Can LLM-based Financial Investing Strategies Outperform the Market in Long Run?

Weixian Waylon Li[1] Hyeonjun Kim[2] Mihai Cucuringu[3,4] Tiejun Ma[1]

[1]University of Edinburgh [2]Sungkyunkwan University [3]UCLA [4]University of Oxford

KDD2026 August 9–13, 2026 Jeju, Korea

The Short Answer

Not under robust evaluation.

Previously reported LLM gains largely disappear when backtests become **longer**, **broader**, and **bias-aware**.

Why Current Evidence Can Be Misleading

Short horizons / Hand-picked winners ▸ Survivorship / Look-ahead / Data snooping ▸ Inflated returns / Fragile conclusions

	Typical claim	FINSABER test
Evaluation horizon	~6 months	2004–2024
Asset universe	4–6 winners	63–91 / selector
Constituents	Today's survivors	Historical + delisted
Market conditions	One short regime	Bull / sideways / bear
Costs and risk	Often incomplete	Explicit + risk-adjusted

What FINSABER Contributes

1. **First comprehensive benchmark** for LLM timing-based investors with two decades of multi-source data and explicit bias mitigation.
2. **Fair strategy coverage:** rule-based, predictor, RL, and LLM methods under the same execution and metric pipeline.
3. **Systematic re-evaluation:** prior LLM advantages deteriorate over broader symbols and longer horizons.
4. **Mechanistic diagnosis:** LLM agents miss upside in bull markets and fail to control downside in bear markets.

Scale At a Glance

20 years
2004–2024 rolling backtests

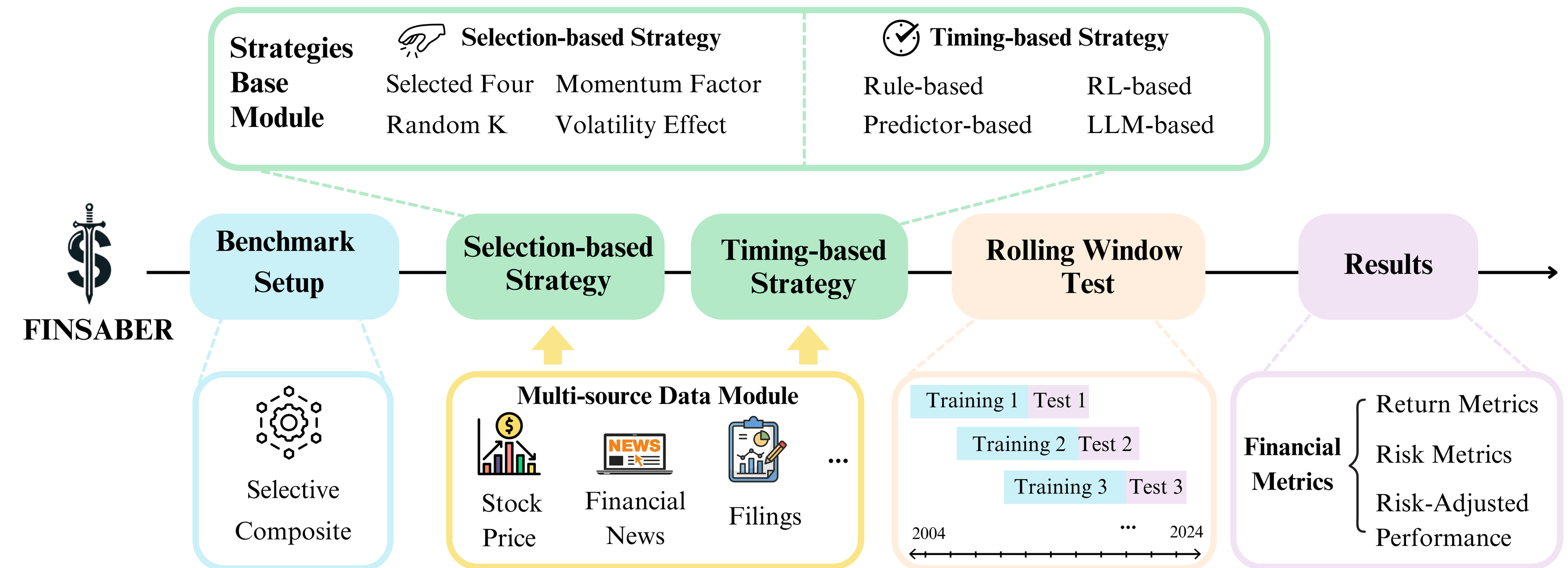
100+
symbols, including delisted stocks

15
timing strategies across 4 families

4
systematic selection configurations

3 BIASES MITIGATED survivorship ▪ look-ahead ▪ data snooping

FINSABER: Bias-aware Evaluation by Design



1. Select assets from a historical universe ▸ 2. Apply timing strategy to each selected asset ▸ 3. Roll through time and aggregate metrics

Coverage: Data, Strategies, and Metrics

MULTI-SOURCE DATA
- Prices and volume
- Financial news
- 10-K and 10-Q filings
- Extensible private sources

STRATEGY BASE
- 7 rule-based
- ARIMA + XGBoost
- A2C / PPO / SAC / TD3
- FinMem + FinAgent

EVALUATION
- Annualised return
- Volatility + drawdown
- Sharpe + Sortino
- Alpha / beta + costs

How Each Bias Is Controlled

Survivorship	Use historically accurate S&P 500 constituents and explicitly include delisted stocks.
Look-ahead	Align every modality to each window; use only information available at the decision point.
Data snooping	Re-select assets across rolling windows and test many symbols over multiple regimes.

Experimental Configurations

Stage	Setup	Scope
1. Reported	4 selected stocks	2022-10 to 2023-04
2. Extended	Same Selected 4	Up to 2004–2024
3. Composite	4 systematic selectors	63–91 symbols / selector
4. Diagnostics	CAPM + regimes + behaviour	Alpha / beta + bull / sideways / bear

Held constant: 15 timing strategies, transaction costs, and return / risk / risk-adjusted metrics.

Finding 1: LLM Advantages Do Not Generalise

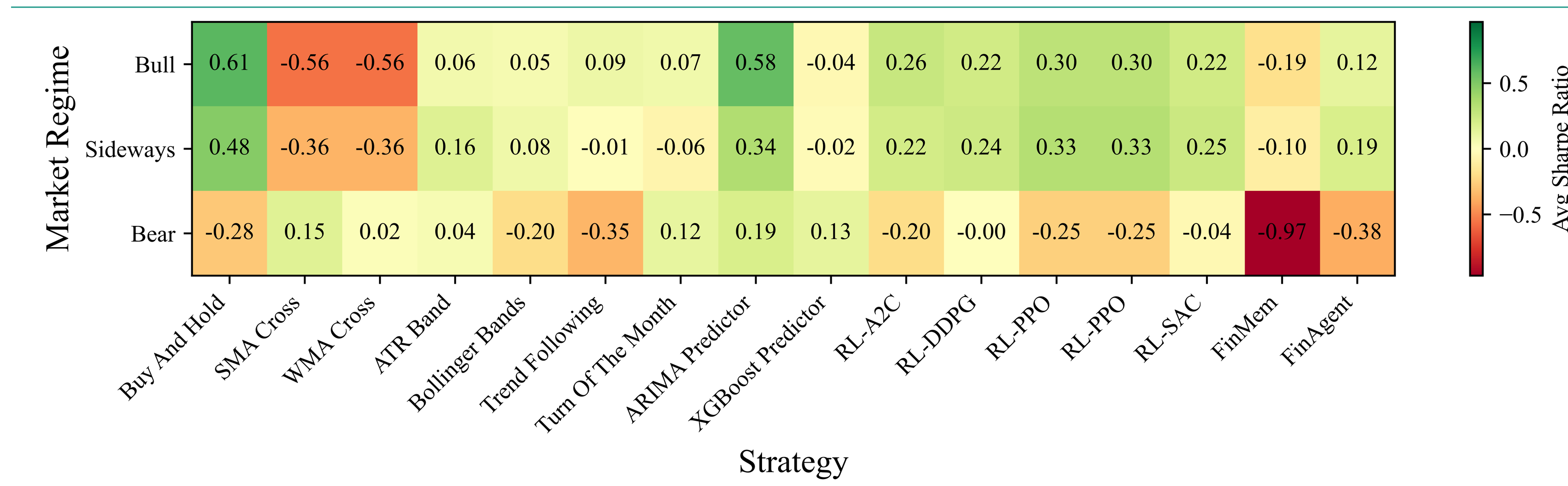
Across configurations: reported gains are inconsistent, fade over the long horizon, and lose their risk-adjusted edge in composite tests. See full tables in the paper.

One quantitative slice: average Sharpe ratio in the 2004–2024 composite backtests.

Timing strategy	Random 5	Momentum	Low-vol	FinCon
Buy & Hold	**0.315**	**0.384**	**0.703**	**0.389**
ARIMA	**0.255**	**0.542**	**0.325**	**0.532**
FinMem	**-0.253**	0.025	**-0.228**	**-0.292**
FinAgent	0.094	0.104	0.241	**-0.076**

Statistical validation: Buy & Hold significantly outperforms both LLM agents in all bias-mitigated setups ($p \leq 0.0117$). Neither LLM agent produces statistically significant alpha ($p > 0.34$).

Finding 2: Agents are Misaligned With Market Regimes



BULL MARKET

B&H: 0.61 FinMem: -0.19
FinAgent: 0.12

Too conservative: agents fail to capture passive equity beta.

BEAR MARKET

B&H: -0.28 FinMem: -0.97
FinAgent: -0.38

Too aggressive: inadequate risk control magnifies losses.

Released Now: FINSABER 2.0

- `pip install finsaber`
- Parquet + custom data
- Explicit `next_open` execution and adjusted OHLC
- Commission, slippage, liquidity caps, and LLM-cost accounting
- Reproducible CSV/JSON run artifacts and manifests

Design priorities for future LLM investors: trend detection and regime-aware risk controls—not merely larger models.

Code Docs Paper Huggingface Data

dl.acm.org/doi/10.1145/3770854.3785702