[24] 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. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. 8048–8057.

[25] Campbell Harvey and Yan Liu. 2013. Backtesting. *SSRN Electronic Journal* 42 (2013).

[26] 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.

[27] 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 (2018), 381–402.

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

[29] 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.

[30] 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.

[31] Kemal Kirtac and Guido Germano. 2024. Sentiment trading with large language models. *Finance Research Letters* 62 (2024), 105227.

[32] 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 on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee (Eds.). 4304–4315.

[33] Weixian Waylon Li and Tiejun Ma. 2025. Learn to Rank Risky Investors: A Case Study of Predicting Retail Traders' Behaviour and Profitability. *ACM Trans. Inf. Syst.* 44, 1, Article 15 (Nov. 2025), 33 pages. doi:10.1145/3768623

[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)*. Article 1, 9 pages.

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

[38] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models.

[39] John J. McConnell and Wei Xu. 2008. Equity Returns at the Turn of the Month. *Financial Analysts Journal* 64, 2 (2008), 49–64.

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

[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.

[42] 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 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). 3970–3977.

[43] Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. LLMFactor: Extracting Profitable Factors through Prompts for Explainable Stock Movement Prediction.

[44] Saizhuo Wang, Hao Kong, Jiadong Guo, Fengrui Hua, Yiyan Qi, Wanyun Zhou, Jiahao Zheng, Xinyu Wang, Lionel M. Ni, and Jian Guo. 2025. QuantBench: Benchmarking AI Methods for Quantitative Investment.

[45] Cole Wilcox, Eric Crittenden, and Blackstar Funds. 2005. Does Trend Following Work on Stocks. In *The Technical Analyst*, Vol. 14. 1–19.

[46] Ruoxu Wu. 2024. Portfolio Performance Based on LLM News Scores and Related Economical Analysis. *SSRN Electronic Journal* (2024).

[47] Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024. TradingAgents: Multi-Agents LLM Financial Trading Framework. *ArXiv preprint* abs/2412.20138 (2024).

[48] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024. FinBen: A Holistic Financial Benchmark for Large Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.).

[49] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *FinLLM Symposium at IJCAI 2023* (2023).

[50] 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.

[51] 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.

[52] 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, Guojun 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 *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.).

[53] Haohan Zhang, Fengrui Hua, Chengjin Xu, Hao Kong, Ruiting Zuo, and Jian Guo. 2023. Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements?

[54] 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. 2024. A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. 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.). 4314–4325.

[55] Jason Zweig. 2019. Where Did This 'Bull Market' Come From, Anyway? *The Wall Street Journal* (2019).

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. Additionally, our dataset includes delisted symbols that were historically part of the S&P 500 index, based on the archived constituent list. 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. [15], 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 via the SEC-API[6], 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

[6]https://sec-api.io/

if available. Researchers may incorporate paid datasets such as premium financial news (e.g., Alpaca Markets[7], Refinitiv[8]), 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* [51] and *FinRobot* [50]. We acknowledge other works, such as *FinCon* [52] and MarketSenseAI [18], 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* [3], *Trend Following* [45], and *Turn of the Month* [39]. 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 (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 [4] and XGBoost [8] in this category and also cover forecasters based on LLMs, but these are not LLM investors.

RL-Based Strategies. We also implement widely used RL algorithms for financial markets, including Advantage Actor-Critic (A2C), Proximal Policy Optimisation (PPO), Twin Delayed Deep Deterministic Policy Gradient (TD3), and Soft Actor-Critic (SAC), utilising the FinRL framework [34, 36]. Each agent learns investing policies by interacting with a simulated trading environment based on the OpenAI Gym API, using real historical market data.

B.2 Selection-based Strategies

This section details the implementation of the primary selection strategies used in our composite backtesting framework. Each selector operates on the historical S&P 500 constituents available at the start of a given rolling-window period to produce a list of tickers for the timing-based strategies.

[7]https://alpaca.markets/
[8]https://www.lseg.com/en

Random Five. This strategy serves as a simple baseline for performance comparison. At the beginning of each evaluation period, it selects five stocks at random, without replacement, from the list of all available historical S&P 500 constituents for that period.

Momentum Factor. Following the well-documented momentum factor [40], this strategy selects the stocks with the highest recent price appreciation. For each candidate stock, we calculate a momentum score based on its historical price data. Specifically, the score is the percentage return over a "momentum period" (e.g., 100 trading days), but we exclude the most recent "skip period" (e.g., 21 trading days) from the calculation. This practice is common in momentum strategies to avoid the "short-term reversal" effect [6]. The score for a given stock is calculated as: Momentum Score = $(\text{Price}_{t-\text{skip\_period}})/(\text{Price}_{t-\text{momentum\_period}})-1$. t is the selection date. All candidate stocks are then ranked in descending order by this score, and the top-k stocks (e.g., $k = 5$) are selected.

Volatility Effect. This strategy is based on the "volatility effect" anomaly, where low-volatility stocks have been empirically shown to generate higher risk-adjusted returns [2]. For each candidate stock, we measure its historical volatility over a recent "look-back period" (e.g., 21 trading days). The volatility is calculated as the standard deviation of its weekly log returns within this period. We use weekly returns ($\ln(P_t/P_{t-5})$) rather than daily returns to smooth out daily noise. Candidate stocks are then ranked in ascending order by their calculated volatility, and the top-k stocks with the lowest volatility are selected for the portfolio.

FinCon Selection Agent. Unlike the single-factor methods above, the FinCon Selection Agent [52] aims to construct a **diversified portfolio** by explicitly considering both performance and inter-stock correlation. Its selection process is more sophisticated:

(1) **Metric Calculation:** For all candidate stocks over a "look-back years" period (e.g., 2 years), the agent calculates daily returns to derive a full correlation matrix and a suite of performance metrics for each stock, including the Sharpe ratio.
(2) **Primary Selection:** The agent first ranks each stock using a combined score that balances risk-adjusted return (Sharpe ratio) and its potential for diversification (low average correlation with all other stocks, $\overline{\rho}$). The score is calculated as: Score = Sharpe Ratio $\times (1 - \overline{\rho})$ The top-k stocks based on this score form the initial portfolio.
(3) **Diversification Check & Fallback:** The agent then assesses the average correlation *within* the selected k-stock portfolio. If this internal correlation is above a predefined threshold (e.g., 0.7), it indicates poor diversification. In this case, the agent discards the initial selection and triggers a fallback algorithm. This second algorithm uses a greedy, diversification-first approach: it starts with the single stock with the highest Sharpe ratio and then iteratively adds the available stock that has the lowest average correlation to the already-selected members until a k-stock portfolio is formed.

C Evaluation Metrics

We group evaluation metrics into three categories, each targeting a distinct aspect of strategy performance. In the following definitions, T represents the total number of trading days, and R_t is the portfolio's return on day t.

C.1 Return Metrics

Annualised Return (AR). Measures the geometric average return of the portfolio on a yearly basis. It is calculated from the total cumulative return C as:

$$R_{\text{annual}} = (1 + C)^{\frac{252}{T}} - 1 \tag{1}$$

where 252 is the approximate number of trading days in a year.

Cumulative Return (CR). Measures the total return of the portfolio over the entire test period. It is calculated as:

$$C = \prod_{t=1}^{T} (1 + S_t \cdot R_{m,t}) - 1 \tag{2}$$

where S_t is the position taken by the strategy on day t (+1 for long, 0 for neutral) and $R_{m,t}$ is the market return of the asset on day t.

C.2 Risk Metrics

Annualised Volatility (AV). Measures the standard deviation of the portfolio's returns, scaled to a yearly figure. It is defined as:

$$\sigma_{\text{annual}} = \sigma_{\text{daily}} \times \sqrt{252} \tag{3}$$

where σ_{daily} is the standard deviation of the portfolio's daily returns, R_t.

Maximum Drawdown (MDD). Measures the largest peak-to-trough decline in portfolio value, representing the worst-case loss from a previous high. It is defined as:

$$\text{MDD} = \max_{t \in [1,T]} \left(\frac{P_t - V_t}{P_t} \right) \tag{4}$$

where V_t is the portfolio value on day t, and P_t is the peak portfolio value recorded up to day t ($P_t = \max_{i \in [1,t]} V_i$).

C.3 Risk-adjusted Performance Metrics

Sharpe Ratio (SPR). Measures the excess return of the portfolio per unit of its total volatility. It is calculated as:

$$\text{SPR} = \frac{\overline{R_t} - R_{f,\text{daily}}}{\sigma_{\text{daily}}} \times \sqrt{252} \tag{5}$$

Sortino Ratio (STR). Similar to the Sharpe ratio, but it only penalises for downside volatility, measuring the excess return per unit of downside risk. It is defined as:

$$\text{STR} = \frac{\overline{R_t} - R_{f,\text{daily}}}{\sigma_{\text{downside}}} \times \sqrt{252} \tag{6}$$

where $\overline{R_t}$ is the average daily portfolio return, $R_{f,\text{daily}}$ is the daily risk-free rate (i.e., the annual rate divided by 252), and σ_{downside} is the standard deviation of only the negative daily returns.

D Extra Results on Selective Symbols

Tables 2 and 7 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 [51] 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. [52], 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 BackTrader[9] and Papers With Backtest[10]. Reinforcement learning-based methods are implemented using FinRL [35]. 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 8 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 search, Bayesian tuning) if desired.

[9] https://www.backtrader.com/
[10] https://paperswithbacktest.com/