



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

An Empirical Reassessment of Automated Trading Agents

KDD 2026

Waylon Li, University of Edinburgh

Hyeonjun Kim, Sungkyunkwan University

Mihai Cucuringu, UCLA & University of Oxford

Tiejun Ma, University of Edinburgh



THE UNIVERSITY
of EDINBURGH



UNIVERSITY OF
OXFORD

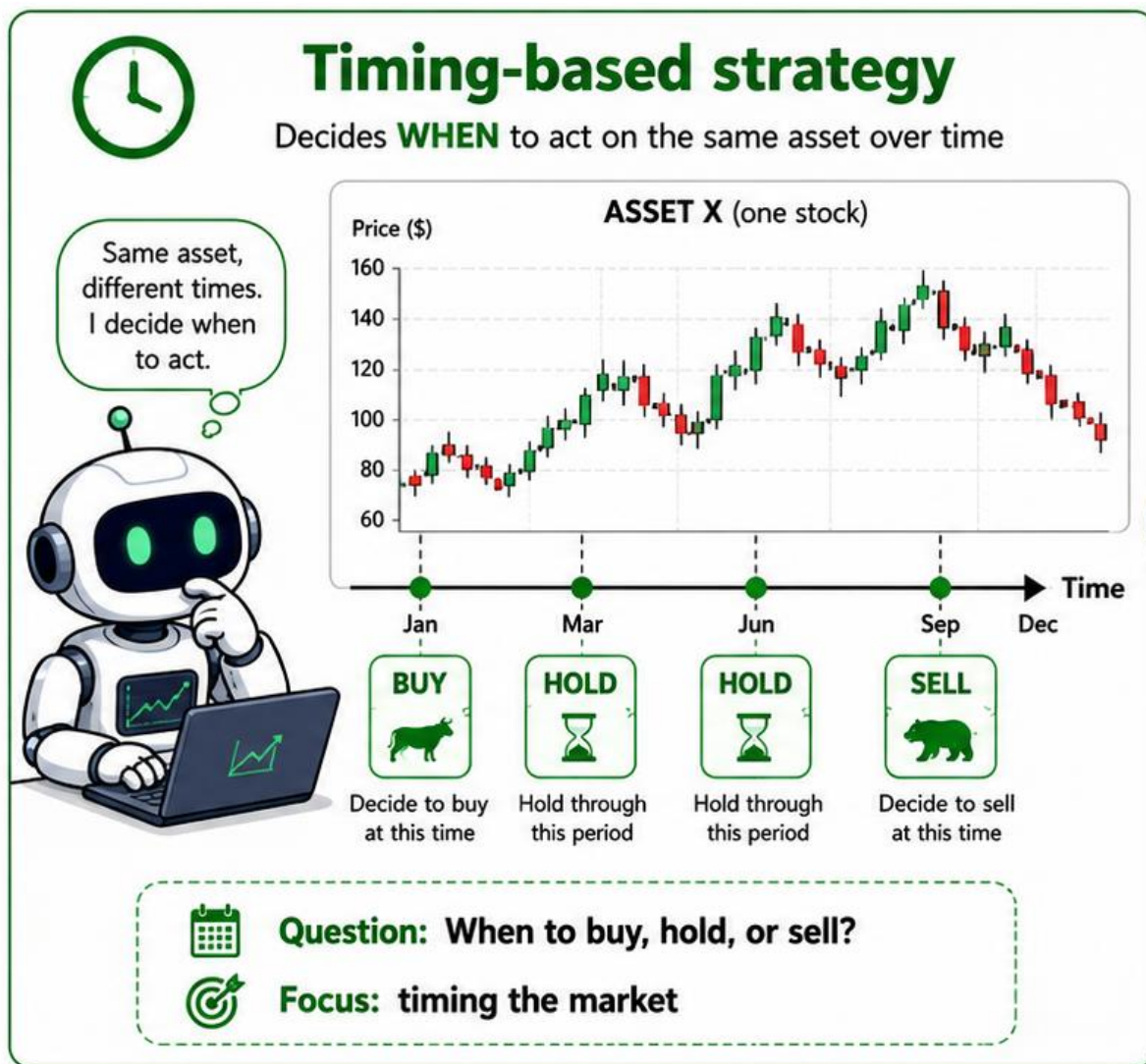
Existing Work



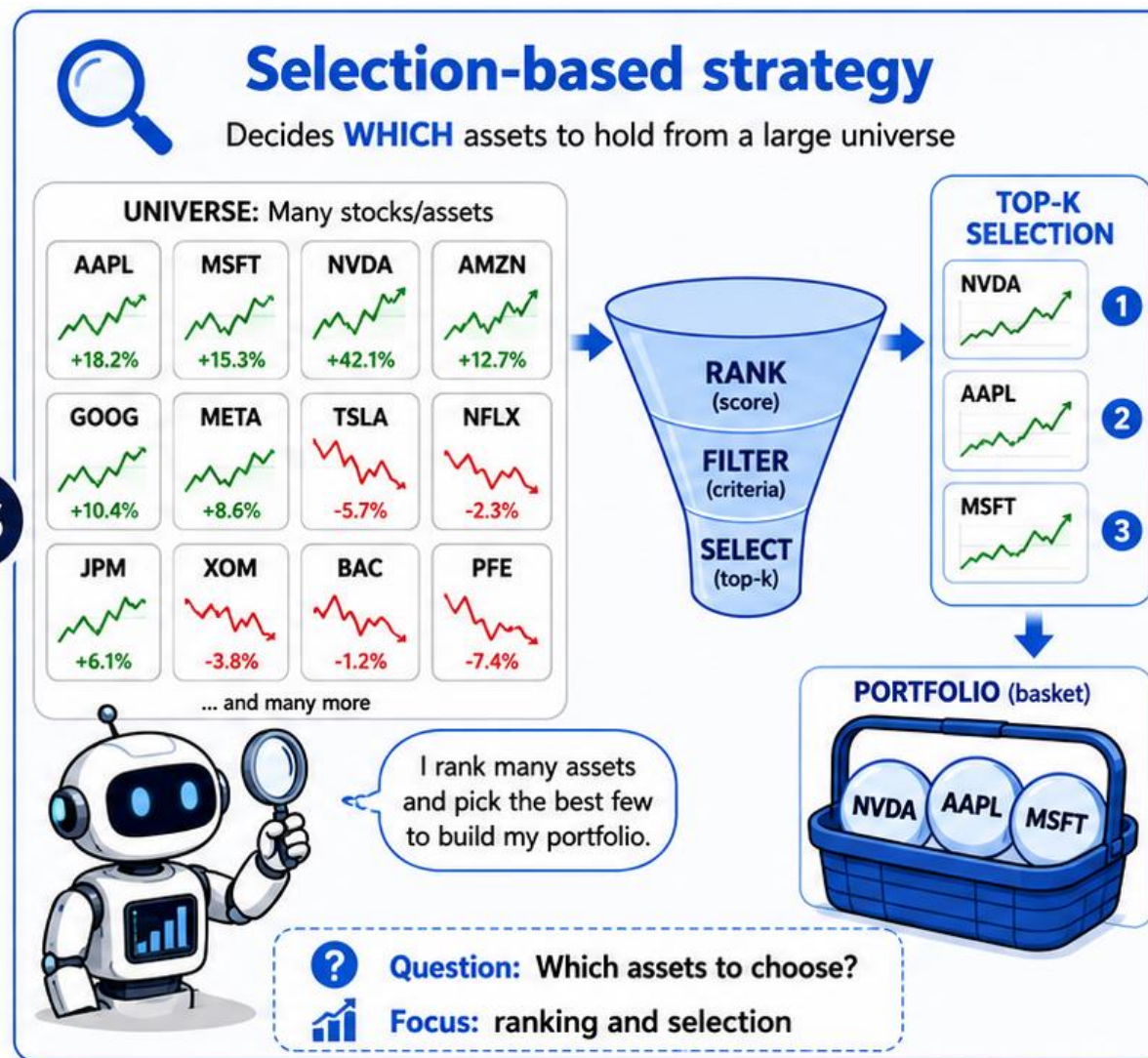
But is it really true?



Definition: Timing-based & Selection-Based Strategies



VS



Timing = when to trade



Selection = what to hold

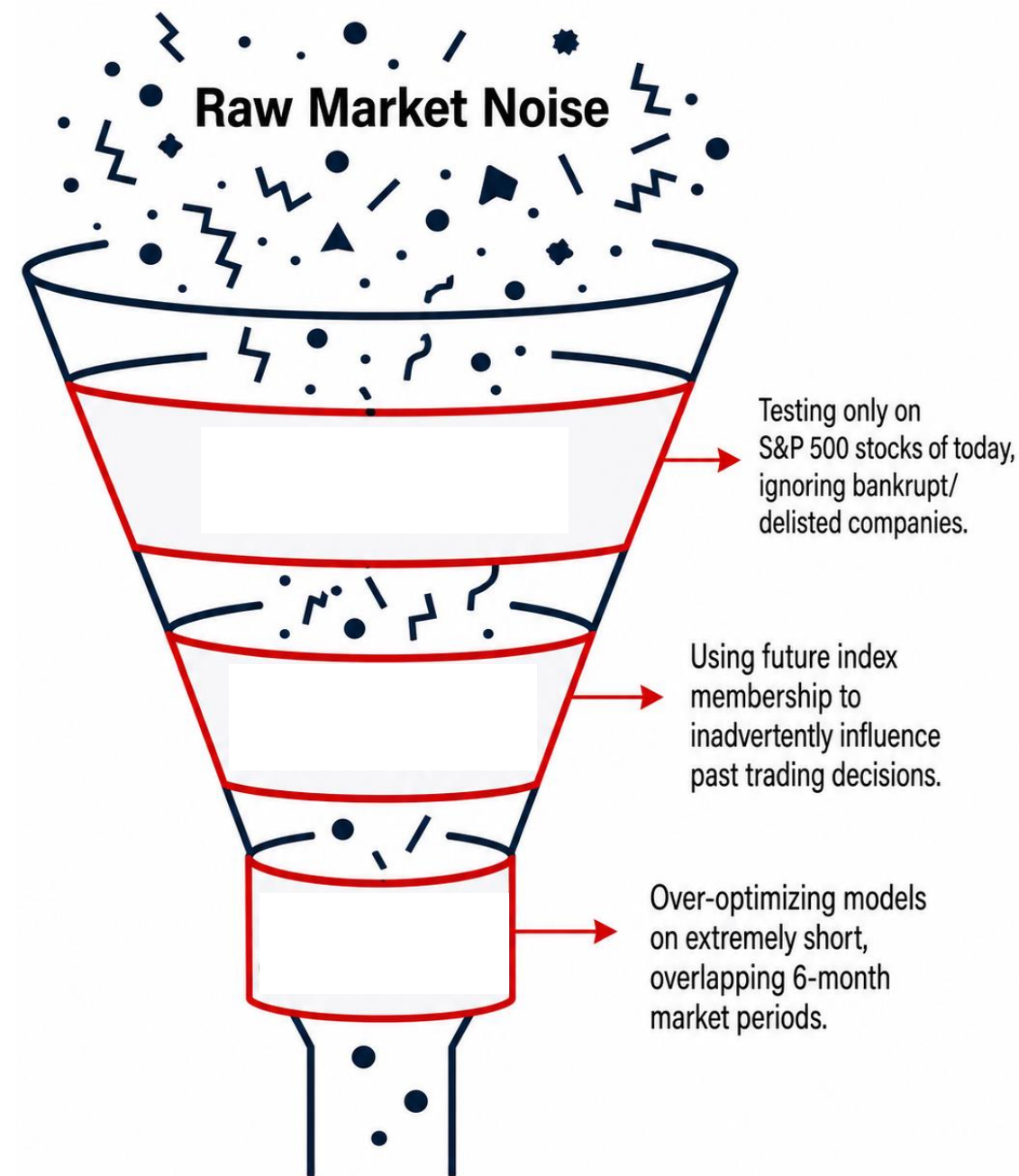
High returns in recent literature are primarily methodological artifacts of narrow evaluation

Method	Eval Period	Eval Symbols	Code
MarketSenseAI	1 year 3 months	100	✗
TradingGPT	N/A	N/A	✗
FinMem	6 months	5	✓
FinAgent	6 months	6	✓
FinRobot	N/A	N/A	✓
TradExpert	1 year	30	✗
FinCon	8 month	8	✗
TradingAgents	3 months	3	✗
MarketSenseAI 2.0	2 years	100	✗

Table 1: Summary of current LLM-based investing strategies.

- Usually < **1 year** of data
- Usually < **10** cherry-picked stocks (e.g. TSLA, AMZN, MSFT etc.)

High returns in these setups might not accurately indicate model skills. It might indicate severe evaluation bias.



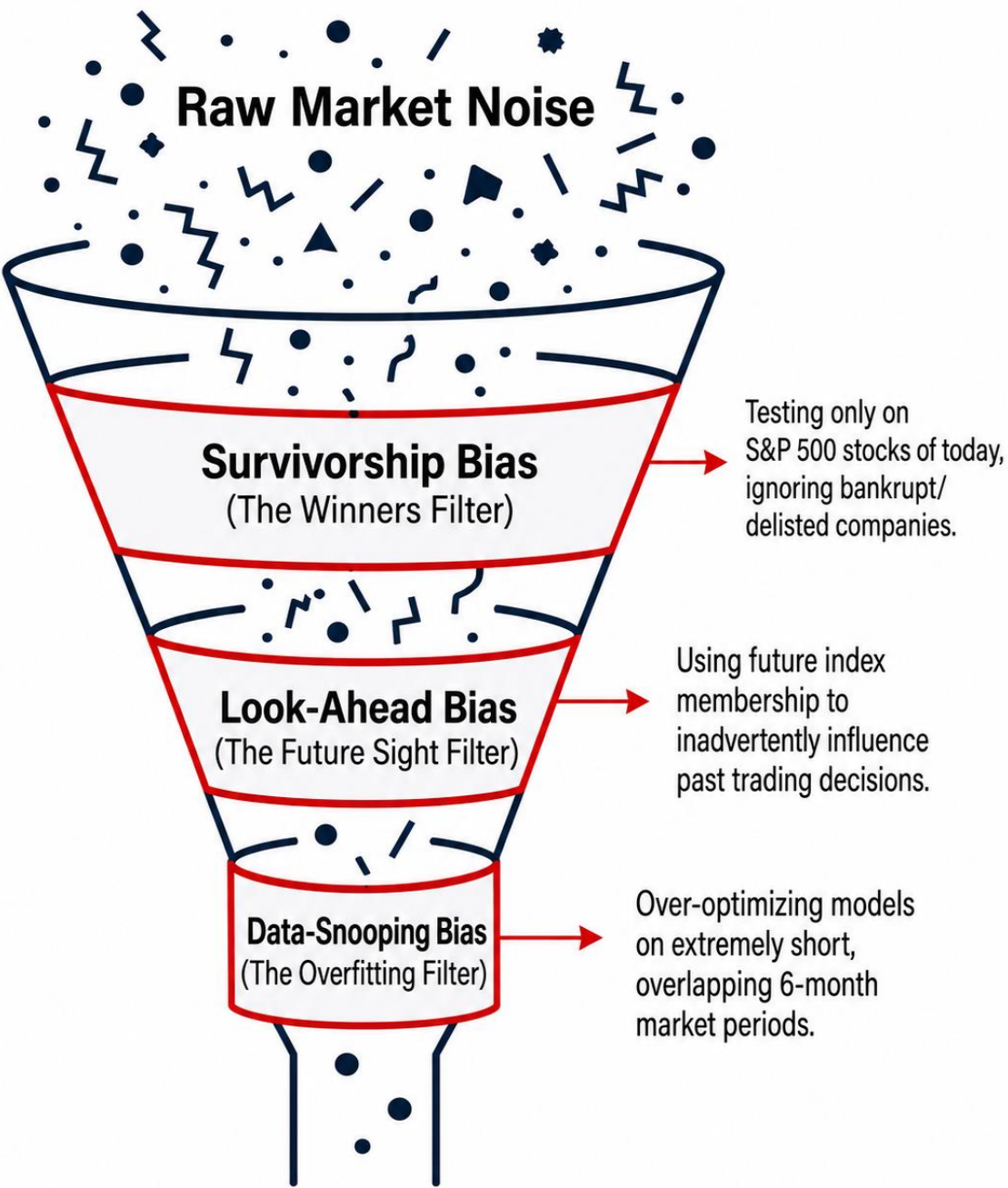
High returns in recent literature are primarily methodological artifacts of narrow evaluation

Method	Eval Period	Eval Symbols	Code
MarketSenseAI	1 year 3 months	100	✗
TradingGPT	N/A	N/A	✗
FinMem	6 months	5	✓
FinAgent	6 months	6	✓
FinRobot	N/A	N/A	✓
TradExpert	1 year	30	✗
FinCon	8 month	8	✗
TradingAgents	3 months	3	✗
MarketSenseAI 2.0	2 years	100	✗

Table 1: Summary of current LLM-based investing strategies.

- Usually < **1 year** of data
- Usually < **10** cherry-picked stocks (e.g. TSLA, AMZN, MSFT etc.)

High returns in these setups might not accurately indicate model skills. It might indicate severe evaluation bias.



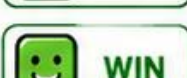
Three Biases Trap in Financial Backtesting

1

Survivorship Bias

Only the winners survive the filter. Losers vanish from the dataset.

All Stocks
(History)



Survivors Only
(Used in Backtest)

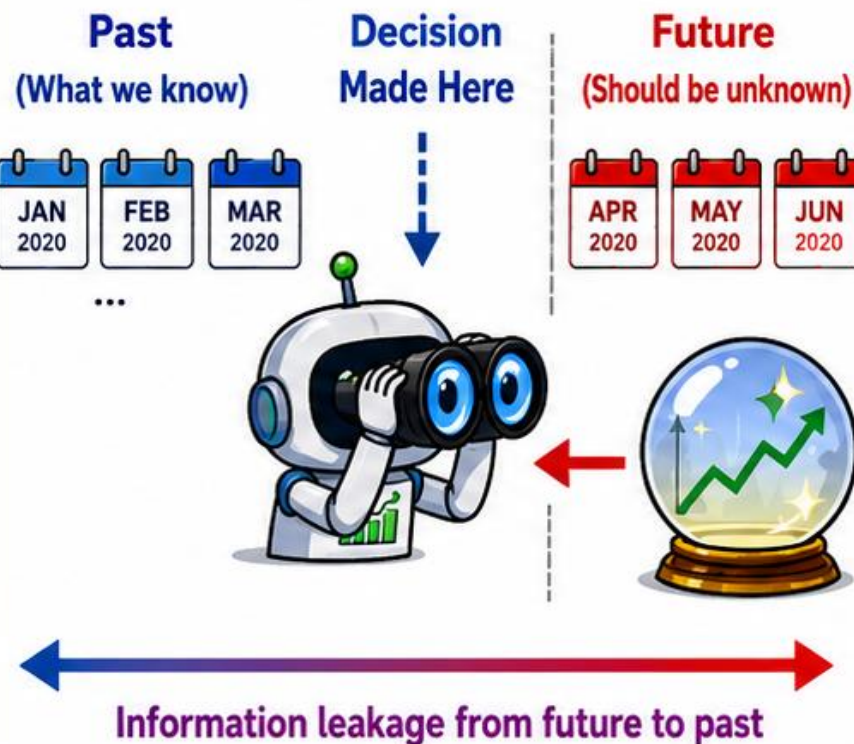


Dropped.
Ignored.
Gone.

2

Look-ahead Bias

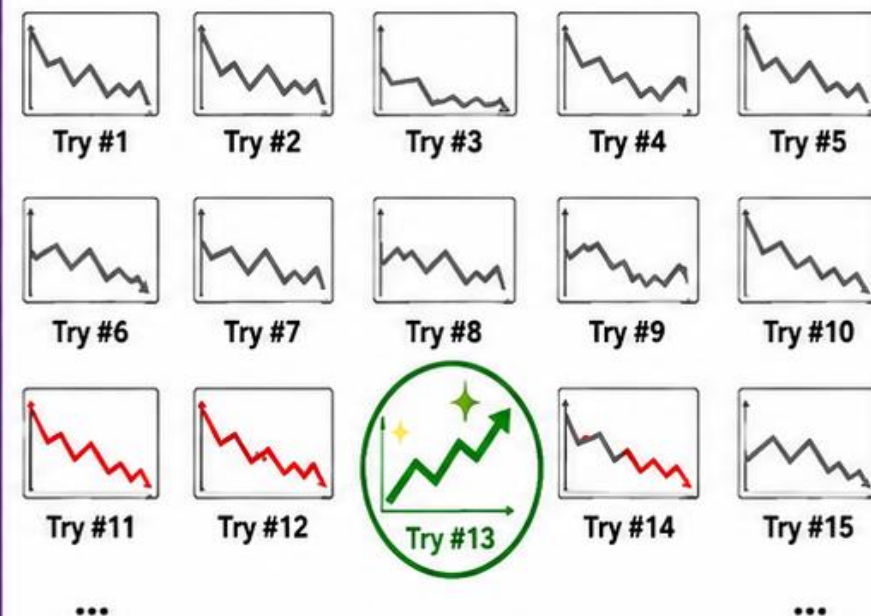
Peeking into the future gives unfair advantages in the past.



3

Data-snooping Bias

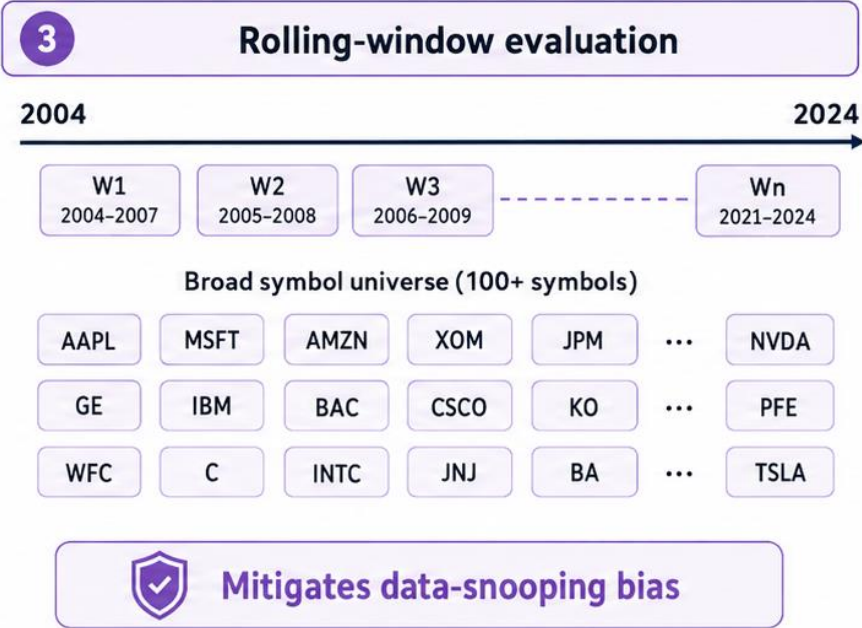
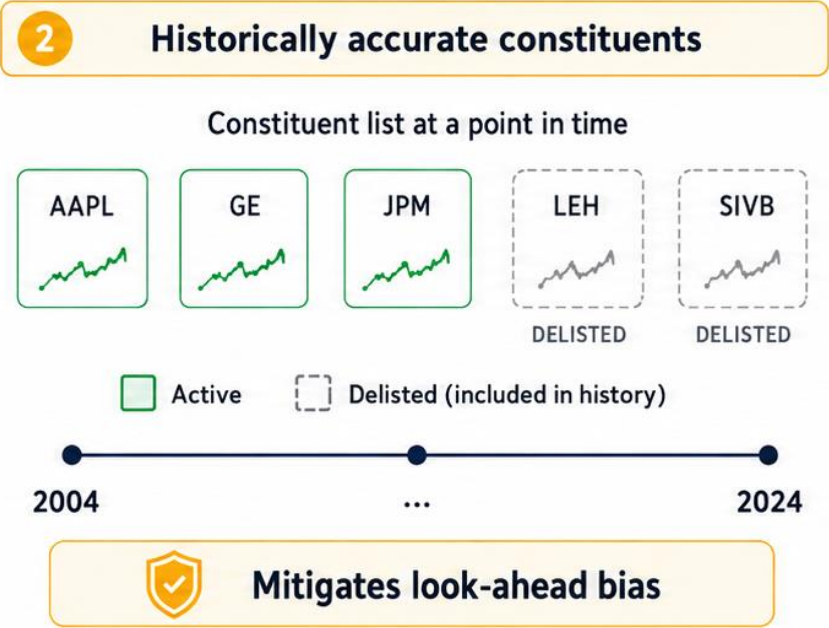
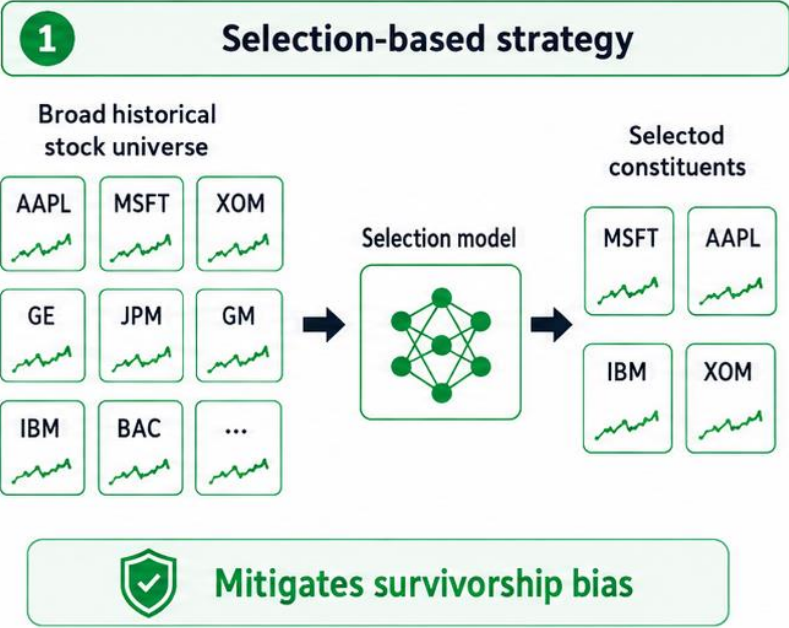
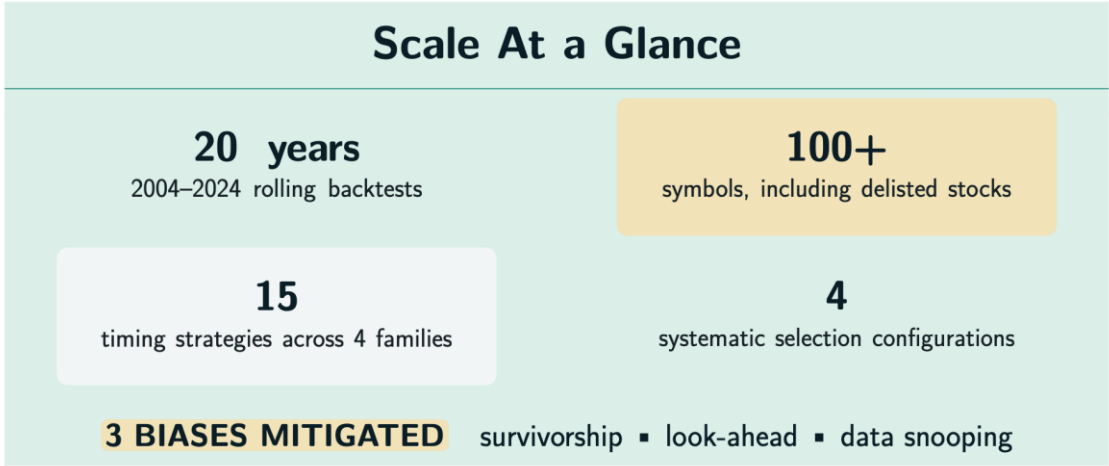
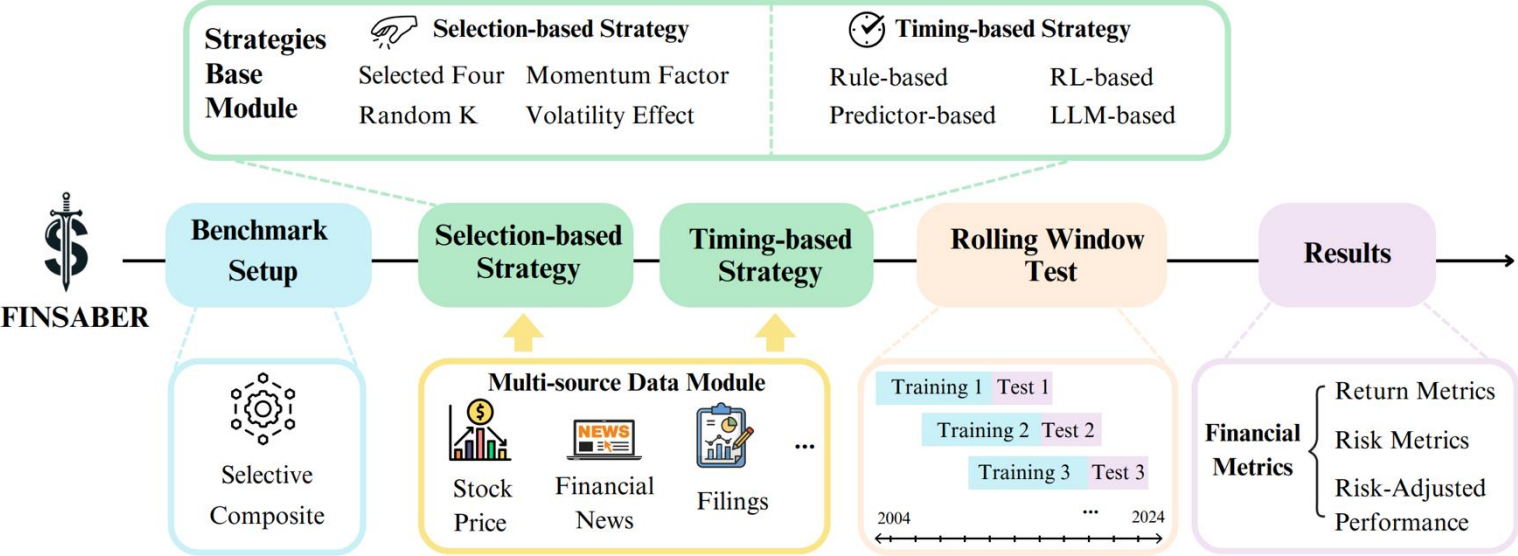
Trying many strategies and settings until a lucky one looks great.



One lucky winner is chosen.
The rest are forgotten.

The Solution: FINSABER

A comprehensive, bias-mitigated evaluation for LLM investors



Key Finding 1: The “Alpha” Vanishes in Robust Settings

Type	Strategy	TSLA					NFLX				
		SPR↑	STR↑	AR↑	MDD↑	AV↓	SPR↑	STR↑	AR↑	MDD↑	AV↓
Rule Based	Buy and Hold	0.630	0.915	37.767	-50.839	45.243	0.622	0.952	23.919	-48.119	41.703
	SMA Cross	0.680	1.013	23.681	-23.707	24.680	0.087	0.160	5.514	-28.689	21.836
	WMA Cross	0.664	0.955	21.158	-25.135	24.087	0.004	0.071	1.447	-32.409	23.074
	ATR Band	0.022	0.066	-0.005	-38.536	26.609	0.186	0.377	2.202	-35.603	23.922
	Bollinger Bands	0.193	0.294	4.282	-37.157	26.267	0.075	0.381	0.286	-34.002	23.088
	Trend Following	0.815	1.356	36.289	-28.113	28.628	0.403	0.646	11.868	-29.179	25.368
	Turn of The Month	0.207	0.353	7.872	-27.902	23.595	0.287	0.487	7.097	-21.646	17.166
Predictor	ARIMA	0.681	1.003	24.138	-30.450	27.612	0.659	1.035	19.022	-27.567	25.514
	XGBoost	0.142	0.370	10.877	-22.901	19.537	0.202	0.355	4.957	-21.301	17.302
RL	A2C	0.172	0.249	3.875	-27.367	22.890	0.171	0.243	4.359	-20.960	16.129
	PPO	0.469	0.663	28.189	-46.810	40.156	0.541	0.814	19.279	-39.615	33.630
	SAC	0.119	0.190	6.654	-11.042	9.902	0.186	0.285	8.397	-9.545	9.216
	TD3	0.417	0.604	23.336	-33.725	30.233	0.291	0.431	10.900	-21.451	19.304
LLM	FinMem	0.641	1.069	42.153	-34.234	35.030	0.293	0.622	12.566	-27.721	26.876
	FinAgent	0.546	0.981	59.835	-36.897	41.579	-0.511	0.459	16.793	-20.903	21.126

Type	Strategy	AMZN					MSFT				
		SPR↑	STR↑	AR↑	MDD↑	AV↓	SPR↑	STR↑	AR↑	MDD↑	AV↓
Rule Based	Buy and Hold	0.551	0.829	15.997	-36.842	30.860	0.461	0.620	11.238	-25.463	21.791
	SMA Cross	0.057	0.205	3.896	-22.096	17.520	-0.263	-0.314	0.192	-17.656	11.840
	WMA Cross	0.175	0.300	5.702	-19.309	17.178	-0.363	-0.437	-1.664	-19.075	11.932
	ATR Band	0.443	0.998	5.452	-19.990	15.130	0.317	0.637	5.725	-11.893	10.885
	Bollinger Bands	0.019	0.125	0.895	-23.757	15.763	-0.054	-0.029	1.578	-16.101	11.931
	Trend Following	0.649	1.111	16.018	-19.120	20.130	0.205	0.321	5.438	-17.515	13.419
	Turn of The Month	-0.029	-0.009	1.534	-20.422	15.728	-0.263	-0.343	-0.177	-14.308	10.438
Predictor	ARIMA	0.339	0.504	7.523	-20.612	19.115	0.304	0.466	8.207	-15.227	13.819
	XGBoost	-0.587	-0.366	1.200	-13.659	11.106	0.171	0.322	5.890	-10.523	10.335
RL	A2C	0.165	0.247	3.925	-14.841	11.654	0.279	0.380	7.478	-13.447	11.933
	PPO	0.505	0.767	13.831	-29.128	24.392	0.344	0.463	8.589	-16.697	14.410
	SAC	0.179	0.257	4.438	-14.093	11.665	0.216	0.288	5.329	-14.866	11.835
	TD3	0.382	0.597	11.738	-21.942	19.149	0.050	0.070	1.405	-9.491	6.648
LLM	FinMem	0.188	0.340	5.695	-28.296	24.786	0.203	0.293	4.567	-19.270	17.891
	FinAgent	0.389	0.622	13.992	-25.082	24.883	0.301	0.513	10.760	-20.877	19.978

Advantage fades with horizon

On the same four symbols extended to 2004–2024, Buy & Hold ranks among the top on NFLX, AMZN and MSFT; LLM investors clearly lead only on TSLA.

Returns come with risk

Even where LLMs win on return, risk is extreme: on TSLA, FinAgent reaches AR 59.8% with 41.6% volatility and -36.9% drawdown; FinMem's Sharpe (0.641) barely surpass B&H (0.630).

Still a biased lens

These are historical bull-market winners — survivorship bias persists and significance is inconsistent across symbols. A broader, unbiased universe is needed.

Takeaway — advantages reported on hand-picked symbols do not survive a two-decade horizon; robust judgement needs an unbiased universe.

Key Finding 1: The “Alpha” Vanishes in Robust Settings

Type	Timing Strategy	RANDOM 5 (91 symbols)					MOMENTUM FACTOR (84 symbols)				
		SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓	SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓
Rule Based	Buy and Hold	0.315	0.456	6.694	-35.130	27.410	0.384	0.694	9.916	-32.596	37.421
	SMA Cross	-0.298	-0.290	0.446	-22.292	15.774	-0.251	0.008	2.109	-19.438	20.050
	WMA Cross	-0.299	-0.305	0.232	-22.754	15.528	-0.169	0.051	3.674	-18.651	20.330
	ATR Band	0.232	0.425	5.119	-21.535	16.113	0.197	0.595	4.314	-19.407	20.038
	Bollinger Bands	0.129	0.288	3.521	-22.487	16.290	0.114	0.702	1.881	-19.451	21.555
	Trend Following	-0.389	-0.198	2.525	-8.587	8.223	0.119	0.531	6.380	-15.726	18.696
	Turn of The Month	0.015	0.072	2.870	-18.582	13.542	0.056	0.662	3.197	-18.108	18.055
Predictor	ARIMA	0.255	0.434	6.928	-21.691	17.504	0.542	1.043	13.257	-18.277	22.892
	XGBoost	-0.055	0.028	3.089	-17.160	13.075	0.094	1.525	6.131	-12.754	17.238
RL	A2C	0.086	0.122	1.902	-9.220	6.887	0.105	0.171	2.488	-14.452	14.815
	PPO	0.179	0.256	3.282	-18.395	13.783	0.185	0.308	1.939	-23.177	25.527
	SAC	0.097	0.142	1.389	-16.058	12.375	0.195	0.321	5.591	-12.235	16.144
	TD3	0.173	0.248	3.682	-14.471	11.565	0.186	0.293	3.464	-14.593	14.953
LLM	FinMem	-0.253	0.114	-0.094	-24.243	21.214	0.025	0.170	3.649	-23.335	28.078
	FinAgent	0.094	0.323	4.477	-28.059	26.387	0.104	0.534	13.950	-20.675	30.635

Type	Timing Strategy	VOLATILITY EFFECT (63 symbols)					FINCON SELECTION AGENT (80 symbols)				
		SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓	SPR ↑	STR ↑	AR ↑	MDD ↑	AV ↓
Rule Based	Buy and Hold	0.703	1.291	7.898	-14.146	14.720	0.389	0.671	6.940	-30.943	41.710
	SMA Cross	-0.568	-0.544	0.781	-9.296	8.665	-0.346	-0.351	-4.187	-21.095	20.765
	WMA Cross	-0.665	-0.348	1.908	-8.481	8.573	-0.176	-0.129	-1.683	-19.432	21.141
	ATR Band	-0.026	0.120	2.798	-8.032	7.951	0.181	0.539	4.469	-18.827	24.820
	Bollinger Bands	-0.077	0.029	2.503	-7.618	7.774	0.116	0.333	7.155	-19.145	27.250
	Trend Following	0.230	0.619	5.503	-8.115	9.297	-0.008	0.189	1.358	-19.500	20.400
	Turn of The Month	-0.156	-0.095	2.881	-6.889	7.233	0.013	0.141	2.020	-15.871	16.862
Predictor	ARIMA	0.325	0.838	4.898	-9.111	9.807	0.532	0.841	10.662	-16.018	19.181
	XGBoost	-0.108	-0.055	2.775	-6.676	7.077	0.116	0.325	8.057	-15.320	18.078
RL	A2C	0.421	0.795	4.620	-4.428	5.149	-0.004	-0.061	0.823	-12.557	11.767
	PPO	0.514	0.972	5.805	-8.757	9.461	0.132	0.147	2.327	-9.744	10.257
	SAC	0.402	0.810	3.527	-4.821	5.030	0.180	0.279	2.661	-11.979	14.210
	TD3	0.269	0.394	4.610	-5.442	5.992	0.130	0.334	0.695	-14.621	21.693
LLM	FinMem	-0.228	0.483	4.061	-10.860	11.641	-0.292	0.135	-1.686	-20.809	24.948
	FinAgent	0.241	0.527	4.954	-10.268	11.502	-0.076	0.381	5.168	-15.563	22.565

Statistically validated

Buy & Hold significantly beats both LLM agents in every bias-mitigated setup (paired t-test, $p \leq 0.0117$).

No genuine alpha

Neither LLM agent produces statistically significant alpha ($p > 0.34$) — consistent with the Efficient Market Hypothesis.

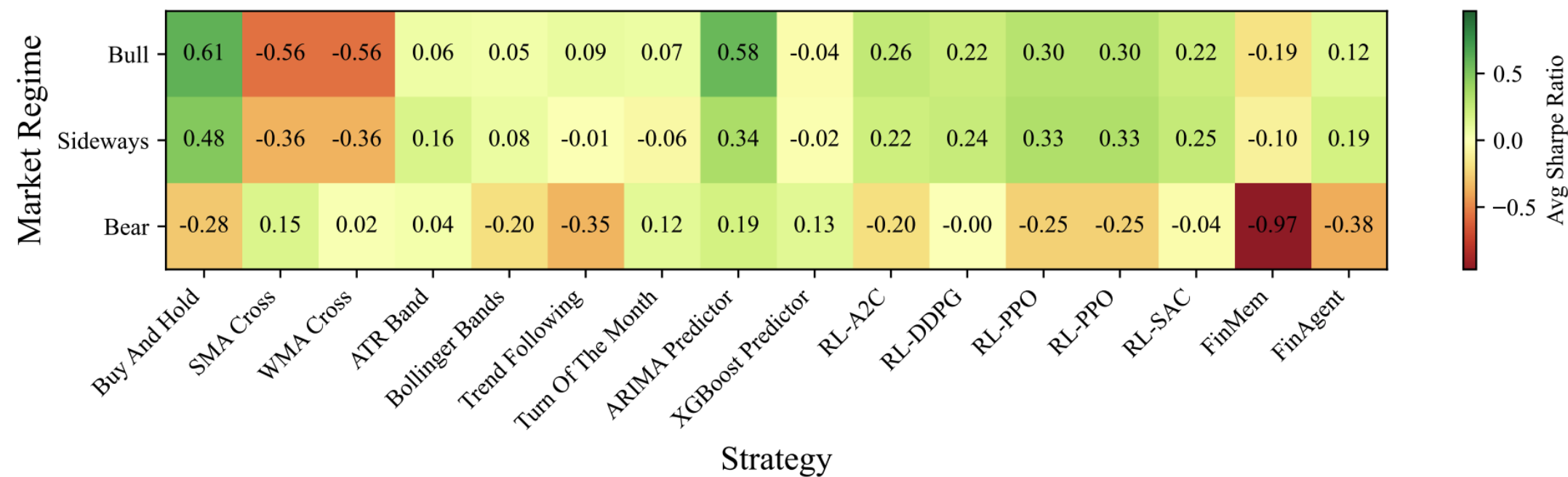
Artefact, not skill

Previously reported gains stem from survivorship and look-ahead biases in narrow, selective setups.

Takeaway — LLM-derived alpha is likely a methodological artefact of biased evaluation, not genuine market skill.

Key Finding 2: Regime Misalignment

Average Sharpe ratio by market regime (Composite setup). A year is bull if S&P 500 return $\geq +20\%$, bear if $\leq -20\%$, else sideways.



Bull — too conservative

B&H 0.61 · FinMem -0.19 · FinAgent 0.12.
No active strategy matches passive beta — LLMs miss the upside.

Bear — too aggressive

B&H -0.28 · FinMem -0.97 · FinAgent -0.38.
Inadequate risk control magnifies downside losses.

Sideways — no edge

Only ATR Band, Turn of the Month and ARIMA stay positive in every regime; FinAgent trails rule-based predictors.

Insight — Agents lack regime awareness: they fail to scale exposure down as volatility and drawdown risk rise.

Conclusion & Future Directions



1. Bias can create fake alpha

- Control survivorship bias
- Control look-ahead bias
- Control data-snooping bias



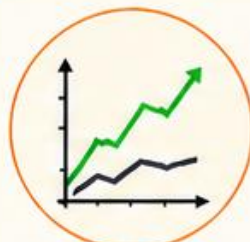
survivorship



look-ahead

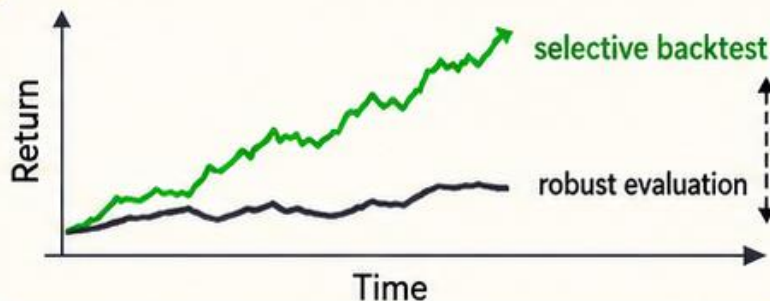


data-snooping



2. Robust tests reduce the hype

- Use long horizons
- Use rolling windows
- Use broad symbol universes



3. Bigger model $\neq$ better investor

- LLM investors do not consistently beat the market
- Better evaluation matters more than bigger claims
- Future focus: regime awareness and risk control

Future directions



Regime-aware trading



Better risk control



Cost-aware and realistic evaluation



Core message: under realistic, bias-mitigated evaluation, apparent outperformance often shrinks or disappears.

Takeaway: robust evaluation is not a detail. It is the result. | Next focus: regime awareness, risk control, and realistic evaluation.

FINSABER has been upgraded to FINSABER V2

The same bias-aware philosophy, re-engineered as an extensible open-source package — bring your own strategies and your own data.



1. Bring your own data

Implement the TradingData interface to plug in private datasets: your own prices, news, filings, or new modalities such as earnings calls and analyst reports.



2. Open data extended to 2026

Curated price, news, and 10-K/10-Q filings now span 2000–2026 in efficient parquet format.



3. Look-ahead safe by design

Explicit next_open execution timing and split-adjusted OHLC.



4. New in v2: realistic frictions

Slippage model, liquidity caps and commission, plus LLM-cost accounting folded into total trading cost.



5. Reproducible run artifacts

Config, manifest, metrics, equity curve, trades and orders exported as CSV/JSON.



Quick start

```
from finsaber import FINSABERbt, FinsaberParquetDataset
from finsaber.strategy.timing import BuyAndHoldStrategy

data = FinsaberParquetDataset("sp500_2000_2026_parquet")
# ...or plug in private data via FinsaberDataset /
# TradingData
config = {
    "data_loader": data,
    "tickers": ["AAPL"],
    "date_from": "2025-01-02",
    "date_to": "2025-01-10",
    "execution_timing": "next_open",
    "slippage_pct": 0.0005,
    "liquidity_cap_pct": 0.025,
}

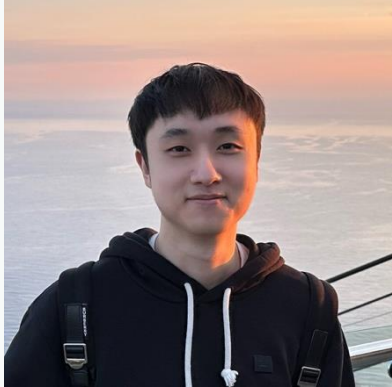
results = FINSABERbt(config).run_iterative_tickers(
    BuyAndHoldStrategy
)
```



FINSABER-1 experiments remain reproducible on the reproduce branch; FINSABER-2 data (2000–2026) is hosted on Hugging Face.

Contributors

The team behind FINSABER — feel free to reach out.



Waylon Li

University of Edinburgh



Hyeonjun Kim

Sungkyunkwan
University



Feiyang Huang

The City University of
Hong Kong



Mihai Cucuringu

UCLA
University of Oxford
Advisor



Tiejun Ma

University of Edinburgh
Advisor

Want to contribute? Issues and PRs are welcome at github.com/waylonli/FINSABER

FINSABER Resources

Scan or click — everything you need to reproduce, reuse and extend the framework.

Code



github.com/waylonli/FINSABER

Dataset

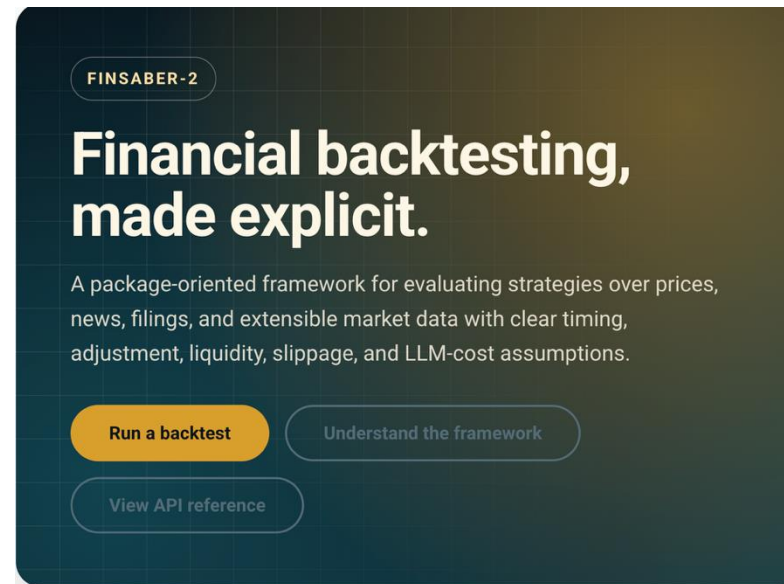


huggingface.co/datasets/finsaber-team/FINSABER-V2-Data

Documentation



waylonli.github.io/FINSABER



TECHNOLOGY • ARTIFICIAL INTELLIGENCE

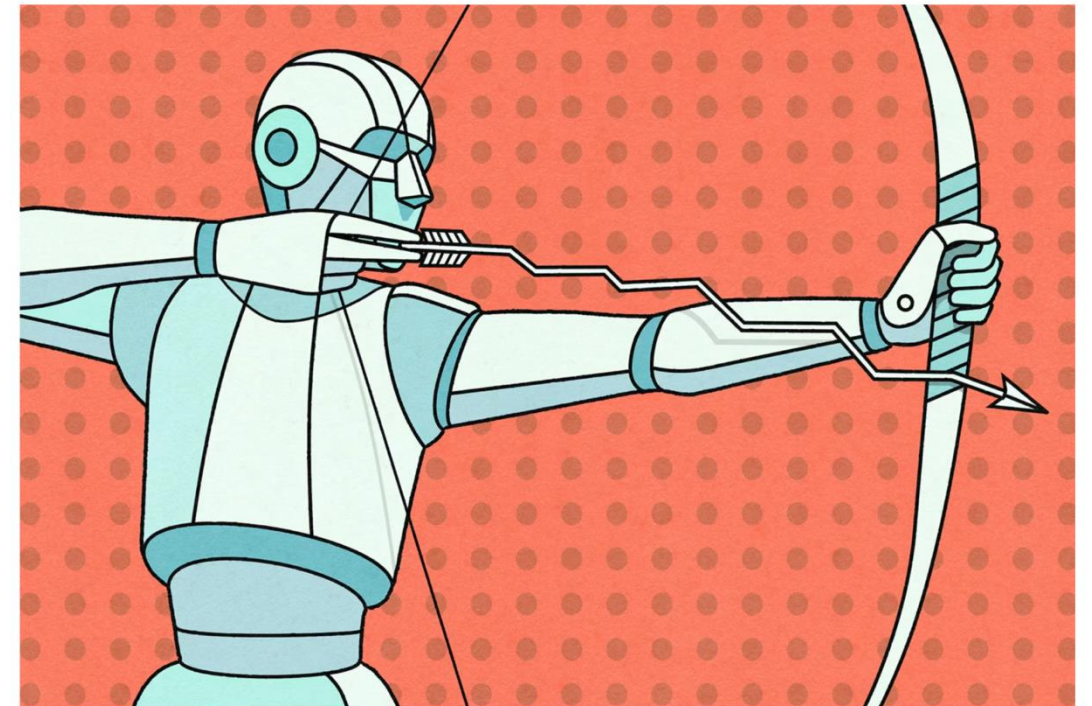
Is AI Good at Stock-Market Timing? A New Study Casts Doubt

Research finds that while large-language models may work well initially, they don't outperform the market over long periods and in changing conditions

By [Andrew Blackman](#)

June 25, 2026 3:00 pm ET

[Share](#) [Aa](#) [Comments](#) 20 | [Listen](#) (1 min) [More](#)



DAN PAGE FOR WSJ

Featured in The Wall Street Journal (Jun 2026).

Follow-up work

Waylon Li, Mengyu Wang, and Tiejun Ma. (2026). Summoning the Oracle to Slay It: Mitigating Look-Ahead Bias in Financial Backtesting with Large Language Models.

<https://arxiv.org/abs/2605.24564>