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 Effect selection (Sharpe 0.703) outperforms FinCon Selection Agent (0.389) and Momentum Factor (0.384), which in turn surpass 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 three. This suggests **RL methods are more dependent on 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 of LLM-based investing strategies.

6.3 Statistical Validation and Behavioural Diagnostics of LLM Agents

To validate our findings from the composite backtests and diagnose the underlying drivers of LLM agent performance, we conduct a unified statistical and behavioural analysis. First, we conduct paired t-tests comparing *Buy and Hold*, *FinMem*, and *FinAgent* across both **Selected 4** (Table 3) and **Composite** (Table 4) setups. Second, we dissect the agents' behavioural characteristics by examining their drawdown profiles, alpha (α) and beta (β) decomposition, and trading turnover across the different selection environments. These metrics are obtained by regressing the strategy's excess returns against the market's excess returns based on the Capital Asset Pricing Model (CAPM) [41]. The model is defined as: $R_s - R_f = \alpha + \beta(R_m - R_f) + \epsilon$, where R_s is the return of the strategy, R_m is the market return, R_f is the risk-free rate, and ϵ is the idiosyncratic residual. In this model, β measures the strategy's systematic risk or volatility relative to the market, while α represents the portion of the return not explained by market exposure, often considered a measure of strategy-specific skill.

Setup	B&H vs FinMem	B&H vs FinAgent	FinMem vs FinAgent
Selective symbols, expanded period (Selected four; Table 3)			
TSLA	0.3643	0.1663	0.2258
NFLX	0.0436	0.0363	0.1493
AMZN	0.0127	0.0984	0.4023
MSFT	0.0005	0.2252	0.5549
Bias-mitigated (Composite; Table 4)			
Random 5	3.0e-6	7.7e-4	4.0e-3
Momentum	4.0e-5	0.0117	0.2001
Volatility Effect	4.0e-6	5.9e-4	3.8e-3

Table 5: Paired t-test p-values comparing *B&H, FinMem,* and *FinAgent* under Selected 4 and Composite setups.

Table 5 reports t-tests and p-values for the previous results, testing the null hypothesis of equal performance distributions. Under the selective period, statistical significance is inconsistent and limited mostly to individual stocks. However, after mitigating biases through the composite setup, the p-values drop substantially, indicating the market baseline (*B&H*) significantly outperforms both LLM strategies across all robust setups. Notably, while *FinAgent* tends to outperform *FinMem* when biases are controlled, both still underperform simple market baselines. Furthermore, the behavioural analysis in Table 6 reveals that this underperformance is rooted in a lack of genuine skill; **neither LLM agent generates statistically significant alpha**, with all measured p-values exceeding 0.34. This finding robustly supports our main thesis that the claimed superiority of these models does not hold under rigorous evaluation, aligning with the Efficient Market Hypothesis [37].

A clear behavioural hierarchy emerges between the two agents. *FinMem* consistently shows a more pathological trading profile, marked by excessive turnover and poor risk management. Its commission ratio is five to nine times higher than *FinAgent*'s across both contexts, and its drawdown durations are substantially longer. This overtrading leads to persistent value destruction, reflected in *FinMem*'s negative alpha in all scenarios. In contrast, *FinAgent* follows a more restrained, though still unskilled, trading strategy. Appendix F provides a comparative analysis with visualisations to further highlight the behavioural differences between *FinMem* and *FinAgent* as supplementary evidence.

These behaviours are directly modulated by the selection strategy, which acts as a powerful environmental filter. The **Momentum selection** strategy elicits the most engaged market posture from the agents, prompting their highest β values. *FinMem*'s performance improves in this context relative to other environments, but it still yields a negative alpha of -1.34%. This is the only scenario where *FinAgent* produces a large positive alpha of **+6.57%**. Although this result lacks statistical significance (p=0.35), it suggests that the LLMs' primary strength may not be in *discovering* novel signals but rather in *exploiting* strong, pre-existing market trends. In contrast, the **Low Volatility** environment takes a risk-averse posture. Here, *FinMem* remains ineffective with a -1.04% alpha and a very low β of 0.20. *FinAgent* also becomes highly conservative, with its risk profile improving (e.g., its average drawdown duration falls to 38.71 days) but at the cost of performance, generating a negative alpha.

In summary, this unified analysis statistically validates the underperformance of LLM agents and reveals that their behaviour is not monolithic. It is highly dependent on the characteristics of the asset universe they operate within, reinforcing the need for bias-mitigated evaluation frameworks like FINSABER.

Strategy	Avg Max Drawdown (Days)	Avg Regular Drawdown (Days)	Alpha (%)	Beta	Alpha p-value
Momentum Factor					
FinMem	210	80	-1.343	0.518	0.477
FinAgent	150	59	6.571	0.758	0.345
Volatility Effect					
FinMem	177	71	-1.036	0.199	0.430
FinAgent	123	39	-0.196	0.354	0.368

Table 6: Behavioural analysis of LLM timing strategies, highlighting drawdown duration, alpha (α) and beta (β) decomposition, and trading turnover (commission ratio).

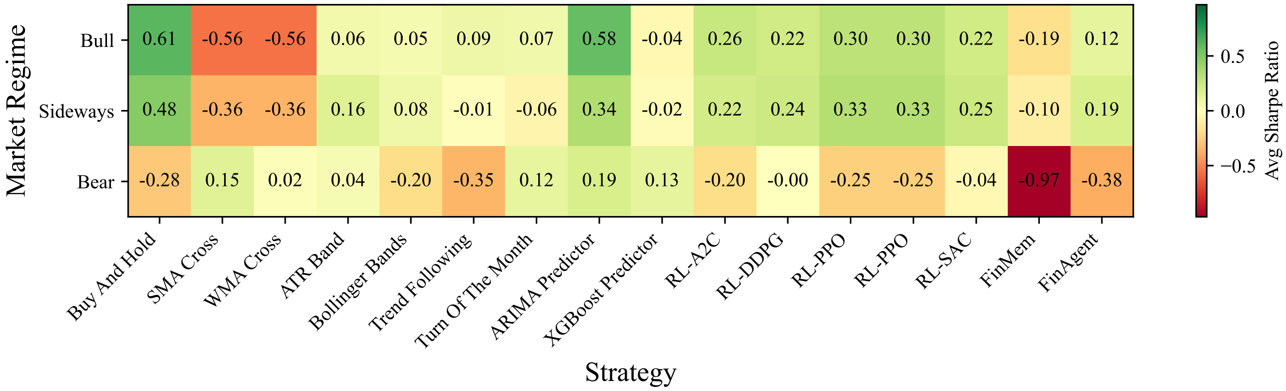

Figure 2: Average Sharpe ratio by regime for all benchmarking strategies. Green = strong, red = weak.

7 Market Regime Analysis

Another key question in evaluating LLM-based investing strategies is whether they adapt appropriately across varying market conditions. Financial markets exhibit time-varying predictability and uncertainty across different economic, and political regimes [30]. 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 [27].

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 ±20% threshold follows standard industry convention [55].

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 {*strategy*, *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.

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

These results suggest two directions for future LLM investors. First, trend-detection capabilities to ensure that the strategy can at least match passive equity beta during upward market phases. Second, incorporating explicit regime-aware risk controls that reduce exposure as volatility or drawdown risk increases. Balancing risk-taking and risk management, rather than simply increasing model size, appears the key to closing the gap with traditional methods.

8 Findings and Takeaways

Our investigation via the FINSABER framework offers several novel findings that challenge the prevailing narrative on LLM-based investors and set a new baseline for future research.

First, we find that **LLM-derived alpha is likely a methodological artefact of narrow, biased evaluations.** The performance advantages reported in short-term, selective studies vanish under our bias-mitigated backtests, which reveal a consistent and statistically significant failure to generate alpha (§6.3). This suggests that current LLMs do not overcome the Efficient Market Hypothesis [17] in reality, and that prior gains stemmed from survivorship and look-ahead biases rather than genuine market inefficiency.

Second, **model complexity does not equate to market competence.** The scaling laws of natural language processing [29] do not translate effectively to financial markets, which impose intrinsic limits on extractable signals [25]. We show that larger models do not reliably outperform smaller ones, and both are consistently bettered by simpler models like ARIMA on risk-adjusted metrics (Table 4). Without encoded financial logic, architectural complexity appears to add noise rather than value.

Third, we diagnose "how" LLM agents fail, revealing a **fundamental misalignment with market regimes.** Our further analysis (§7, Appendix F) shows that agents are pathologically miscalibrated: they are too conservative in bull markets and too aggressive in bear markets. This behavioural flaw contradicts the Adaptive Markets Hypothesis [37], shifting the issue from merely a lack of profitability to a more profound failure in the agents' decision-making policies.

Synthesising these points, our work establishes that the primary barrier to successful LLM investors is not model scale, but a **lack of domain-aware financial logic**. The path forward is designing smarter, more adaptive agents, and FINSABER provides the framework to rigorously test such designs, moving the field beyond flawed evaluations toward practical and robust financial agents.

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. 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 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.

A remaining limitation is potential data leakage, as some evaluation data may have been included in the pretraining corpora of proprietary LLMs and cannot be fully verified. However, any such leakage would bias results in favour of LLMs and therefore does not alter our central findings.

Finally, our cost analysis (Appendix G) 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.

Acknowledgements

We thank the reviewers and the area chair for their useful feedback. The authors acknowledge the use of resources provided by the Edinburgh Compute and Data Facility[5] (ECDF).

References

[1] 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)* (2015).

[2] David Blitz and Pim Vliet. 2007. The Volatility Effect: Lower Risk without Lower Return. *The Journal of Portfolio Management* 34 (2007).

[3] J. Bollinger. 2002. *Bollinger on Bollinger Bands.*

[4] George Edward Pelham Box and Gwilym Jenkins. 1990. *Time Series Analysis, Forecasting and Control.*

[5] 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 (1992), 553–580.

[6] Mark M. Carhart. 1997. On Persistence in Mutual Fund Performance. *The Journal of Finance* 52, 1 (1997), 57–82.

[7] E.P. Chan. 2021. *Quantitative Trading: How to Build Your Own Algorithmic Trading Business.*

[8] 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, San Francisco, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). 785–794.

[9] Rama Cont. 2001. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* 1, 2 (2001), 223–236.

[10] 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 (2007), 1915–1953.

[11] Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large Language Model Agent in Financial Trading: A Survey.

[12] Qianggang Ding, Haochen Shi, and Bang Liu. 2024. TradExpert: Revolutionizing Trading with Mixture of Expert LLMs.

[13] 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.

[14] Binh Do and Robert Faff. 2010. Does simple pairs trading still work? *Financial Analysts Journal* 66, 4 (2010), 83–95.

[15] 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 2024, Barcelona, Spain, August 25-29, 2024*, Ricardo Baeza-Yates and Francesco Bonchi (Eds.). 4918–4927.

[16] Edwin J Elton, Martin J Gruber, and Christopher R Blake. 1996. Survivor bias and mutual fund performance. *The review of financial studies* 9, 4 (1996), 1097–1120.

[17] Eugene F Fama. 1970. Efficient capital markets. *Journal of Finance* 25, 2 (1970), 383–417.

[18] George Fatouros, Kostas Metaxas, John Soldatos, and Manos Karathanassis. 2025. MarketSenseAI 2.0: Enhancing Stock Analysis through LLM Agents.

[19] 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.

[20] 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, Article 27 (2019), 30 pages.

[21] CB Garcia and FJ Gould. 1993. Survivorship bias. *Journal of Portfolio Management* 19, 3 (1993), 52.

[22] 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 (2006), 797–827.

[23] Mark Grinblatt and Sheridan Titman. 1989. Mutual fund performance: An analysis of quarterly portfolio holdings. *Journal of business* (1989), 393–416.

[5] http://www.ecdf.ed.ac.uk/