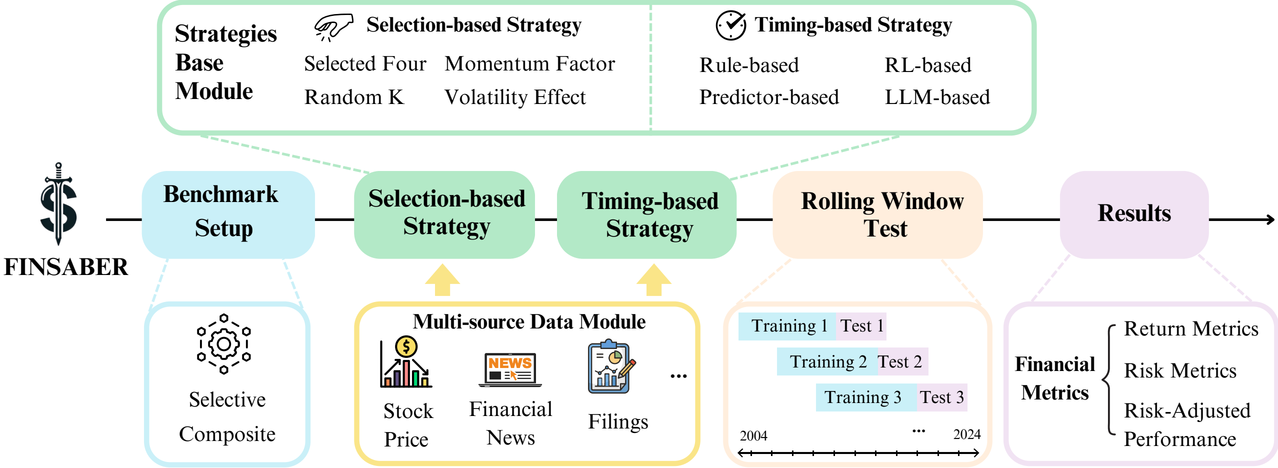

**Figure 1: Overview of the FINSABER Backtest Framework. The central pipeline illustrates the backtesting process. The framework includes a Strategies Base Module (green), which covers both selection-based and timing-based strategies, and a Multi-source Data Module (yellow), integrating diverse financial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

## 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, FINSABER[2], which systematically addresses biases and meets 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 strategies 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 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).

*Strategies Base.* We incorporate a comprehensive collection of strategies across multiple paradigms to ensure robust benchmarking. The *timing-based strategies* include open-source LLM investors (FinMem [53], FinAgent [56]), traditional rule-based approaches (Buy and Hold, Moving Average Crossover, Bollinger Bands [4], Trend Following [46]), ML/DL forecaster-based methods (ARIMA, XGBoost), and RL-based strategies (A2C, PPO, TD3, SAC implemented via FinRL [36] framework). Selection-based strategies encompass random K selection, Momentum Factor Selection (based on past returns), Volatility Effect Selection (selecting low-volatility stocks), and the stocks selection agent from the FinCon [54] framework. This diverse strategy base enables comprehensive performance comparison across different methodological approaches while maintaining extensibility for custom implementations. More technical details of the strategies are available in Appendix B.

*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 [8]. These metrics are standard in the literature [10, 11] and widely used in recent LLM-based investing benchmarks [54, 56]. Formal definitions and formulas are provided in Appendix C.

[2] **F**inancial **IN**vesting **S**trategy **A**ssessment with **B**ias mitigation, **E**xpanded time, and **R**ange of symbols