

- [47] Ruoxu Wu. 2024. Portfolio Performance Based on LLM News Scores and Related Economical Analysis. *SSRN Electronic Journal* (01 2024). doi:10.2139/ssrn.4709617
- [48] Hongjie Xia, Huijie Ao, Long Li, Yu Liu, Sen Liu, Guangnan Ye, and Hongfeng Chai. 2024. CI-STHPAN: Pre-trained Attention Network for Stock Selection with Channel-Independent Spatio-Temporal Hypergraph. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 8 (Mar. 2024), 9187–9195. doi:10.1609/aaai.v38i8.28770
- [49] Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024. TradingAgents: Multi-Agents LLM Financial Trading Framework. *arXiv preprint arXiv:2412.20138* (2024).
- [50] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyang Xiong, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haoqiang 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*. A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 95716–95743. [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/ad1d9fa8be4576d28703b396b82ba1b-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/ad1d9fa8be4576d28703b396b82ba1b-Paper-Datasets_and_Benchmarks_Track.pdf)
- [51] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *FinLLM Symposium at IJCAI 2023* (2023).
- [52] 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. *arXiv:2405.14767 [q-fin.ST]* <https://arxiv.org/abs/2405.14767>
- [53] Yangyang Yu, Haoqiang 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. *arXiv:2311.13743 [q-fin.CP]* <https://arxiv.org/abs/2311.13743>
- [54] Yangyang Yu, Zhiyuan Yao, Haoqiang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhao 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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=dG1HwKMYbC>
- [55] Haoqiang Zhang, Fengrui Hua, Chengjin Xu, Hao Kong, Ruiting Zuo, and Jian Guo. 2024. Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements? *arXiv:2306.14222 [cs.CL]* <https://arxiv.org/abs/2306.14222>
- [56] Wentao Zhang, Lingxuan Zhao, Haoqiang 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. *arXiv:2402.18485 [q-fin.TR]* <https://arxiv.org/abs/2402.18485>
- [57] Jason Zweig. 2019. Where Did This ‘Bull Market’ Come From, Anyway? *The Wall Street Journal* (2019). <https://www.wsj.com/articles/where-did-this-bull-market-come-from-anyway-11564766200>

## 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 list<sup>5</sup>. 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. [16], comprises 15.7 million records pertaining to 4,775 S&P 500 companies, spanning the years 1999 to 2023. We have

<sup>5</sup><https://github.com/fja05680/sp500>

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<sup>6</sup>, 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 Markets<sup>7</sup>, Refinitiv<sup>8</sup>), 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* [53] and *FinRobot* [52]. We acknowledge other works, such as *FinCon* [54] and *MarketSenseAI* [19], 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* [4], *Trend Following* [46], 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—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 [5] and XGBoost [9] in this category and also cover forecasters based on LLMs, differentiating them from the LLM investors.

<sup>6</sup><https://sec-api.io/>

<sup>7</sup><https://alpaca.markets/>

<sup>8</sup><https://www.lseg.com/en>

**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.

- **Environment design.** The trading process is modelled as a Markov Decision Process (MDP), where the agent observes a market state, selects an action, and receives a reward at each step. The environment simulates live stock markets and follows the time-driven simulation principle.
- **State space.** At each timestep, the agent receives a state vector capturing account balance, current shares held for each asset, stock prices (close, open, high, low), trading volumes, and technical indicators (e.g., MACD, RSI). The feature set is configurable and can be aggregated at daily, hourly, or minute frequency.
- **Action space.** The action space allows buying, holding, or selling any subset of assets. Typically, actions are discrete values representing the number of shares to transact per asset (e.g.,  $a \in \{-k, \dots, -1, 0, 1, \dots, k\}$  for each stock), where negative, zero, and positive values denote selling, holding, and buying, respectively.
- **Reward function.** The primary reward is the change in portfolio value after each trading action, i.e.,  $r(s, a, s') = v' - v$ , where  $v$  and  $v'$  are the portfolio values at states  $s$  and  $s'$ .

## 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.

**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 [7]. The score for a given stock is calculated as shown in Equation 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.

$$\text{Momentum Score} = \frac{\text{Price}_{t-\text{skip\_period}}}{\text{Price}_{t-\text{momentum\_period}}} - 1 \quad (1)$$

**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 [3]. 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 [54] 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,  $\bar{\rho}$ ). The score is calculated as:

$$\text{Score} = \text{Sharpe Ratio} \times (1 - \bar{\rho})$$

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.

This two-stage process ensures that the final selection prioritises a balance of individual stock quality and overall portfolio diversification.

## 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:

$$R_{\text{annual}} = (1 + C)^{\frac{252}{T}} - 1 \quad (2)$$

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:

$$C = \prod_{t=1}^T (1 + S_t \cdot R_{m,t}) - 1 \quad (3)$$

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:

$$\sigma_{\text{annual}} = \sigma_{\text{daily}} \times \sqrt{252} \quad (4)$$

where  $\sigma_{\text{daily}}$  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:

$$\text{MDD} = \max_{t \in [1, T]} \left( \frac{P_t - V_t}{P_t} \right) \quad (5)$$

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:

$$\text{SPR} = \frac{\overline{R_t} - R_{f, \text{daily}}}{\sigma_{\text{daily}}} \times \sqrt{252} \quad (6)$$

*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:

$$\text{STR} = \frac{\overline{R_t} - R_{f, \text{daily}}}{\sigma_{\text{downside}}} \times \sqrt{252} \quad (7)$$

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  $\sigma_{\text{downside}}$  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 [53] 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. [54], 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<sup>9</sup> and Papers With Backtest<sup>10</sup>. 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.

## F Comparative Drawdown Analysis via Underwater Plots

This appendix provides a visual analysis of strategy risk profiles through **underwater plots**. An underwater plot visualises the drawdown of a portfolio over time, offering an intuitive way to assess the depth, duration, and frequency of its losses.

The plots are derived by calculating the percentage loss of a portfolio’s equity curve from its running maximum value (its previous peak). At any given point in time, the drawdown  $D_t$  is calculated as:

$$D_t = \frac{\text{Current Value}_t - \text{Previous Peak}_t}{\text{Previous Peak}_t} \quad (8)$$

A value of 0% indicates the portfolio is at a new all-time high, while a negative value shows how far it is “underwater”. When interpreting the plots, two key features should be considered:

- **Depth**: The magnitude of the drawdown, indicated by how low the line drops on the y-axis. Deeper drawdowns represent larger losses and greater risk.

<sup>9</sup><https://www.backtrader.com/>

<sup>10</sup><https://paperswithbacktest.com/>