broad market exposure while controlling concentration risk, often outperforming naive buy-and-hold in diversified universes due to the rebalancing premium.

3.4.3. Momentum (3-Month, Top Half)

A classic trend-following strategy that, at each monthly rebalance, ranks tickers by their past 3-month total return and allocates equally to the top half (or all tickers equally if insufficient history). This exploits persistence in intermediate-term price trends and is one of the most robust quantitative factors historically.

3.4.4. Mean Reversion (3-Month, Bottom Half)

The contrarian counterpart to momentum: at each monthly rebalance, the strategy ranks tickers by past 3-month return and allocates equally to the bottom half of performers (or all equally if insufficient history). It bets on price reversals and tends to perform well in range-bound or high-dispersion regimes but can suffer in strong trends.

3.4.5. Moving Average Crossover (50/100-Day, Daily Rebalance)

A trend-following rule that holds a ticker only when its 50-day simple moving average is above its 100-day simple moving average. Positions are rebalanced daily: capital is equally allocated to all tickers meeting the condition (cash otherwise). This classic dual-moving-average system aims to ride trends while avoiding major drawdowns.

3.4.6. Random Noise (Control)

A deliberately naive control strategy that, each day, applies random weights (±40% long, ±40% short, 20% zero per ticker, normalized) to daily ticker returns. This "monkey with a dartboard" baseline helps gauge whether observed performance exceeds what pure randomness could achieve in the same universe.

3.5. LLM-based Trading Agents

We implement the LLM-based trading agents using the open-source AI Hedge Fund framework (Singh, 2025), a popular proof-of-concept system (45.3k stars on GitHub as of January 2026) for exploring AI-driven investment decisions. The repository is available at `https://github.com/virattt/ai-hedge-fund`.

The framework employs a multi-agent architecture where specialized LLM-powered agents—including a Valuation Agent, Sentiment Agent, Fundamentals Agent, Technicals Agent, and others modeled after prominent investors (e.g., Warren Buffett)—analyze market data and generate trading signals. A Risk Manager agent computes position limits based on risk metrics, and a Portfolio Manager agent synthesizes the signals into final allocation decisions. Trades are simulated only (no live execution), making it ideal for research and benchmarking.

Key features relevant to our evaluation include:

- Support for multiple LLM providers via API (OpenAI, Groq, Anthropic, DeepSeek)
- Built-in access to historical price data for popular tickers (AAPL, GOOGL, MSFT, NVDA, TSLA) without requiring an external API key—aligning precisely with our trading universe.

In our experiments, we adapt the framework to a consistent monthly rebalancing protocol (aligned with the quantitative baselines): at the start of each month, the agents are prompted with historical data up to that point, generate target weights for the five stocks, and the weights are normalized and executed (fractional shares allowed, equal initial capital).

We evaluate the following models within this agentic workflow:

Standard Foundation Models Widely used open-source LLMs with potential training data contamination extending into 2023–2024:

- Meta Llama 3.1 (8B and 70B parameters), with cutoff date of December 2023
- DeepSeek 3.2, with cutoff date of July 2024

Point-in-Time (PiT) Models The proprietary Pitinf family from PiT-Inference (PiT-Inference, 2026), designed with effective temporal cutoffs of January 2020 to eliminate lookahead bias:

- Pitinf-Small (<10B parameters) — suited for low-latency tasks.
- Pitinf-Medium (<100B parameters).
- Pitinf-Large (>500B parameters) — for frontier-level reasoning performance.

4. Experimental Results

Table 1 summarizes the performance of quantitative baselines and AI models across both periods.

4.1. Key Insights

Relative to Quant Strategies:

Table 1. Performance Comparison across quantitative baselines and AI models. Note the ”Scaling Paradox” in P2: while Standard models degrade as they scale (due to stronger false priors), PiT models improve as they scale (due to cleaner reasoning).

Model / Strategy	**Variant**	**P1 Return (%)**	**P1 Alpha (pp)**	**P2 Return (%)**	**P2 Alpha (pp)**	**Alpha Decay**
QUANT STRATEGIES (Baseline)						
Buy & Hold	Passive	+25.32	+0.00	+24.75	+0.00	+0.00
Equal Weight	Systematic	+25.68	+0.36	+22.33	-2.42	-2.78
Momentum (3M)	Systematic	+33.28	+7.96	+30.50	+5.75	-2.21
Mean Reversion	Systematic	+20.70	-4.62	+9.35	-15.40	-10.78
MA Crossover	Trend	-2.46	-27.78	+6.91	-17.84	+9.94
Random Noise	Control	+11.43	-13.89	+0.56	-24.19	-10.30
AI MODELS (Agents)						
Llama 3.1 8B	Standard	+39.13	+13.81	+21.33	-3.42	-17.23
Llama 3.1 70B	Standard	+44.59	+19.27	+28.77	+4.02	-15.25
DeepSeek 3.2	Standard	+46.05	+20.73	+23.71	-1.04	-21.77
Pitinf-Small	PiT	+25.07	-0.25	+24.81	+0.06	+0.31
Pitinf-Medium	PiT	+27.76	+2.44	+28.04	+3.29	+0.85
Pitinf-Large	PiT	+31.34	+6.02	+32.07	+7.32	+1.30

- **Momentum Robustness:** The best quantitative strategy (Momentum) generated +7.96pp alpha in P1 and maintained +5.75pp in P2, showing only mild decay (-2.21pp).

- **Mean Reversion Failure:** The worst strategy (Mean Reversion) decayed significantly (-10.78pp), highlighting the risk of overfitting strategies to specific mean-reverting regimes in trending markets.

The Scaling Paradox and Inverse Scaling: Our results highlight a critical divergence in how model scaling affects performance in the presence of lookahead bias. We observe what we term the ”Scaling Paradox”:

- **Standard Models (The ”Memory Trap”):** For standard models, scaling size actually *hurt* P2 performance relative to the peak. While DeepSeek 3.2 achieved the highest P1 Alpha (+20.73pp) due to superior memorization of 2021 data, it suffered the most severe collapse in P2 (-21.77pp decay).

- *Interpretation:* This aligns with the ”Inverse Scaling” phenomenon (McKenzie et al., 2023), where performance on specific tasks degrades as model scale increases. In financial forecasting, larger models develop stronger, more brittle priors based on training data. As noted by Lopez-Lira et al. (2025c), larger models are more adept at recalling exact historical values (perfect memory), but this ”photographic” recall becomes a liability during regime shifts. When deployed in P2 (2024), these strong internal priors conflict with reality, leading to confident hallucinations that override immediate contextual signals.

- **Point-in-Time Models (The ”Reasoning Dividend”):** Conversely, the Pitinf family demonstrates a positive scaling law in the absence of bias. Once the distractor of ”future memory” is removed, scaling the model size improves genuine financial reasoning—such as sentiment interpretation and macro-analysis—rather than merely scaling the capacity to memorize historical price paths.

5. Conclusion and Future Work

This paper is a proof-of-concept of how Look-Ahead-Bench can serve as a diagnostic tool for look-ahead bias in financial LLMs. Standard foundation models, while powerful, suffer from severe alpha decay when moved from memorized training windows to new data. In contrast, Pitinf models demonstrate better generalization.

However, more experimentations would be needed, such as:

5.1. Expand Experimental Scope

The current evaluation is restricted to five large-cap US technology stocks. To ensure robustness, more diverse backtesting settings are needed, with at least 20–30 stocks across multiple sectors, including:

- Small-cap stocks: These are less likely to be memorized in training corpora.

- Non-tech sectors: E.g., Healthcare, Industrials.

- Diverse Assets: Commodities or FX, though LLM applicability there differs.

- Multiple Time Periods: Testing 4+ periods of varying lengths.

5.2. Expand Models and Agents

Look-ahead Bench needs to be enriched with a broader range of trading agents, such as FinMem (Yu et al., 2023), FinGPT (Yang et al., 2023), FinRL-DeepSeek (Benhenda, 2025), TradingAgents (Xiao et al., 2025), and Hedge-Agents (Li et al., 2025a).

Additionally, the scope of autonomous trading architectures should be expanded by integrating specialized agents including FinAgent (Zhang et al., 2024b), FinRobot (Yang et al., 2024), and StockAgent (Zhang et al., 2024a). Furthermore, QuantAgent (Xiong et al., 2025) needs to be incorporated for quantitative investment strategies, while CryptoTrade (Li et al., 2024) should be added to extend coverage to cryptocurrency markets.

To assess fundamental analysis and predictive capabilities, models focusing on factor discovery and market forecasting should be considered. This should include LLMFactor (Cao, 2025) for quantitative factor mining, as well as predictive agents such as TradingGPT (Li et al., 2023), AlphaGPT (Wang et al., 2025), MarketGPT (Wheeler & Varner, 2024), and StockGPT (Mai, 2024). Evaluation frameworks like PIXIU (Xie et al., 2023) and FinPos (Bijia Liu and Ronghao Dang, 2025) should also be utilized for their comprehensive datasets and insights into position sizing.

Finally, to address complex market dynamics involving interaction and high-frequency trading, the benchmark needs to be extended to support multi-agent reinforcement learning environments. This should involve frameworks such as Language Model Guided RL (Darmanin & Vella, 2025). Recent tools like AutoTrader (RexiaAI, 2025) should also be examined to cover the spectrum from research prototypes to deployable trading systems.

5.3. Expand Backtesting Methodologies

The benchmark should aim to move beyond historical price paths by employing advanced validation techniques, including:

- Synthetic Data and Counterfactuals: Testing how models react to market events that could have happened but didn't.

- Rademacher Anti-Serum (RAS): Utilizing methods like those proposed by Paleologo (2025) to rigorously test for backtesting overfitting.

References

Benhenda, M. FinRL-DeepSeek: LLM-infused risk-sensitive reinforcement learning for trading agents, 2025. URL https://arxiv.org/abs/2502.07393.

Bijia Liu and Ronghao Dang. FinPos: A position-aware trading agent system for real financial markets, 2025. URL https://arxiv.org/abs/2510.27251.

Cao, L. Chain-of-alpha: Unleashing the power of large language models for alpha mining in quantitative trading, 2025. URL https://arxiv.org/abs/2508.06312.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *International Conference on Machine Learning (ICML)*, 2022.

Darmanin, A. and Vella, V. Language model guided reinforcement learning in quantitative trading, 2025. URL https://arxiv.org/abs/2508.02366.

Drinkall, F., Rahimikia, E., Pierrehumbert, J. B., and Zohren, S. Time Machine GPT, 2024. URL https://arxiv.org/abs/2404.18543.

Gao, Z., Jiang, W., and Yan, Y. A test of lookahead bias in LLM forecasts, 2025. URL https://arxiv.org/abs/2512.23847.

Glasserman, P. and Lin, C. Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis, 2023. URL https://arxiv.org/abs/2309.17322.

He, S., Lv, L., Manela, A., and Wu, J. Chronologically consistent large language models, 2025. URL https://arxiv.org/abs/2502.21206.

Kong, Y., Hwang, Y., Kaiser, M., Vryonides, C., Oomen, R., and Zohren, S. Fusing narrative semantics for financial volatility forecasting, 2025. URL https://arxiv.org/abs/2510.20699.

Levy, B. Caution Ahead: Numerical Reasoning and Look-ahead Bias in AI Models. *SSRN Electronic Journal*, 2025. URL https://ssrn.com/abstract=5082861.