Are AI Trading Agents Skilful or Lucky?

A Statistical Audit of Published Results for Trading Systems Built on Large Language Models

Navid Abdollahzadeh

Working paper · October 2026

Abstract

Since 2023, dozens of papers have reported that trading agents built on large language models (LLMs) earn high risk-adjusted returns, often with Sharpe ratios above 2. This paper asks whether the published evidence can support those claims. I audit 48 papers posted to arXiv between 2023 and September 2026: the eight most-cited systems and a random sample of 40 from the remaining eligible literature. For each paper I extract the test window, the reported Sharpe ratios, the number of configurations the authors report trying, the baselines, transaction costs, the number of runs, and the LLM's knowledge cutoff. The median out-of-sample test lasts 0.68 years and 69% of tests are shorter than one year. At that length a strategy with a true Sharpe ratio of 1 has only a 19% chance of producing a significant result. Half of the reported Sharpe ratios are no larger than the expected best of seven strategies with zero skill. Correcting for the configurations the authors themselves report trying lowers the share of papers with significant results from 43% to 29%, and this is a lower bound on the correction. Proposed methods beat their best baseline in 75% of comparisons, but by more than sampling error in only 13%. Only 15% of papers report any significance test, 17% report more than one run, 25% account for transaction costs, and 42% evaluate after the LLM's knowledge cutoff. Finally, four landmark systems were re-run by other authors in 11 later papers: in those 63 re-tests, the mean Sharpe ratio fell from 6.73 to -0.03 for TradingAgents, from 1.32 to 0.02 for FinCon, from 1.42 to 0.34 for FinMem and from 1.63 to 0.64 for FinAgent. The evidence is consistent with a literature in which reported performance is dominated by noise and selection. I close with a short reporting standard that would make future claims testable.

Keywords: large language models; LLM trading agents; backtest overfitting; Sharpe ratio; Deflated Sharpe Ratio; statistical power; replication; look-ahead bias; meta-research.

JEL classification: C12, C18, C45, C52, G11, G17.

Extraction records, analysis code and figures: https://github.com/Navid-A-08/LLM-Trading-Audit

1 Introduction

Large language models (LLMs) have moved quickly from reading financial text to making trading decisions. Systems such as FinMem [22], FinAgent [24], FinCon [23] and TradingAgents [20] wrap one or more LLMs in an agent architecture that reads prices, news and filings and then buys, sells or holds. Their papers report impressive numbers. TradingAgents, for example, reports annualized Sharpe ratios of 8.21, 6.39 and 5.60 on three technology stocks; for comparison, a Sharpe ratio above 1 sustained over many years is rare even among the best professional managers.

A reported Sharpe ratio is an estimate, and its precision depends mainly on how long the test lasts. The standard error of an annualized Sharpe ratio estimated from T years of data is roughly $1/\sqrt{T}$ [15]. Three months of data therefore gives a standard error near 2, so a strategy with no skill at all will often show a Sharpe ratio of ± 2 by chance. Researchers also try several models, prompts, assets and settings before reporting one, and the best of several noisy estimates is biased upward [3, 17]. A companion paper studies this *false strategy problem* in machine-learning pipelines [1]. LLM agents add a third problem: a model trained on text written after the start of the test period may already "know" what happened, which inflates backtests in ways that cannot occur in live trading [7, 16].

This paper audits the evidence rather than any single system. I build a sample of 48 papers, extract every number needed to judge statistical support, and ask five questions:

1. Are the test windows long enough to detect the effects being claimed?
2. How many of the reported results survive a correction for the number of configurations tried?
3. Do proposed methods beat their baselines by more than noise?
4. How common are basic safeguards: significance tests, repeated runs, transaction costs and tests after the LLM's knowledge cutoff?
5. When other authors re-run a published system, do they get the same answer?

The answers are mostly negative. Tests are short (median 0.68 years), power is low, a large share of reported Sharpe ratios are within the range expected from luck, and safeguards are rare. The most direct evidence comes from re-tests: the four landmark systems that other papers re-ran lost most or all of their reported performance. I do not claim that LLM trading agents cannot work. The evidence is also not uniform: several recent papers use post-cutoff test windows, multiple seeds, transaction costs and formal tests, and I highlight them. The claim is narrower: *the published record, as a whole, cannot distinguish skill from luck*.

My aim is to criticise evidence, not people. Many of the issues below are standard practice in a young field where tests are expensive (some systems make more than ten LLM calls per trading decision), and several authors acknowledge them openly. Where a paper reports something unusually well, I say so.

2 Related work

Backtest overfitting. Bailey et al. [4] show that the best of many backtests is expected to look good even when no strategy has skill, and López de Prado and Bailey [17] give the expected maximum Sharpe ratio among K unskilled strategies (the False Strategy Theorem). The Probabilistic Sharpe Ratio (PSR) of Bailey and López de Prado [2] measures the probability that a true Sharpe ratio exceeds a benchmark given the sample length; the Deflated Sharpe Ratio (DSR) of Bailey and López de Prado [3] sets that benchmark to the expected maximum under K trials. Harvey et al. [8] apply multiple-testing thresholds to hundreds of published factors, and Hou et al. [9] and Jensen et al. [11] study whether published anomalies replicate.

Power and meta-research. Low statistical power together with selective reporting produces a literature in which many published positive results are false [10]; Button et al. [5] document this in

neuroscience. Funnel plots and the regression test of Egger et al. [6] are standard tools for detecting small-study effects in a body of results.

LLM trading agents and their evaluation. Recent surveys catalogue the field [19]. Li et al. [14] re-test FinMem and FinAgent over longer horizons and more assets and find that their advantages largely disappear. Ye et al. [21] argue in a position paper that reported alpha from LLM trading agents should not be treated as deployable, and list the threats to validity, but do not quantify them across papers. Glasserman and Lin [7] show how look-ahead bias can affect LLM-based return predictions, and Lopez-Lira and Tang [16] restrict their sample to the period after the model's knowledge cutoff for this reason.

Contribution. To my knowledge this is the first paper to (i) extract test lengths, Sharpe ratios and reported trial counts from a defined sample of LLM trading papers, (ii) compute PSR, DSR, power and luck-equivalent trial counts for each, (iii) code knowledge-cutoff overlap systematically, and (iv) assemble the cross-paper re-tests of landmark systems that are scattered across later papers' baseline tables.

3 Data

3.1 Sample construction

I searched the arXiv API for papers posted between 1 January 2023 and 30 September 2026 using six keyword queries that combine *trading*, *portfolio*, *stock* or *investment* with *large language model*, *LLM*, *GPT* or *agent*. This returned 1,349 records. Because keyword search missed several well-known systems, I added every arXiv identifier cited in four recent surveys of the field. A title and abstract screen kept papers that (a) use an LLM to make or drive trading or portfolio decisions and (b) report out-of-sample trading performance. I excluded benchmarks whose main contribution is a dataset, pure prediction-accuracy studies, factor-mining papers that do not report a trading result, surveys and critiques. This left 73 eligible papers.

The eight most-cited systems were included with certainty: FinMem, FinAgent, FinCon, TradingAgents, QuantAgent, FLAG-Trader, CryptoTrade and HedgeAgents. From the remaining 65 papers I drew a simple random sample of 40 with a fixed seed (20261006), giving 48 papers in total. Random sampling means the results describe the eligible literature, not a hand-picked subset. Appendix A lists every paper.

3.2 Extraction

For each paper I read the full text and recorded, from the paper's own tables and text:

- the asset class, data frequency, and start and end of the out-of-sample test window;
- each reported Sharpe ratio of the proposed method, and whether it is annualized;
- the Sharpe ratio of buy-and-hold (or the market index) and of the best baseline in the same table;
- the number of configurations of the proposed method reported on the test period (backbone LLMs, prompts, ablations, hyperparameter settings), used as K below;
- whether transaction costs are deducted, how many runs or seeds are reported, and whether any significance test is reported;
- the backbone LLM(s) and whether the configuration was chosen using test-period results;
- any re-run of another paper's system reported in the baseline tables.

When a paper reports several assets, each asset is a separate unit; paper-level figures average over the units of the headline method. This gives 93 units from 35 papers with a Sharpe ratio and a known test length. Of the remaining 13 papers, nine report no Sharpe ratio at all and four lack a usable test length or report Sharpe ratios in unknown units.

3.3 Knowledge-cutoff coding

For the main backbone LLM I compared the provider's published training-data cutoff with the test window and classified each paper as *none* (the test starts after the cutoff), *partial*, *full* (the whole test window precedes the cutoff), *ambiguous* (the model version is not stated and the plausible versions differ) or *unknown* (no backbone or no test dates are given). Overlap does not prove that a result is contaminated; it means the test cannot rule it out.

4 Methods

Let $\widehat{SR}$ be an annualized Sharpe ratio estimated from T observations at q observations per year (252 for daily stock data, 365 for daily crypto data, 52 weekly, 12 monthly). Write $sr = \widehat{SR}/\sqrt{q}$ for the per-period value.

Probabilistic Sharpe Ratio. Following Bailey and López de Prado [2], under approximately normal returns,

$$\mathrm{PSR}(SR^*) = \Phi\left(\frac{(sr - sr^*)\sqrt{T-1}}{\sqrt{1 + sr^2/2}}\right), \tag{1}$$

the probability that the true Sharpe ratio exceeds SR^*. With $SR^* = 0$ this is a standard one-sided test; I call a result significant if PSR > 0.95. Papers rarely report skewness or kurtosis, so I use the normal case, which favours the papers: fat tails would widen the intervals.

Deflated Sharpe Ratio. If K configurations are evaluated on the same test window and the best is reported, the expected maximum annualized Sharpe ratio among K unskilled configurations is approximately

$$E[\max_K] = \sqrt{q/T}\left((1-\gamma)\Phi^{-1}\left(1 - \tfrac{1}{K}\right) + \gamma\Phi^{-1}\left(1 - \tfrac{1}{Ke}\right)\right), \tag{2}$$

with $\gamma \approx 0.5772$ [17]. The DSR is $\mathrm{PSR}(E[\max_K])$ [3]. I set K to the number of configurations the paper itself reports on the test period. Unreported trials, correlated variants and the choice of assets and dates are not counted, so this K is a *lower bound* on the true number of trials and the DSR an upper bound on the true deflated significance.

Luck-equivalent trials. For each unit I solve $E[\max_K] = \widehat{SR}$ for K. The result is the number of zero-skill strategies one would need to test, over the same window, for the best of them to be expected to match the reported Sharpe ratio. A small luck-equivalent K means a result could easily be produced by ordinary trial and error.

Power. For a strategy with true annualized Sharpe ratio S, the power of a one-sided 5% test over T/q years is approximately $\Phi(S\sqrt{T/q} - 1.645)$. I also compute the test length needed for 80% power at each unit's own reported Sharpe ratio, $((1.645 + 0.842)/\widehat{SR})^2$ years.

Comparison with baselines. The standard error of the difference between two Sharpe ratios depends on the correlation ρ between the strategies [12, 13]. Papers do not report ρ, so I assume $\rho = 0.5$, which is generous for strategies trading the same asset over the same days, and approximate the standard error of the difference by $\sqrt{2(1-\rho)q/T}$.

Small-study effects. I plot reported Sharpe ratios against test length (a funnel plot) and run the regression test of Egger et al. [6]: regress $\widehat{SR}/\mathrm{SE}$ on $1/\mathrm{SE}$, with $\mathrm{SE} = 1/\sqrt{\text{years}}$. A positive intercept indicates that short, imprecise tests report systematically larger effects, the pattern expected under selective reporting.

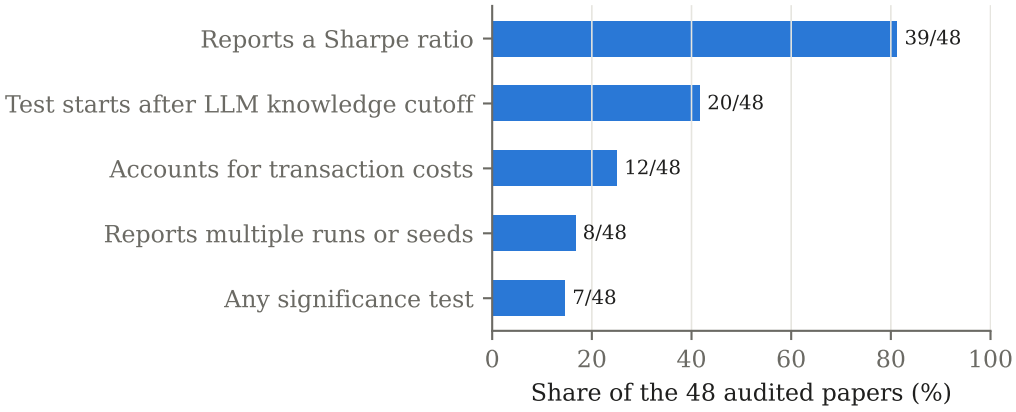


Figure 1: Share of the 48 audited papers that adopt each safeguard. "Accounts for transaction costs" includes papers that deduct costs or report a cost sensitivity analysis. "Any significance test" counts tests on trading performance (for example factor-model alphas, bootstrap tests of Sharpe-ratio differences, or Wilcoxon tests), not tests of prediction accuracy.

5 Results

5.1 Reporting practices

Figure 1 summarises five basic safeguards. Most papers report a Sharpe ratio (39 of 48), but only 7 report any significance test for their trading results, 8 report more than one run or random seed, and 12 deduct transaction costs or report how results change with them. Only 20 of 48 evaluate entirely after the knowledge cutoff of their backbone LLM; in 19 the test window overlaps the cutoff fully or partly, and in 9 the overlap cannot be determined. Two papers choose their reported configuration by comparing results *on the test period*, which turns the test set into a validation set; FinMem, for example, states that its reported results use whichever of three risk settings achieved the highest cumulative return in the testing phase. Several others show a "representative" configuration without saying how it was chosen. Practices have improved: the 2026 papers in the sample account for five of the seven significance tests and seven of the twelve papers that account for transaction costs.

5.2 Test length and power

Out-of-sample tests are short. Across the 35 papers with usable data, the median test lasts 0.68 years (interquartile range 0.48 to 1.48), and 24 of 35 (69%) last less than one year. Several headline results come from one to three months: TradingAgents uses 61 trading days, Janus-Q 57, Trading-R1 65 and Fin-Analyst's live result 33.

Figure 2 shows what this means for power. At the median test length, a strategy with a true Sharpe ratio of 1 has a 19% chance of producing a significant result, and a strategy with a true Sharpe ratio of 0.5 has a 10% chance. Even a true Sharpe ratio of 2, an exceptional level, would be detected less than half the time. For 80% of the units, the test is shorter than the length needed to have 80% power at the Sharpe ratio the paper itself reports; the median required length is 2.1 years.

Low power matters even for positive results. When power is low, the results that do clear a significance threshold are disproportionately those with large positive errors, so their reported effects are inflated [10, 5].

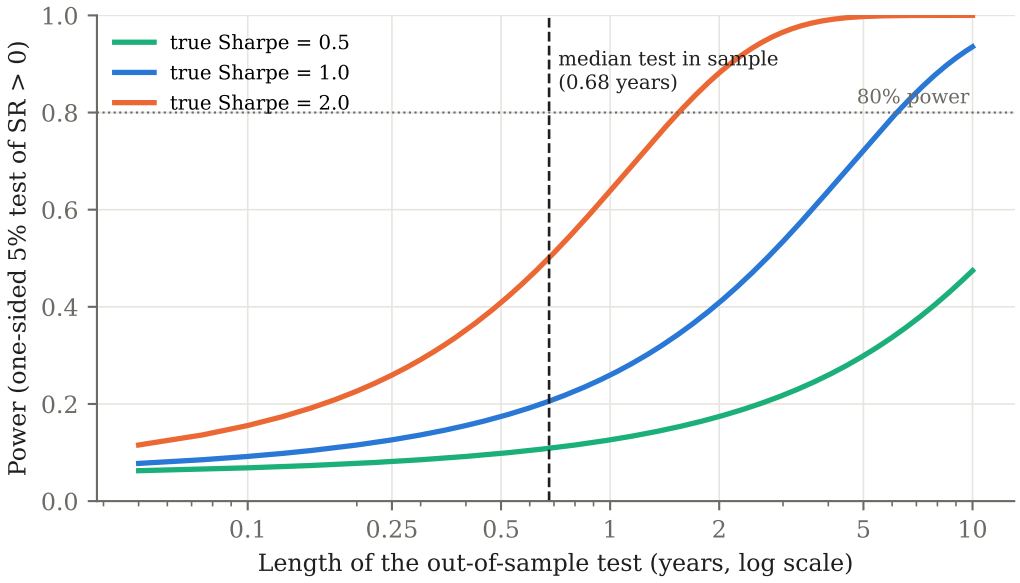


Figure 2: Power of a one-sided 5% test that the Sharpe ratio is positive, as a function of test length, for three true Sharpe ratios. The dashed line marks the median test length in the audited sample.

Table 1: Papers with the largest drop from PSR to DSR. K is the number of configurations the paper reports on the test period (a lower bound). Sharpe ratios are annualized averages over the headline units.

Paper	Sharpe	Years	K	PSR	DSR
Strat-LLM	2.19	0.33	36	0.89	0.19
FinMem	1.42	0.50	15	0.80	0.26
FinRankGRPO	0.64	6.04	24	0.94	0.34
MASS (2023 test)	1.73	0.96	10	0.93	0.55
Janus-Q	1.31	0.23	9	0.73	0.19
Regime-aware RAG	1.19	1.48	12	0.91	0.40
Black-Litterman with LLM views	1.23	0.79	4	0.86	0.52
Trading-R1	1.59	0.26	3	0.78	0.48

5.3 How much could be luck?

The median headline Sharpe ratio is 1.73. Taken one at a time and before any correction, 15 of 35 papers (43%) and 36 of 93 units (39%) are significant at the 5% level. Correcting for the number of configurations each paper reports trying lowers this to 10 of 35 papers (29%) and 24 of 93 units (26%). The median reported K is only 4, and ten papers report a single configuration, so this correction is mild; the true number of trials, including unreported ones, is almost certainly larger.

The luck-equivalent trial count makes the same point more directly. For half of the positive units, the reported Sharpe ratio is no larger than the expected best of **7** zero-skill strategies tested over the same window; 49 of 88 need 10 or fewer and 70 of 88 need 100 or fewer. A research process that tries a handful of models and prompts on a few assets can easily reach these numbers.

Table 1 shows the papers where the correction matters most. Strat-LLM evaluates 36 configurations (12 model variants in three autonomy modes) on a four-month window; its best Sharpe ratio of 2.19 has a PSR of 0.89 but a DSR of 0.19. FinMem reports 15 configurations and selects among them on the test period; its DSR is 0.26.

Some papers make the selection problem visible in their own sensitivity analyses. In one paper on LLM-driven strategy discovery, small changes in two loss weights or in network hyperparameters move

the test-period Sharpe ratio between -8.0 and $+13.3$. The reported configuration is one point in that range. In such a setting the reported number says more about which configuration was chosen than about the method.

5.4 Do proposed methods beat their baselines?

Papers almost always report that their method wins. The proposed method has a higher Sharpe ratio than buy-and-hold in 64 of 73 units (88%) and than the best baseline in its own table in 59 of 79 (75%). The margins are small relative to sampling error, however. With $\rho = 0.5$, the difference from the best baseline exceeds 1.645 standard errors in only 10 of 79 units (13%). Several margins are far inside the noise: AlphaCrafter beats its closest competitor by 0.01–0.03, the best of eleven MarketSenseAI portfolios beats the S&P 100 index by 0.04, and the regime-aware RAG method ties or trails the best single expert in all three of its markets.

Some papers report losses to simple benchmarks clearly, which is to their credit. In Fin-Analyst's ten-month offline Bitcoin backtest, a random strategy (Sharpe 1.13) and buy-and-hold (0.96) both beat the agent (0.73). In the Black-Litterman study, two of four LLMs underperform an equal-weight portfolio. FS-ReasoningAgent and CryptoTrade do not beat buy-and-hold in Bitcoin bull markets.

5.5 Small-study effects

Figure 3 plots each paper's headline Sharpe ratio against the length of its test. The shaded region is the 90% range of Sharpe ratios that a strategy with zero true skill would produce in a single test of that length. Many points lie inside or near it, and the most extreme values come from short tests: TradingAgents (6.73 over 0.24 years) and Fin-Analyst (4.10 over 0.13 years). The Egger regression gives an intercept of 0.84 (standard error 0.42, $t = 2.01$, $n = 35$), weak evidence that shorter tests report larger effects. The rank correlation between Sharpe ratio and test length is close to zero ($\rho = -0.06$, $p = 0.71$). With 35 papers this test has limited power, so I treat the evidence for small-study effects as suggestive rather than established.

5.6 Knowledge-cutoff overlap

Of the 48 papers, 20 test entirely after the LLM's knowledge cutoff, 9 overlap partly, 10 test entirely before it, 2 do not identify the model version and 7 do not give enough information. Examples of full overlap include a 2019 test of a system built on Llama 3.3 (trained to December 2023), FinMem's October 2022 to April 2023 test with GPT-4-Turbo (cutoff April 2023), and Trading-R1, whose June to August 2024 test window lies inside its own training-data span of January 2024 to May 2025. Among papers with usable Sharpe ratios, the median is 1.69 with no overlap, 1.92 with partial overlap and 1.66 with full overlap. I find no detectable difference, but with 6 to 18 papers per group this comparison has little power, and contamination may act through which assets and periods are chosen rather than through the level of the Sharpe ratio.

5.7 What happens when others re-run the systems

The strongest evidence comes from re-tests. Later papers often include earlier systems as baselines, re-running them on their own data. I collected every such re-run of the landmark systems in the sample: 63 asset-level results from 11 papers. Figure 4 and Table 2 compare them with the original papers.

All four systems lose most of their reported performance. Every one of TradingAgents' 11 re-test results is below the lowest Sharpe ratio in its own paper; on Chinese stocks in 2023 its Sharpe ratios are -1.71, -3.02 and -2.82. FinCon re-ran FinMem on an overlapping period and got -1.25 on MSFT, where FinMem's own paper reports $+1.44$. These re-tests are not exact replications: they use other

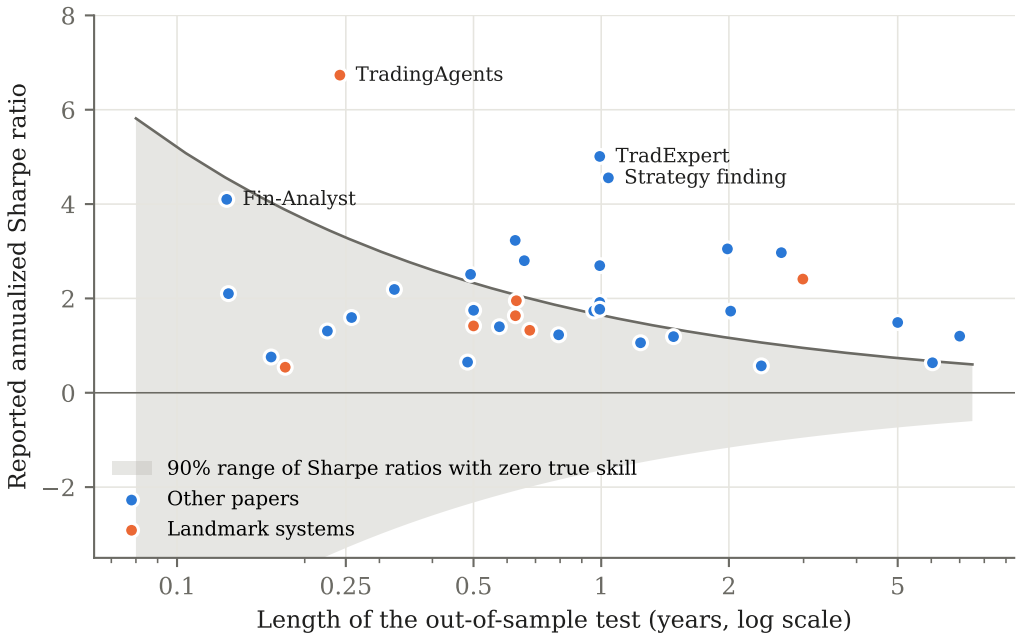


Figure 3: Reported annualized Sharpe ratio against test length, one point per paper. The shaded band is the 90% range of the Sharpe ratio of a strategy with zero true skill in a single test; it ignores selection across configurations, which would raise the upper edge.

Table 2: Original and re-tested Sharpe ratios of four landmark systems.

System	Original mean	Re-test mean	Re-test median	Re-testing papers	Re-test results
TradingAgents	6.73	−0.03	0.66	4	11
FinCon	1.32	0.02	0.38	4	13
FinMem	1.42	0.34	0.54	6	27
FinAgent	1.63	0.64	0.73	3	12

periods, markets and sometimes other LLMs, and some authors may have implemented baselines less carefully than their own method. Yet that is exactly the situation a user of these systems would face, and the direction of the difference is the same in all four cases.

6 Discussion

6.1 Interpretation

The results fit together. Short tests give imprecise Sharpe ratios. Researchers try several configurations and report the best, which selects positive errors. Few papers report significance tests or repeated runs, so readers cannot see the noise. And when other teams re-run the systems on new data, much of the performance disappears. This is the pattern expected if a substantial part of the reported performance is noise and selection.

Two alternative explanations deserve mention. First, markets change, so a system might work in its original period and fail later. That is possible, but it implies that the original result does not tell users what to expect, which is the practical point. Second, re-implementations by competing authors may be weaker than the originals. This would bias re-tests downward, but it cannot explain why the original results are so often statistically indistinguishable from luck to begin with.

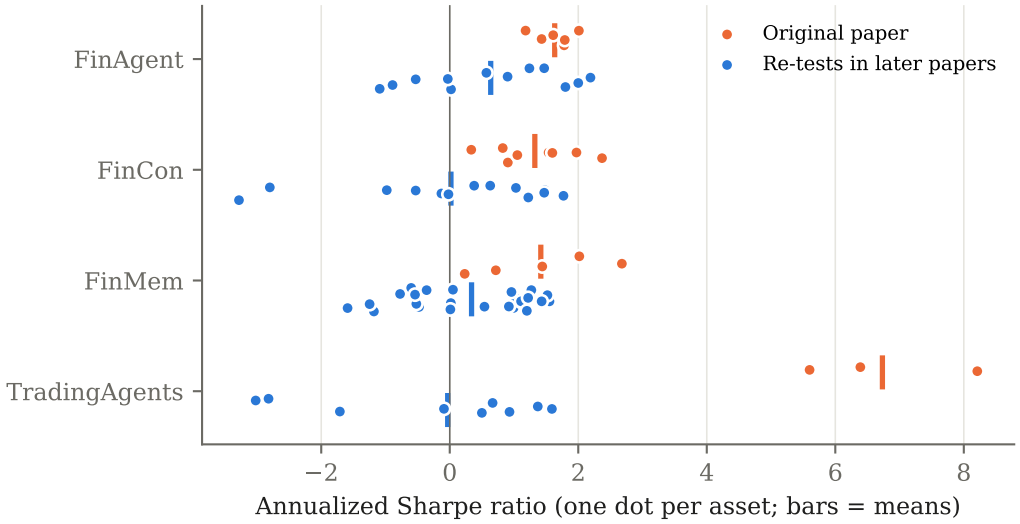


Figure 4: Annualized Sharpe ratios of four landmark systems as reported in their own papers (orange) and as re-run by other papers in the sample (blue). Re-tests use different periods, assets and sometimes backbone LLMs, so they are out-of-sample evaluations rather than exact replications.

6.2 Good practice in the sample

Several papers show that better evidence is achievable. Lopez-Lira and Tang [16] restrict their sample to the period after the model's cutoff, report factor-model alphas with t-statistics, and show how costs change the result: at 10 basis points per round trip, their headline Sharpe ratio of 2.97 falls to 1.29. BlindTrade anonymises tickers so the LLM cannot recall company histories, runs 20 seeds and deducts costs. The Russell 1000 nowcasting study runs live, in real time. The commodity study uses a stationary bootstrap [18] to test Sharpe-ratio differences, finds that its LLM agents beat an inverse-volatility portfolio by 0.05 with $p = 0.26$, and says so. The debate study reports standard errors over 35 scenarios and finds no relation between reasoning quality and Sharpe ratio ($r = 0.07$, $p = 0.29$). These papers report smaller effects, which is what one expects when the noise is measured honestly.

6.3 A minimum reporting standard

I suggest that papers claiming trading performance for an LLM system report the following:

1. **Test after the cutoff.** The test window should start after the knowledge cutoff of every LLM used, or inputs should be anonymised.
2. **Count every trial.** Report the number of configurations evaluated on any data used for selection, and compute the DSR with that number.
3. **Test long enough.** State the power of the test for a plausible effect, and prefer test windows of at least two years for daily strategies.
4. **Report the noise.** Give results over multiple seeds and samples, with standard errors, and test Sharpe-ratio differences against the strongest baseline, including buy-and-hold.
5. **Deduct costs.** Report results after realistic transaction costs and report turnover.
6. **Release the code and decisions.** Publish code, prompts and the full decision log so that others can re-run the system.

6.4 Limitations

The audit has several limitations. Extraction was done by one author with AI assistance, from the papers' arXiv HTML versions. To check it, eight papers (four landmark cases and four others) were re-extracted from their PDF versions. Every test date and Sharpe ratio matched; the re-check changed three coded fields (one trial count, one selection flag and one transaction-cost flag), which are corrected in the results above and listed in the repository. Because the re-check was not done by an independent second person, extraction errors in the remaining papers cannot be ruled out. Paper-level summaries average over assets, which treats correlated assets as separate pieces of evidence. The PSR assumes normal returns, and the baseline comparison assumes a correlation of 0.5. The reported K is a lower bound, so the DSR results understate the correction. Several papers give approximate test dates, and two do not identify the model version. The sample covers arXiv only; journal and conference versions may differ. Finally, a random sample of 40 from 65 papers gives an imprecise estimate of the publication-bias test.

7 Conclusion

The published evidence for LLM trading agents is weaker than its headline numbers suggest. Typical tests are too short to detect even exceptional skill, many reported Sharpe ratios are within the range produced by trying a handful of unskilled strategies, safeguards are rare, and the most-cited systems lose most of their performance when others re-run them. None of this shows that LLMs cannot help in trading. It shows that the current literature cannot tell. The good news is that the fixes are cheap compared with the cost of the experiments themselves: test after the cutoff, count every trial, test for longer, report the noise, deduct costs and release the code.

Data and code availability

The extraction records for all 48 papers, the knowledge-cutoff coding, the analysis code and the figure scripts are available at https://github.com/Navid-A-08/LLM-Trading-Audit.

Use of AI tools

The author used the assistance of AI (Claude, by Anthropic) while preparing this paper and its code, and takes full responsibility for the content.

References

[1] Abdollahzadeh, N. (2026). The false strategy problem in machine learning for finance: Why it is dangerously easy to discover "profitable" rules that do not exist. Working paper. https://doi.org/10.5281/zenodo.23085798

[2] Bailey, D. H., & López de Prado, M. (2012). The Sharpe ratio efficient frontier. *Journal of Risk*, 15(2), 3–44.

[3] Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5), 94–107.

[4] Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5), 458–471.

[5] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.

[6] Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634.

[7] Glasserman, P., & Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. arXiv:2309.17322.

[8] Harvey, C. R., Liu, Y., & Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.
[9] Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, 33(5), 2019–2133.
[10] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
[11] Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance*, 78(5), 2465–2518.
[12] Jobson, J. D., & Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, 36(4), 889–908.
[13] Ledoit, O., & Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15(5), 850–859.
[14] Li, W. W., Kim, H., Cucuringu, M., et al. (2025). Can LLM-based financial investing strategies outperform the market in long run? arXiv:2505.07078.
[15] Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4), 36–52.
[16] Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv:2304.07619.
[17] López de Prado, M., & Bailey, D. H. (2021). The false strategy theorem: A financial application of experimental mathematics. *American Mathematical Monthly*, 128(9), 825–831.
[18] Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313.
[19] Xia, Y., You, P., Wang, T., et al. (2026). Agentic trading: When LLM agents meet financial markets. arXiv:2605.19337.
[20] Xiao, Y., Sun, E., Luo, D., & Wang, W. (2024). TradingAgents: Multi-agents LLM financial trading framework. arXiv:2412.20138.
[21] Ye, Y., Han, J., Hu, A., et al. (2026). The alpha illusion: Reported alpha from LLM trading agents should not be treated as deployable. arXiv:2605.16895.
[22] Yu, Y., Li, H., Chen, Z., et al. (2023). FinMem: A performance-enhanced LLM trading agent with layered memory and character design. arXiv:2311.13743.
[23] Yu, Y., Yao, Z., Li, H., et al. (2024). FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. arXiv:2407.06567.
[24] Zhang, W., Zhao, L., Xia, H., et al. (2024). A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. arXiv:2402.18485.

A The 48 audited papers

Table 3: Audited papers. †: landmark system (included with certainty). Sharpe: annualized headline Sharpe ratio averaged over units. Years: test length. K: configurations reported on the test period. Overlap: knowledge-cutoff coding. Flags: C = transaction costs deducted, S = significance test reported, R = multiple runs or seeds reported.

arXiv	System	First author	Sharpe	Years	K	DSR	Overlap	Flags
2304.07619	Can ChatGPT forecast stock prices	Lopez-Lira et al.	2.97	2.66	12	1.00	none	CS
2311.13743	FinMem†	Yu et al.	1.42	0.50	15	0.26	full	SR
2401.03737	MarketSenseAI	Fatouros et al.	2.70	0.99	11	0.85	ambiguous	–
2402.03755	QuantAgent†	Wang et al.	–	–	–	–	full	–
2402.18485	FinAgent†	Zhang et al.	1.63	0.63	1	0.90	none	–
2407.06567	FinCon†	Yu et al.	1.32	0.68	1	0.83	partial	R
2407.09546	CryptoTrade†	Li et al.	0.54	0.18	2	0.40	ambiguous	–
2407.18103	Fine-tuning LLMs for return predic	Guo et al.	1.49	5.00	6	0.97	full	–
2409.06289	Automate strategy finding with LLM	Kou et al.	4.56	1.04	16	1.00	partial	–
2410.12464	FS-ReasoningAgent	Wang et al.	2.10	0.13	4	0.49	none	–
2411.00782	TradExpert	Ding et al.	5.01	0.99	5	1.00	none	–
2411.11059	News-driven LLM RL for portfolio m	Unnikrishnan	–	–	–	–	unknown	–
2412.19245	Sentiment trading with LLMs	Kirtac et al.	3.05	1.98	12	1.00	none	C
2412.20138	TradingAgents†	Xiao et al.	6.73	0.24	1	1.00	none	–
2502.07393	FinRL-DeepSeek	Benhenda	–	–	–	–	full	–
2502.11433	FLAG-Trader†	Xiong et al.	1.95	0.63	1	0.91	full	R
2502.13165	HedgeAgents†	Li et al.	2.41	2.99	13	0.99	partial	–
2503.09647	LLM top-down sector allocation	Heng et al.	2.51	0.49	1	0.96	full	C
2504.14345	LLM-enhanced Black-Litterman	Lee et al.	1.23	0.79	4	0.52	none	–
2505.10278	MASS	Guo et al.	1.73	0.96	10	0.55	full	–

arXiv	System	First author	Sharpe	Years	K	DSR	Overlap	Flags		
2506.09080	FinHEAR	Chen et al.	1.92	0.99	1	0.97	partial	C
2507.18560	HARLF	Coriat et al.	1.20	7.00	11	0.93	none	–
2507.20474	MountainLion	Wu et al.	–	–	–	–	unknown	–
2508.11152	AlphaAgents	Zhao et al.	–	–	–	–	unknown	–
2509.11420	Trading-R1	Xiao et al.	1.59	0.26	3	0.48	full	–
2510.04787	TiMi	Song et al.	–	–	–	–	unknown	–
2510.14264	AlphaQuanter	Deng et al.	0.65	0.48	2	0.47	none	–
2510.15949	ATLAS	Papadakis et al.	–	–	–	–	none	R
2510.17393	3S-Trader	Chen et al.	1.73	2.02	3	0.92	partial	–
2510.23032	P1GPT	Lu et al.	2.80	0.66	1	0.98	unknown	–
2510.26228	ChatGPT in systematic investing	Anic et al.	1.06	1.24	1	0.88	none	C
2511.12599	FinRS	Liu et al.	0.76	0.17	1	0.62	none	–
2511.13614	Market-dependent communication in	Shi et al.	–	–	–	–	none	R
2601.11958	Agentic AI nowcasting	Chen et al.	3.23	0.63	12	0.81	none	CS
2602.00082	LLM multi-agent REIT trading	Li	1.77	0.99	2	0.89	none	C
2602.19919	Janus-Q	Li et al.	1.31	0.23	9	0.19	none	–
2603.17692	BlindTrade	Jeon et al.	1.40	0.58	3	0.58	none	CSR
2603.22567	TrustTrade	Li et al.	–	–	–	–	full	–
2604.17327	Signal or noise in multi-agent LLM	Fatouros et al.	–	–	–	–	unknown	S
2605.05580	AlphaCrafter	Yuan et al.	–	–	–	–	partial	CR
2605.06024	Strat-LLM	Huang et al.	2.19	0.33	36	0.19	partial	–
2606.03704	DOSS	Sakurai et al.	–	–	–	–	full	–
2606.08283	Macro economists in the machine	Wang et al.	0.57	2.38	4	0.43	none	CS
2607.12233	Fin-Analyst	Rashid et al.	4.10	0.13	5	0.61	none	–
2608.05668	F2Agent	Liu et al.	1.75	0.50	1	0.88	none	C
2608.28252	Regime-aware RAG expert selection	Asadi et al.	1.19	1.48	12	0.40	unknown	–
2609.24175	FinRankGRPO	Deng et al.	0.64	6.04	24	0.34	partial	C
2609.29701	Multi-agent debate for explainable	Huang et al.	–	–	–	–	partial	SR