MASFIN: A Multi-Agent System for Decomposed Financial Reasoning and Forecasting

Marc S. Montalvo*
Computer Science Department
Rochester Institute of Technology
ssm7830@rit.edu

Hamed Yaghoobian
Department of Math, Computer Science & Statistics
Muhlenberg College
hamedyaghoobian@muhlenberg.edu

Abstract

Recent advances in large language models (LLMs) are transforming data-intensive domains, with finance representing a high-stakes environment where transparent and reproducible analysis of heterogeneous signals is essential. Traditional quantitative methods remain vulnerable to survivorship bias, while many AI-driven approaches struggle with signal integration, reproducibility, and computational efficiency. We introduce MASFIN, a modular multi-agent framework that integrates LLMs with structured financial metrics and unstructured news, while embedding explicit bias-mitigation protocols. The system leverages GPT-4.1-nano for reproducability and cost-efficient inference and generates weekly portfolios of 15–30 equities with allocation weights optimized for short-term performance. In an eight-week evaluation, MASFIN delivered a 7.33% cumulative return, outperforming the S&P 500, NASDAQ-100, and Dow Jones benchmarks in six of eight weeks, albeit with higher volatility. These findings demonstrate the promise of bias-aware, generative AI frameworks for financial forecasting and highlight opportunities for modular multi-agent design to advance practical, transparent, and reproducible approaches in quantitative finance.

1 Introduction

Short-term stock prediction is difficult due to volatility, non-stationarity, and the need to integrate quantitative and qualitative signals [Schwert, 1989]. Recent advances in large language models (LLMs) have renewed interest in this task, with multi-agent frameworks offering modularity, interpretability, and task decomposition [Li et al., 2025, Yu et al., 2024, Liu et al., 2022]. Yet, existing systems remain limited: many rely on single-modality inputs, focus on either metrics or sentiment, or depend on proprietary data such as BloombergGPT, reducing reproducibility [Joshi, 2025, Li, 2025]. Sentiment-driven approaches show promise [Mun and Kim, 2025] but often lack safeguards against common pitfalls in financial research. Traditional quantitative models remain prone to survivorship bias (excluding delisted firms) [Brown et al., 1992], hindsight bias (using future information) [Biais and Weber, 2009], and overfitting [Aliferis and Simon, 2024], limiting the transparency and robustness of financial AI.

We introduce MASFIN (**M**ulti-**A**gent **S**ystem for **Fin**ancial Forecasting), a modular five-stage framework designed to address these shortcomings. MASFIN integrates structured financial metrics from Finnhub and market data from Yahoo Finance with unstructured news sentiment, embedding explicit safeguards against survivorship, hindsight [Liu et al., 2022], and overfitting bias. The system is implemented on CrewAI, with agents organized into sequential roles: **Postmortem** (accounting for delisted firms to prevent survivorship bias), **Screening** (selecting candidate firms using contemporaneous

*This research was conducted while the author was an undergraduate student at Muhlenberg College.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

data to prevent hindsight bias), **Analysis** (combining financial ratios with sentiment under feature constraints to reduce overfitting), **Timing** (validating signals with human-in-the-loop oversight), and **Portfolio** (allocating weights under risk-adjusted constraints). By combining generative reasoning with openly available financial data and bias-aware design, MASFIN provides a transparent, reproducible, and low-cost alternative to proprietary pipelines. Our contributions are:

1. **MASFIN Framework**: We present a five-stage, multi-agent pipeline that integrates Finnhub and Yahoo Finance data with news sentiment and statistical insights, while embedding explicit safeguards against survivorship, hindsight, and overfitting bias. The framework incorporates human-in-the-loop (HITL) validation to mitigate hallucinations, ensures reproducibility with open data and code, and is evaluated in live-market conditions against major benchmarks.
2. **Design Principles for Multi-Agent Systems**: Using finance as a high-stakes testbed, we demonstrate how task decomposition, HITL oversight, and bias-aware modular design improve the reliability, interpretability, and affordability of multi-agent generative systems.

By situating finance as a demanding environment, MASFIN contributes a transparent, bias-aware framework that advances the goals of reliability, reproducibility, and robustness in generative AI, with lessons transferable to other high-stakes domains. All implementation details, Python and reproducibility files are available at `github.com/mmontalvo9/MASFIN`. The remainder of this paper is organized as follows: Section 2 details MASFIN's architecture and methodology, Section 3 presents results, Section 4 discusses limitations and avenues for future work, and Section 5 concludes.

2 MASFIN System Design

MASFIN is a multi-agent sequential pipeline for short-term stock forecasting, organized into five crews of 3-5 LLM-based agents. Each contributes to constructing a 15–30 stock portfolio optimized for short-term returns. Except for the Portfolio Crew, all crews include a Summary Agent that synthesizes outputs into a structured report, functioning as the manual handoff in the HITL workflow [Buckley et al., 2021]. Figure 1 illustrates the pipeline, with agents color-coded by crew, a human figure for manual handoffs, and data sources (Finnhub news and Yahoo Finance) annotated for each agent.

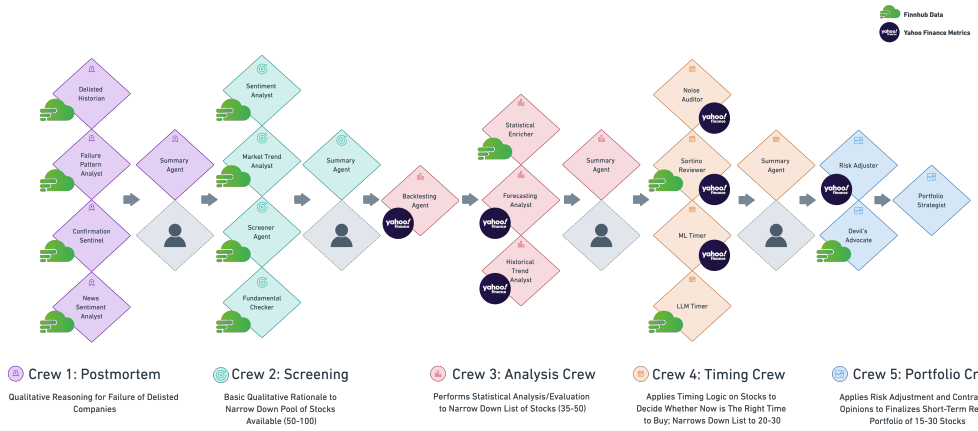


Figure 1: MASFIN's five-stage pipeline with sequential HITL processing.

2.1 Postmortem Crew

The Postmortem Crew, MASFIN's first stage, identifies risks, failure patterns, and sentiment signals from delisted or at-risk firms while mitigating cognitive and data-driven biases [Schropp et al., 2024]. Its inputs are Finnhub headlines from eighteen such companies spanning multiple sectors (Table 1). Agents derive structured features such as historical risk factors, sentiment dynamics, and bias indicators of corporate collapse. By analyzing delisted firms, this stage directly counters survivorship bias, a common omission in financial modeling. A consolidated structured report of failure signals, rationale, mitigation strategies, and references is passed to the Screening Crew as input.

Table 1: Examples of delisted or at-risk firms reviewed by the Postmortem Crew.

Sector	Ticker	Reason	Date
EV / Auto	NKLA, RIDE, ZEV	Bankruptcy or listing failure	2023–2025
Biotech / Pharma	ADMP, SBBP, CNSP, BLUE	Acquisition, merger, or rule violation	2021–2024
Consumer / Retail	BBBYQ, REV, GNLN	Bankruptcy or non-compliance	2022–2024
Cannabis	AGFY, HEXO	Price or listing failure	2023–2024
Tech & Other	FRSX, GPRO, SIEB, HYMC	Price or filing failure	2020–2025

2.2 Screening Crew

The Screening Crew, MASFIN's second stage, filters the market to a shortlist of 50–100 tickers for downstream analysis. Inputs include real-time Finnhub headlines and contextual insights from the Postmortem Crew. Agents perform sentiment evaluation, market trend analysis, and rule-based screening to capture complementary perspectives on candidate stocks [Li, 2025]. A structured, bias-aware shortlist with rationale is passed to the Analysis Crew.

2.3 Analysis Crew

The Analysis Crew, MASFIN's third stage, evaluates tickers from the Screening Crew and prior-week holdings with a quantitative framework to identify 35–50 short-term outperformers while minimizing biases, such as hindsight bias [Biais and Weber, 2009]. Inputs include screened tickers, prior-week survivors, weekly snapshots, and percentage changes. Metrics are derived from Yahoo Finance to ensure consistency, while Finnhub headlines provide qualitative context without introducing real-time data leakage. Agents assess indicators such as multi-horizon returns, volatility, Sharpe and Sortino ratios, maximum drawdown, momentum, beta, alpha, return z-scores, volume trends, and moving-average deviations (Table 2). To ensure bias reduction, all analysis metrics are computed using fixed historical windows and contemporaneous data snapshots, preventing look-ahead bias and ensuring temporal alignment across tickers. The crew outputs a validated shortlist enriched with sectoral context and thematic notes, which is then passed to the Timing Crew.

Table 2: Key MASFIN metrics. The *Global Mean Benchmarking* method contextualizes metrics against the cohort average.

Category	Metrics
Return-based	21D Return, 5D Return, Momentum (21D price change)
Risk / Adjusted	Volatility (annualized std.), Max Drawdown, Sharpe, Sortino, Beta, Alpha
Technical	RSI-14, 5D Return Z-Score, Volume Trend, Price vs. 5D MA
Benchmarking	*Global Mean:* $G_m = \frac{1}{N}\sum_{i=1}^{N} M_{i,m}$, where M is a metric for ticker i

2.4 Timing Crew

The Timing Crew, MASFIN's fourth stage, assesses whether candidate tickers from the Analysis Crew are appropriately timed for near-term entry. Inputs include the validated shortlist, Yahoo Finance metrics, and macroeconomic or firm-level context from Finnhub. Agents rely strictly on historical data to prevent hindsight bias and ensure decisions reflect only information available ex-ante [Biais and Weber, 2009, Levy, 2024]. Using metrics such as Sortino ratio, return z-score, momentum, regression slope, and trading volume (Table 2), agents generate buy, hold, or sell decisions. A unified decision schema flags inconsistencies or elevated risks, yielding a refined list of 20–30 candidates with justified timing for portfolio entry.

2.5 Portfolio Crew

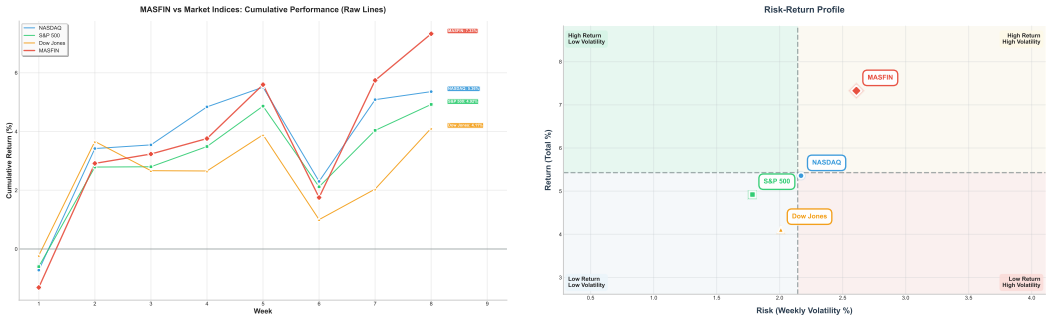
The Portfolio Crew, MASFIN's final stage, consolidates prior outputs into a 15–30 stock portfolio with allocation weights. It resolves conflicts, challenges weak or bias-prone selections, and ensures diversification. Agents use Yahoo Finance for quantitative measures (volatility, Sharpe, Sortino, drawdowns (Table 2)) and Finnhub for qualitative validation via sentiment and macro context. The

portfolio emphasizes systematic bias control, resisting overfitting and hindsight bias while balancing risk, limiting concentration, and grounding each inclusion in both statistical evidence and external signals [Roberts and Henneberry, 2007, Biais and Weber, 2009, Aliferis and Simon, 2024].

3 MASFIN Performance Analysis

We evaluated MASFIN over an eight-week horizon with weekly rebalancing, benchmarking against the S&P 500 (SPY), NASDAQ-100 (QQQ), and Dow Jones Industrial Average (DIA). As shown in Figure 2a, MASFIN achieved a cumulative return of 7.33%, surpassing the NASDAQ (5.36%), S&P 500 (4.92%), and Dow Jones (4.11%). It delivered positive returns in six of eight weeks, a 75% win rate comparable to the NASDAQ and S&P 500, and consistently outperformed in total return.

This outperformance was accompanied by higher risk. As shown in Figure 2b, MASFIN's weekly volatility was 2.61%, the highest among all benchmarks, placing it in the high-risk, high-return quadrant. Nevertheless, the magnitude of gains suggests a favorable risk-adjusted profile. MASFIN's strong correlations with the S&P 500 (0.97) and NASDAQ (0.95) suggests amplified performance within existing market trends.



(a) Cumulative returns over 8 weeks.

(b) Risk–return profile.

Figure 2: MASFIN performance over an 8-week live evaluation.

4 Limitations

End-to-end architectures ingesting all metrics, headlines, and analyses proved impractical, exceeding LLM context limits and reducing interpretability [Hosseini et al., 2025]. Fully automated execution also raised our API costs. To address these constraints, we adopted a human-in-the-loop (HITL) workflow with manual handoffs [Buckley et al., 2021], enabling validation of outputs, mitigation of hallucinations and bias, and adjustment of agent behavior.

Currently, MASFIN produces weekly portfolios and evaluates them with short-term metrics but lacks a learning mechanism and statistical inference tools such as confidence intervals or hypothesis testing [De Prado, 2018]. Restricting evaluation to eight weeks limits overfitting to specific market regimes while ensuring reproducible testing. These trade-offs emphasize interpretability over automation. Future work includes comparing MASFIN with other AI-based financial systems to better contextualize its performance.

5 Conclusion

This paper introduced MASFIN, a modular multi-agent system that integrates generative AI with financial metrics and news for short-term portfolio construction under explicit bias-mitigation protocols. In an eight-week evaluation, MASFIN outperformed benchmarks in six of eight weeks, albeit with higher volatility. Its modular design and HITL workflow balanced interpretability and reproducibility against the challenges of context length, computational cost, and lack of formal validation. These results suggest that bias-aware, modular frameworks can enhance the reliability and accessibility of generative AI in high-stakes domains, with financial forecasting serving as a representative testbed for how such systems may evolve [Joshi, 2025].

Acknowledgments

This research was supported by the Department of Mathematics, Computer Science, and Statistics and by a Summer Research Grant from the Dean of Academic Life at Muhlenberg College.

References

Constantin Aliferis and Gyorgy Simon. Overfitting, underfitting and general model overconfidence and underperformance pitfalls and best practices in machine learning and ai. *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, pages 477–524, 2024.

Bruno Biais and Martin Weber. Hindsight bias, risk perception, and investment performance. *Management Science*, 55(6):1018–1029, 2009.

Stephen J Brown, William Goetzmann, Roger G Ibbotson, and Stephen A Ross. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4):553–580, 1992.

Ross P Buckley, Dirk A Zetzsche, Douglas W Arner, and Brian W Tang. Regulating artificial intelligence in finance: Putting the human in the loop. *Sydney Law Review, The*, 43(1):43–81, 2021.

Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

Peyman Hosseini, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver. Efficient solutions for an intriguing failure of LLMs: Long context window does not mean LLMs can analyze long sequences flawlessly. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1880–1891, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.coling-main.128/`.

Satyadhar Joshi. A comprehensive review of gen ai agents: Applications and frameworks in finance, investments and risk domains. *International Journal of Innovative Science and Research Technology*, pages 1339–1355, 2025.

Bradford Levy. Caution ahead: Numerical reasoning and look-ahead bias in ai models. *Available at SSRN 5082861*, 2024.

Jiayi Li. Research on financial investment strategy optimization with the aid of large language model. *International Journal of Global Economics and Management*, 6(3):119–125, Apr. 2025. doi: 10.62051/ijgem.v6n3.17. URL `https://ijgem.org/index.php/ojs/article/view/43`.

Weixian Waylon Li, Hyeonjun Kim, Mihai Cucuringu, and Tiejun Ma. Can llm-based financial investing strategies outperform the market in long run? *arXiv preprint arXiv:2505.07078*, 2025.

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Wang, Zhaoran Wang, and Jian Guo. Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1835–1849, 2022.

Yejoon Mun and Namhyoung Kim. Leveraging large language models for sentiment analysis and investment strategy development in financial markets. *Journal of Theoretical & Applied Electronic Commerce Research*, 20(2), 2025.

Claire Roberts and John Henneberry. Exploring office investment decision-making in different european contexts. *Journal of Property Investment & Finance*, 25(3):289–305, 2007.

Theresa Constanze Schropp, Melanie Martini, Stephan Kaiser, and Marcus John. Cognitive biases in data-driven decision-making – a literature review. In *Proceedings of the XXXV ISPIM Innovation Conference*, Tallinn, Estonia, June 9–12 2024. ISBN 978-952-65069-6-8. Presented at the XXXV ISPIM Innovation Conference.

G William Schwert. Why does stock market volatility change over time? *The journal of finance*, 44(5): 1115–1153, 1989.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024.