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 *Buy and Hold (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 (*Buy and Hold*) significantly outperforms both LLM strategies across all robust setups. Notably, while *FinAgent* tends to outperform *FinMem* when biases are controlled, both still significantly 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 also emerges between the two agents. ***FinMem* consistently exhibits a more pathological trading profile**, characterised by hyperactive turnover and poor risk management. Its commission ratio is approximately 5 to 9 times higher than *FinAgent*'s across both contexts, and its drawdown durations are substantially longer. This excessive, costly trading correlates with consistent value destruction, as evidenced by *FinMem*'s negative alpha in all scenarios. In contrast, *FinAgent* demonstrates a more controlled, albeit still unskilled, approach. Appendix F provides a comparative analysis with visualisations to further highlight the behavioural differences between *FinMem* and *FinAgent*.

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 prompts 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).

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, financial, and political regimes [31]. 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 [28].

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 [57].

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.

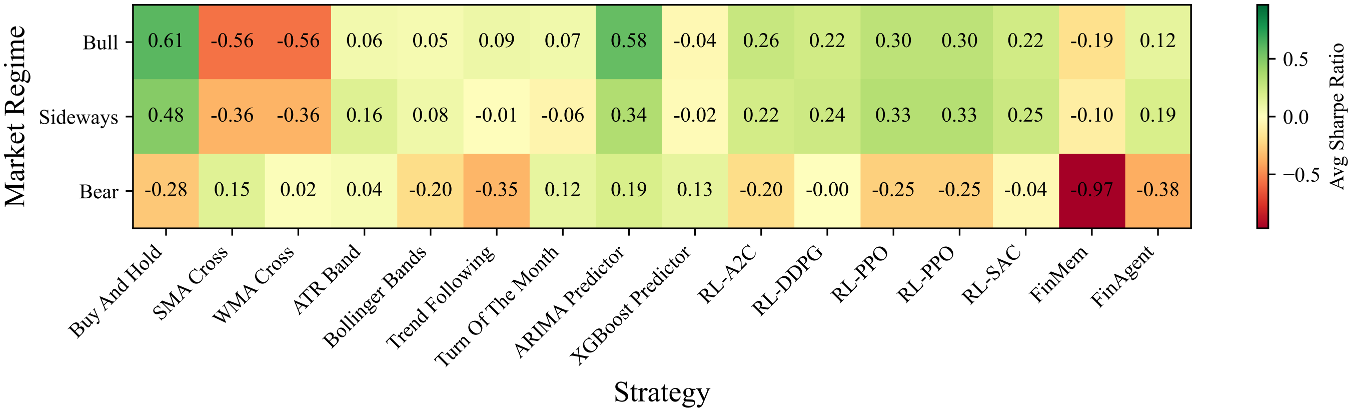

Figure 2: Average Sharpe ratio by regime for all benchmarking strategies. Green = strong, red = weak.

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. 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

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 (Section 6.3). This suggests that current LLMs do not overcome the Efficient Market Hypothesis [18] in realistic conditions, 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 [30] do not translate effectively to financial markets, which impose intrinsic limits on extractable signals [26]. 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 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.

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.

References

[1] 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, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv:2309.16609 [cs.CL] https://arxiv.org/abs/2309.16609

[2] 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). https://api.semanticscholar.org/CorpusID:14849749

[3] David Blitz and Pim Vliet. 2007. The Volatility Effect: Lower Risk without Lower Return. *The Journal of Portfolio Management* 34 (08 2007). doi:10.3905/jpm.2007.698039

[4] J. Bollinger. 2002. *Bollinger on Bollinger Bands*. McGraw-Hill Education. https://books.google.com.sg/books?id=RxkVMFg-KIIC

[5] George Edward Pelham Box and Gwilym Jenkins. 1990. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., USA.

[6] 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.

[7] Mark M. Carhart. 1997. On Persistence in Mutual Fund Performance. *The Journal of Finance* 52, 1 (1997), 57–82. doi:10.1111/j.1540-6261.1997.tb03808.x

[8] E.P. Chan. 2021. *Quantitative Trading: How to Build Your Own Algorithmic Trading Business*. Wiley. https://books.google.co.uk/books?id=j70yEAAAQBAJ

[9] 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, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. doi:10.1145/2939672.2939785

[10] Rama Cont. 2001. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* 1, 2 (2001), 223–236. doi:10.1080/713665670

[11] 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 (12 2007), 1915–1953. doi:10.1093/rfs/hhm075

[12] Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large Language Model Agent in Financial Trading: A Survey. arXiv:2408.06361 [q-fin.TR] https://arxiv.org/abs/2408.06361

[13] Qianggang Ding, Haochen Shi, and Bang Liu. 2024. TradExpert: Revolutionizing Trading with Mixture of Expert LLMs. arXiv:2411.00782 [cs.AI] https://arxiv.org/abs/2411.00782

[14] 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. https://ideas.repec.org/p/arx/papers/2310.05627.html

[15] Binh Do and Robert Faff. 2010. Does simple pairs trading still work? *Financial Analysts Journal* 66, 4 (2010), 83–95. doi:10.2469/faj.v66.n4.1

[16] 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* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 4918–4927. doi:10.1145/3637528.3671629

[17] 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.

[18] Eugene F Fama. 1970. Efficient capital markets. *Journal of Finance* 25, 2 (1970), 383–417. doi:10.1111/jofi.12365

[19] George Fatouros, Kostas Metaxas, John Soldatos, and Manos Karathanassis. 2025. MarketSenseAI 2.0: Enhancing Stock Analysis through LLM Agents. arXiv:2502.00415 [q-fin.CP] https://arxiv.org/abs/2502.00415

[20] 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. arXiv:2401.03737 [q-fin.CP] https://arxiv.org/abs/2401.03737

[21] 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 (March 2019), 30 pages. doi:10.1145/3309547

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

[23] 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 (02 2006), 797–827. doi:10.1093/rfs/hhj020

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

[25] 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. arXiv:2402.01680 [cs.CL] https://arxiv.org/abs/2402.01680

[26] Campbell Harvey and Yan Liu. 2013. Backtesting. *SSRN Electronic Journal* 42 (01 2013). doi:10.2139/ssrn.2345489

[27] 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. arXiv:2502.18834 [cs.CE] https://arxiv.org/abs/2502.18834

[28] 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 (09 2018), 381–402. doi:10.3846/ijspm.2018.5222

[29] Jacques Joubert, Dragan Sestovic, Illya Barziy, Walter Distaso, and Marcos Lopez de Prado. 2024. The three types of backtests. *Available at SSRN* (2024).

[30] 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. arXiv:2001.08361 [cs.LG] https://arxiv.org/abs/2001.08361

[31] 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 (2011), 868–879. doi:10.1016/j.jempfin.2011.08.002

[32] Kemal Kirtac and Guido Germano. 2024. Sentiment trading with large language models. *Finance Research Letters* 62 (2024), 105227. doi:10.1016/j.frl.2024.105227

[33] 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* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 4304–4315. doi:10.1145/3589334.3645611

[34] Xiao-Yang Liu, Ziyi Xia, Hongyang Yang, Jiechao Gao, Daochen Zha, Ming Zhu, Christina Dan Wang, Zhaoran Wang, and Jian Guo. 2024. Dynamic Datasets and Market Environments for Financial Reinforcement Learning. *Machine Learning - Springer Nature* (2024).

[35] 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)* (2021).

[36] 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* (Virtual Event) *(ICAIF '21)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. doi:10.1145/3490354.3494366

[37] Andrew Lo. 2004. The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. *The Journal of Portfolio Management* 30, 5 (10 2004), 15–29.

[38] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv:2304.07619 [q-fin.ST] https://arxiv.org/abs/2304.07619

[39] John J. McConnell and Wei Xu. 2008. Equity Returns at the Turn of the Month. *Financial Analysts Journal* 64, 2 (March 2008), 49–64. doi:10.2469/faj.v64.n2.11

[40] C Muller and M Ward and. 2010. Momentum Effects in Country Equity Indices. *Studies in Economics and Econometrics* 34, 1 (2010), 111–127. doi:10.1080/03796205.2010.12129444

[41] William F. Sharpe. 1964. CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK. *The Journal of Finance* 19, 3 (1964), 425–442. doi:10.1111/j.1540-6261.1964.tb02865.x

[42] 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. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971

[43] 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*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3970–3977. doi:10.24963/ijcai.2022/551 Main Track.

[44] Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. LLMFactor: Extracting Profitable Factors through Prompts for Explainable Stock Movement Prediction. arXiv:2406.10811 [cs.CL] https://arxiv.org/abs/2406.10811

[45] Saizhuo Wang, Hao Kong, Jiadong Guo, Fengrui Hua, Yiqi Qi, Wanyun Zhou, Jiahao Zheng, Xinyu Wang, Lionel M. Ni, and Jian Guo. 2025. QuantBench: Benchmarking AI Methods for Quantitative Investment. arXiv:2504.18600 [q-fin.CP] https://arxiv.org/abs/2504.18600

[46] Cole Wilcox, Eric Crittenden, and Blackstar Funds. 2005. Does Trend Following Work on Stocks. In *The Technical Analyst*, Vol. 14. 1–19. https://api.semanticscholar.org/CorpusID:166585921