

17 Oct 2026 · Study: Backtest versus live evidence · Research note

LLM trading agents vs buy-and-hold: what the 2025-26 benchmarks show

Huyen X. Vu · Quantex Research, AlgoTrade Pte. Ltd., Singapore · ORCID 0000-0002-2209-842X

Abstract

Four published benchmarks let large language models make trading decisions and compared them with buy-and-hold or an index. We set out what each tested and found. Results are mixed: agents won in some markets and lost in others, and in the one benchmark that ran the same models in three markets no model beat the baseline in all three. Comparisons lasted two weeks to about four months, each in one market period, and none trades gold (XAUUSD) live on MetaTrader 5.

Keywords: large language models; trading agents; benchmarks; buy-and-hold

JEL: G11; G17; C45

Highlights

- **The results are mixed.** LLM agents beat buy-and-hold or an index in some markets and lost in others. In AI-Trader, the one benchmark that ran the same models in three markets, no model beat the baseline in all three.
- **The tests are short.** The comparisons with buy-and-hold or an index ran for two weeks to about four months, each in one market period.
- **None trades gold live on MT5.** All four benchmarks use stocks, crypto or both. None trades gold (XAUUSD) live on MetaTrader 5.

1. Introduction

Several research groups have now let large language models (LLMs) make trading decisions and compared them with simply buying and holding. This note sets out what four published benchmarks tested, what they found against buy-and-hold or an index, and what none of them covers yet.

2. Data and method

Method notes and limits

- We read the four papers on arXiv and used only figures printed in them. We did not rerun any test.
- The baselines differ. StockBench and Agent Market Arena use buy-and-hold of the same assets. AI-Trader uses an index (QQQ, SSE-50 and a crypto index), so its baseline is close to, but not the same as, buy-and-hold of the assets the agents traded.
- Each paper sets its own rules for trading costs, position size and trading frequency. We did not check how each one models costs, so the rows are not directly comparable.
- In Agent Market Arena, one agent design (HedgeFundAgent) gave identical results on all five LLMs. The paper does not explain this.
- StockBench's abstract and its results table point in different directions. We report the table's numbers and quote the abstract.

- We left out Alpha Arena by Nof1, a public contest where models are reported to trade real money. We could not open Nof1's own results page, so we give no figures for it.

3. Results

What each benchmark tests

"Buy-and-hold" means buying the assets at the start and holding them to the end. Two benchmarks compare with an index instead, which works in a similar way. Returns are cumulative over the test period, as printed in each paper.

AI-Trader

Who	University of Hong Kong (Chao Huang's group)
Assets and models	Nasdaq-100 stocks (hourly), SSE-50 stocks (daily), 10 crypto pairs (daily). 6 LLMs
Live or backtest	Live
Period	1 Oct to 7 Nov 2025 (stocks); 1 to 14 Nov 2025 (crypto)
Result vs buy-and-hold or index	US: 3 of 6 beat QQQ (+1.87%); best +9.56%. China A-shares: 0 of 6 beat the SSE-50 (+1.65%). Crypto: 1 of 6 beat the index (−14.30%), at −12.18%.

StockBench

Who	Tsinghua University; Beijing University of Posts and Telecommunications
Assets and models	20 Dow Jones stocks with the largest index weights. 13 LLMs
Live or backtest	Backtest on historical data
Period	3 Mar to 30 Jun 2025 (82 trading days)
Result vs buy-and-hold or index	Equal-weight buy-and-hold: +0.4%, max drawdown −15.2%. 8 of 13 models returned more (best +2.5%). All 13 had a smaller max drawdown.

Agent Market Arena

Who	The Fin AI, with university co-authors
Assets and models	TSLA, BMRN, BTC, ETH. 4 agent designs, each run on 5 LLMs
Live or backtest	Live
Period	1 Aug to 30 Sep 2025
Result vs buy-and-hold or index	Best agent vs buy-and-hold: BMRN +23.70% vs −6.89%; BTC +5.02% vs +0.66%; ETH +39.66% vs +18.56%; TSLA +40.83% vs +46.88% (buy-and-hold ahead).

AlphaForgeBench

Who	NTU Singapore; HKUST (Guangzhou); Hong Kong Polytechnic University
Assets and models	BTC, ETH, AAPL, GOOGL, MSFT, NVDA, TSLA. 6 LLMs design factor-based strategies, not trade-by-trade decisions
Live or backtest	Backtest
Period	1 Jan 2021 to 1 Jan 2026
Result vs buy-and-hold or index	No buy-and-hold or index comparison. Models are ranked against each other.

"Best" means the highest return among the models or agent and model pairs tested on that market. All figures are as printed in the papers listed under Sources.

4. Discussion

What the results suggest

The data. In AI-Trader, agents beat the index in the US market, but none did in China A-shares and only one did in crypto, where every agent lost money. In StockBench, most models made a little more than buy-and-hold over four months, and all had smaller drawdowns. Even so, the paper's abstract says "most models struggle to outperform the simple buy-and-hold baseline". In Agent Market Arena, the best agent beat buy-and-hold on three of four assets, but not on TSLA, which rose 46.88% in two months.

Our reading.

- **The margins are small or narrow.** In StockBench, the best model returned 2.5% against 0.4% for buy-and-hold. A gap of about two percentage points over 82 days, in one market period, is weak evidence of skill.
- **"Best of many" flatters the result.** Agent Market Arena tested 20 agent and model pairs per asset. The best of 20 can beat buy-and-hold by luck alone.
- **The market matters more than the model.** The same six models in AI-Trader did well in US stocks and badly in crypto. A strong run can say more about the period than about the agent.
- **Lower drawdown is the most consistent finding.** Agents that trade in and out are often less exposed than a fully invested buy-and-hold position. In StockBench that showed up in all 13 models.

None of the four benchmarks shows an LLM agent beating buy-and-hold across markets over a long period. The tests are too short and too few to settle the question either way.

The gap: no long live gold record

None of these benchmarks trades gold (XAUUSD), and none runs on MetaTrader 5 (MT5), the platform our own gold Expert Advisors (EAs) run on. The live tests ran for weeks, not years. A long live gold control series would add three things:

- **A different market.** Gold moves in a different way from single stocks and crypto, with its own sharp one-day moves. See note 05, Gold's largest one-day moves.
- **A rule-based control.** A fixed EA run on the same broker and the same prices gives agents a baseline that does not change its mind, alongside buy-and-hold.
- **Time.** Many months of live results cover calm and volatile periods, not one short window.

Why the difference between a test and a live record matters: see Why an EA backtest is profitable but the live account loses and Case study 01.

References

- Fan, Yang, Jiang, Zhang, Chen, Huang, "AI-Trader: Benchmarking Autonomous Agents in Real-Time Financial Markets", arXiv 2512.10971 (Dec 2025); abstract, full text (Table 1)
- Chen, Yao, Liu, Xin, Ye, Yu, Hou, Li, "StockBench: Can LLM Agents Trade Stocks Profitably In Real-world Markets?", arXiv 2510.02209; abstract, full text (Table 2)
- "When Agents Trade: Live Multi-Market Trading Benchmark for LLM Agents" (Agent Market Arena), arXiv 2510.11695; abstract, full text (Table 1)
- Zhang, Zhao, Gao, You, Jia, Zhao, An, Sun, "AlphaForgeBench: Benchmarking End-to-End Trading Strategy Design with Large Language Models", arXiv 2602.18481; abstract, full text

Declarations

- **Author:** Huyen X. Vu, Quantex Research, AlgoTrade Pte. Ltd., Singapore (ORCID 0000-0002-2209-842X).
- **Data availability:** Based on the published benchmarks listed under Sources.
- **Use of AI tools:** Data processing, analysis code and drafting were assisted by AI tools (Claude, Anthropic). The author checked the results and is responsible for the content.
- **Conflict of interest:** The author is the founder of AlgoTrade Pte. Ltd., which develops and runs systematic trading models.
- **Funding:** No external funding.
- **Peer review:** Not yet peer reviewed. Comments are welcome at info@quantex.com.sg.
- **How to cite:** Vu, H. X. (2026). LLM trading agents vs buy-and-hold: what the 2025-26 benchmarks show. Quantex Research, AlgoTrade Pte. Ltd. <https://quantex.com.sg/research/2026-10-17-llm-agents-vs-buy-and-hold/>