

From Algorithms to Agents: A Survey of LLM-Based Trading Systems

YEBO FENG, Nanyang Technological University, Singapore

YICHEN LUO, UCL, Centre for Blockchain Technologies, United Kingdom and The University of Hong Kong, FinTech Academy, China

JINGYANG DAI, Nanyang Technological University, Singapore

CHIEN-CHIH CHEN, University of Waterloo, Canada

GUANGYAN GAN*, Nanyang Technological University, Singapore

JIAHUA XU, UCL, Centre for Blockchain Technologies DLT Science Foundation, United Kingdom

LIN WILLIAM CONG, Nanyang Technological University, Singapore, ABFER, Singapore, CEPR, United Kingdom, and NBER, USA

YANG LIU, Nanyang Technological University, Singapore

Large language models (LLMs) now drive trading systems, from signal extractors to autonomous agents, but the literature is fragmented across markets, architectures, and evaluation practices. This survey traces the path from quantitative trading to LLM agents, organizes the literature by capability (perception, reasoning, action, memory, collaboration) and by six markets from equities to prediction markets, and maps 47 systems across both to show where evidence is thin. We audit evaluation protocols and consolidate head-to-head evidence against traditional quantitative baselines. Four findings emerge. First, the margins of independently re-run agents shrink or vanish, though the re-runs cover few systems and many reported margins rest on windows of eight months or less, where an annualized Sharpe ratio has a standard error above one. Second, the best-supported hypothesis places the LLM upstream of the trade (text signals, factor and model code, and bounded execution tilts), although most upstream evidence comes from single author-run studies. Third, controlled comparisons do not establish a multi-agent advantage, though risk controls help where ablated. Fourth, structure and input design matter alongside the backbone, and the best structure depends on the backbone. Open challenges span reliability, security, deployment, regulation, market stability, reproducibility, and human-AI collaboration.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Multi-agent systems**; • **Information systems** → *Data mining*.

Additional Key Words and Phrases: Large Language Models, Autonomous Agents, Financial Trading, Quantitative Trading, Multi-Agent Systems, Cryptocurrency, Survey

1 Introduction

Trading systems have moved from rule-based strategies and statistical arbitrage through deep learning and reinforcement learning (RL) [81, 151, 171] to large language models (LLMs), which now power signal extractors and agents that trade with diminishing human oversight [213, 232, 237]. LLM agents read text [134] and chart images [142], reason into natural-language theses [200, 212],

*Corresponding author.

Authors' Contact Information: Yebo Feng, yebo.feng@ntu.edu.sg, Nanyang Technological University, Singapore; Yichen Luo, UCL, Centre for Blockchain Technologies, London, United Kingdom and The University of Hong Kong, FinTech Academy, Hong Kong SAR, China, yichen.luo.22@ucl.ac.uk; Jingyang Dai, Nanyang Technological University, Singapore, jingyang.dai@ntu.edu.sg; Chien-Chih Chen, University of Waterloo, Canada, j2255che@uwaterloo.ca; Guangyan Gan, Nanyang Technological University, Singapore, GUANGYAN001@e.ntu.edu.sg; Jiahua Xu, UCL, Centre for Blockchain Technologies and DLT Science Foundation, London, United Kingdom, jiahua.xu@ucl.ac.uk; Lin William Cong, Nanyang Technological University, Singapore and ABFER, Singapore and CEPR, London, United Kingdom and NBER, Cambridge, MA, USA, Will.Cong@ntu.edu.sg; Yang Liu, Nanyang Technological University, Singapore, yangliu@ntu.edu.sg.

and organize as multi-agent firms that reportedly surpass single-agent [232] and rule-based [213] baselines.

The topic matters now because deployment is outpacing evidence. In a survey by the International Organization of Securities Commissions (IOSCO), 63.3% of industry respondents report using AI for algorithmic trading [86]. Supervisors expect autonomous AI to move quickly into trading and warn that agents responding alike could amplify volatility in stress [20, 58]. The research that should inform these decisions is hard to trust: in 164 main-conference papers, no finance-specific evaluation bias is discussed by more than 28% [100], and an audit of 185 studies finds protocol reporting improving but still incomplete [210]. Practitioners and regulators need a synthesis that separates reported from reproducible performance.

Existing surveys do not provide it. Since late 2022, LLM trading papers have proliferated [44, 119, 167, 214, 232, 237], but surveys cover LLMs in finance [104, 120, 149], LLM agents generally [198, 208], or, concurrently, LLM trading agents [26, 40, 44]; none combines broad market coverage, a capability-centric agent taxonomy, and systematic cross-paradigm comparison. Drawing on 246 works, 45 of them primary studies in our evidence tables, we make three contributions:

- **Paradigm-transition analysis.** We trace the move from quantitative trading to LLM agents, showing what is new and where traditional methods remain competitive.
- **Capability $\times$ market map.** We cross a capability-centric taxonomy, adapted from general agent frameworks [179, 198], with six markets from equities to cryptocurrency and decentralized finance (DeFi), forex, derivatives, and prediction markets, and code 47 systems into the resulting map (Figure 3). Its empty cells locate the gaps: every memory design comes from an equity trader, and the other markets rest mostly on few systems or on bounded LLM components.
- **Evidence audit and cross-paradigm comparison.** We set reported results against same-window passive baselines, audit their protocols, grade multi-agent claims by how well they control confounds, examine why author-run (self-evaluated) and independent results diverge, and consolidate 25 head-to-head studies against quantitative methods.

The audit supports four findings: independently re-run margins shrink or vanish (§7.4); the best-supported current hypothesis places the LLM upstream of the trade, in text signals, factor and model code, and bounded execution tilts (§8.6); a multi-agent advantage is not established (§4.3); and structure and input design matter alongside the backbone, with the best structure depending on the backbone (§4.3, §5.2). Our evidence tables are purposive and over-represent highly cited 2023–2024 systems, and 20 of the 45 primary studies are preprints (Appendix A11). The appendices give a glossary, the methodology, and the full evidence tables.

2 Related Surveys

Table 1 scores coverage overlap; the surveys fall into three lines, each missing a different part of the problem. Pre-LLM surveys of machine, deep, and reinforcement learning for trading [75, 81, 151, 158] supply the baselines LLM agents must beat but no agent architecture. General agent surveys [179, 198, 208] contribute the module decompositions our taxonomy adapts, yet assume domain-agnostic environments without adversarial participants, regulation, or real-time execution. Finance surveys either address LLMs and financial NLP rather than trading systems [104, 120, 149], or review trading agents with limited market breadth: Ding et al.’s Trader/Alpha Miner taxonomy [40] and Cao et al.’s review [26] cover equities only, and Mahdavi et al. [143] restate self-reported returns without discussing contamination or comparability. The newest surveys audit rather than catalog: Xia et al. [210] assess protocol reporting across 185 studies and propose a reporting checklist but decline to compare performance, and Nguyen and Pham [148] hypothesize, without testing it, that coordination protocol matters more than backbone.

Table 1. Comparison with related surveys (○: not included; ◐: partially included; ●: included). #Refs: number of references; † number of included studies instead, as the reference count was not determined. The six subject dimensions map to core sections of this survey: Paradigm Transition (§3), Agent Architecture (§4), Multi-Agent (§4.3), Multi-Market Applications (§6), Evaluation & Benchmarks (§7), and Cross-Paradigm Comparison (§8). Because the dimensions are this survey’s organizing axes, the table shows overlap of coverage, not quality; §2.1 states the substantive differences.

Survey	Year	#Refs	Description	Subjects Covered					
				Paradigm Transition	Agent Architecture	Multi-Agent	Multi-Market Applications	Evaluation & Benchmarks	Cross-Paradigm Comparison
Traditional ML/DL Trading Surveys									
Ozbayoglu et al. [151]	2020	202	Deep learning (CNN, RNN, LSTM) for stock trading, portfolio management, and risk.	🟡	🟢	🟢	🟡	🟢	🟢
Sezer et al. [171]	2020	220	Systematic review of deep learning for financial time series forecasting, 2005–2019.	🟡	🟢	🟢	🟡	🟣	🟢
Cao [27]	2023	221	AI in finance: classic statistical methods versus modern deep learning across tasks.	🟣	🟢	🟢	🟡	🟢	🟡
Hambly et al. [75]	2023	246	RL in mathematical finance: execution, portfolio optimization, option pricing.	🟡	🟢	🟢	🟡	🟢	🟢
Sun et al. [180]	2023	165	RL for quantitative trading: algorithmic trading, portfolio management, execution, market making.	🟡	🟢	🟢	🟡	🟣	🟢
Pippas et al. [158]	2025	356	Evolution of RL in quantitative finance, incl. transfer, meta-, and multi-agent RL.	🟣	🟢	🟡	🟡	🟣	🟡
LLM-Based Agent Surveys									
Xi et al. [208]	2025	675	LLM agent framework (brain, perception, action) with multi-agent interaction across domains.	🟢	🟣	🟣	🟢	🟡	🟢
Guo et al. [72]	2024	88	Multi-agent LLM systems: communication protocols, role specialization, coordination.	🟢	🟣	🟣	🟢	🟢	🟢
Sumers et al. [179]	2024	235	CoALA: agent design grounded in cognitive science (memory, action, decision).	🟢	🟣	🟢	🟢	🟢	🟢
Wang et al. [198]	2024	185	Unified four-module agent framework (profiling, memory, planning, action).	🟢	🟣	🟡	🟢	🟢	🟢
LLM Agents for Finance Surveys									
Li et al. [120]	2023	68	LLM adoption guide for finance with a four-level framework (zero-shot to pretraining) and costs.	🟡	🟢	🟢	🟢	🟢	🟢
Lee et al. [104]	2025	54	Evolution from pre-trained language models to financial LLMs; five training techniques compared.	🟡	🟢	🟢	🟢	🟣	🟢
Nie et al. [149]	2024	318	LLM application areas in finance: linguistic tasks, sentiment, time series, reasoning, agent-based modeling.	🟡	🟡	🟡	🟡	🟣	🟢
Ding et al. [40]	2026	60	First LLM trading-agent survey; "Trader vs. Alpha Miner" taxonomy; equities only.	🟢	🟣	🟡	🟢	🟡	🟢
Dong et al. [44]	2025	68	Maps LLM agents to five financial-institution divisions with a benchmark catalog.	🟢	🟣	🟣	🟡	🟣	🟢
Cao et al. [26]	2025	186	DL-to-LLM evolution along the alpha pipeline (data, prediction, optimization, execution).	🟣	🟡	🟡	🟢	🟡	🟣
Mahdavi et al. [143]	2025	—	LLM-for-investment work in four technique buckets: descriptive, equity-centric.	🟢	🟡	🟡	🟢	🟢	🟢
Fu [61]	2025	—	Task-centered survey of LLMs for equity return prediction and trading; time-safe, economically grounded evaluation.	🟡	🟡	🟢	🟢	🟣	🟡
Gong [66]	2026	—	Conceptual "agentic finance" framework and systemic-risk model of agent markets.	🟡	🟡	🟢	🟡	🟢	🟢
Nguyen & Pham [148]	2026	—	Workshop survey of 12 multi-agent trading systems: design taxonomy, five evaluation failures, coordination-primacy hypothesis.	🟢	🟣	🟣	🟢	🟣	🟢
Xia et al. [210]	2026	185 [‡]	Audit of 185 studies (80 closed-loop) to Sep. 2026: protocol coverage, reproducibility tiers, reporting checklist.	🟡	🟣	🟣	🟡	🟣	🟡
Ours	2026	246	Capability + scenario dual-dimensional LLM trading agent survey with cross-paradigm comparison.	🟣	🟣	🟣	🟣	🟣	🟣

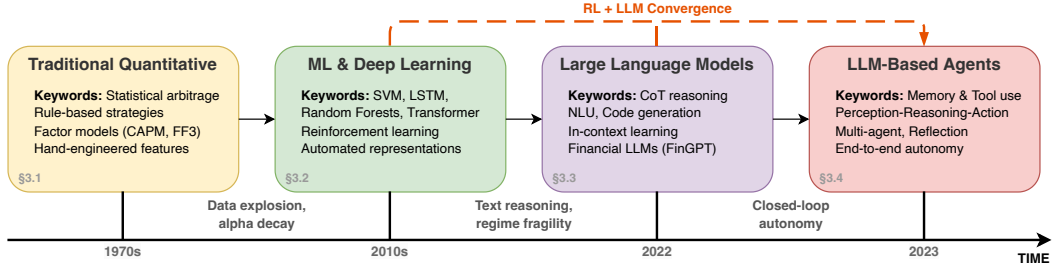


Fig. 1. Paradigm evolution in quantitative trading across four eras, with representative methods and the drivers of each transition.

2.1 Positioning of This Survey

Our contributions (§1) differ from these surveys in four respects rather than in coverage of Table 1’s dimensions: (i) market breadth beyond equities, with the capability $\times$ market map (§4.1) and each market’s evidence gaps (§6); (ii) evidence grading of multi-agent claims, separating comparisons with published systems, within-system ablations, and matched single-agent controls (§4.3); (iii) a consolidation of independent re-evaluations that sets author-run margins against re-run ones (§7); and (iv) a consolidation of head-to-head evidence against traditional baselines that keeps signal-level and agent-level studies apart (§8).

3 Background: Paradigm Evolution in Quantitative Trading

Quantitative trading has passed through four paradigms (Figure 1), each relaxing one binding constraint. Appendix A10 defines the finance and ML terms used throughout.

3.1 Traditional Quantitative Trading

Traditional quantitative trading lets humans specify signals and statistical models estimate them [27], through rule-based strategies [21], statistical arbitrage [62], and factor models from CAPM [173] to Fama-French [51]. Its binding constraint was human feature design, and even text entered this way, as dictionary sentiment scores built on a fixed vocabulary [136, 183].

3.2 Machine Learning and Deep Learning

Machine learning and deep learning moved feature construction to the model [70, 151, 171] but not past the low signal-to-noise ratio of financial data, and capacity did not guarantee gains: Feng et al.’s selected deep factor models reach out-of-sample Sharpe ratios of 2.49 and 3.00, below the 3.67 of a five-factor instrumented principal component analysis (IPCA) model [55]. Attention, the architecture behind LLMs, is likewise not automatically superior on numerical sequences; on limit-order-book prediction, LSTM variants beat Informer, Autoformer, and FEDformer on price-change and direction tasks [16]. Evaluation reported movement accuracy on disparate data sets rather than profitability under common protocols [246], a fragmentation that LLM agent studies inherited (§7.5). Reinforcement learning made the position the output [75, 180], and FinRL standardized environments [130], but sample efficiency and the sim-to-real gap remain open. Learned encoders such as FinBERT [6, 84] replaced text dictionaries but were each trained for one labeled target.

3.3 Large Language Models

LLMs made unstructured text a trading input through zero-shot reading of filings and news [206], chain-of-thought reasoning [204], in-context learning [22], and code generation. Yet the most cited

early results cast the LLM as analyst or signal generator, from headline sentiment trading [134] to alpha mining [202], because a standalone LLM is stateless, passive, and disconnected from live data and execution.

3.4 From LLM to Agent

Agents close this loop by embedding the LLM in an architecture of perception, reasoning, action [225], memory [231], and multi-agent collaboration [213, 232], the five dimensions of our taxonomy (§4). Each transition changed one component: ML automated features and RL replaced prediction error with sequential reward, LLMs widened the information set to text, and agents widened the action space to tool calls, retrieval, and memory updates. What stayed fixed are the environment and its testing: noisy, non-stationary data, regimes that break fitted models, and evaluation by bespoke backtests. §7.5 and §8 test whether text, reasoning, and memory overcome these constants or merely add ways to overfit them.

4 Architecture of LLM Trading Agents

Five capability dimensions frame LLM trading agents: perception, reasoning, action, and memory within one agent (§4.1), and collaboration among several (§4.3). Three functional paradigms, Trader, Alpha Miner, and bounded component, place the LLM at different points of the pipeline (§4.2).

4.1 Overall Architecture and Information Flow

An LLM agent is a language model in a closed perceive–reason–act loop with memory and tools [175, 179, 225]. Trading keeps this loop but changes its constraints: perception fuses narratives with numerical series [134, 237]; actions are irreversible and subject to slippage, latency, and market impact [25, 153]; memory must survive regime shifts [231, 232]; and reasoning is judged on risk-adjusted performance, not task completion [14, 217]. The reasoning core is a prompted frontier LLM, a finance-specialized LLM [206, 214], or a smaller open-weight model trained end-to-end for trading [212, 245]. Four modules surround this core (Figure 2), each with an open challenge (techniques in §5).

- **Perception** treats text, images, and on-chain data as first-class signals, and retrieval decides what reaches the reasoner [107, 134, 237]. Open challenges are selectivity under cost limits and robustness to fabricated inputs (§9.2).
- **Reasoning and planning** externalize the decision trajectory in auditable natural language [175, 204]. Open challenges are building risk limits into the plan rather than post-hoc filters, and bounding deliberation cost.
- **Action** trades expressiveness for safety: a buy/hold/sell vocabulary limits strategies, while code- and tool-based spaces enlarge the surface for errors and manipulation [2, 203]. An emerging remedy is a deterministic execution layer that bounds what the LLM can emit [139, 174].
- **Memory** moves to retrievable stores because effective context is shorter than nominal [126]; trading systems add layered hierarchies and verbal reinforcement (feedback stored as text) [231, 232]. The trading-specific challenge is *regime robustness*: keeping long-horizon patterns while discounting stale experience.
- **Tool use** spans all four modules and mirrors the action trade-off: fixed-schema function calling is easy to validate [156, 164], whereas code-as-action is Turing-complete [203].

The literature populates these capabilities unevenly across markets (Figure 3). Equities hold 32 of 47 systems and all nine memory designs. Crypto is perception-heavy (8 of 10 systems), though crypto-only systems propose no action or memory design; futures and options are action-heavy (4 of 6) with no memory design; and forex is nearly empty [89, 116]. By functional paradigm,

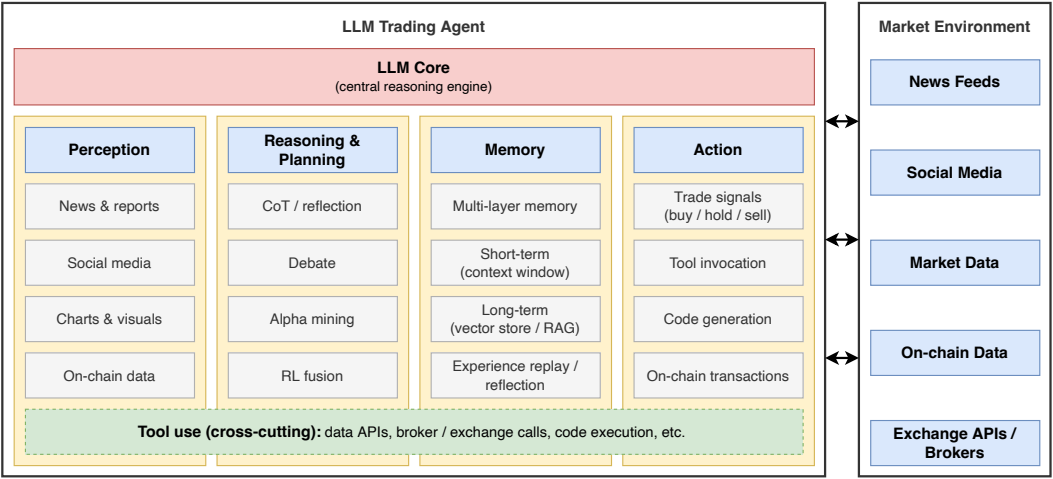


Fig. 2. Architecture of an LLM trading agent. The LLM core coordinates perception, reasoning and planning, memory, and action modules through a cross-cutting tool-use layer that connects to the market environment.

a Systems with a distinctive contribution in each capability

	Equities (n=32)	Crypto/ DeFi (n=10)	Forex (n=2)	Futures/ options (n=6)	Prediction markets (n=3)	Cross- market (n=1)
Perception	16	8	1	3	3	—
Reasoning	19	8	1	1	1	1
Action	15	3	1	4	1	1
Memory	9	2	1	—	—	1
Collaboration	20	7	1	3	1	1

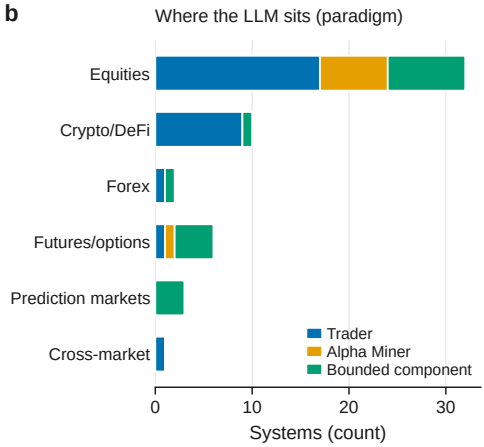


Fig. 3. Capability × market map of the 47 surveyed LLM trading systems. (a) Systems for which a capability is a distinctive contribution, per market (“—” marks an empty cell); (b) functional paradigm (§4.2) per market. Coding rules are in Appendix A4.1.

Traders dominate crypto (9 of 10), Alpha Miners appear only in equities and futures, and bounded components dominate futures and prediction markets.

Takeaway. Each module’s open problem stems from a market constraint. For action and tool use, the surveyed systems respond by bounding the LLM at its interfaces; how tight the bounds can be before language reasoning stops paying off is unresolved.

Table 2. Design-level contrast between the three functional paradigms.

Dimension	Agent as a Trader	Agent as an Alpha Miner	LLM as a bounded component
Output	Buy/hold/sell orders with position sizes.	Alpha factors, ranking scores, or signal definitions.	A signal, view, parsed query, or bounded tilt.
Scope	Full perceive–reason–act–memory loop in-agent.	Upstream of quant pipeline; delegates execution and risk.	One input to a classical module (optimizer, RL policy, execution schedule, rule engine) that owns the decision.
Supervision	Portfolio-level: Sharpe, drawdown, cumulative return.	Signal-level: Information Coefficient, Information Ratio.	Module-level: the host module’s objective (e.g., execution price vs. TWAP, query validity).
Architecture	Role-specialized multi-agent debate; reflection over trade history.	Writer–critic inner loop; backtest-in-the-loop outer loop.	Typed or bounded interface checked or executed by deterministic code.
Failures	Execution friction; decision-time reasoning errors.	Alpha decay [146]; factor crowding; look-ahead bias; data snooping across candidates [9, 76, 205].	Gain depends on the host module [14]; look-ahead in the LLM’s input.
Examples	FinAgent [237], FinMem [231], FinCon [232], TradingAgents [213], ContestTrade [239].	Alpha-GPT [202] and Alpha-GPT 2.0 [233], QuantAgent [201], Kou et al. [101], R&D-Agent-Quant [121], AlphaAgent [182].	PACE [174], OQL [139], LLM Black-Litterman [105], FinRL-DeepSeek [14], Macro Economists [63].

4.2 Functional Paradigms: Trader, Alpha Miner, and Bounded Component

Published systems cluster into two end-to-end *functional paradigms* [40] plus a third in which the LLM is only a bounded component of a classical system (Table 2). Traders are 22 of the 47 systems in Figure 3 and the main target of independent re-evaluation.

- **Agent as a Trader.** The agent emits trading decisions and owns the full loop, either alone (FinAgent [237], FinMem [231]) or with modules assigned to specialized roles (FinCon [232], TradingAgents [213]).
- **Agent as an Alpha Miner.** The agent produces alpha factors or signal definitions for a downstream pipeline [51, 70], from Alpha-GPT [202] to backtest-in-the-loop systems [121, 201]. The LLM adds economic rationale, which AlphaAgent turns into a constraint: its factors hold an information coefficient (IC) near 0.02 on CSI 500 through 2024 while Alpha158, genetic-programming, and RSI factors decay to near zero [182]. Proposing code rather than trades only partly limits replay of memorized prices: in Profit Mirage’s post-cutoff re-runs, Wang et al.’s QuantAgent decayed nearly as much as the Traders (\$7.4) [117].
- **LLM as a bounded component.** The LLM supplies one bounded input (a signal, a view, a parsed query, or a tilt) to a classical module that owns the decision, such as an execution schedule [174], an option-strategy engine [139], a Black-Litterman optimizer [105], or a reinforcement-learning policy [14]. This functional paradigm underlies the division of labor of §8.6.

The two end-to-end paradigms diverge because their outputs differ in the cost of being wrong. A factor is cheap to re-score on history, suiting writer-critic iteration [201, 202]; a trade has immediate monetary consequence, so Traders justify it through role-specialized debate [213] and reflection over trade histories [231]. Hybrids that feed LLM signals to reinforcement-learning policies blur the boundary [61, 217], so the practical design question is which modules to assign to LLM reasoning and which to classical infrastructure.

Table 3. Representative multi-agent LLM trading systems and the strongest evidence each offers for its multi-agent design, graded as in §4.3. Roles, aggregation rules, and figures are in Table 9.

System	Market	Organization	Strongest evidence for the multi-agent design
TradingGPT [122]	US equities	Peer debate	None (design paper)
FinCon [232]	US stocks	Hierarchical	vs. prior (FinMem, FinAgent); ablation of risk control, not of agents
TradingAgents [213]	Three US stocks	Hierarchical, two debate stages	Rule-based baselines only
HedgeAgents [116]	BTC, DJ30, USD/CNY	Hierarchical, conference-based	Ablation of each conference; no single-agent row
QuantAgents [118]	NASDAQ-100; live A/HK	Hierarchical, meeting-based	vs. prior (HedgeAgents, FinAgent); ablation of each meeting
ContestTrade [239]	Chinese A-shares	Internal contest	Ablation: removing contests and deep research cuts Sharpe 3.12 to 0.07
AlphaAgents [240]	15 US tech stocks	Debate to consensus	vs. single (figure only, one run); all portfolios lose money
Macro Economists [63]	15 commodity ETFs	Persona debate	vs. single with bootstrap: no Sharpe gain over the better persona
Expert Teams [147]	TOPIX 100	Hierarchical, three levels	Leave-one-out (50 runs): four of five specialists lower Sharpe
MASS [74]	SSE 50, CSI 300, ChiNext	Population ensemble	Ablation of aggregation; IC rises with agent count (figure only)
Crypto portfolio MAS [141]	15 layer-1 coins, weekly	Hierarchical, collaborative, debate	vs. single at matched prompting (one run each): gain in 10 of 12 configurations
TradExpert [42]	DOW 30	Mixture of experts	Ablation of each expert
FS-ReasoningAgent [200]	BTC, ETH, SOL	Fact/subjectivity pipeline	Ablation (one run)
CryptoTrade [119]	BTC, ETH, SOL	Modular pipeline	Ablation of reflection and inputs
MarketSenseAI 2.0 [53]	S&P 100 and 500	Modular pipeline	Index benchmarks only
Meme-coin copy trading [140]	Solana meme coins	Independent judges	Ablation: removing wallet or coin agent turns return negative

Takeaway. The functional paradigm decides whether LLM error reaches the order book directly or is first filtered by re-scoring and classical code. No surveyed study compares the paradigms’ net performance under matched conditions (Profit Mirage runs both on the same windows only to measure post-cutoff decay [117]).

4.3 Single-Agent vs. Multi-Agent Organization

Multi-agent systems are justified by specialization, heterogeneous backbones, diversity of opinion, and checks and balances [82, 108, 213, 232], whereas single agents [134, 231] are simpler but struggle to cover all the information a trading decision needs. Six organizational patterns recur (Table 3): hierarchical coordination [116, 118, 232]; debate [213, 240]; modular pipelines [53, 119]; competitive selection [239]; mixtures of fine-tuned experts [42]; and population ensembles of independent agents, as in MASS, whose rank IC rises with population size [74]. With roles fixed, hierarchy gives the best overall risk-adjusted return under all three backbones of the crypto portfolio multi-agent system (MAS), though not in every capability configuration [141].

The evidence comes in three grades (Figure 4a), and none establishes a multi-agent advantage. Prior-system comparisons (*vs. prior*) re-run others’ published systems, confounding organization with data, LLM version, prompting, and window; *ablations* remove one component within the same system; and matched single-agent controls (*vs. single*) give a single agent the same inputs, LLM, and window. Over short windows without paired tests (\$7.5), the gaps are neither established nor refuted. Prior-system gains are large and ablations credit the ablated mechanism, but the three matched single-agent controls disagree: AlphaAgents’ multi-agent portfolio beats its single agents under risk-neutral prompts, though under risk-averse prompts every variant trails the benchmark [240]; Geng et al.’s Macro Economists debate agent adds no Sharpe over the better single persona [63];

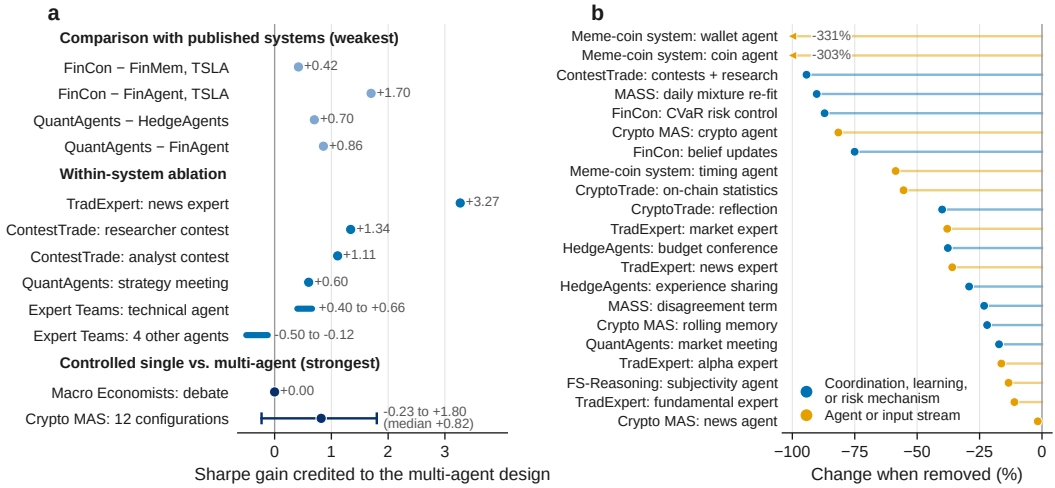


Fig. 4. Evidence for the multi-agent designs in Table 3: (a) Sharpe-ratio gain credited to the multi-agent design, by evidence grade; (b) change in headline return when one component is removed, as a percentage of the full system’s value [42, 63, 74, 116, 118, 119, 140, 141, 147, 200, 232, 239]. No gap carries a standard error.

and the crypto portfolio MAS’s headline gain is the best of twelve single-run configurations, while two of its three chain-of-thought architectures trail the matched single agent [141].

One reading reconciles them: multiple agents help when they change how information is processed, not when they multiply opinions on identical inputs. The Macro Economists’ personas share inputs, so debate can only average two priors, whereas the crypto portfolio MAS’s agents split modalities. General-domain results agree: in a common harness no debate protocol reliably beats self-consistency [178], and majority voting over independent samples outranks debate [111]. In Agent Market Arena’s two months of live paper trading [161], architecture rather than backbone drove outcome variation, yet the debate hierarchy was the most backbone-sensitive (−38.7% to +21.9% on one stock).

Gains trace to structure, input design, and risk mechanisms, and the best choice depends on the backbone. ContestTrade’s gains largely vanish without its selection contests [239]. In Miyazaki et al.’s Expert Teams, the cleanest agent-level ablation (fifty post-cutoff runs per configuration on TOPIX 100), analysts given pre-computed indicators and explicit criteria beat analysts given raw data, and removing any of four non-technical agents *raises* the Sharpe ratio, by up to 0.5 [147]; in the crypto portfolio MAS, pre-computed indicators suit GPT-4o, whereas GPT-5 favors chain-of-thought and Claude retrieval [141]. Removing FinCon’s conditional value-at-risk control, HedgeAgents’ extreme-market conference, or QuantAgents’ risk-alert meeting causes some of the largest ablation losses [116, 118, 232].

Costs and error propagation are under-measured: costs scale with universe, frequency, and debate rounds (\$8.5), and debate can converge on a confidently wrong consensus [46]. Nguyen and Pham price coordination at roughly \$0.50–2.00 per daily decision for seven agents and argue from an uncalibrated breakeven test that it pays only when the half-spread is below half the expected execution-price improvement, so mainly liquid instruments qualify [148]. Outside finance, 1,642 annotated multi-agent traces attribute failures to system design (44%), inter-agent misalignment (32%), and missing verification (24%), and an objective-level verification step raised task success

by 15.6% [30]. No surveyed trading paper analyzes convergence or disagreement handling, and AlphaAgents’ risk-seeking and risk-neutral agents behaved indistinguishably [240].

Takeaway. Multi-agent organization is an engineering pattern for partitioning information and inserting checks, not yet shown to beat a single agent given the same inputs, backbone, and window. The open question is which input partition and which check buy the most risk-adjusted return per token, which needs multi-seed, matched-input ablations on independent benchmarks.

5 Key Techniques for LLM Trading Agents

We organize this section by technique rather than module, because most techniques serve several capabilities of §4; collaboration is treated in §4.3.

5.1 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) counters hallucination and stale training data with external knowledge [107], but its evidence is strong for factual question answering and weak for trading. Retrieval quality bounds the gains: GPT-4-Turbo fails most FinanceBench questions from a shared vector store yet answers most once the relevant filing is in context [88], and hallucination detectors lose much of their discriminative power once noisy triplets enter retrieval [102], so retrieval relocates the reliability problem rather than removing it (§9.1, §9.2). Trading systems bundle retrieval with fine-tuning or multi-agent design [115], and only FactFin [117] isolates retrieval’s contribution to returns (8.8% to 13.4% on AAPL). Prices are better served by tool calls, which answered stock-price queries with 100% accuracy [93].

5.2 Prompting and Reasoning Strategies

General reasoning capability does not predict trading performance: structure and input design matter alongside the backbone, and the best structure depends on the backbone. Domain blueprints help weak models: FinCoT lifts Qwen3-8B-Base from 63.2% to 80.5% on CFA-style questions, but the gain vanishes or reverses for models with strong built-in financial reasoning [150]. In direct cryptocurrency trading, stronger LLMs sometimes underperform weaker ones by over-weighting facts relative to sentiment, until FS-ReasoningAgent’s fact–subjectivity split largely restores the expected ordering [200]. Nor do reasoning-tuned variants reliably beat their instruction-tuned counterparts on return or drawdown [33]; LMArena rank does not predict live portfolio performance [229], and a 135M-parameter trading-trained model beats GPT-4 agents (§5.3). Reasoning distillation reliably lifts financial reasoning *benchmarks* [131, 162, 244], though Fin-R1’s ablation shows that RL without a supervised stage yields incoherent reasoning even as accuracy rises, and general reasoning models regress on financial tasks relative to their instruction-tuned bases [131]. General leaderboards are therefore a poor guide to backbone choice: only Trading-R1 [212] backtests a reasoning-trained trading decision, and no surveyed study varies scale with structure held fixed.

5.3 Fine-Tuning and Domain Adaptation

Fine-tuning pays at the perception and action ends of the pipeline and fails out of distribution. At the perception end, LLMs fine-tuned on newsflow beat conventional sentiment scores as long-short signals [73], and FinLoRA adapters on an 8B model lift an aggregate financial-task score from 37 to 75 (zero-shot GPT-4o: 63) [197]. At the action end, FLAG-Trader trains SmoLLM2-135M as a reinforcement-learning (PPO) policy on Sharpe-ratio change and beats GPT-4 and GPT-4o agents for four of six assets [217]. Out of distribution, Palmyra-Fin-70B leads InvestorBench on ETFs yet loses 20.8% on Bitcoin [109], two finance-tuned models fall to 10–12% accuracy on FinCall-Surprise by over-predicting negative labels and breaking the answer format [176], and tuning on historical news raises in-distribution accuracy by about twenty points while cutting unseen-data accuracy by

18–22% relative [117]. An agent that operates across regimes therefore cannot treat in-distribution gains as skill.

5.4 Tool Use and External Integration

Tools help most when they move computation into a deterministic, verifiable engine. The Option Query Language (OQL) [139] *narrows* the action space to typed queries that an engine validates against the option chain (§6.3), whereas LLM-built stochastic-differential-equation models *widen* it, lifting an LLM trader’s mean Sharpe ratio from 0.88 to 1.40 across four stocks (as reported; level with buy-and-hold) while hurting weaker models (Table 12) [47]. Typed interfaces thus lower the skill the model needs, whereas open-ended tool construction pays only for strong models; in both cases numerical errors survive delegation (§9.1).

5.5 Memory Architectures

Memory’s trading evidence comes from ablating the *update loop*, not storage. Agents use layered memory (FinMem [231]), reflection on prior outcomes (CryptoTrade [119]), or verbal reinforcement (FinCon [232]). Removing reflection cuts CryptoTrade’s bull-market return from 28.5% to 17.1% (Table 9), and FinCon’s belief updates add 25–85 percentage points of cumulative return. Layered tiers amount to paging policies [152] with hand-set decay constants, yet no surveyed trading agent tests whether they beat a flat store, and no open benchmark evaluates memory modules directly [238]. The main open problem is learning when to forget: pruning stale market beliefs while keeping risk principles [85].

5.6 Perception Techniques: Sentiment, Multimodal, and Alternative Data

Every ablated perception modality helps, and its mechanism predicts its durability. GPT-4 headline sentiment earns a pre-cost long-short Sharpe ratio of 2.97, falling from 6.54 in late 2021 to 1.22 by early 2024 as LLM adoption spread [134], and on close to a million articles OPT sentiment beats FinBERT and the Loughran–McDonald dictionary after costs (long-short Sharpe 3.05 vs. 2.07 and 1.23) [98] (head-to-head evidence in §8.1). Adding charts and reflection raises GPT-4o’s return in a virtual market from 33.7% to 47.7% (charts alone: 35.8%); the authors attribute the chart gain to weak reading of numerical text [142]. Dropping on-chain statistics costs CryptoTrade more return than dropping news (28.5% falls to 12.7% versus 19.7%) [119]. Sentiment alpha thus decays with adoption, chart gains compensate for a numerical-reading deficiency that better tooling should erode, and on-chain data exploit crypto’s transparency, but their persistence is unmeasured. No surveyed system models each source’s staleness, and perception remains the least benchmarked module.

6 Applications Across Financial Markets

Markets differ in price drivers, action space, microstructure, and available test data; within each market we group systems by how they respond to these constraints. Fixed income and market making are omitted because no surveyed LLM-agent study trades bonds or makes markets; market makers appear only as simulated personas [133].

6.1 Equity Markets

Equities pair the richest text with the simplest action space and are the default testbed [40] (Table 10). Traders differ mainly in where they adapt (memory [231], organization [213, 232], or weights [217]), and three regularities follow. First, richer inputs and rationales beat their own ablations without lifting direction calls far above chance: training SEP to explain lifts accuracy over a binary-only variant (40.8% to 44.1%), yet the full model reaches only 51.4% (Matthews correlation 0.03) [99],

and extra specialist analysts can lower Sharpe (\$4.3). Second, the locus of adaptation sets cost: TradingAgents makes 11 LLM calls and 20+ tool calls per decision [213], whereas weight-trained designs pay once in training and infer with 7B or 135M models [99, 217]. Third, rigor tracks portfolio scope: no single-stock system prices transaction costs, and their tests span 3–8 months (Table 11), whereas portfolio systems such as MarketSenseAI 2.0 [53] more often state costs and run longer. Equities thus supply nearly all architectural variety but mostly short-window evidence. The one leakage-controlled, fifty-run study, Expert Teams [147], trails its index on net Sharpe while every blend with the index beats both, so its demonstrated value may be diversification rather than alpha.

6.2 Cryptocurrency and DeFi

Crypto adds 24/7 trading, a public ledger as data, and direct protocol actions. Exchange agents resemble equity agents: CryptoTrade beats buy-and-hold in only four of nine asset-regime cells [119], and the strongest post-cutoff result is fragile. Over 2025, after all backbones’ cutoffs, the best configuration of the crypto portfolio MAS returns 133.5% net of costs against −3.4% for holding Bitcoin, but with a deeper drawdown (39.3% vs. 29.7%) and one run per configuration [141]. On-chain perception is further along than execution: the TIM framework infers DeFi transaction intent better than the best supervised classifier [145], whereas Alqithami advises high-value systems to stop short of autonomous on-chain signing [2]. Deployed practice runs ahead of this caution: ElizaOS lets agents hold wallets and swap tokens [196], yet an instruction planted in its shared memory can trigger a 50x-leverage copy-trade [157] (\$9.2). Maximal extractable value (MEV) is an exposure, not a strategy: we find no LLM-driven extraction [67], and since replay front-running needs no insight into the victim’s logic [163], reasoning cannot protect agents that submit swaps. In crypto, LLMs thus add the clearest value in reading the ledger, while on-chain autonomy lags by design; where deployed frameworks bypass that design, shared memory becomes the attack surface.

6.3 Emerging Entry Points: Forex and Derivatives

Apart from HedgeAgents’ USD/CNY sleeve [116] and QuantHarness’s cost-free direction calls on futures [216], no end-to-end LLM agent trades forex or derivatives in the studies we found; LLMs enter at the signal, factor, allocation, or order level. We conjecture that forex agents are scarce because high-frequency institutions dominate at latencies far below LLM inference times and major pairs’ lower volatility limits sentiment-driven alpha. In forex, GPT-4o labels of 174,820 G10 macroeconomic releases yield a long-short portfolio that beats carry and value on Sharpe, though without an agent loop or costs [89] (Table 12). In futures, forty GPT-generated factors earn significant alpha on Chinese futures [35], but against non-LLM rules on identical inputs the edge is thin (\$8.1): QuantHarness’s agents trail the better of linear regression and XGBoost on Sharpe for every instrument with baselines [216]. In options, OQL [139] translates rather than trades, lifting win rate to 60.9% from 52.8% for Text-to-SQL. Every entry point thus leaves pricing, Greeks, and margin arithmetic to deterministic code (\$5.4), and no derivatives system shows a robust, cost-adjusted gain over a non-LLM rule where one is tested.

6.4 Prediction Markets

Prediction markets reward LLMs’ text synthesis, yet agents lose when forecasts must beat prices. Forecasting approached superforecaster level within about eighteen months. On ForecastBench’s unresolved questions, frontier LLMs in July 2024 matched an inexperienced crowd (Brier 0.122 vs. 0.121) but significantly trailed superforecasters (0.096) [94]; by late 2025 the AIA Forecaster statistically matched superforecasters yet trailed the market price on 1,610 liquid-market questions, and only its ensemble with the market beat the market [3] (Table 12). Trading profit is harder:

in Prediction Arena all six frontier models lost real money on curated Kalshi markets [236], and PolyBench’s two profitable models lose their edge at larger order sizes [34]. Positive results exploit structure: an agent that recovers logical relations among Kalshi contracts nets 14.1% in two months of fee-aware arbitrage, though five of 46 trades supply 76% of the profit [29]. Because the price already aggregates what an LLM reads, agents profit only from information the market lacks or from unpriced structure.

6.5 Summary

Maturity falls from equities through crypto to derivatives, prediction markets, and forex (32, 10, 6, 3, and 2 of the 47 systems in Figure 3; multi-market systems count in each), and one rule explains the ordering: LLM agents go furthest where text moves prices and actions are simple, and elsewhere they only interpret (releases, news, intents, contract relations) or propose (factors, views, queries) while deterministic code executes [29, 105, 116, 139]. As in prediction markets, credible profits come from unpriced structure (contract relations, manipulation patterns) rather than better prediction [3, 29, 140]. Cross-market work confirms the rule: HedgeAgents [116] allocates across Bitcoin, Dow Jones stocks, and USD/CNY around a classical allocator, with the LLM supplying views and hedging triggers (Table 10). Adding asset classes is easy; robustness is not, since an agent beating a US index can trail the SSE 50 and lose in crypto [52].

7 Evaluation Frameworks and Empirical Analysis

What a reported result can claim is bounded by the instruments and protocol behind it, so we examine metrics, environments, and benchmarks first, and then the results and the protocol pitfalls that explain why author-run and independent results diverge.

7.1 Evaluation Metrics

The standard metrics measure return and risk but omit what decides whether a return is deployable (Table 8 defines the terms). Papers converge on cumulative and annualized return, the Sharpe ratio, and maximum drawdown; Alpha Miners add the information coefficient and its variants [201, 239]. Among the surveyed studies, turnover appears in only two [101, 113], capacity in one [134], and factor exposure in three [53, 89, 113], while system metrics (token cost, latency) rarely appear beside returns. Without turnover, factor exposure, and cost beside returns, most headline Sharpe ratios in this literature cannot be distinguished from a cost-free market or momentum bet.

7.2 Backtesting Frameworks and Simulation Environments

Environments trade execution fidelity against accessibility. FinRL [129, 130] and FinRobot [222] model no execution, ABIDES [25] simulates exchanges for traditional agents, and StockSim [153] pairs limit-order-book execution with LLM agents. Shared-market simulators [133, 142, 224, 234] capture the price impact that single-agent backtests ignore: TwinMarket’s persona-driven agents reproduce fat tails and volatility clustering better than classical heterogeneous-agent models [224]. Macro-level realism, however, does not imply human-like micro behavior. LLM populations in experimental markets price near fundamentals where human subjects produce bubbles [80], and persona-conditioned agents reproduce loss aversion but never migrate toward fundamentals as prices diverge from value [123].

Takeaway. No environment yet combines execution friction, multi-agent price impact, and multimodal inputs. LLM simulators reproduce what prompts and personas encode rather than human behavior, so without per-model behavioral validation they cannot replace human-subject experiments in market design.

7.3 Major Benchmarks and Datasets

Benchmarks form four groups, the first three progressively closer to trading, and success in one does not carry to the next. **Knowledge and reasoning benchmarks** such as FinBen [214] involve no trading and can mislead: on FinCall-Surprise’s skewed labels, most models exceed 70% accuracy yet score below 0.55 in macro-F1 [176]. **Agentic research benchmarks** locate the bottleneck in retrieval: on Finance Agent Benchmark the best model reaches 46.8% accuracy at \$3.79 per question [15], and models without search score zero on FinSearchComp’s time-sensitive questions [83]. **Trading-specific benchmarks** evaluate decisions under market conditions: StockBench [33] and InvestorBench [109] backtest, LiveTradeBench [229] and AI-Trader [52] trade on live data, and the Agent Market Arena [161] places published architectures under one live paper-trading protocol. **Domain-specific benchmarks** target segments such as on-chain tasks [71] and prediction markets, from scoring unresolved questions [94] to replaying order books [34] and trading real capital [236].

Takeaway. Neither QA performance nor forecasting accuracy reliably predicts trading profit [34, 236], and trading benchmarks cover few asset classes over short periods, so an integrated, cost-aware evaluation of the full perception–reasoning–action loop against market prices is still missing.

7.4 Empirical Results Summary

Author-run comparisons in Table 10 favor multi-agent and multimodal designs, whereas independent evaluations find thin, regime-dependent edges. FinCon [232] and TradingAgents [213] beat their baselines on return and Sharpe ratio, and MarketSenseAI 2.0 [53] returns 125.9% on the S&P 100. By contrast, 8 of StockBench’s 13 generic agents beat buy-and-hold, mainly through lower drawdowns, yet all trail it in the downturn window [33], and none of AI-Trader’s live agents profits in crypto [52]. Re-running published systems gives the sharpest contrast: FINSABER, Profit Mirage, and Alpha Illusion shrink or erase the author-run margins [113, 117, 228]. In Profit Mirage’s matched windows on either side of the training cutoff, five agents, including Wang et al.’s QuantAgent [201] (the only Alpha Miner re-run, about 51%), lose 51–62% of their Sharpe ratios [117].

Read beside its baseline (Figure 5), each headline conceals two problems. First, the largest margins come in windows where the passive benchmark lost money or barely gained (FinMem’s 61.8% on TSLA against a –18.6% buy-and-hold), so they largely reflect loss avoidance; in rising markets, MarketSenseAI 2.0 draws down deeper than the index and Expert Teams trails the TOPIX 100 [53, 147, 231]. Second, costs are often undisclosed: only 8 of the 19 author-run systems in Table 10 state a transaction-cost value, all from 2025–2026 (Appendix A11).

Takeaway. Where published agents have been independently re-run over longer, post-cutoff, or cost-charged windows, their margins shrink or vanish. The finding is narrow. The re-runs cover a handful of early, mostly single-stock systems. Two of the three re-evaluations are preprints by authors proposing competing designs, so independent does not mean disinterested, and re-evaluators target the most prominent results, inviting regression to the mean. Because many original margins lie within the sampling error of their short windows (\$7.5), the re-runs show that margins are not reproduced, not that a real edge decayed. Neither author-run nor independent results yet establish that LLM agents beat passive strategies, or that they cannot.

7.5 Evaluation Protocols and Their Pitfalls

We audit the protocols behind this divergence on six dimensions (Table 11). Five pitfalls recur; correcting the three that re-evaluations have tested lowers headline Sharpe ratios (Figure 6), and the evaluation mode bounds the execution realism of any result (Figure 7).

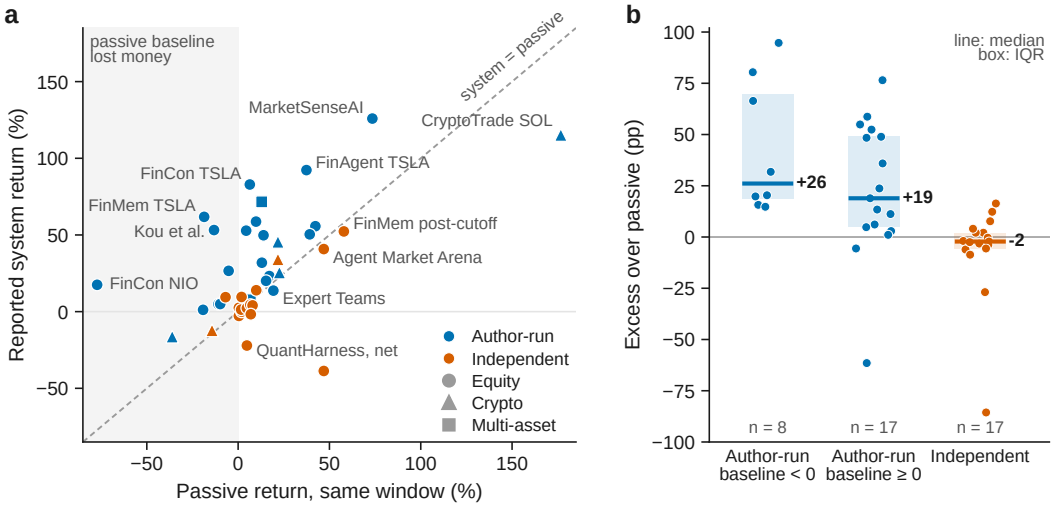


Fig. 5. Headline returns against the passive baseline each paper reports over the same window (Table 10). (a) Points above the dashed line beat their baseline, and shading marks windows in which the baseline lost money; (b) excess return by evaluator and baseline sign, with medians that are descriptive only.

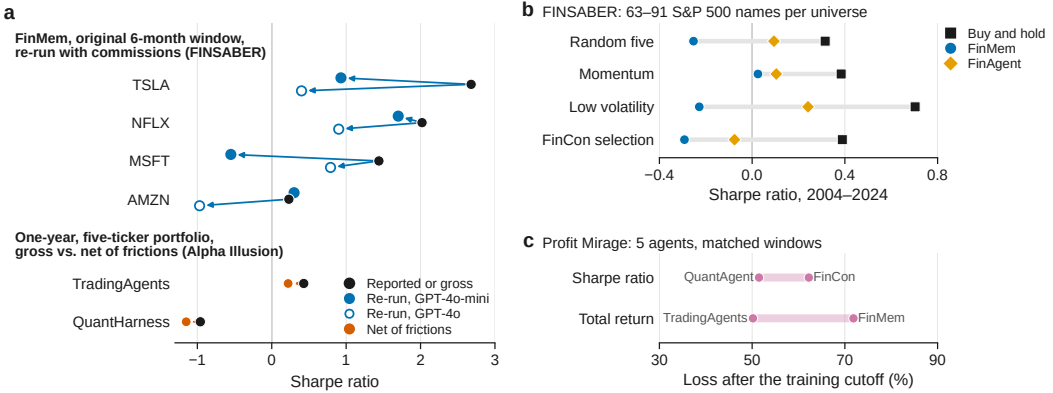


Fig. 6. Headline Sharpe ratios under independent re-evaluation. (a) As reported versus re-run with costs [113, 228]; (b) FINSABER's 2004–2024 composites under four stock-selection rules [113]; (c) the range of Profit Mirage's post-cutoff decay across five agents [117].

Short windows and small universes. Eight of the seventeen dated author-run evaluations in Table 11 last eight months or less, several in rallies, and single-stock agents trade three to eight news-rich large caps, which reveals nothing about breadth or capacity. Lengthening the window erases the edge: over 2004–2024 with commissions on survivorship-free S&P 500 universes, FinMem’s 2.68 Sharpe on TSLA becomes 0.93 or 0.40 depending on the backbone, and buy-and-hold beats both FinMem and FinAgent on every composite universe [113]. TradingAgents’ authors themselves flag their 5.6–8.2 Sharpe ratios as artifacts of a period with few pullbacks [213].

Unmodeled costs. Costs can reverse conclusions. Charging full frictions over one year on five tickers cuts TradingAgents’ Sharpe from 0.43 to 0.22, below equal-weight buy-and-hold, and 35 of the 40 system-by-friction cells in that audit of five systems are unmodeled [228]. Every reported gross Sharpe is therefore an upper bound on deployable net Sharpe; ML return predictability in traditional finance likewise attenuates under realistic costs [7]. Even stated rates are flat charges that ignore how impact grows with capital.

Contamination and look-ahead. An LLM pre-trained over the test window has already read the news, and possibly the prices, it trades on. Profit Mirage traces its post-cutoff decay to recall: backbones recall dated prices, trends, and event reactions with 85–93% accuracy [117]. Yet only 26.8% of 164 main-conference LLM-finance papers discuss look-ahead bias [100]. Instructions to ignore the future and anonymization both fail across models and tasks [114, 135, 168]: GPT-4o told to ignore post-2010 data forecasts GDP direction as well after that date as before (98.0% vs. 97.6%, against 40% past its real cutoff) [135], and stronger models re-identify 70% of name-censored earnings calls [168]. Only post-cutoff evaluation, live evaluation, and chronologically consistent backbones [79] control leakage by construction (mitigations in Appendix A7.5).

Single runs and self-selected baselines. Decoding is stochastic, yet only 5 of the 28 studies in Table 11 report repeated runs, and two runs of one FinRL-DeepSeek configuration reorder the agents [14]. Baselines are self-selected, and raw buy-and-hold is the wrong benchmark for agents that vary their exposure (§8.4); the baseline should match the agent’s average beta. The fifty-run rank tests of Expert Teams [147] and the block bootstrap of Macro Economists [63] are the first significance tests on return differences among the surveyed studies.

Statistical power and selection. Even a correctly run backtest over these windows cannot separate most agents. Under iid returns, the standard error of an annualized Sharpe ratio from T years sampled q times a year is about $\sqrt{(1 + SR^2/2q)/T}$ [132], roughly 1.4 for six months of daily returns; FinMem’s 2.68 on TSLA over six months thus carries a 95% interval of roughly -0.1 to 5.5 , which already contains FINSABER’s re-run values, and serial correlation and fat tails widen it further [10, 132]. A margin over a baseline, being a difference of two such estimates, has a standard error of about 2.0 over six months for uncorrelated series; by our calculation, 9 of the 14 author-run Sharpe gaps over a passive baseline in Table 10 lie within two such standard errors. The same noise bounds the corrective evidence. Published equity anomalies lose 26% of their returns out of sample, an upper bound on data mining, and 58% after publication [146]. Profit Mirage’s 51–62% decay exceeds the former and matches the latter, but it is one correlated draw of five agents on one pair of six-month windows, so decay from post-publication arbitrage cannot be separated from the loss of memorized information. Selection compounds the noise, because the expected maximum Sharpe ratio of N zero-skill configurations rises with N [10]. For example, a seasonal rule optimized on a pure random walk reaches an in-sample Sharpe ratio of 1.27 that a single-trial test calls significant, yet its probability of backtest overfitting (PBO) is 55% [9]. The deflated Sharpe ratio and PBO correct for both effects only given the number and returns of all configurations tried [9, 10], which no system in Table 11 reports.

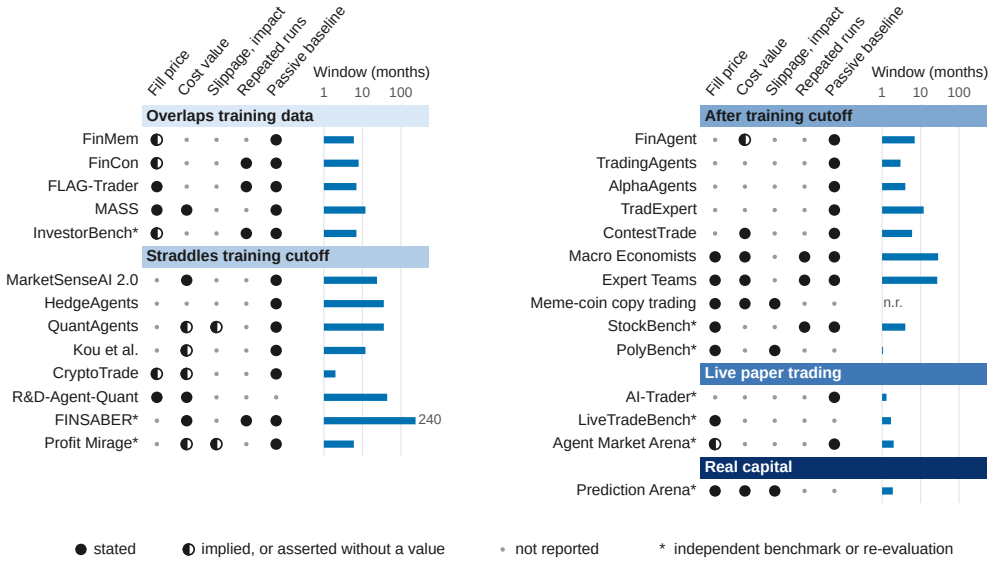


Fig. 7. Execution disclosure by evaluation mode for the systems of Table 11, ordered by execution realism. Marks record whether fill price, cost, slippage, repeated runs or tests, and a passive baseline are stated (filled), inferred or asserted without a value (half), or absent (empty); bars give test windows on a log scale.

Evaluation mode and fill realism. Evaluation modes form a ladder (Figure 7) from pre-cutoff backtests through post-cutoff backtests and live paper trading to real capital, which only 1 of the 28 studies in Table 11 reaches. In Prediction Arena, orders were rejected for lack of counterparties, and spread and fee drag were large enough that skill cannot be separated from execution [236]. Paper trading removes look-ahead but not execution: it fills at closing prices or scoring signals without sizing or costs [161, 229]. Live paper trading nonetheless reorders agents: risk control rather than reasoning quality separates robust from fragile agents in AI-Trader, and general-capability rank does not predict live portfolio quality in LiveTradeBench [52, 229]. Below the live rungs, only 6 of the 18 author-run systems state a fill rule, and none reports fill rates or partial fills; in PolyBench’s order-book replay, both profitable models’ alpha collapses as order size grows from \$10 toward \$1,000 [34].

Corrections and audits agree. Each correction in Figure 6 leaves at least one corrected agent trailing buy-and-hold, though their combined effect is unknown, and Xia et al.’s audit of 185 studies [210] finds reporting improving but incomplete, so the proportions in this section describe the most-cited early systems rather than current practice. Published checklists [61, 100, 210, 228] converge on a minimum: a post-cutoff or live window spanning a downturn, the full universe with capacity analysis, explicit cost and slippage, run counts and dispersion, the number of configurations tried, and exposure-matched passive plus tuned active baselines.

Takeaway. The headline returns in Figure 5 are existence proofs under favorable conditions, not estimates of expected performance; even live returns are upper bounds conditional on full fills, and no live window yet spans a market cycle. The infrastructure exists (§7.3); what is missing is a norm that labels claims below this minimum as exploratory.

Table 4. LLM trading agents vs. traditional quantitative trading across five dimensions, with the figures reported in the surveyed studies. Costs are LLM API spend as stated by each paper.

Dimension	Traditional Quant	LLM Trading Agents
Data	Structured price, volume, order-book, and factor data; text enters through hand-engineered sentiment features that do not significantly predict next-day returns ($t = 1.87$) where LLM scores do (Loughran-McDonald [136] long-short Sharpe 1.23 vs. 3.05 for OPT) [98].	Text, numbers, and chart images without task-specific features [237]; headline sentiment alone yields a pre-cost Sharpe of 2.97 over 2021–2024, falling from 6.54 (2021Q4) to 1.22 (2024) [134]. Weak at exact arithmetic: agents miscalculate share quantities and read absolute levels over percentage changes [33, 142].
Interpret.	Factor strategies are transparent by construction; DRL policies (A2C, PPO, SAC) are opaque.	Natural-language theses [212], but 69–82% of decisions survive counterfactual changes to events, prices, and fundamentals, so stated rationales often do not drive the decision [117].
Adapt.	Regime-specific calibration; buy-and-hold Sharpe 0.61 in bull years, -0.28 in bear years over 2004–2024 [113].	In-context adaptation and reflection [119, 200], yet FinAgent earns 0.12 in bull years and -0.38 in bear years on the same sample [113]; model rankings reorder between downturn and upturn windows [33]. Robustness in extreme markets comes from hard-coded conferences and thresholds [116, 118].
Latency and cost	Order-book time scales, simulated at nanosecond granularity [25]; execution schedulers such as TWAP and Almgren-Chriss run at minute granularity [174]; no per-call inference fee.	Seconds to minutes per decision: 4.7 s per stock on one GPU [42]; 11 LLM and 20+ tool calls per decision [213]; one decision per 15–45 minutes in live prediction-market agents [236]. API spend: \$15 (three asset classes) and \$180 (NASDAQ-100) over three-year daily backtests [116, 118]; \$0.68–2.27 per trading day for 50–300 stocks [74]; \$198 vs. \$32 for a twenty-year composite re-run of FinAgent vs. FinMem [113]; about \$30 for 1,680 execution orders [174].
Scale	Linear to thousands of instruments; FINSABER re-runs agents on historical S&P 500 universes of 63–91 names including delisted stocks [113].	The most-cited single-stock agents (FinMem, FinCon, FinAgent, TradingAgents) trade three to eight names (Table 11); returns fall and volatility rises from 10 to 30 stocks, though 5-stock portfolios fare worst [33]; FinCon’s authors note hallucination becomes likelier as inputs lengthen [232]; a 512-agent ensemble exceeds the context window on CSI 300 without per-agent stock subsets [74]. Exceptions cover the S&P 500 with monthly signals [53].

8 LLM Agents vs. Traditional Quantitative Trading

This section compares LLM agents with traditional quantitative trading on reported evidence rather than claimed capability. Table 4 contrasts the two approaches on five engineering trade-offs; because it draws mainly on LLM agent papers, we first consolidate the head-to-head record and then judge each dimension (\$8.2–\$8.5).

8.1 Head-to-Head Evidence

Signal-level evidence is the strongest; author-run agent leads are common but rarely net of costs. LLM text signals beat dictionary [136] and FinBERT [6, 84] sentiment on return predictability and Sharpe ratio in most cells across groups, data sets, and model families (Table 5) [73, 98, 134], though the edge decays with adoption (§5.6). Of the 17 agent-level studies, 11 lead their strongest traditional baseline, 4 are mixed, and 2 trail. Only four leads hold at a stated cost rate (three of them by Alpha Miners or bounded components, one not significant), and only three studies report a significance test (Appendix A8.1). Where agents lose, the winner is often a simple rule or passive holding [119, 217]: QuantHarness, for example, trails linear regression or XGBoost on Sharpe for all seven instruments [216]. Wins over deep reinforcement learning (DRL) baselines are confounded, as several A2C and PPO baselines see only prices and use default settings.

Independent evaluation shrinks the end-to-end edge. Re-runs of published systems shrink or erase author-run margins over passive baselines [113, 117, 228] (§7.4). Contamination-controlled benchmarks agree: generic single-agent workflows beat passive baselines by at most about two points in StockBench and lose in its downturn [33], and in AI-Trader they beat no index in A-shares and make no profit in crypto [52]. Alpha Miners are not exempt: the one re-run decayed past

Table 5. Head-to-head outcomes of LLM-based approaches against the strongest non-LLM baseline each paper reports (passive included), on its headline metric (risk-adjusted if reported, else return; mean over backbones if the paper designates none). Codes use point estimates over cells (assets, universes, settings): *ahead* wins every cell, *behind* none, *mixed* (k/n) wins k of n ; n.s.: the paper’s own test finds the margin insignificant at 5%. Cost flags (*net*: the code holds at the stated rate; *gross*; *exec.*: the metric is itself an execution cost; *unspec.*) do not change the code; n.c.: not coded; n.r.: not reported. Functional paradigms (A, B) follow Figure 3b; rows are author-run unless marked ext. (external scoring). Windows, baselines, and figures: Table 12.

Study	Market	Traditional baselines	Code	Costs	Note
<i>Signal-level studies (LLM as perception or predictor)</i>					
Lopez-Lira and Tang [134]	US stocks	FinBERT [6], market	Ahead (1/1)	net	10 bps; unprofitable at 20 bps; edge decays
Kirtac and Germano [98]	US stocks	Dictionary [136], FinBERT [84], market	Ahead (1/1)	net	10 bps
Guo and Hauptmann [73]	NA, EU, EM equities	FinBERT [6], FinVader, universe	Mixed (4/6)	gross	Loses NA and EM long-only to FinBERT
Alpha-GPT [202]	Quant competitions	Human quant teams	Behind (0/2)	n.r.	ext.; top ten but below each winner
Izadyar [89]	G10 currencies	Carry, value, momentum	Ahead (1/1)	gross	No surprise-rule baseline
Makri et al. [144]	ETH	PatchTST, LSTM, MLP	Ahead (2/2)	–	Forecast only; no trading test
AIA Forecaster [3]	Forecasting questions	Superforecasters, market prices	Mixed (2/4), n.s.	–	Forecast only; ext.; trails superforecasters on FB-Market, the market on liquid markets
<i>Agent-level studies (Trader; A: Alpha Miner; B: bounded component)</i>					
FinMem [231]	5 US stocks	Buy-and-hold; DRL	Ahead (5/5)	gross	Risk setting chosen on test window
FinCon [232]	8 US stocks	DRL; Markowitz	Ahead (10/10)	gross	
FinAgent [237]	5 US stocks, ETH	Rule; ML; DRL	Mixed (3/6)	gross	Ahead on return in 6 of 6
TradExpert [42]	DOW 30	ML; DRL; index	Ahead (1/1)	gross	
CryptoTrade [119]	BTC, ETH, SOL	Rules; deep forecasters	Mixed (1/9)	unspec.	Tops buy-and-hold in 4 of 9 cells
R&D-Agent-Quant (A) [121]	CSI 300/500, NASDAQ 100	Factor libraries; 16 models	Ahead (3/3)	net	No passive benchmark
MASS (A) [74]	Chinese indices	Rule; index; ML; DL	Ahead (3/3)	net	Rank IC agrees
Kou et al. (A) [101]	SSE 50	ML; DRL; index	Ahead (1/1)	unspec.	Turnover-capped; half-years vs. index not coded
ContestTrade [239]	A-shares	Rules; ML; DRL	Ahead (1/1)	net	Larger drawdown than a rule
HedgeAgents [116]	BTC, DJ30, FX	DRL; momentum; buy-and-hold	Ahead (1/1)	gross	2024 hold-out not coded
QuantAgents [118]	NASDAQ-100	DRL; mean-variance; index	Ahead (1/1)	gross	Live track not coded
FLAG-Trader [217]	5 US stocks, BTC	Buy-and-hold	Mixed (5/6)	gross	Ahead on return in 6 of 6; no active non-LLM baseline
FinRL-DeepSeek (B) [14]	NASDAQ-100	PPO; CVaR-PPO; index	Behind (0/1)	gross	Two runs reorder the agents
Macro Economists (B) [63]	Commodity ETFs	Same-input rule; passive	Ahead (1/1), n.s.	net	+0.04 Sharpe; $p \approx 0.07$ –0.09 unadjusted
QuantHarness [216]	Index futures, crude, BTC	Linear regression; XGBoost	Behind (0/7)	gross	Ahead on directional accuracy in 8 of 8
Emmanoulopoulos et al. [47]	4 US stocks	Buy-and-hold	Mixed (2/4)	net	Mean over backbones; best backbone wins 4 of 4
PACE (B) [174]	SZSE execution	TWAP; Almgren–Chriss; ML	Ahead (2/2)	exec.	Execution cost; sig. vs. TWAP only; no market impact
<i>Not coded (abstract only)</i>					
Cheng et al. [35]	Chinese futures	IPCA factors	n.c.	n.r.	Reports significant alpha; magnitudes n.r.

the backbone's cutoff like the Traders, while R&D-Agent-Quant's out-of-sample lead over factor libraries and sixteen machine-learning and deep models [121] remains a single author-run study. Because Alpha Miners score many candidates and report survivors, they automate the multiple-testing ("factor zoo") problem, under which a new factor needs a t -ratio near 3.0, not 2.0 [76, 205]; yet no Alpha Miner study we reviewed reports its candidate count with a selection correction (§7.5). Independent and author-run results agree only that LLM agents, holding less exposure, show shallower drawdowns in tested windows [33, 116]; in sustained declines they do not protect capital better than passive exposure [33, 113].

Takeaway. The LLM results not yet contradicted all feed or bound a conventional pipeline. With inputs matched, the decision layer adds little: Geng et al.'s Macro Economists beat a same-input rule by an economically small margin, significant only at the 10% level [63], and PACE beats TWAP by about one basis point (95% CI 0.15–2.12) by tilting a classical schedule [174]. Each such result yields a score, formula, or bounded tilt that a conventional backtester or matcher evaluates. Autonomous LLM traders beat their chosen baselines in most author-run tests, but only one of their six leads in Table 5 is net of costs, and they lose ground under independent re-evaluation.

8.2 Data Dependency and Information Processing

LLMs win on text and traditional systems on numbers; the text edge is well evidenced but perishable. LLM agents take text, numbers, and images directly [237], and the edge is largest when the LLM extracts features rather than decides: embeddings of full news articles fed to a conventional return regression beat past-return, dictionary, and word-embedding signals across 16 markets and 13 languages [32], although FinBERT still wins on emerging-market long-only Sharpe [73]. On numerical forecasting the ranking reverses because of text pretraining rather than the transformer architecture: models pretrained on time series beat prompted or reprogrammed text LLMs in zero-shot forecasting [5, 38], and replacing the LLM in Time-LLM [91] with one attention layer beats it in all 26 dataset-metric cases [181]. Base LLMs prompted with digit strings forecast competitively zero-shot, but the chat-aligned models that agents run on do worse [69].

8.3 Strategy Interpretability and Transparency

LLM explanations are the most legible but not demonstrably faithful, so only an inspectable strategy makes them trustworthy. Unlike opaque DRL policies, LLM agents produce structured theses stating *why* a trade is recommended [212]. Yet 69–82% of decisions survive counterfactual changes to their inputs [117], chain-of-thought explanations can rationalize answers driven by input features they never mention (only 1 of 426 explanations supporting a biased answer names the bias) [186], and an answer's dependence on its stated reasoning often falls as models grow [103]. Alpha-GPT [202, 233] sidesteps the problem: the LLM writes alpha expressions, so auditors inspect a formula, not a narrative. Whether any faithfulness test can certify an end-to-end trader remains open.

8.4 Market Adaptability and Generalization

No evidence shows LLM agents adapting better to regime change; the strongest evidence shows them adapting worse than passive exposure. LLM agents can in principle interpret novel situations and reflect [119, 200], yet CryptoTrade's reflective agent leads the best rule or forecaster in only one of nine asset-regime cells, and over twenty years FINSABER finds LLM timing agents "too conservative in bull markets and too aggressive in bear markets" [113]. Reported extreme-market robustness comes from explicit triggers and hard thresholds, not LLM reasoning [116, 118], so abrupt dislocations still require hard-coded controls.

Table 6. A reference architecture for the division of labor: for each LLM role, the artifact it emits, the deterministic gate it passes, the bound on its action, what to monitor, and the current evidence. Gates, bounds, and monitoring are our recommendations; controls common to every role (testing before deployment and after substantial or autonomous model changes, pre-trade limits, kill functionality, algorithm identifiers [50, 59, 160, 190]) are not repeated.

LLM role	Artifact	Validation gate before use	Bound on action	Monitoring	Evidence
Text and event signals	Score or label per document	Out-of-sample IC or long-short Sharpe net of costs on post-cutoff data, incremental to existing sentiment and price factors, with a capacity curve (§7.5)	Input to a conventional model; no order authority	Rolling net Sharpe by year [134]; return captured at the realized decision lag; drift after backbone updates	Signal-level, long-sample; costed in 2 of 3 [32, 98, 134]
Factor and strategy code (Alpha Miner)	Formula or code	Backtest that counts every candidate tried: deflated Sharpe or PBO [9, 10], a t -ratio near 3 [76]; capacity curve; human review of the formula	Factor enters an optimizer under position limits	Live versus backtest IC; retire decayed factors	Single author-run [121, 182]; the one independent re-run decayed [117]
View or tilt on an optimizer or allocator	Return views with confidence; signals to a policy	Comparison with the same optimizer without views, on post-cutoff data	Optimizer shrinks views toward the prior; position limits	Backbone-specific bias in views [105]	Single author-run, mixed [14, 63, 105]
Execution tilt	Size adjustment around a schedule	Shared-matcher test against TWAP and Almgren–Chriss on identical orders, with an impact model; then transaction-cost analysis (TCA) against arrival price in a capped live pilot	Planner shifts $\leq 30\%$ toward its own weights; Executor sizes within $\pm 50\%$ of each slice [174]	Slippage against the schedule; completion	Single author-run, no market impact [174]
Query or intent parsing	Typed query (e.g., OQL)	Schema and constraint check by a deterministic engine [139]	Engine selects among templated strategies	Validity rate of parsed queries	Single author-run benchmark, no costs [139]
Research synthesis for humans	Report or thesis	Human sign-off; plain description of intended function [59]	No order authority	Faithfulness checks [100]; reliance on unflagged errors [154]	Field evidence on analysts [172]

8.5 Latency, Cost, and Scalability

Traditional systems win decisively below the minute scale and across broad universes; LLM agents are viable only at minute-scale or slower frequencies. Simulators resolve high-frequency trading at nanosecond granularity [25], whereas one LLM inference takes seconds [42, 148] and multi-agent pipelines take minutes, with MASS measuring 125–378 s per trading day [74] (Figure 8b). The practical floor is the minute: PACE tilts each minute over 10–60 minute windows [174], and QuantHarness degrades at 1–15 minute bars [216]. Vendor-reported typed decision models could lower this floor. TypeSafe’s Jev returns a typed decision with a probability distribution and a confidence score in a reported 70–500 ms [1, 187]; if this holds for financial tasks, schema-constrained outputs would remove malformed-order errors (§9.1) and bounded components could move toward intraday frequencies. However, its calibration is untested against realized market outcomes, its training cutoff is undisclosed, no trading evaluation is published, and it forgoes the textual rationale that makes LLM theses legible.

Cost and breadth narrow the niche further: API charges accumulate with calls per decision and with breadth (Figure 8a) [74, 116, 118, 213], and where traditional systems scale linearly to thousands of instruments, LLM agents hit context limits [74]. For text signals, API spend is not the binding cost; capacity, market impact, and the signal’s decay relative to pipeline latency are. Lopez-Lira and Tang’s signal is unprofitable at 20 bps [134], and no surveyed paper reports capacity curves or impact-adjusted returns.

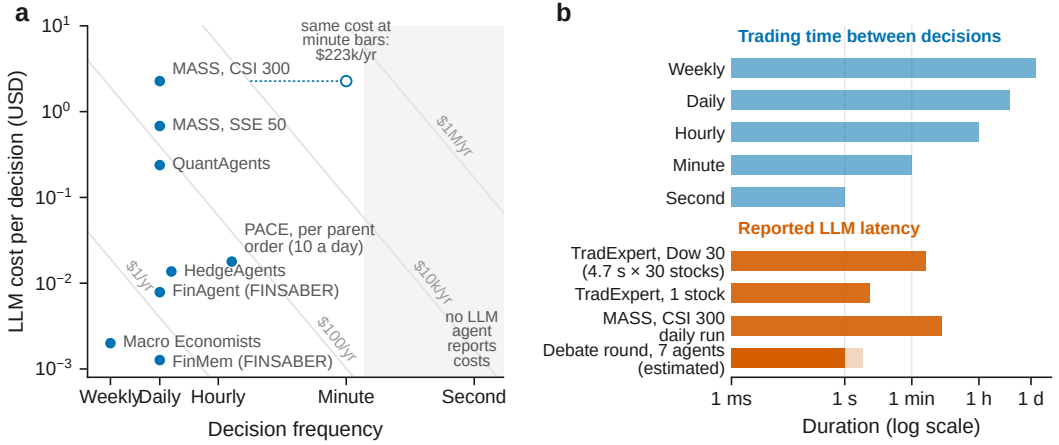


Fig. 8. Cost and latency against decision frequency. (a) Reported LLM cost per decision, with iso-lines of annual spend for one continuously running system; (b) trading time between decisions against reported or estimated LLM latency [42, 63, 74, 113, 116, 118, 148, 174].

8.6 Synthesis

The best-supported current hypothesis is a division of labor: *LLMs produce artifacts that a quantitative system can test and bound; quantitative systems make, size, and execute the trade*. Table 6 turns this division into deployable roles. Support for its three upstream outputs (text signals, factor and model code, and bounded execution tilts) is uneven: text signals rest on long, mostly costed samples, whereas factor and model code validated in a backtester [121] and bounded execution tilts [174] rest on single author-run studies. Within author-run studies, Alpha Miners and bounded components lead their strongest baseline in 5 of 6 studies and Traders in 6 of 11 (Table 5), too few to separate position from baseline choice. We nonetheless conjecture that the pattern is positional: the LLM’s strengths (reading text, interpreting novel situations, writing legible theses and code) lie upstream of the decision, while its liabilities (memorized prices, imprecise arithmetic, unfaithful rationales, regime-blind conservatism, second-scale latency, context-bound breadth) bind at the decision itself.

Being upstream is not sufficient: the Macro Economists gain only marginally over a same-input rule, FinRL-DeepSeek’s LLM signals help one policy and hurt another [14], and upstream outputs carry their own look-ahead and data-snooping risks (§7.5). LLM agents should therefore own information-intensive perception and strategy ideation, and traditional systems should own latency-sensitive execution, numerical computation, risk control, portfolio optimization, and breadth. The hypothesis would be falsified by an independently re-run autonomous trader that holds a costed edge over a passive baseline matched to its market exposure. The decisive missing test is a controlled comparison of hybrid, pure-LLM, and pure-quant architectures.

9 Open Challenges and Future Directions

Seven challenges block reliable production deployment of LLM trading agents. Table 7 merges them with the open questions of §4–§8 into one research agenda for the whole survey.

Table 7. Research agenda consolidating the open questions of §4–§9. Each row pairs an open problem with a concrete next step or measurement.

Open problem	Concrete next step or measurement	§
Hybrid vs. pure-LLM vs. pure-quant designs	One protocol (same data, window, backbone, costs, and exposure-matched baselines) comparing the three LLM paradigms with non-LLM pipelines; the decisive test of Table 6	4.2, 8.6
Value of multi-agent structure	Matched-input, multi-seed ablations on an independent benchmark	4.3
Component gains in a trading loop	Domain-specific vs. generic retrieval, layered vs. flat memory with learned forgetting, and distilled reasoning, each tested in a costed trading loop	5.1–5.5
Capacity and latency of upstream signals	Capacity curves and impact-adjusted returns; returns measured at the agent’s realized decision lag	8.5
Calibrated typed decision models	Test whether vendor-reported calibration [1] holds against realized market outcomes, post-cutoff and net of costs, against LLM verbal confidence and classical classifiers	8.5
Numerical errors that reach orders	Order-level audits of how often errors and skipped tool calls reach executed orders; typed action interfaces validated deterministically against holdings and liquidity; trajectory-level extensions of CNFinBench [41] and HalluBench [102] under noisy retrieval	9.1
Attack success versus portfolio damage	Closed-loop red-teaming with mixed, adaptive, and cross-architecture attacks, scored by Net Resilient Performance (NRP) [235] extended with financial risk measures [219]	9.2
Integrity of tools and shared memory	Authenticated, provenance-tracked tool channels; memory reconciled against an independent ledger; memory-injection tests on trading agents with cross-platform memory [157]	9.2
Production readiness unreported	Staged deployment to capped real capital under the controls of §9.4, reporting monitoring, failover, and API-outage behavior alongside returns	9.3
Text-centric perception	Real-time multimodal and audio ingestion (order books, press conferences) within decision latency	9.3
Validating non-deterministic, vendor-updated models	Re-testing and re-authorization triggered by backbone updates [50, 160]; compliance modules implementing the IOSCO and FSB examination items [58, 87]; audit trails independent of possibly post-hoc reasoning traces (§8.3)	9.4
Liability and scale of autonomous trading	Liability rules for intent-free “coordination outcomes” [125, 165]; measuring how much autonomous LLM trading occurs	9.4
Monoculture and cascades	Cross-agent signal-correlation studies in securities markets varying backbone, architecture, and information set; overlap of LLM-mined factors with published and peer factors; behavior under a common backbone outage or update; LLM-populated agent-based simulations (§7.2) of the cascade to price impact	9.5
Look-ahead and version drift	Shared point-in-time datasets including delisted securities; chronologically consistent backbones whose pretraining stops before each test year [79, 168], scaled up and instruction-tuned to drive agents; reproducibility checklists (§7.5) as submission requirements	9.6
Human-agent division of labor	Joint metrics including tail errors [28, 195]; human-owned mechanical aggregation [170]; trust signals tied to informative agent confidence [220]; sign-off designs tested for automation bias [154]; autonomy calibrated to regime and position size	9.7

9.1 Hallucination and Reliability

Numerical hallucination matters most for trading because it hides in fluent reasoning and costs money once acted on. RAG and tool learning are the most effective mitigations across financial tasks [93], but grounding inherits retrieval quality: noisy knowledge-graph triplets cut detectors’ Matthews correlation by 44–84% for most detector families [102], and with the filing in context GPT-4-class models still give 17–21% incorrect FinanceBench answers [88]. Inside agents the errors reach the order: StockBench agents miscalculate share quantities or emit malformed JSON actions [33]. Delegation through typed queries, schema-constrained relations, code-based calibration, or LLMs that choose only trade direction [29, 47, 77, 139] fixes the steps it covers, but every step around the tool remains a language step. **Takeaway.** Grounding and delegation shrink numerical error without eliminating it, and no surveyed paper measures how often an error reaches an executed order, because question-answering benchmarks score answers and trading benchmarks score returns.

9.2 Adversarial Robustness and Security

LLM agents add attack surfaces along the perceive–reason–act loop, from prompt injection to tool, memory, and protocol exploits [57]. Agent Security Bench reports 84.3% average success for mixed attacks against largely ineffective defenses [235], and on Web3 tasks memory injection beats prompt injection on every tested backbone, while fine-tuning on attack examples cuts its success

from 85.1% to 1.7% [157]. TradeTrap [219] shows that success rates understate financial damage: a reverse-expectation prompt injection cuts an adaptive agent's Sharpe ratio from 5.72 to 0.29, tool hijacking and memory poisoning make it trade a phantom portfolio, and procedural agents lose less only because rigid pipelines cap upside as well as downside. Because attack success rates do not predict portfolio damage, generic agent benchmarks cannot certify a trading agent, and no surveyed study tests combined, adaptive, or cross-backbone attacks in a trading loop.

9.3 Engineering Challenges of Live Deployment

Continuous live operation demands engineering that published systems rarely report. Inference cost and latency confine current agents to low-frequency trading on small universes (\$8.5), and multi-agent designs compound both: TradingAgents makes 11 LLM calls and 20+ tool calls per decision [213]. No surveyed study reports monitoring, failover, or degradation under API outages, and most systems remain text-centric, lacking low-latency multimodal ingestion. Because the cost, latency, data, and fault-tolerance engineering that live trading requires is largely unreported, production readiness cannot be judged from published results.

9.4 Regulatory Compliance and Legal Risks

Existing algorithmic-trading rules already cover LLM agents run by regulated firms; LLMs add uncertainty about when an algorithm has changed and what it is meant to do. EU RTS 6 requires pre-deployment testing, order and position limits, real-time monitoring, and a kill functionality [50], and US rules and guidance add pre-trade controls, change management, and a quick way to disable an algorithm [59, 190]; agents run by individuals or on crypto venues largely fall outside them. Liability is less settled. MiFID II rather than the EU AI Act governs AI trading [8], and US manipulation law requires scienter (intent), yet DRL agents have learned manipulative trading unprompted [165]. Supervisors now specify kill switches, action scope, and human-oversight ladders for agents [20, 58, 87]. Model-risk management fits least, because sampling non-determinism and silent vendor updates mean the validated configuration may not be the one running. Model-risk frameworks nonetheless translate into concrete practice: pin backbone versions and decoding settings, place prompts, memory stores, and tool schemas under change control as model components, and re-run the previously validated configuration as a challenger after each update [18, 160]. In the US, however, SR 26-2, which replaced SR 11-7 in 2026, sidesteps the issue by excluding generative and agentic AI [19]. **Takeaway.** The controls supervisors request are already law for regulated algorithmic traders; what remains unresolved is how to validate non-deterministic, vendor-updated models, who is liable for intent-free manipulation, and how AI claims to retail clients are policed [193].

9.5 Market Stability and Systemic Risk

Automated trading has already produced crises without intent: the August 2007 quant unwind through common factor positions [96], the 2010 Flash Crash [97, 189], and Knight Capital's faulty deployment, which lost over \$460 million in 45 minutes [191]. Agents sharing similar LLMs could reach correlated decisions, but no study has yet shown the full cascade, and regulators find herding evidence thin [86]. Evidence on which kind of diversity matters is emerging: tacit collusion among identical LLM pricing agents [60] falls from about 22% price elevation to 2–10% when objectives, data access, or algorithms differ, but stays near 21% under model-size heterogeneity [95]. **Takeaway.** Crowding, interacting algorithms, and faulty updates can each destabilize markets without intent; diversity in objectives, data, or algorithms decorrelates agents whereas capability asymmetry does not, but this rule comes from pricing games and stylized models and is untested in securities markets.

9.6 Reproducibility and Open-Source Ecosystem

Many systems rely on version-unstable proprietary APIs and unreplicable real-time data, and protocol reporting remains incomplete (§7.5). Open-source frameworks [130, 221, 222] and open-weight models [223] reduce API dependence, yet 74% of surveyed practitioners rate ready-to-use evaluation tools scarce or non-existent, and half call this the main bottleneck to bias mitigation [100]; public leaderboards and re-run harnesses are only beginning [161, 228]. The bottleneck has thus shifted from model access to shared evaluation tooling.

9.7 Human-AI Collaborative Trading

The evidence favors complementarity in finance because humans hold information the model lacks. Human-AI teams usually trail the better of the two unless the human alone beats the AI [195], and finance fits the exception: an ML analyst beats 54.5% of sell-side return forecasts, but humans keep the edge on illiquid, intangible-heavy, and distressed firms, and a model given analysts' forecasts avoids about 90% of their extreme errors [28]. Bank customers followed human-AI advice more often than an ML advisor's (79.2% vs. 68.3%) and earned 44.9% more, though bankers' revisions left accuracy unchanged [220]. Oversight ladders assume the human catches the agent's errors, yet users of imperfect decision aids miss unflagged errors, more so when attention is split across many agents [154]. **Takeaway.** Humans add private information and tail-error control, but trust does not track accuracy and automation bias weakens human sign-off as a safeguard; no surveyed study tests this division of labor in live trading.

10 Conclusion

This survey traced LLM trading agents from their predecessors, mapped 47 systems by capability and market (Figure 3), and audited the evidence behind their results. Four findings follow. First, independently re-run margins shrink or vanish, though the re-runs cover few systems and the original margins rest on short, noisy windows (§7.4, §7.5). Second, the best-supported current hypothesis places the LLM upstream of the trade, in text signals, factor and model code, and bounded execution tilts (§8.6). Third, controlled comparisons do not establish a multi-agent advantage (§4.3). Fourth, structure and input design matter alongside the backbone, and the best structure depends on the backbone; no study varies scale with structure held fixed (§4.3, §5.2).

Table 6 turns this division of labor into controls for practitioners, giving each LLM role a deterministic gate, a bound on its action, and a monitoring target. Table 7 lists the open gaps, including unmeasured numerical errors in executed orders and the scarcity of live, real-capital evaluation [236]. Until a backtested Sharpe ratio above the passive baseline survives a bias-mitigated re-run, it remains exploratory.

Acknowledgments

This research is partially supported by the National Research Foundation, Singapore, and the Cyber Security Agency under its National Cybersecurity R&D Programme (NCRP25-P04-TAICeN), and by Ripple under the University Blockchain Research Initiative (UBRI) [56]. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore, and Ripple.

References

- [1] Diogo Almeida. 2026. Introducing System One Models & Jev. TypeSafe AI blog, <https://typesafe.ai/blog/introducing-system-one-models-and-jev>. Vendor announcement; accessed 25 September 2026.

- [2] Saad Alqithami. 2026. Autonomous Agents on Blockchains: Standards, Execution Models, and Trust Boundaries. *arXiv preprint arXiv:2601.04583* (2026).
- [3] Rohan Alur, Bradly C. Stadie, Daniel Kang, Ryan Chen, Matt McManus, Michael Rickert, Tyler Lee, Michael Federici, Richard Zhu, Dennis Fogerty, Hayley Williamson, Nina Lozinski, Aaron Linsky, and Jasjeet S. Sekhon. 2025. AIA Forecaster: Technical Report. *arXiv preprint arXiv:2511.07678* (2025).
- [4] Siyu An, Qin Li, Junru Lu, Di Yin, and Xing Sun. 2024. FinVerse: An Autonomous Agent System for Versatile Financial Analysis. *arXiv preprint arXiv:2406.06379* (2024).
- [5] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research (TMLR)* (2024).
- [6] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063* (2019).
- [7] Doron Avramov, Si Cheng, and Lior Metzker. 2023. Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability. *Management Science* 69, 5 (2023), 2587–2619. doi:10.1287/mnsc.2022.4449
- [8] Alessio Azzutti. 2024. AI Governance and Algorithmic Trading: Some Regulatory Insights from the EU AI Act. *Banking & Finance Law Review* 41, 1 (2024), 133–168.
- [9] David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. 2017. The Probability of Backtest Overfitting. *The Journal of Computational Finance* 20, 4 (2017), 39–69. doi:10.21314/JCF.2016.322
- [10] David H. Bailey and Marcos López de Prado. 2014. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *The Journal of Portfolio Management* 40, 5 (2014), 94–107. doi:10.3905/jpm.2014.40.5.094
- [11] Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM Hallucination Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. 24128–24156. doi:10.18653/v1/2025.acl-long.1176
- [12] Bank for International Settlements. 2022. *OTC Foreign Exchange Turnover in April 2022*. Triennial Central Bank Survey. Bank for International Settlements. <https://www.bis.org/publications/202210-commentary-otc-derivatives>
- [13] Rajat M. Barot and Arjun S. Borkhatariya. 2026. PolySwarm: A Multi-Agent Large Language Model Framework for Prediction Market Trading and Latency Arbitrage. *arXiv preprint arXiv:2604.03888* (2026).
- [14] Mostapha Benhenda. 2025. FinRL-DeepSeek: LLM-Infused Risk-Sensitive Reinforcement Learning for Trading Agents. *arXiv preprint arXiv:2502.07393* (2025).
- [15] Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. 2025. Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks. *arXiv preprint arXiv:2508.00828* (2025).
- [16] Paul Bilokon and Yitao Qiu. 2023. Transformers versus LSTMs for Electronic Trading. *arXiv preprint arXiv:2309.11400* (2023).
- [17] Fischer Black and Myron Scholes. 1973. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy* 81, 3 (1973), 637–654. doi:10.1086/260062
- [18] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. 2011. *Supervisory Guidance on Model Risk Management*. Supervision and Regulation Letter SR 11-7. Board of Governors of the Federal Reserve System, Washington, DC. <https://web.archive.org/web/2025/https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm> Issued 4 April 2011; co-issued by the OCC as Bulletin 2011-12. Superseded by SR 26-2 on 17 April 2026.
- [19] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, and Federal Deposit Insurance Corporation. 2026. *Revised Guidance on Model Risk Management*. Supervision and Regulation Letter SR 26-2. Board of Governors of the Federal Reserve System, Washington, DC. <https://www.federalreserve.gov/supervisionreg/srletters/sr2602.htm> Issued 17 April 2026; supersedes SR 11-7 and SR 21-8.
- [20] Sarah Breeden. 2026. Agents of Change. Speech at the ECB Forum on Central Banking, Sintra. Bank of England. <https://www.bankofengland.co.uk/speech/2026/june/sarah-breeden-panel-at-the-european-central-bank-forum-on-central-banking-2026>
- [21] William Brock, Josef Lakonishok, and Blake LeBaron. 1992. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance* 47, 5 (1992), 1731–1764. doi:10.1111/j.1540-6261.1992.tb04681.x
- [22] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

- [23] David Byrd. 2025. The Accidental Pump and Dump: When Agentic AI Meets Autonomous Trading. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF)*. 88–95. doi:10.1145/3768292.3770424
- [24] David Byrd. 2025. Exploring Sentiment Manipulation by LLM-Enabled Intelligent Trading Agents. *arXiv preprint arXiv:2502.16343* (2025).
- [25] David Byrd, Maria Hybinette, and Tucker Balch. 2020. ABIDES: Towards High-Fidelity Multi-Agent Market Simulation. In *Proceedings of the ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (SIGSIM-PADS)*. 11–22. doi:10.1145/3384441.3395986
- [26] Bokai Cao, Saizhuo Wang, Xinyi Lin, Xiaojun Wu, Haohan Zhang, Lionel M. Ni, and Jian Guo. 2025. From Deep Learning to LLMs: A Survey of AI in Quantitative Investment. *arXiv preprint arXiv:2503.21422* (2025).
- [27] Longbing Cao. 2023. AI in Finance: Challenges, Techniques, and Opportunities. *Comput. Surveys* 55, 3 (2023), 64:1–64:38. doi:10.1145/3502289
- [28] Sean Cao, Wei Jiang, Junbo Wang, and Baozhong Yang. 2024. From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. *Journal of Financial Economics* 160 (2024), 103910. doi:10.1016/j.jfineco.2024.103910
- [29] Agostino Capponi, Alfio Gliozzo, and Brian Zhu. 2025. Agentic AI for Clustering, Relationship Discovery, and Semantic Trading in Prediction Markets. *arXiv preprint arXiv:2512.02436* (2025).
- [30] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why Do Multi-Agent LLM Systems Fail?. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*. doi:10.52202/085713-4082
- [31] Luyang Chen, Markus Pelger, and Jason Zhu. 2024. Deep Learning in Asset Pricing. *Management Science* 70, 2 (2024), 714–750. doi:10.1287/mnsc.2023.4695
- [32] Yifei Chen, Bryan T. Kelly, and Dacheng Xiu. 2022. Expected Returns and Large Language Models. Working paper.
- [33] Yanxu Chen, Zijun Yao, Yantao Liu, Amy Xin, Jin Ye, Jianing Yu, Lei Hou, and Juanzi Li. 2025. StockBench: Can LLM Agents Trade Stocks Profitably In Real-world Markets? *arXiv preprint arXiv:2510.02209* (2025).
- [34] Pu Cheng, Juncheng Liu, and Yunshen Long. 2026. PolyBench: Benchmarking LLM Forecasting and Trading Capabilities on Live Prediction Market Data. *arXiv preprint arXiv:2604.14199* (2026).
- [35] Yuhan Cheng, Yanchu Liu, and Heyang Zhou. 2026. Large Language Models and Futures Price Factors in China. *Journal of Futures Markets* 46, 2 (2026), 262–282. doi:10.1002/fut.70061
- [36] Dehao Dai, Ding Ma, Dou Liu, Kerui Geng, and Yiqing Wang. 2026. Beyond Polarity: Multi-Dimensional LLM Sentiment Signals for WTI Crude Oil Futures Return Prediction. *arXiv preprint arXiv:2603.11408* (2026).
- [37] Philip Daian, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels. 2020. Flash Boys 2.0: Frontrunning in Decentralized Exchanges, Miner Extractable Value, and Consensus Instability. In *Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 910–927. doi:10.1109/SP40000.2020.00040
- [38] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A Decoder-Only Foundation Model for Time-Series Forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML) (PMLR, Vol. 235)*. 10148–10167.
- [39] Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. 2025. AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. *Comput. Surveys* 57, 7 (2025), 1–36. doi:10.1145/3716628
- [40] Han Ding, Yinheng Li, Junhao Wang, Hang Chen, Doudou Guo, and Yunbai Zhang. 2026. Large Language Model Agent in Financial Trading: A Survey. In *Proceedings of the 2026 9th International Conference on Computers in Management and Business (ICCMB)*. 76–83. doi:10.1145/3802463.3802475
- [41] Jinru Ding, Chao Ding, Yidong Jiang, Wenrao Pang, Boyi Xiao, Zhiqiang Liu, Jiayuan Chen, Yun Zhong, Tiantian Yuan, Junming Guan, Dawei Cheng, and Jie Xu. 2026. Beyond Knowledge to Agency: Evaluating Expertise, Autonomy, and Integrity in Finance with CNFinBench. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD)*. 8765–8776. doi:10.1145/3770855.3817482
- [42] Qianggang Ding, Haochen Shi, Jiadong Guo, and Bang Liu. 2025. TradExpert: Revolutionizing Trading with Mixture of Expert LLMs. In *Advances in Financial AI Workshop at ICLR 2025*.
- [43] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep Learning for Event-Driven Stock Prediction. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*. 2327–2333.
- [44] Yifei Dong, Fengyi Wu, Kunlin Zhang, Yilong Dai, Sanjian Zhang, Wanghao Ye, Sihao Chen, and Zhi-Qi Cheng. 2025. Large Language Model Agents in Finance: A Survey Bridging Research, Practice, and Real-World Deployment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 17889–17907. doi:10.18653/v1/2025.findings-emnlp.972
- [45] Winston Wei Dou, Itay Goldstein, and Yan Ji. 2025. *AI-Powered Trading, Algorithmic Collusion, and Price Efficiency*. NBER Working Paper 34054. National Bureau of Economic Research. doi:10.3386/w34054

- [46] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML) (PMLR, Vol. 235)*. 11733–11763.
- [47] Dimitrios Emmanoulopoulos, Ollie Olby, Justin Lyon, and Namid R. Stillman. 2025. To Trade or Not to Trade: An Agentic Approach to Estimating Market Risk Improves Trading Decisions. *arXiv preprint arXiv:2507.08584* (2025).
- [48] Robert F. Engle and C. W. J. Granger. 1987. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica* 55, 2 (1987), 251–276. doi:10.2307/1913236
- [49] Alexander Erlei and Lukas Meub. 2026. LLM-Agent Interactions on Markets with Information Asymmetries. *arXiv preprint arXiv:2603.08853* (2026).
- [50] European Commission. 2017. Commission Delegated Regulation (EU) 2017/589 of 19 July 2016 Supplementing Directive 2014/65/EU of the European Parliament and of the Council with Regard to Regulatory Technical Standards Specifying the Organisational Requirements of Investment Firms Engaged in Algorithmic Trading. Official Journal of the European Union, L 87, 31 March 2017, p. 417. https://eur-lex.europa.eu/eli/reg_del/2017/589/oj RTS 6; applies from 3 January 2018.
- [51] Eugene F. Fama and Kenneth R. French. 1993. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* 33, 1 (1993), 3–56. doi:10.1016/0304-405X(93)90023-5
- [52] Tianyu Fan, Yuhao Yang, Yangqin Jiang, Yifei Zhang, Yuxuan Chen, and Chao Huang. 2025. AI-Trader: Benchmarking Autonomous Agents in Real-Time Financial Markets. *arXiv preprint arXiv:2512.10971* (2025).
- [53] George Fatouros, Kostas Metaxas, John Soldatos, and Manos Karathanassis. 2025. MarketSenseAI 2.0: Enhancing Stock Analysis through LLM Agents. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*. 883–892. doi:10.1109/ICDMW69685.2025.00105
- [54] George Fatouros, Kostas Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2025. Can Large Language Models Beat Wall Street? Evaluating GPT-4’s Impact on Financial Decision-Making with MarketSenseAI. *Neural Computing and Applications* 37, 30 (2025), 24893–24918. doi:10.1007/s00521-024-10613-4
- [55] Guanhao Feng, Jingyu He, Nicholas G. Polson, and Jianeng Xu. 2024. Deep Learning in Characteristics-Sorted Factor Models. *Journal of Financial and Quantitative Analysis* 59, 7 (2024), 3001–3036. doi:10.1017/S0022109023000893
- [56] Yebo Feng, Jiahua Xu, and Lauren Weymouth. 2022. University Blockchain Research Initiative (UBRI): Boosting Blockchain Education and Research. *IEEE Potentials* 41, 6 (2022), 19–25. doi:10.1109/MPOT.2022.3198929
- [57] Mohamed Amine Ferrag, Norbert Tihanyi, Djallel Hamouda, Leandros Maglaras, Abderrahmane Lakas, and Merouane Debbah. 2026. From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows. *ICT Express* 12, 2 (2026), 353–383. doi:10.1016/j.icte.2025.12.001
- [58] Financial Stability Board. 2026. *Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation Report*. Technical Report. Financial Stability Board, Basel. <https://www.fsb.org/uploads/P100626.pdf>
- [59] FINRA. 2015. *Guidance on Effective Supervision and Control Practices for Firms Engaging in Algorithmic Trading Strategies*. Regulatory Notice 15-09. Financial Industry Regulatory Authority, Washington, DC. https://www.finra.org/sites/default/files/notice_doc_file_ref/Notice_Regulatory_15-09.pdf
- [60] Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. 2026. Algorithmic Collusion by Large Language Models. In *Proceedings of the 27th ACM Conference on Economics and Computation (EC)*.
- [61] Weilong Fu. 2025. The New Quant: A Survey of Large Language Models in Financial Prediction and Trading. *arXiv preprint arXiv:2510.05533* (2025).
- [62] Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. 2006. Pairs Trading: Performance of a Relative-Value Arbitrage Rule. *The Review of Financial Studies* 19, 3 (2006), 797–827. doi:10.1093/rfs/hhj020
- [63] Kerui Geng, Yiqing Wang, Dehao Dai, and Ding Ma. 2026. Macro Economists in the Machine: A Multi-Agent LLM Framework for Commodity-Related ETF Portfolio Construction. *Cogent Economics & Finance* 14, 1 (2026), 2726071. doi:10.1080/23322039.2026.2726071
- [64] Mohammed-Khalil Ghali, Cecil Pang, Oscar Molina, Carlos Gershenson-Garcia, and Daehan Won. 2025. Forecasting Commodity Price Shocks Using Temporal and Semantic Fusion of Prices Signals and Agentic Generative AI Extracted Economic News. *arXiv preprint arXiv:2508.06497* (2025).
- [65] Paul Glasserman and Caden Lin. 2024. Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis. *The Journal of Financial Data Science* 6, 1 (2024), 25–42. doi:10.3905/jfds.2023.1.143
- [66] Hui Gong. 2026. AI Agents in Financial Markets: Architecture, Applications, and Systemic Implications. *FinTech* 5, 2 (2026), 34. doi:10.3390/fintech5020034
- [67] Vincent Gramlich, Dennis Jelito, and Johannes Sedlmeir. 2024. Maximal Extractable Value: Current Understanding, Categorization, and Open Research Questions. *Electronic Markets* 34, 1 (2024), 49. doi:10.1007/s12525-024-00727-x
- [68] Arnav Grover. 2025. FinRLlama: A Solution to LLM-Engineered Signals Challenge at FinRL Contest 2024. *arXiv preprint arXiv:2502.01992* (2025). Competition track, FinRL Contest 2024, ACM ICAIF 2024.

- [69] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large Language Models Are Zero-Shot Time Series Forecasters. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. 19622–19635. doi:10.52202/075280-0861
- [70] Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies* 33, 5 (2020), 2223–2273. doi:10.1093/rfs/hhaa009
- [71] Jiacheng Guo, Suozhi Huang, Zixin Yao, Yifan Zhang, Yifu Lu, Jiashuo Liu, Zihao Li, Nicholas Deng, Qixin Xiao, Jia Tian, Kanghong Zhan, Tianyi Li, Xiaochen Liu, Jason Ge, Chaoyang He, Kaixuan Huang, Lin Yang, Wenhao Huang, and Mengdi Wang. 2025. CryptoBench: A Dynamic Benchmark for Expert-Level Evaluation of LLM Agents in Cryptocurrency. *arXiv preprint arXiv:2512.00417* (2025).
- [72] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*. 8048–8057.
- [73] Tian Guo and Emmanuel Hauptmann. 2024. Fine-Tuning Large Language Models for Stock Return Prediction Using Newsflow. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 1028–1045. doi:10.18653/v1/2024.emnlp-industry.77
- [74] Taian Guo, Haiyang Shen, Jinsheng Huang, Zhengyang Mao, Junyu Luo, Binqi Chen, Zhuoru Chen, Luchen Liu, Bingyu Xia, Xuhui Liu, Yun Ma, and Ming Zhang. 2025. MASS: Multi-Agent Simulation Scaling for Portfolio Construction. *arXiv preprint arXiv:2505.10278* (2025).
- [75] Ben Hambly, Renyuan Xu, and Huining Yang. 2023. Recent Advances in Reinforcement Learning in Finance. *Mathematical Finance* 33, 3 (2023), 437–503. doi:10.1111/mafi.12382
- [76] Campbell R. Harvey, Yan Liu, and Heqing Zhu. 2016. ... and the Cross-Section of Expected Returns. *The Review of Financial Studies* 29, 1 (2016), 5–68. doi:10.1093/rfs/hhv059
- [77] Ryuji Hashimoto, Takehiro Takayanagi, Masahiro Suzuki, and Kiyoshi Izumi. 2026. Agent-Based Simulation of a Financial Market with Large Language Models. In *PRIMA 2025: Principles and Practice of Multi-Agent Systems, 26th International Conference (Lecture Notes in Computer Science, Vol. 16366)*, Catalin Dima, Angelo Ferrando, and Vadim Malvone (Eds.). Springer, 20–28. doi:10.1007/978-3-032-13562-9_2
- [78] Feng He, Tianqing Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S. Yu. 2025. The Emerged Security and Privacy of LLM Agent: A Survey with Case Studies. *Comput. Surveys* 58, 6 (2025), 1–36. doi:10.1145/3773080
- [79] Songrun He, Linying Lv, Asaf Manela, and Jimmy Wu. 2026. Chronologically Consistent Large Language Models. *Journal of Financial Economics* 185 (2026), 104374. doi:10.1016/j.jfineco.2026.104374
- [80] Thomas Henning, Siddhartha M. Ojha, Ross Spoon, Jiatong Han, and Colin F. Camerer. 2025. LLM Agents Do Not Replicate Human Market Traders: Evidence From Experimental Finance. *arXiv preprint arXiv:2502.15800* (2025).
- [81] Bruno Miranda Henrique, Vinicius Amorim Sobreiro, and Herbert Kimura. 2019. Literature Review: Machine Learning Techniques Applied to Financial Market Prediction. *Expert Systems with Applications* 124 (2019), 226–251. doi:10.1016/j.eswa.2019.01.012
- [82] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiaowu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *International Conference on Learning Representations (ICLR)*.
- [83] Liang Hu, Jianpeng Jiao, Jiashuo Liu, Dongyuan Mutu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, Yali Liao, Zaiyuan Wang, Jingkai Liu, Daibin Sun, Ziqing Zeng, Zhiyuan Zeng, Chenghao Yang, Qianyu Yang, Mingren Yin, Ge Zhang, Xinyi Zhang, Xiyang Zhao, Zhenwei Zhu, Hongseok Namkoong, and Wenhao Huang. 2026. FinSearchComp: Towards a Realistic, Expert-Level Evaluation of Financial Search and Reasoning. In *The Fourteenth International Conference on Learning Representations*.
- [84] Allen H. Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research* 40, 2 (2023), 806–841. doi:10.1111/1911-3846.12832
- [85] Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, Shanglin Wu, Ruiyao Xu, Liangwei Yang, Rui Yang, Woosong Yang, Chin-Yuan Yeh, Hanrong Zhang, Haozhen Zhang, Siqi Zhu, Henry Peng Zou, Wanjia Zhao, Song Wang, Wujiang Xu, Zixuan Ke, Zheng Hui, Dawei Li, Yaozu Wu, Langzhou He, Chen Wang, Xiong Xiao Xu, Baixiang Huang, Juntao Tan, Shelby Heinecke, Huan Wang, Caiming Xiong, Ahmed A. Metwally, Jun Yan, Chen-Yu Lee, Hanqing Zeng, Yinglong Xia, Xiaokai Wei, Ali Payani, Yu Wang, Haitong Ma, Wenya Wang, Chengguang Wang, Yu Zhang, Xin Eric Wang, Yongfeng Zhang, Jiaxuan You, Hanghang Tong, Xiao Luo, Xue Liu, Yizhou Sun, Wei Wang, Julian McAuley, James Zou, Jiawei Han, Philip S. Yu, and Kai Shu. 2026. A Survey of Agent Memory in the Second Half: Towards Self-Evolving and Long-Horizon Agents. *Transactions on Machine Learning Research* (2026). <https://arxiv.org/abs/2602.06052>
- [86] IOSCO. 2025. *Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges*. Consultation Report CR/01/2025. International Organization of Securities Commissions, Madrid. <https://www.iosco.org/library/pubdocs/>

pdf/IOSCOPD788.pdf

- [87] IOSCO. 2026. *Supervisory Toolkit for AI Use in Capital Markets*. Final Report FR/02/2026. International Organization of Securities Commissions, Madrid. <https://www.iosco.org/library/pubdocs/pdf/IOSCOPD823.pdf>
- [88] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A New Benchmark for Financial Question Answering. *arXiv preprint arXiv:2311.11944* (2023).
- [89] Amin Izadyar. 2026. AI and Exchange Rate Predictability. *arXiv preprint arXiv:2608.00761* (2026).
- [90] Jingwen Jiang, Bryan Kelly, and Dacheng Xiu. 2023. (Re-)Imag(in)ing Price Trends. *The Journal of Finance* 78, 6 (2023), 3193–3249. doi:10.1111/jofi.13268
- [91] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuanfang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *International Conference on Learning Representations (ICLR)*.
- [92] Kiran Kamble, Melisa Russak, Dmytro Mozolevskyi, Muayad Ali, Mateusz Russak, and Waseem AlShikh. 2025. Expect the Unexpected: FailSafe Long Context QA for Finance. *arXiv preprint arXiv:2502.06329* (2025).
- [93] Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. *arXiv preprint arXiv:2311.15548* (2023).
- [94] Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. 2025. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- [95] Jussi Keppo, Yuze Li, Gerry Tsoukalas, and Nuo Yuan. 2026. On the Fragility of AI Agent Collusion. *arXiv preprint arXiv:2603.20281* (2026).
- [96] Amir E. Khandani and Andrew W. Lo. 2011. What Happened to the Quants in August 2007? Evidence from Factors and Transactions Data. *Journal of Financial Markets* 14, 1 (2011), 1–46. doi:10.1016/j.finmar.2010.07.005
- [97] Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The Flash Crash: High-Frequency Trading in an Electronic Market. *The Journal of Finance* 72, 3 (2017), 967–998. doi:10.1111/jofi.12498
- [98] Kemal Kirtac and Guido Germano. 2024. Sentiment Trading with Large Language Models. *Finance Research Letters* 62 (2024), 105227.
- [99] Kelvin J. L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to Generate Explainable Stock Predictions using Self-Reflective Large Language Models. In *Proceedings of the ACM Web Conference 2024 (WWW)*. ACM, 4304–4315. doi:10.1145/3589334.3645611
- [100] Yaxuan Kong, Hoyoung Lee, Yoontae Hwang, Alejandro Lopez-Lira, Bradford Levy, Dhagash Mehta, Qingsong Wen, Chanyeol Choi, Yongjae Lee, and Stefan Zohren. 2026. Position: Evaluating LLMs in Finance Requires Explicit Bias Consideration. In *Proceedings of the 43rd International Conference on Machine Learning*. Position paper. arXiv:2602.14233.
- [101] Zhizhuo Kou, Holam Yu, Junyu Luo, Jingshu Peng, Xujia Li, Chengzhong Liu, Juntao Dai, Lei Chen, Sirui Han, and Yike Guo. 2025. Automate Strategy Finding with LLM in Quant Investment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 18517–18533. doi:10.18653/v1/2025.findings-emnlp.1005
- [102] Mahesh Kumar, Bhaskarjit Sarmah, and Stefano Pasquali. 2026. FinReflectKG–HalluBench: GraphRAG Hallucination Benchmark for Financial Question Answering Systems. *arXiv preprint arXiv:2603.20252* (2026).
- [103] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702* (2023).
- [104] Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large Language Models in Finance (FinLLMs). *Neural Computing and Applications* 37, 30 (2025), 24853–24867. doi:10.1007/s00521-024-10495-6
- [105] Youngbin Lee, Yejin Kim, Juhyeong Kim, Suin Kim, and Yongjae Lee. 2025. LLM-Enhanced Black-Litterman Portfolio Optimization. *arXiv preprint arXiv:2504.14345* (2025). Presented at the CIKM 2025 Workshop on Advances in Financial AI (non-archival).
- [106] Yuchen Lei, Yuexin Xiang, Rafael Dowsley, Tsz Hon Yuen, Andreas Deppeler, Jiangshan Yu, Qin Wang, and Kim-Kwang Raymond Choo. 2025. Large Language Models for Cryptocurrency Transaction Analysis: A Bitcoin Case Study. *arXiv preprint arXiv:2501.18158* (2025).
- [107] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 9459–9474.

- [108] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems (NeurIPS)*. 51991–52008. doi:10.52202/075280-2264
- [109] Haochang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K. P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. 2025. InvestorBench: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 2509–2525. doi:10.18653/v1/2025.acl-long.126
- [110] Junjie Li, Yang Liu, Weiqing Liu, Shikai Fang, Lewen Wang, Chang Xu, and Jiang Bian. 2025. MarS: A Financial Market Simulation Engine Powered by Generative Foundation Model. In *International Conference on Learning Representations (ICLR)*.
- [111] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More Agents Is All You Need. *Transactions on Machine Learning Research* (2024). <https://openreview.net/forum?id=bgzUSZ8aeg>
- [112] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. 15523–15536. doi:10.18653/v1/2024.acl-long.829
- [113] Weixian Waylon Li, Hyeonjun Kim, Mihai Cucuringu, and Tiejun Ma. 2026. Can LLM-based Financial Investing Strategies Outperform the Market in Long Run?. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '26)*. ACM, Jeju Island, Republic of Korea, 2711–2722. doi:10.1145/3770854.3785702
- [114] Weixian 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. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*. To appear. arXiv:2605.24564.
- [115] Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024. AlphaFin: Benchmarking Financial Analysis with Retrieval-Augmented Stock-Chain Framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. 773–783.
- [116] Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025. HedgeAgents: A Balanced-aware Multi-agent Financial Trading System. In *Companion Proceedings of the ACM Web Conference 2025*. 296–305. doi:10.1145/3701716.3715232
- [117] Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025. Profit Mirage: Revisiting Information Leakage in LLM-based Financial Agents. *arXiv preprint arXiv:2510.07920* (2025).
- [118] Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025. QuantAgents: Towards Multi-agent Financial System via Simulated Trading. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 17438–17464. doi:10.18653/v1/2025.findings-emnlp.945
- [119] Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. 2024. CryptoTrade: A Reflective LLM-based Agent to Guide Zero-shot Cryptocurrency Trading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1094–1106. doi:10.18653/v1/2024.emnlp-main.63
- [120] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. In *Proceedings of the 4th ACM International Conference on AI in Finance (ICAIF)*. 374–382. doi:10.1145/3604237.3626869
- [121] Yuante Li, Xu Yang, Xiao Yang, Xisen Wang, Weiqing Liu, and Jiang Bian. 2025. R&D-Agent-Quant: A Multi-Agent Framework for Data-Centric Factors and Model Joint Optimization. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. 131781–131830. doi:10.52202/085713-3967
- [122] Yang Li, Yangyang Yu, Haochang Li, Zhi Chen, and Khaldoun Khashanah. 2023. TradingGPT: Multi-Agent System with Layered Memory and Distinct Characters for Enhanced Financial Trading Performance. *arXiv preprint arXiv:2309.03736* (2023).
- [123] Zeping Li, Guancheng Wan, Keyang Chen, Yu Chen, Yiwen Zhao, Philip Torr, Guangnan Ye, Zhenfei Yin, and Hongfeng Chai. 2026. Behavioral Consistency Validation for LLM Agents: An Analysis of Trading-Style Switching through Stock-Market Simulation. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, San Diego, California, USA, 40356–40370. doi:10.18653/v1/2026.findings-acl.2006
- [124] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s Verify Step by Step. In *International Conference on Learning Representations (ICLR)*.
- [125] Xiang-Li Lim, Puja Singh, and Richard Stobo. 2025. *Regulatory Considerations Regarding Accelerated Use of AI in Securities Markets*. IMF Technical Notes and Manuals TNM/2025/016. International Monetary Fund, Washington, DC. doi:10.5089/9798229026888.005
- [126] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics (TACL)* 12 (2024), 157–173. doi:10.1162/tacl_a_00638

- [127] Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou, Man Lan, and Yang Chong. 2025. FinDABench: Benchmarking Financial Data Analysis Ability of Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*. 710–725.
- [128] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AgentBench: Evaluating LLMs as Agents. In *The Twelfth International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=zAdUB0aCTQ>
- [129] Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Dan Wang, Zhaoran Wang, and Jian Guo. 2022. FinRL-Meta: Market Environments and Benchmarks for Data-Driven Financial Reinforcement Learning. In *NeurIPS 2022 Datasets and Benchmarks Track*. 1835–1849. doi:10.52202/068431-0134
- [130] Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. 2021. FinRL: Deep Reinforcement Learning Framework to Automate Trading in Quantitative Finance. In *Proceedings of the Second ACM International Conference on AI in Finance (ICAIF '21)*. ACM, 1–9. doi:10.1145/3490354.3494366
- [131] Zhaowei Liu, Xin Guo, Zhi Yang, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Mengping Li, Qi Qi, Zhiqiang Liu, Yiyang Han, Dongpo Cheng, Ronghao Chen, Huacan Wang, Xingdong Feng, Huixia Judy Wang, Chengchun Shi, and Liwen Zhang. 2025. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. *arXiv preprint arXiv:2503.16252* (2025).
- [132] Andrew W. Lo. 2002. The Statistics of Sharpe Ratios. *Financial Analysts Journal* 58, 4 (2002), 36–52. doi:10.2469/faj.v58.n4.2453
- [133] Alejandro Lopez-Lira. 2025. Can Large Language Models Trade? Testing Financial Theories with LLM Agents in Market Simulations. *arXiv preprint arXiv:2504.10789* (2025).
- [134] Alejandro Lopez-Lira and Yuehua Tang. 2026. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *Journal of Financial Economics* 184 (2026), 104335. doi:10.1016/j.jfineco.2026.104335
- [135] Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. 2025. The Memorization Problem: Can We Trust LLMs' Economic Forecasts? *arXiv preprint arXiv:2504.14765* (2025). doi:10.2139/ssrn.5217505 SSRN Working Paper 5217505.
- [136] Tim Loughran and Bill McDonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66, 1 (2011), 35–65. doi:10.1111/j.1540-6261.2010.01625.x
- [137] Guilong Lu, Xuntao Guo, Rongjunchen Zhang, Wenqiao Zhu, and Ji Liu. 2025. BizFinBench: A Business-Driven Real-World Financial Benchmark for Evaluating LLMs. *arXiv preprint arXiv:2505.19457* (2025).
- [138] Janna Lu. 2025. Evaluating LLMs on Real-World Forecasting Against Expert Forecasters. *arXiv preprint arXiv:2507.04562* (2025).
- [139] Haochen Luo, Zhengzhao Lai, Junjie Xu, Yifan Li, Pok Hin Tang, Yuan Zhang, and Chen Liu. 2026. From Natural Language to Executable Option Strategies via Large Language Models. *arXiv preprint arXiv:2603.16434* (2026).
- [140] Yichen Luo, Yebo Feng, Jiahua Xu, and Yang Liu. 2026. Resisting Manipulative Bots in Meme Coin Copy Trading: A Multi-Agent Approach with Chain-of-Thought Reasoning. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*. ACM.
- [141] Yichen Luo, Yebo Feng, Jiahua Xu, Paolo Tasca, and Yang Liu. 2025. LLM-Powered Multi-Agent System for Automated Crypto Portfolio Management. *arXiv preprint arXiv:2501.00826* (2025).
- [142] Tianmi Ma, Jiawei Du, Wenxin Huang, Wenjie Wang, Liang Xie, Xian Zhong, and Joey Tianyi Zhou. 2025. Agent Trading Arena: A Study on Numerical Understanding in LLM-Based Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 5496–5514. doi:10.18653/v1/2025.findings-emnlp.294
- [143] Sedigheh Mahdavi, Jiating Chen, Pradeep Kumar Joshi, Lina Huertas Guativa, and Upmanyu Singh. 2025. Integrating Large Language Models in Financial Investments and Market Analysis: A Survey. *arXiv preprint arXiv:2507.01990* (2025).
- [144] Eftychia Makri, Georgios Palaiokrassas, Sarah Bouraga, Antigoni Polychroniadou, and Leandros Tassioulas. 2025. Ethereum Price Prediction Employing Large Language Models for Short-term and Few-shot Forecasting. In *2025 7th Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*. IEEE, 1–10. doi:10.1109/BRAINS67003.2025.11302944
- [145] Qian'ang Mao, Yuxuan Zhang, Jiaman Chen, Wenjun Zhou, and Jiaqi Yan. 2025. Know Your Intent: An Autonomous Multi-Perspective LLM Agent Framework for DeFi User Transaction Intent Mining. *arXiv preprint arXiv:2511.15456* (2025).
- [146] R. David McLean and Jeffrey Pontiff. 2016. Does Academic Research Destroy Stock Return Predictability? *The Journal of Finance* 71, 1 (2016), 5–32. doi:10.1111/jofi.12365
- [147] Kunihiro Miyazaki, Takanobu Kawahara, Stephen Roberts, and Stefan Zohren. 2026. Toward Expert Investment Teams: A Multi-Agent LLM System with Fine-Grained Trading Tasks. *The Journal of Financial Data Science* 8, 3 (2026), 168–202. doi:10.3905/jfds.2026.008

- [148] Phat Nguyen and Thang Pham. 2026. Toward Reliable Evaluation of LLM-Based Financial Multi-agent Systems: Taxonomy, Coordination Primacy, and Cost Awareness. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2026 Workshops (Lecture Notes in Computer Science)*, Cheng Long, Jieming Shi, Yuxuan Liang, Xiao Huang, Byron Choi, Yuanfeng Song, and Man Lung Yiu (Eds.). Springer, 324–336. doi:10.1007/978-981-92-2014-4_26
- [149] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *arXiv preprint arXiv:2406.11903* (2024).
- [150] Natapong Nitarach, Warit Sirichotedumrong, Panop Pitchayarthorn, Pittawat Taveekitworachai, Potsawee Manakul, and Kunat Pipatanakul. 2025. FinCoT: Grounding Chain-of-Thought in Expert Financial Reasoning. In *Proceedings of the 10th Workshop on Financial Technology and Natural Language Processing (FinNLP)*. 93–123. doi:10.18653/v1/2025.finnlp-2.8
- [151] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. 2020. Deep Learning for Financial Applications: A Survey. *Applied Soft Computing* 93 (2020), 106384. doi:10.1016/j.asoc.2020.106384
- [152] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560* (2023).
- [153] Charidimos Papadakis, Giorgos Filandrianos, Angeliki Dimitriou, Maria Lymperaioi, Konstantinos Thomas, and Giorgos Stamou. 2025. StockSim: A Dual-Mode Order-Level Simulator for Evaluating Multi-Agent LLMs in Financial Markets. *arXiv preprint arXiv:2507.09255* (2025).
- [154] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors* 52, 3 (2010), 381–410. doi:10.1177/0018720810376055
- [155] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*. 1–22. doi:10.1145/3586183.3606763
- [156] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2024. Gorilla: Large Language Model Connected with Massive APIs. In *Advances in Neural Information Processing Systems (NeurIPS)*. 126544–126565. doi:10.52202/079017-4020
- [157] Atharv Singh Patlan, Peiyao Sheng, S. Ashwin Hebbar, Prateek Mittal, and Pramod Viswanath. 2025. Real AI Agents with Fake Memories: Fatal Context Manipulation Attacks on Web3 Agents. *arXiv preprint arXiv:2503.16248* (2025).
- [158] Nikolaos Pippas, Elliot A. Ludvig, and Cagatay Turkay. 2025. The Evolution of Reinforcement Learning in Quantitative Finance: A Survey. *Comput. Surveys* 57, 11 (2025), 295:1–295:51. doi:10.1145/3733714
- [159] Polymarket. 2024. Polymarket Agents: Trade Autonomously on Polymarket Using AI Agents. Open-source software, MIT license. <https://github.com/Polymarket/agents>
- [160] Prudential Regulation Authority. 2023. *Model Risk Management Principles for Banks*. Supervisory Statement SS1/23. Bank of England, London. <https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss> Published 17 May 2023 with PS6/23; effective 17 May 2024.
- [161] Lingfei Qian, Xueqing Peng, Hanley Smith, Yi Han, Yueru He, Haochang Li, Yupeng Cao, Yangyang Yu, Guojun Xiong, Peng Lu, Yan Wang, Vincent Jim Zhang, Huan He, Alejandro Lopez-Lira, Jimin Huang, Jian-Yun Nie, and Sophia Ananiadou. 2026. When Agents Trade: Live Multi-Market Trading Arena for LLM Agents. In *Proceedings of the ACM Web Conference 2026 (WWW ’26)*. ACM.
- [162] Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Yilun Zhao, Jimin Huang, Qianqian Xie, and Jian-Yun Nie. 2025. Fino1: On the Transferability of Reasoning-Enhanced LLMs and Reinforcement Learning to Finance. *arXiv preprint arXiv:2502.08127* (2025).
- [163] Kaihua Qin, Liyi Zhou, and Arthur Gervais. 2022. Quantifying Blockchain Extractable Value: How Dark Is the Forest?. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 198–214. doi:10.1109/SP46214.2022.9833734
- [164] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In *International Conference on Learning Representations (ICLR)*.
- [165] Gabriel V. Rauterberg. 2025. How Artificial Intelligence Will Shape Securities Regulation. *University of Chicago Law Review Online* (Jan. 2025). <https://lawreview.uchicago.edu/online-archive/how-artificial-intelligence-will-shape-securities-regulation>
- [166] Oriol Saguillo, Vahid Ghafouri, Lucianna Kiffer, and Guillermo Suarez-Tangil. 2025. Unravelling the Probabilistic Forest: Arbitrage in Prediction Markets. In *7th Conference on Advances in Financial Technologies (AFT) (LIPIcs, Vol. 354)*. 27:1–27:24. doi:10.4230/LIPIcs.AFT.2025.27
- [167] Preetha Saha, Jingrao Lyu, Arnav Saxena, Tianjiao Zhao, and Dhagash Mehta. 2025. Large Language Model Agents for Investment Management: Foundations, Benchmarks, and Research Frontiers. In *Proceedings of the 6th ACM*

- International Conference on AI in Finance (ICAIF)*. 736–744. doi:10.1145/3768292.3770387
- [168] Suproteem K. Sarkar and Keyon Vafa. 2024. Lookahead Bias in Pretrained Language Models. *SSRN Working Paper 4754678* (2024). doi:10.2139/ssrn.4754678
 - [169] Philipp Schoenegger, Peter S. Park, Ezra Karger, Sean Trott, and Philip E. Tetlock. 2025. AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy. *ACM Transactions on Interactive Intelligent Systems* 15, 1 (2025), 4:1–4:25.
 - [170] Philipp Schoenegger, Indre Tuminauskaitė, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. 2024. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. *Science Advances* 10, 45 (2024), eadp1528. doi:10.1126/sciadv.adp1528
 - [171] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2020. Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. *Applied Soft Computing* 90 (2020), 106181.
 - [172] Devin Shanthikumar and Il Sun Yoo. 2026. Beyond Automation: AI and the Human Value of Sell-Side Analysts. *Journal of Accounting Research* (2026). doi:10.1111/1475-679x.70083 Early view.
 - [173] William F. Sharpe. 1964. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance* 19, 3 (1964), 425–442.
 - [174] Zane Shen, Xinli Xu, Guangyi Zhang, Jialong Chen, Jinsong Zhou, Cong Chen, Guibao Shen, Dongyu Yan, Luozhou Wang, and Zhen Yang. 2026. Can Large Language Models Execute Parent Orders? *arXiv preprint arXiv:2607.28410* (2026).
 - [175] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. 8634–8652.
 - [176] Dong Shu, Yanguang Liu, Huopu Zhang, and Mengnan Du. 2026. FinCall-Surprise: A Large Scale Multi-modal Benchmark for Earning Surprise Prediction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13357–13370. doi:10.18653/v1/2026.acl-long.610
 - [177] Aadi Singhi. 2025. An Adaptive Multi Agent Bitcoin Trading System. *arXiv preprint arXiv:2510.08068* (2025).
 - [178] Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should We Be Going MAD? A Look at Multi-Agent Debate Strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning (ICML) (PMLR, Vol. 235)*. 45883–45905.
 - [179] Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research (TMLR)* (2024).
 - [180] Shuo Sun, Rundong Wang, and Bo An. 2023. Reinforcement Learning for Quantitative Trading. *ACM Transactions on Intelligent Systems and Technology* 14, 3 (2023), 1–29.
 - [181] Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. 2024. Are Language Models Actually Useful for Time Series Forecasting?. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 37. 60162–60191. doi:10.52202/079017-1922
 - [182] Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsan Zheng, Keze Wang, Jinrui Chen, and Liang Lin. 2025. AlphaAgent: LLM-Driven Alpha Mining with Regularized Exploration to Counteract Alpha Decay. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. ACM, 2813–2822. doi:10.1145/3711896.3736838
 - [183] Paul C. Tetlock. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance* 62, 3 (2007), 1139–1168. doi:10.1111/j.1540-6261.2007.01232.x
 - [184] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).
 - [185] Kentaro Toyoda, Xiao Wang, Mingzhe Li, Bo Gao, Yuan Wang, and Qingsong Wei. 2024. Blockchain Data Analysis in the Era of Large-Language Models. *arXiv preprint arXiv:2412.09640* (2024).
 - [186] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. 74952–74965. doi:10.52202/075280-3275
 - [187] TypeSafe AI. 2026. Introduction. TypeSafe AI documentation, <https://docs.typesafe.ai/introduction>. Accessed 25 September 2026.
 - [188] Ananya Unnikrishnan. 2024. Financial News-Driven LLM Reinforcement Learning for Portfolio Management. *arXiv preprint arXiv:2411.11059* (2024).
 - [189] U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission. 2010. *Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Technical Report. U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission, Washington, DC. <https://www.sec.gov/news/studies/2010/marketevents-report.pdf> 30 September

- 2010.
- [190] U.S. Securities and Exchange Commission. 2010. Risk Management Controls for Brokers or Dealers with Market Access. Final rule, Release No. 34-63241, File No. S7-03-10; 75 Fed. Reg. 69792 (15 November 2010). <https://www.sec.gov/rules/final/2010/34-63241.pdf> Adopting Exchange Act Rule 15c3-5 (the Market Access Rule); dated 3 November 2010.
 - [191] U.S. Securities and Exchange Commission. 2013. In the Matter of Knight Capital Americas LLC: Order Instituting Administrative and Cease-and-Desist Proceedings. Exchange Act Release No. 70694, Administrative Proceeding File No. 3-15570. <https://www.sec.gov/litigation/admin/2013/34-70694.pdf> 16 October 2013.
 - [192] U.S. Securities and Exchange Commission. 2023. Conflicts of Interest Associated with the Use of Predictive Data Analytics by Broker-Dealers and Investment Advisers. Proposed rule, Release Nos. 34-97990, IA-6353, File No. S7-12-23; 88 Fed. Reg. 53960 (9 August 2023). <https://www.federalregister.gov/documents/2023/08/09/2023-16377>
 - [193] U.S. Securities and Exchange Commission. 2024. SEC Charges Two Investment Advisers with Making False and Misleading Statements About Their Use of Artificial Intelligence. Press Release 2024-36. <https://www.sec.gov/news/press-release/2024-36> 18 March 2024. Orders: In the Matter of Delphia (USA) Inc., Advisers Act Release No. 6573; In the Matter of Global Predictions, Inc., Advisers Act Release No. 6574.
 - [194] U.S. Securities and Exchange Commission. 2025. Withdrawal of Proposed Regulatory Actions. Release Nos. 33-11377, 34-103247, IA-6885, IC-35635; 90 Fed. Reg. 25531 (17 June 2025). <https://www.federalregister.gov/documents/2025/06/17/2025-11110> Withdraws, among others, the predictive data analytics proposal (88 Fed. Reg. 53960).
 - [195] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303. doi:10.1038/s41562-024-02024-1
 - [196] Shaw Walters, Sam Gao, Shakker Nerd, Feng Da, Warren Williams, Ting-Chien Meng, Amie Chow, Hunter Han, Frank He, Allen Zhang, Ming Wu, Timothy Shen, Maxwell Hu, and Jerry Yan. 2025. Eliza: A Web3 Friendly AI Agent Operating System. *arXiv preprint arXiv:2501.06781* (2025).
 - [197] Dannong Wang, Jaisal Patel, Daochen Zha, Steve Y. Yang, and Xiao-Yang Liu. 2025. FinLoRA: Benchmarking LoRA Methods for Fine-Tuning LLMs on Financial Datasets. *arXiv preprint arXiv:2505.19819* (2025).
 - [198] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science* 18, 6 (2024), 186345. doi:10.1007/s11704-024-40231-1
 - [199] Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. LLMFactor: Extracting Profitable Factors through Prompts for Explainable Stock Movement Prediction. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 3120–3131. doi:10.18653/v1/2024.findings-acl.185
 - [200] Qian Wang, Yuchen Gao, Zhenheng Tang, Bingqiao Luo, Nuo Chen, and Bingsheng He. 2025. Exploring LLM Cryptocurrency Trading Through Fact-Subjectivity Aware Reasoning. In *Advances in Financial AI Workshop at ICLR 2025*.
 - [201] Saizhuo Wang, Hang Yuan, Lionel M. Ni, and Jian Guo. 2024. QuantAgent: Seeking Holy Grail in Trading by Self-Improving Large Language Model. *arXiv preprint arXiv:2402.03755* (2024).
 - [202] Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2025. Alpha-GPT: Human-AI Interactive Alpha Mining for Quantitative Investment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 196–206. doi:10.18653/v1/2025.emnlp-demos.14
 - [203] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024. Executable Code Actions Elicit Better LLM Agents. In *Proceedings of the 41st International Conference on Machine Learning (ICML) (PMLR, Vol. 235)*. 50208–50232.
 - [204] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*. 24824–24837.
 - [205] Halbert White. 2000. A Reality Check for Data Snooping. *Econometrica* 68, 5 (2000), 1097–1126. doi:10.1111/1468-0262.00152
 - [206] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv preprint arXiv:2303.17564* (2023).
 - [207] Siyi Wu, Junqiao Wang, Zhaoyang Guan, Leyi Zhao, Xinyuan Song, Xinyu Ying, Dexu Yu, Jinhao Wang, Hanlin Zhang, Michele Pak, Yangfan He, Yi Xin, Jianhui Wang, and Tianyu Shi. 2025. MountainLion: A Multi-Modal LLM-Based Agent System for Interpretable and Adaptive Financial Trading. *arXiv preprint arXiv:2507.20474* (2025).
 - [208] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu,

- Xuanjing Huang, Qi Zhang, and Tao Gui. 2025. The Rise and Potential of Large Language Model Based Agents: A Survey. *Science China Information Sciences* 68, 2 (2025), 121101. doi:10.1007/s11432-024-4222-0
- [209] Yunfeng Xia, Chao Li, Lei Li, Chenhao Zhang, Li Duan, Runhua Xu, and Wei Wang. 2026. Toward Web 4.0: Bidirectional Trust between AI Agents and Blockchain. *Blockchain* (2026), 1–58. doi:10.55092/blockchain20260008
- [210] Yihan Xia, Panpan You, Taotao Wang, Fang Liu, Han Qi, Xiaoxiao Wu, and Shengli Zhang. 2026. Agentic Trading: When LLM Agents Meet Financial Markets. *arXiv preprint arXiv:2605.19337* (2026).
- [211] Yuxin Xiang, Yuchen Lei, Yuanzhe Zhang, Qin Wang, Tsz Hon Yuen, Andreas Deppeler, and Jiangshan Yu. 2025. Leveraging Large Language Models to Bridge Cross-Domain Transparency in Stablecoins. *arXiv preprint arXiv:2512.02418* (2025).
- [212] Yijia Xiao, Edward Sun, Tong Chen, Fang Wu, Di Luo, and Wei Wang. 2025. Trading-R1: Financial Trading with LLM Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2509.11420* (2025).
- [213] Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2025. TradingAgents: Multi-Agents LLM Financial Trading Framework. In *AAAI 2025 Workshop on Multi-Agent AI in the Real World (MARW)*. Oral.
- [214] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyang Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024. FinBen: A Holistic Financial Benchmark for Large Language Models. In *NeurIPS 2024 Datasets and Benchmarks Track*. 95716–95743.
- [215] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*. 33469–33484.
- [216] Fei Xiong, Xiang Zhang, Aosong Feng, Siqi Sun, and Chenyu You. 2025. QuantHarness: Price-Driven Multi-Agent LLMs for High-Frequency Trading. *arXiv preprint arXiv:2509.09995* (2025).
- [217] Guojun Xiong, Zhiyang Deng, Keyi Wang, Yupeng Cao, Haohang Li, Yangyang Yu, Xueqing Peng, Mingquan Lin, Kaleb E Smith, Xiao-Yang Liu, Jimin Huang, Sophia Ananiadou, and Qianqian Xie. 2025. FLAG-TRADER: Fusion LLM-Agent with Gradient-based Reinforcement Learning for Financial Trading. In *Findings of the Association for Computational Linguistics: ACL 2025*. 13921–13934. doi:10.18653/v1/2025.findings-acl.716
- [218] Yumo Xu and Shay B. Cohen. 2018. Stock Movement Prediction from Tweets and Historical Prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1970–1979. doi:10.18653/v1/P18-1183
- [219] Lewen Yan, Jilin Mei, Tianyi Zhou, Lige Huang, Jie Zhang, Dongrui Liu, and Jing Shao. 2025. TradeTrap: Are LLM-based Trading Agents Truly Reliable and Faithful? *arXiv preprint arXiv:2512.02261* (2025).
- [220] Cathy (Liu) Yang, Kevin Bauer, Xitong Li, and Oliver Hinz. 2026. My Advisor, Her AI, and Me: Evidence from a Field Experiment on Human–AI Collaboration and Investment Decisions. *Management Science* 72, 1 (2026), 242–264. doi:10.1287/mnsc.2022.03918
- [221] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. In *FinLLM Symposium at IJCAI 2023*. arXiv:2306.06031.
- [222] Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and Christina Dan Wang. 2024. FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models. *arXiv preprint arXiv:2405.14767* (2024).
- [223] Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning. *arXiv preprint arXiv:2309.13064* (2023).
- [224] Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. 2025. TwinMarket: A Scalable Behavioral and Social Simulation for Financial Markets. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. 70830–70880.
- [225] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.
- [226] Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024. MIRAI: Evaluating LLM Agents for Event Forecasting. *arXiv preprint arXiv:2407.01231* (2024).
- [227] Junyi Ye, Bhaskar Goswami, Jingyi Gu, Ajim Uddin, and Guiling Wang. 2024. From Factor Models to Deep Learning: Machine Learning in Reshaping Empirical Asset Pricing. *arXiv preprint arXiv:2403.06779* (2024).
- [228] Yuxuan Ye, Jun Han, Ao Hu, Juncheng Bu, Yiyi Chen, Liangjian Wen, Danilo Mandic, Danny Dongning Sun, Yinghui Xu, and Zenglin Xu. 2026. The Alpha Illusion: Reported Alpha from LLM Trading Agents Should Not Be Treated as Deployment Evidence. *arXiv preprint arXiv:2605.16895* (2026).

- [229] Haofei Yu, Fenghai Li, and Jiaxuan You. 2025. LiveTradeBench: Seeking Real-World Alpha with Large Language Models. *arXiv preprint arXiv:2511.03628* (2025).
- [230] Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. 2023. Generating Synergistic Formulaic Alpha Collections via Reinforcement Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 5476–5486. doi:10.1145/3580305.3599831
- [231] Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W. Suchow, Denghui Zhang, and Khaldoun Khashanah. 2025. FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design. *IEEE Transactions on Big Data* 11, 6 (2025), 3443–3459. doi:10.1109/TBDDATA.2025.3593370
- [232] Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making. In *Advances in Neural Information Processing Systems (NeurIPS)*. 137010–137045.
- [233] Hang Yuan, Saizhuo Wang, and Jian Guo. 2024. Alpha-GPT 2.0: Human-in-the-Loop AI for Quantitative Investment. *arXiv preprint arXiv:2402.09746* (2024).
- [234] Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, Sujian Li, Mengnan Du, and Yongfeng Zhang. 2026. When AI Meets Finance (StockAgent): A Benchmark for Simulating Large Language Model Behaviors in Controlled Trading Environments. *ACM Transactions on Intelligent Systems and Technology* (2026). doi:10.1145/3844605 Article 3844605, published online 2 September 2026.
- [235] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. 2025. Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents. In *International Conference on Learning Representations (ICLR)*.
- [236] Jaden Zhang, Gardenia Liu, Oliver Johansson, Hileamlak Yitayew, Kamryn Ohly, and Grace Li. 2026. Prediction Arena: Benchmarking AI Models on Real-World Prediction Markets. *arXiv preprint arXiv:2604.07355* (2026).
- [237] Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. 2024. A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4314–4325. doi:10.1145/3637528.3671801
- [238] Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A Survey on the Memory Mechanism of Large Language Model-Based Agents. *ACM Transactions on Information Systems* 43, 6 (2025), 1–47. doi:10.1145/3748302
- [239] Li Zhao, Rui Sun, Zuoyou Jiang, Bo Yang, Yuxiao Bai, Mengting Chen, Jing Li, and Zuo Bai. 2025. ContestTrade: A Multi-Agent Trading System Based on Internal Contest Mechanism. *arXiv preprint arXiv:2508.00554* (2025).
- [240] Tianjiao Zhao, Jingrao Lyu, Stokes Jones, Harrison Garber, Stefano Pasquali, and Dhagash Mehta. 2025. AlphaAgents: Large Language Model Based Multi-Agents for Equity Portfolio Constructions. *arXiv preprint arXiv:2508.11152* (2025).
- [241] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Yiwen Hu, Jian-Yun Nie, and Ji-Rong Wen. 2026. A Survey of Large Language Models. *Frontiers of Computer Science* 20, 12 (2026), 2012627. doi:10.1007/s11704-026-60308-3
- [242] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing Large Language Models with Long-Term Memory. In *AAAI Conference on Artificial Intelligence*, Vol. 38. 19724–19731. doi:10.1609/aaai.v38i17.29946
- [243] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations (ICLR)*.
- [244] Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025. DianJin-R1: Evaluating and Enhancing Financial Reasoning in Large Language Models. *arXiv preprint arXiv:2504.15716* (2025).
- [245] Jie Zhu, Yuanchen Zhou, Shuo Jiang, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2026. Fin-PRM: A Domain-Specialized Process Reward Model for Financial Reasoning in Large Language Models. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*. 6136–6144. doi:10.24963/ijcai.2026/683
- [246] Jinan Zou, Qingying Zhao, Yang Jiao, Haiyao Cao, Yanxi Liu, Qingsen Yan, Ehsan Abbasnejad, Lingqiao Liu, and Javen Qinfeng Shi. 2022. Stock Market Prediction via Deep Learning Techniques: A Survey. *arXiv preprint arXiv:2212.12717* (2022).

Appendices

The appendices extend the main text. Appendices A2–A9 annotate Sections 2–9 with further studies, caveats on reported results, coding rules, and the construction of figures; each of their subsections carries the number and title of the subsection it annotates, prefixed with A (Appendix A5.3 annotates Section 5.3), and subsections without notes are omitted. Appendix A10 is a glossary, Appendix A11 describes the survey methodology, and Appendix A12 gives the full evidence tables behind the condensed tables and figures of the main text.

A2 Related Surveys

Notes to Section 2.

Surveys not discussed in the main text. Among pre-LLM surveys, only Cao [27] and Pipas et al. [158] partially compare paradigms, while Sezer et al. [171] and Sun et al. [180] cover deep learning for forecasting and reinforcement learning for trading, respectively. Guo et al. survey interaction and communication in multi-agent LLM systems [72]. Agent-security surveys catalog prompt injection, memory poisoning, tool misuse, and inter-agent attacks, but do not consider market-specific adversaries [39, 78]. Dong et al. organize finance agents by institutional role, sacrificing depth on architectures and markets for breadth [44]. Gong’s stylized model links agent autonomy, heterogeneity, coupling, and infrastructure concentration to systemic risk [66].

Xia et al.’s two cohorts. Xia et al.’s protocol audit finds reporting far more complete in its 2026 extension cohort than in its earlier cohort [210]. Appendix A11 reconciles their counts with those of our evidence tables.

A3 Background: Paradigm Evolution in Quantitative Trading

Notes to Section 3. In Figure 1, the dashed arc marks the convergence of reinforcement learning and LLMs in hybrid systems such as FLAG-Trader [217].

A3.1 Traditional Quantitative Trading

Further roots and caveats. The modern roots of quantitative trading include 1970s options pricing [17]. Statistical arbitrage builds on cointegration [48] and therefore assumes a stationarity that often fails in practice. Machine-learned return signals earn most of their profits in stocks that are hard to arbitrage [7].

A3.2 Machine Learning and Deep Learning

Deep asset pricing. Further deep asset-pricing work imposes no-arbitrage constraints through adversarial test assets [31] and traces a lineage from the CAPM through instrumented principal components analysis (IPCA) [227]. Qualifying §3.2: Feng et al.’s selected deep factor models reach out-of-sample Sharpe ratios below the 3.67 of five-factor IPCA, yet the authors themselves describe the two as comparable [55].

Reinforcement learning and pre-LLM text models. FinRL-Meta extended standardized RL environments to stocks, cryptocurrencies, and forex [129], and reward shaping remains an open problem alongside sample efficiency [158]. Before FinBERT, event embeddings [43] and generative models of tweets and prices [218] already learned dense text representations for return prediction.

A3.3 Large Language Models

Domain adaptation. Domain pre-training or fine-tuning (FinGPT [221], FinMA [215], InvestLM [223]) improves downstream financial performance, and finance-specialized LLMs often

beat general models of comparable size on in-domain financial NLP tasks [206, 214]. Several of these models are fine-tuned from open-weight families such as LLaMA [184]; Zhao et al. survey LLM development more broadly [241].

A3.4 From LLM to Agent

Origins of agent memory. The short- and long-term memory designs of trading agents (§5.5) derive from general-purpose agents, notably the memory stream of generative agents [155] and MemoryBank’s long-term memory [242].

A4 Architecture of LLM Trading Agents

Notes to Section 4.

A4.1 Overall Architecture and Information Flow

Further findings by module. *Reasoning:* process-level rewards substantially outperform outcome-only supervision on rigorous reasoning tasks [124]; finance-specific process rewards are discussed in Appendix A5.2. *Action:* LLMs remain far from fluent at long-horizon action in complex interfaces [243]. *Tools:* trading tools range from indicator calculators [237] through multi-agent data platforms [53, 222] to sandboxed code generation [4], and tool selection, sandboxing, and latency remain open problems.

Coding rules for Figure 3. The figure codes the 47 surveyed trading systems, excluding benchmarks, re-evaluations, and studies without a trading outcome. A capability counts only when it is a *distinctive contribution*, that is, proposed by the paper or varied as a main experimental factor rather than inherited as default scaffolding. Debate counts as collaboration. Memory requires cross-episode storage with retrieval, so rolling context windows (e.g., the four weeks of the crypto portfolio MAS [141]) do not count. A system counts in every market it covers, and futures and options include commodity ETFs [63].

Counts not visible in Figure 3. Seven of the nine memory designs are equity-only. Crypto perception relies chiefly on on-chain data, charts, and retrieval. HedgeAgents [116] is the only system that allocates jointly across markets.

A4.2 Functional Paradigms: Trader, Alpha Miner, and Bounded Component

Alpha Miner lineage and boundary cases. AlphaGen’s reinforcement-learning factor search, which uses no language semantics, preceded the LLM miners and reached a CSI 300 test IC of 0.073 against 0.040 for XGBoost [230]. Alpha-GPT 2.0 extends mining to modeling and analysis and therefore shades into the Trader paradigm [233].

A4.3 Single-Agent vs. Multi-Agent Organization

Naming. QuantAgents [118] (Table 9) is unrelated to Wang et al.’s QuantAgent [201].

Construction of Figure 4. In panel (a), the vs.-prior points are re-runs of FinMem, FinAgent, and HedgeAgents reported by FinCon and QuantAgents [118, 232]. The ablation points take the full system minus the ablated variant (Expert Teams over portfolios of 20–50 stocks) [42, 118, 147, 239]. The vs.-single points come from Macro Economists and from the crypto portfolio MAS, whose twelve configurations are each compared with the single agent under the same prompting, one GPT-4o run each [63, 141]. AlphaAgents reports its controlled comparison only in a figure and is therefore not plotted [240]. Panel (b) uses rank IC instead of return for MASS. Short windows

make each Sharpe ratio noisy [132], but the standard error of a gap shrinks as the two return series become more correlated, as they are in ablations.

Further evidence on multi-agent designs. Outside trading, debate over shared inputs beats a single agent on checkable reasoning tasks [46], but it costs more and is more sensitive to hyperparameters than self-consistency [178], and majority voting alone improves with agent count [111]. In Agent Market Arena, the persona-based agent was identical across backbones [161]. FS-ReasoningAgent’s fact–subjectivity split does not fully restore the expected model ordering, since GPT-4 still trails GPT-3.5 on SOL [200]. In a single-asset crypto system, the aggregating agent earns less than either of its specialists [177]. In financial compliance, the single reasoning model DianJin-R1 matches an eight-call workflow (96.0 vs. 95.0 accuracy), which suggests that some decompositions are scaffolding rather than a source of capability [244].

A5 Key Techniques for LLM Trading Agents

Notes to Section 5.

A5.1 Retrieval-Augmented Generation

Domain-specific retrieval. MarketSenseAI 2.0 [53] retrieves from institutional macroeconomic reports using semantic chunking and hypothetical-document embeddings. Whether such domain-specific retrieval beats off-the-shelf vector search inside a trading loop has not been tested.

Caveat on FactFin. FactFin’s retrieval gain [117] is measured inside a pipeline that generates strategy code, not in an agent that trades directly.

A5.2 Prompting and Reasoning Strategies

Caveat on Trading-R1. Trading-R1’s [212] backtest covers a short window and only six assets.

Process reward models. Fin-PRM [245] supervises financial reasoning at both the step and the trajectory level and verifies steps against curated financial facts. Used for supervised fine-tuning, it gains 12.9 points on CFLUE over a Qwen2.5-7B base and also beats general mathematical process reward models, majority voting, and a rule-based reward. Neither the RL-trained trading systems [14, 217] nor the reasoning-distilled agents [212] have yet combined such a specialized reward model with policy training.

A5.3 Fine-Tuning and Domain Adaptation

Fine-tuning at the perception end. When fine-tuned on newsflow, decoder-only models (Mistral, Llama 3) outperform encoder-only ones, most clearly in the large North American universe [73]. TradExpert [42] LoRA-tunes (Table 8) four LLaMA-2-7B experts, one per modality (news, market data, alpha factors, and fundamentals), whose mixture reaches a Sharpe ratio of 5.01 on the DOW 30 (Table 12).

Lighter variants of “LLM as sensor, RL as executor”. FLAG-Trader (§5.3) trains the LLM itself as the policy. Lighter variants keep a conventional RL agent and only feed it LLM output: one puts GPT-derived sentiment into a PPO agent’s state, action scaling, and reward [188], and FinRLlama rewards a small model’s headline sentiment when it matches realized returns [68]. The former beats a price-only baseline only in-sample, and FinRLlama shows only lower return dispersion than its untuned base model, so both remain design patterns rather than evidence.

Further out-of-distribution evidence. Finance-tuned models lead on only one of the 29 CNFinBench subtasks [41].

A5.4 Tool Use and External Integration

Code as action in finance. FinVerse [4] retrieves API documentation to generate code over 642 financial APIs in a Docker sandbox. Its supervised fine-tuned open models match GPT-4 on robustness but trail it by about 20 points on best-answer rate.

Levels of on-chain integration. Alqithami [2] orders agent–blockchain integration on five levels, from read-only analytics through delegated execution via ERC-4337 account abstraction to fully autonomous multi-agent workflows, and finds that most production systems sit at the intermediate levels.

A5.5 Memory Architectures

Position in general memory taxonomies. Zhang et al. classify agent memory by source, form, and operation [238]; in their terms, nearly every trading agent uses textual, cross-trial memory accessed through retrieval. Huang et al. [85] classify finance as a memory-critical domain and name its low signal-to-noise ratio as the main open problem.

A5.6 Perception Techniques: Sentiment, Multimodal, and Alternative Data

Sentiment by regime. Removing FS-ReasoningAgent’s subjectivity channel hurts most in bull markets, and removing its statistics channel hurts most in bear markets [200].

Charts. FinAgent [237] was, by its authors’ account, the first trading agent to feed K-line charts to a vision-language model (GPT-4V). In the Agent Trading Arena [142], GPT-4o’s return gain from 33.7% to 47.7% includes reflection; charts alone raise it only to 35.8%. A convolutional network trained on images of stock-level price charts out-predicts momentum and reversal signals and transfers across countries [90]; whether a zero-shot vision-language model can extract the same information remains open.

On-chain data. CryptoTrade’s on-chain ablation [119] is run on an ETH bull-market window, so its gain may not hold in other regimes.

A6 Applications Across Financial Markets

Notes to Section 6.

A6.1 Equity Markets

Further equity systems. MarketSenseAI’s GPT-4 stock selection, using chain-of-thought and in-context learning, yielded 10–30% excess return [54]. LLMFactor [199] predicts price direction from GPT-extracted news factors and company relations plus a textualized five-day price path. Its 60–66% accuracy is measured on 2014–2021 data that overlap its backbones’ training data, it reports no transaction costs, and its Chinese templates reach an MCC of 0.245 on CSI 300 constituents. The best configurations of the LLM-enhanced Black-Litterman model [105], whose LLM sees only prices and company metadata, beat equal weighting over a 10-month test.

Chinese A-shares. Kou et al.’s alpha-mining pipeline [101] beats a falling SSE 50 index as well as SEP and FinCon over the same window (Table 10), and it generalizes to the CSI 300 and S&P 500.

A6.2 Cryptocurrency and DeFi

Centralized exchange (CEX) trading. The gains that the crypto portfolio MAS obtains from indicator skills reverse in bear weeks [141]. MountainLion [207] adds GraphRAG and adaptive weighting of LLM and classical machine-learning forecasts to four specialist agents. Makri et al. [144]

tune only the embeddings, normalization, and input and output layers of frozen Llama-3 and GPT-2 backbones as time-series models; these beat PatchTST, LSTM, and MLP baselines on next-day Ethereum price error but are not evaluated in trading.

Meme coins and copy trading. In meme-coin copy trading on Solana, XGBoost ranks wallets slightly better than an LLM multi-agent system (AUC 0.727 vs. 0.716), but only the multi-agent system yields positive copier returns (about 3% per position) once the imitation penalty of bonding curves is modeled [140]. Whether copying wallets flagged as smart money is fair dealing is unaddressed, and such venues sit largely outside the algorithmic-trading rules discussed in §9.4.

On-chain perception. Frontier LLMs read node facts from Bitcoin transaction graphs almost perfectly but classify address types with only 40–50% accuracy, limited by capability and token budgets; compressed graph formats and connectivity-preserving sampling ease both limits [106]. A GPT-5 agent reconciles stablecoin reserve attestations with circulation data through a Model Context Protocol tool layer and flags anomalous USDT and USDC months [211]. Toyoda et al. [185] classify such uses as input enhancement, output interpretation, or direct prediction.

Deployed frameworks and agent tokens. ElizaOS bots reportedly manage over \$25M [157], yet the framework’s own paper reports only general agent ability (GAIA average 19.42 vs. 29 for GPT-4o) and no trading results [196]. In February 2026, one protocol (Virtuals) held roughly 90% of the market capitalization of tokenized agents, only 2 of 13 agent-related Ethereum standards were final, and an industry survey of 306 projects counted 13% trading and analytics agents against 44% meme and sentiment agents [209].

Maximal extractable value (MEV). MEV searchers are deterministic and optimized for latency [37], and block-time latency makes LLM reasoning about transaction ordering unlikely, though not impossible. The nearest LLM-adjacent case we found is off-chain latency arbitrage against stale on-chain prediction-market prices, which is proposed without results [13].

A6.3 Emerging Entry Points: Forex and Derivatives

Why forex agents are absent. Forex is the largest market (about \$7.5 trillion in daily turnover in April 2022 [12]), yet forex agents are also absent from Ding et al.’s review [40] and Xia et al.’s evidence map [210]. We conjecture two reasons: high-frequency institutions dominate at latencies far below LLM inference times, and the lower volatility of major pairs limits sentiment-driven alpha.

Caveat on Izadyar. Because the study has no actual-minus-consensus benchmark, the LLM’s increment over a mechanical surprise rule remains open [89].

Futures and commodities. The forty LLM-generated futures factors [35] come from a single unreplicated study and face the multiple-testing hurdle of factor research [76]. GPT-4o scores of energy news on five sentiment dimensions predict weekly WTI direction at an AUC of 0.634 (0.652 with FinBERT), against 0.569 for a vendor score, although with about 300 weekly observations the differences may lie within noise [36]. Ghali et al. [64] fuse agent-written yearly news summaries with World Bank prices and raise the AUC for annual commodity price spikes from at most 0.57 to 0.94, although retrospective summaries may encode hindsight. Macro Economists’ LLM agents keep their edge under 30 bps one-way costs, whereas the edge of the z-score rule vanishes near 5 bps [63]. QuantHarness’s agents lead in three-candle directional accuracy on four-hour bars despite trailing on Sharpe ratio [216].

Caveat on the Option Query Language (OQL). OQL [139] lacks portfolio state, position management, closed-loop evaluation, and non-LLM baselines.

A6.4 Prediction Markets

Venues and tooling. Polymarket Agents [159] is an open-source toolkit covering the path from news retrieval to execution. Using LLMs to mine Polymarket for combinatorial relations between markets, Saguillo et al. [166] estimate that about \$40M has already been extracted through rebalancing and combinatorial arbitrage.

Further forecasting results. By mid-2025, o3 reached a Brier score of 0.135 on 334 Metaculus questions, whereas hired expert forecasters reached a median of 0.023 on a 157-question subset, and every model forecast politics better than economics and finance [138]. Tool-using GPT-4o agents reach an F1 near 33 on geopolitical events in MIRAI [226]. Across 991 participants, LLM assistants improved human forecasters' accuracy by 24–28% regardless of forecaster skill [169].

Further trading results. In Prediction Arena [236], losses over the same final month were far smaller on Polymarket, where agents discovered markets themselves, than on Kalshi's curated markets (about -1% vs. -22.6%), and outcomes tracked initial accuracy and adding to correct positions rather than research volume or compute. In PolyBench [34], all seven LLMs state confidence near 0.85 regardless of domain, and crypto questions are deeply unprofitable.

A6.5 Summary

Cross-market systems. LLM agents address multi-asset allocation but not cross-market arbitrage, whose latency-sensitive detection of price gaps remains deterministic. Besides HedgeAgents, FinAgent [237] also spans stocks and cryptocurrencies.

A7 Evaluation Frameworks and Empirical Analysis

Notes to Section 7.

A7.1 Evaluation Metrics

Secondary metrics and scorecards. Beyond return and Sharpe ratio, some papers report the Sortino ratio (StockBench [33]), the Calmar ratio (TradeTrap [219]), portfolio entropy and the PRUDEX-Compass scorecard (HedgeAgents and QuantAgents [116, 118]), or, for prediction tasks, accuracy and the Matthews correlation coefficient [42, 215]. Agent Security Bench's Net Resilient Performance trades task utility off against resilience to attack [235].

A7.2 Backtesting Frameworks and Simulation Environments

Further environments. The 2024 FinRL Contest added a task on LLM-engineered signals [68]. MarS [110] trains a generative market model (not an LLM) on order data to produce realistic order flows for evaluating agents. EconAgent's LLM agents reproduce macroeconomic phenomena more realistically than rule-based agents [112], and an RL-controlled LLM learns to manipulate sentiment through social posts, up to pump-and-dump schemes [23, 24]. In FCLAgent [77], which leaves price and volume formation to calibrated rules, a handful of LLM agents among a thousand reproduce the all-time-high anomaly that rule-based populations miss.

Behavioral validity. In credence-goods markets, GPT-5.1 agents follow their prompts and ignore the incentives that make verifiability and reputation effective among humans [49], a caution for simulations that use LLM agents as proxies for human participants.

A7.3 Major Benchmarks and Datasets

Further benchmarks and results. On FinDABench’s [127] structured data analysis tasks, GPT-4 reaches only 32.37% zero-shot. On BizFinBench [137], the strongest models average 71–74 overall but score only 52–56 on stock-price prediction. In InvestorBench, proprietary models beat buy-and-hold on only two to four of seven stocks [109]. On CryptoBench, the best models retrieve on-chain facts well but predict poorly, with top overall accuracy near 44% [71]. AgentBench [128] supplies a general template for agent evaluation and a failure taxonomy that transfers to trading.

A7.4 Empirical Results Summary

Counting rule. Table 11 omits Capponi et al.’s prediction-market study, so it lists 18 rather than the 19 author-run systems of Table 10.

Construction of Figure 5. The figure plots one or two representative rows per paper from Table 10: the best and worst asset where both are reported, and the net figure where an independent paper reports one. The medians in panel (b), over 25 author-run and 17 independent points, pool cumulative, annualized, and total returns, count several points per paper, and compare windows of unequal length. The 17 independent points mix 7 re-runs of published systems with 10 points from benchmarks of generic base LLMs.

A7.5 Evaluation Protocols and Their Pitfalls

Construction of Figure 6. Panel (a) pairs FinMem’s per-stock Sharpe ratios as reported [231] with FINSABER’s re-run over the same window with commissions and two backbones [113]. It also pairs TradingAgents and QuantHarness (Xiong et al. [216]) as reported with Ye et al.’s re-runs under full frictions [228]. In panel (b), ARIMA beats buy-and-hold under two of the four selection rules [113].

Windows spanning a downturn. Only HedgeAgents [116] and QuantAgents [118] test over three years that include both bull and bear phases, and R&D-Agent-Quant’s 2017–2020 window spans the 2018 drawdown [121].

Contamination channels. FinMem and FinCon test through April and June 2023, respectively, with GPT-4-Turbo, whose training data extend into 2023. Alpha Miners add a second channel: their weighting layers are trained on data adjacent to the test window [101, 239].

Mitigations of look-ahead bias. Anonymizing headlines does not necessarily lower returns: Glasserman and Lin’s anonymized GPT-3.5 headlines earn *more* in-sample, suggesting that firm knowledge distracts more than look-ahead helps [65]. FinCAD subtracts an elicited memory prior from the logits; this cuts the in-sample mega-cap returns of the high-memorization Qwen2.5-14B by 45% but leaves post-cutoff 2025 results essentially unchanged [114]. Post-cutoff evaluation controls leakage only if the post-cutoff sample is the primary test [135]. Chronologically consistent backbones match Llama-3.1-8B’s news-based long-short Sharpe ratio over 2008–2023 (4.92 vs. 4.90), but at 1.5B parameters or fewer they cannot yet drive agents [79]. Further controls include Izadyar’s four look-ahead tests [89] and Emmanoulopoulos et al.’s synthetic market with fictional news [47].

Repeated runs and significance tests. The five studies with repeated runs are FinCon, InvestorBench, StockBench, FLAG-Trader, and Expert Teams. FinCon ran a Wilcoxon signed-rank test against its strongest comparator but reported no p -values [232].

Power calculation. For return correlation ρ between an agent and its comparator, the standard error of a Sharpe-ratio margin is roughly $\sqrt{2(1-\rho)}$ times that of a single ratio. No surveyed study reports ρ or a paired test. The main text’s count, 9 of 14 gaps within two standard errors, assumes

$\rho = 0$; the five exceptions are TradingAgents (flagged as anomalous by its authors), TradExpert, MASS, HedgeAgents, and QuantAgents. At $\rho = 0.5$, the count falls to 8 of 14.

Construction of Figure 7. Each system sits on the rung of its primary backbone, and Alpha Illusion is omitted because its backbone is unreported. FinMem is placed by the backbone of its preprint (GPT-4-Turbo, tested in-sample); the GPT-4 run of its journal version would sit after the cutoff. Half marks for fill price are inferred from the return formula (close-to-close for FinMem, FinCon, and InvestorBench), and half marks for cost and slippage denote assertions without values (Profit Mirage and QuantAgents).

A8 LLM Agents vs. Traditional Quantitative Trading

Notes to Section 8.

A8.1 Head-to-Head Evidence

Coding of the 17 agent-level studies. Of the agent-level studies in Table 12, only FinCon, Macro Economists, and PACE report a significance test. Coded by baseline class, the agents lead deep reinforcement learning (DRL) in 7 of the 9 studies that test against it and lead rule-based or statistical baselines in 8 of 11, so the weak spots are specific to individual studies rather than to a baseline class.

Alpha Miner results depend on design. Scored by AlphaAgent’s authors over 2021–2024, an earlier LLM research-and-development framework earns under 2% annualized excess return on the CSI 500 and S&P 500, against 11.0% and 8.7% after costs for AlphaAgent [182].

A8.3 Strategy Interpretability and Transparency

Diagnostics for unfaithful theses. An LLM agent’s natural-language thesis may not reflect what actually drives its decision, and two diagnostics target this failure. Kong et al. [100] call it “narrative bias” and detect it with entity substitution and negative controls. Ye et al. [228] add “parametric prior lock-in”, a factor exposure acquired in pre-training that a fluent thesis masks, and diagnose it with a reversed-evidence test.

A8.4 Market Adaptability and Generalization

Regime dependence. Rankings of LLM backbones reorder between downturns and upturns [33] and across markets [52], and the evidence channel that deserves more weight (sentiment or factual statistics) flips between bull and bear markets [200].

A8.5 Latency, Cost, and Scalability

Construction of Figure 8. In panel (a), the cost per decision divides each paper’s reported total by the number of decisions implied by its trading frequency. This is one decision per trading day for daily systems, with 3×365 days for HedgeAgents, whose crypto and FX sleeves trade daily, and 3×252 days for QuantAgents; one per symbol-day for FINSABER’s multi-symbol runs; and one per one-minute slice for PACE [63, 74, 113, 116, 118, 174]. Grey diagonals are iso-lines of annual spend for one continuously running system, and an open circle moves MASS’s CSI 300 cost per decision, unchanged, to minute bars ($252 \text{ days} \times 390 \text{ minutes}$). Panel (b) assumes 6.5-hour trading sessions; its debate-round latency is Nguyen and Pham’s estimate [148], and its Dow 30 bar assumes that TradExpert scores its 30 stocks sequentially [42].

Cost of self-improvement. Wang et al.’s QuantAgent [201] refines answers in an inner writer–judge loop of T iterations nested in an outer loop of K backtest-feedback rounds, with planning

horizon H . Its reported token cost of order KT^2H grows quadratically in the inner iterations, which makes long self-improvement runs expensive.

A9 Open Challenges and Future Directions

Notes to Section 9.

A9.1 Hallucination and Reliability

Robustness benchmarks. On CNFinBench [41], top LLMs that refuse single-turn violations collapse under multi-turn attacks, with violations rising by 159% by the second round, so reliability estimates need dynamic test sets [11]. FailSafeQA [92] finds that robustness to degraded inputs trades off against refusal when context is missing, with reasoning-tuned models fabricating answers in 41–70% of unanswerable cases. Abstention should therefore be scored as a valid response rather than a missed answer [100].

A9.2 Adversarial Robustness and Security

Procedural agents. Rigid pipelines are not immune to attack: in TradeTrap, procedural agents lose about 61% once their position state is tampered with [219].

A9.4 Regulatory Compliance and Legal Risks

Provisions not stated in the main text. The EU’s RTS 6 also requires annual validation of trading algorithms [50], and the pre-trade controls required by the SEC’s Market Access Rule must be under the broker-dealer’s direct and exclusive control [190]. The model-risk guidance of SR 11-7 and SS1/23 does not bind asset managers directly, since it addresses US banks and UK firms with internal-model approval, respectively [18, 160].

Supervisory positions. The Financial Stability Board’s (FSB) 2026 consultation names “agentic memory poisoning” as a risk and notes that there is little evidence of fully autonomous generative-AI trading [58]. A Bank of England deputy governor stated that human review of every action “is unlikely to be realistic” and that market-wide guardrails similar to circuit breakers may be needed [20].

Liability proposals. Azzutti proposes importing the EU AI Act’s lifecycle duties into trading regulation, scaled to materiality [8], and Rauterberg’s proposed FAIRR Act would drop the scienter requirement for deployers who fail to take reasonable preventive steps [165]. The empirical case for such rules rests on LLM pricing games [60], on RL trading algorithms that sustain collusive profits without communication or intent [45], and on bundle bots in about 25% of meme-coin projects [140], not on deployed LLM traders. For retail protection, an SEC rule requiring firms to neutralize conflicts of interest arising from predictive analytics used with investors was proposed in 2023 [192] and withdrawn in June 2025 [194].

A9.5 Market Stability and Systemic Risk

Further studies. In StockAgent [234], GPT-3.5 and Gemini backbones behave differently, one herding tightly and the other dispersing. In Gong’s stylized market [66], the correlation of actions across agents rises with execution coupling and infrastructure concentration and falls with heterogeneity. AlphaAgent’s authors attribute the decay of LLM-mined alpha to crowding [182], the mechanism behind the quant unwind of August 2007 [96].

A9.7 Human-AI Collaborative Trading

Further evidence on human-AI teams. LLM assistants raise the accuracy of human forecasters [169], but LLMs shown the crowd median under-update and trail a simple average of their own forecast with that median [170]; in that study, mechanical aggregation therefore beat LLM aggregation. In the field, more AI hiring at sixteen investment banks goes with faster, more accurate, and bolder analyst forecasts after 10-K filings, and Morgan Stanley’s GPT-4 assistant made analysts timelier by freeing time for access to management [172]. Two findings further weaken human sign-off as a safeguard: agents follow instructions even into losses [133], and automation bias affects experts as well as novices and persists despite training [154].

A10 Glossary

Table 8. Glossary of finance and machine-learning terms used in this survey.

Term	Definition
Sharpe ratio (annualized)	Mean return in excess of the risk-free rate divided by its standard deviation, scaled to one year (by $\sqrt{252}$ for daily returns if they are serially uncorrelated); its estimate is noisy over short windows [132].
Sortino ratio	As the Sharpe ratio, but dividing by downside deviation, the volatility of returns below a target.
Maximum drawdown (MDD)	Largest peak-to-trough fall in portfolio value over the test window, as a fraction of the peak.
CR / TR / AR	Cumulative or total return: change in portfolio value from start to end of the window; annualized return: the same compounded to a one-year rate.
IC, rank IC, ICIR	Information coefficient: per-period Pearson (IC) or Spearman (rank IC) correlation between a signal and next-period returns across assets; ICIR: mean IC divided by its standard deviation over time.
Alpha	Return not explained by exposure to benchmark risk factors (e.g., market, size, value); the intercept of a factor regression. Also used loosely for any predictive signal (an “alpha factor”).
Buy-and-hold / passive baseline	Holding the traded asset, or an index of the traded universe, over the same window without trading; the minimum benchmark for an active strategy.
Basis point (bp); transaction costs	1 bp = 0.01%. Costs are commissions, fees, and the bid-ask spread paid per trade, usually quoted in bp per side.
Slippage; market impact	Slippage: gap between the price at decision time and the executed price. Market impact: the part of it caused by one’s own order moving the price, which grows with order size relative to liquidity.
Look-ahead bias	Using, at decision time, information that was not yet available.
Survivorship bias	Testing only on assets that survive to the end of the sample, omitting delisted or failed ones, which inflates returns.
Training cutoff; contamination	Cutoff: the date after which a model’s pretraining data end. Contamination: test-period prices, news, or outcomes present in pretraining data, so a backtest before the cutoff can replay memorized history.
Backtest / paper / live trading	Simulation on historical data / simulated orders placed in real time without capital / real orders with capital.
CVaR	Conditional value at risk: expected loss in the worst α fraction of outcomes (e.g., the worst 5%).
DRL (PPO, A2C, DQN)	Deep reinforcement learning: a policy learned from reward by interacting with an environment. PPO (proximal policy optimization) and A2C (advantage actor-critic) are policy-gradient methods; DQN (deep Q-network) learns action values.
GRPO	Group relative policy optimization: RL fine-tuning of an LLM that scores each sampled response against the mean reward of other responses to the same prompt, without a value network.
LoRA / PEFT	Parameter-efficient fine-tuning: training a small set of added parameters while the base model stays frozen; LoRA adds trainable low-rank matrices to the weight updates.
RAG	Retrieval-augmented generation: retrieving documents relevant to a query and conditioning the LLM’s output on them [107].
Chain of thought (CoT)	Prompting or training a model to write intermediate reasoning steps before its answer [204].
MEV	Maximal (originally miner) extractable value: profit obtainable by including, excluding, or reordering transactions within a block, e.g., by front-running or sandwiching pending trades [37].
IPCA	Instrumented principal component analysis: a latent-factor return model whose factor loadings are linear functions of observable firm characteristics.
TWAP; Almgren–Chriss	Time-weighted average price: an order split into equal slices over a fixed interval. Almgren–Chriss: an execution schedule that trades expected market-impact cost against price risk.
Deflated Sharpe ratio; PBO	Deflated Sharpe ratio: a Sharpe ratio corrected for the number of configurations tried and for non-normal returns. Probability of backtest overfitting: the probability that the configuration best in sample ranks below the median out of sample [9, 10].

A11 Survey Methodology

Search. We built the corpus by keyword search and citation tracking. Queries combined {LLM, large language model, GPT, foundation model} with {trading, portfolio, quantitative, alpha, stock, cryptocurrency, forex, options, futures, prediction market} and {agent, multi-agent} over Google Scholar, Semantic Scholar, the ACM Digital Library, IEEE Xplore, arXiv, and SSRN. From the results and foundational papers [134, 198, 232, 237] we tracked citations forward and backward, which also brought in seminal pre-LLM work. Because each tracking round seeded new queries, we report the final corpus instead of a PRISMA flow. This survey cites 246 works, including regulatory documents and vendor sources (204 from 2023 onward and 42 earlier). Of these, 45 distinct primary studies (25 peer-reviewed and 20 preprints) populate the evidence tables, 21 prior surveys are compared in Table 1, and the rest supply methods, benchmarks, and context. Every entry was verified against the publication record and read before citation.

Inclusion criteria. A study is included if an LLM perceives, reasons about, or produces trading decisions, signals, factors, or execution actions and is evaluated empirically on market data or in a market simulator. LLM signal studies count as perception-level evidence, distinct from agent-level systems (Table 12 separates the two). Benchmarks and independent re-evaluations are included because they supply most of the corrective evidence; surveys, position papers, and NLP benchmarks without a trading or forecasting outcome serve only for positioning or capability context. Fixed income and credit, market making, and LLM portfolio construction outside agent systems are out of scope unless an agent system touches them.

Selection rules for the evidence tables. Table 9 lists 16 multi-agent systems. Table 10 (29 rows) and its protocol audit, Table 11 (28 rows), include four groups. The first is the most-cited agent-level system of each design family (single-agent memory, multimodal, hierarchical, debate, contest, population, mixture-of-experts, RL-trained, and alpha mining). The second is the multi-agent systems of Table 9 that report numerical returns, except FS-ReasoningAgent and the crypto portfolio MAS, whose headlines are best-of-several configurations and which serve as controlled comparisons instead (§4.3). The third is two systems from families the list otherwise misses (Kou et al.’s index-benchmarked factor pipeline [101]; Capponi et al.’s fee-net prediction-market arbitrage [29]). The fourth is every independent benchmark or re-evaluation we found that runs LLM agents on market data and reports a return. Table 12 purposively selects 25 studies (24 coded, one abstract-only) comparing an LLM approach with at least one non-LLM baseline on the same window and universe; further comparisons are cited in the text. Figure 3 codes the 47 systems with a trading outcome that appear in these tables or in Sections 4–6 and the appendices, excluding benchmarks and re-evaluations.

Reconciliation with Xia et al.’s audit. Split by period, our cost-disclosure rate reconciles with Xia et al.’s audit of 80 closed-loop studies [210]. None of our seven 2023–2024 author-run systems states a cost rate, close to 1 of 19 in their pre-March-2026 cohort, whereas 8 of our 12 systems from 2025–2026 do, against 46 of 61 in their extension cohort (2 of 19 and 48 of 61 under time-consistent splits). Xia et al. attribute the rise partly to search composition and coding depth.

Limitations of this survey. Four limits bound our findings. First, the purposive tables over-represent the highly cited 2023–2024 systems that seeded our citation tracking. Second, 20 of the 45 primary studies, including most of the corrective evidence, are preprints; where one later appeared at a venue, we cite the formal version. Third, the LLM corpus runs from 2023 to September 2026, and claims of absence refer to it. Fourth, independent does not mean disinterested: two of

the three re-evaluations, Profit Mirage and Alpha Illusion, also advocate upstream designs of their own [117, 228], so they are not neutral on the upstream hypothesis of §8.6.

A12 Full Evidence Tables

This appendix collects four tables: the organization of multi-agent systems (Table 9), the reported results of LLM trading systems against same-window passive baselines (Table 10), the audit of their evaluation protocols (Table 11), and the head-to-head comparisons with traditional baselines (Table 12).

Table 9. Organization of representative multi-agent LLM trading systems. Evidence grades (§4.3): *vs. single*, a single agent receives the same inputs and LLM and runs over the same window; *ablation*, a role or coordination mechanism is removed within the same system, LLM, and window; *vs. prior*, systems published by others are re-run by the authors, confounding the comparison with data, LLM version, prompting, and window. AR, ARR = annualized return; CR = cumulative return; SR = Sharpe ratio.

System	Market	Organization	Roles	Decision aggregation and learning	Evidence for the multi-agent design
TradingGPT [122]	US equities	Peer debate	Three traders: risk-seeking, risk-neutral, risk-averse	Debate records kept in layered memory; no learning beyond memory decay	None (design paper)
FinCon [232]	US stocks; three-stock portfolios	Hierarchical	Seven analyst types (news, filings, earnings calls, data, selection); one manager	Manager decides alone; within-episode CVaR alerts; verbal reinforcement updates prompts across episodes	vs. prior on the same window: TSLA 82.9% CR, Sharpe 1.97 vs. 34.6%, 1.55 FinMem and 12.0%, 0.27 FinAgent; ablation of risk control, not of agents: without CVaR control the portfolio falls 113.8% to 14.7%, without belief updates to 28.4%
TradingAgents [213]	Three US stocks	Hierarchical, two debate stages	Four analysts, bull and bear researchers, trader, three risk debaters, fund manager	Facilitator picks prevailing view after n rounds; manager approves; no cross-episode learning	Rule-based baselines only (best: KDJ+RSI 2.1% vs. 26.6% on AAPL); no single-agent LLM baseline and no ablation; 11 LLM and 20+ tool calls per decision
HedgeAgents [116]	BTC, DJ30, USD/CNY	Hierarchical, conference-based	Three asset-class analysts, one fund manager	Budget conference (mean-variance, CVaR); extreme-market conference (hedging); experience-sharing conference (shared memory)	Ablation: no budget conference, ARR 71.6% to 44.6% (drawdown 12.0%); no extreme-market conference, drawdown 14.2% to 24.4%; no experience sharing, ARR to 50.7%; no single-agent row
QuantAgents [118]	NASDAQ-100; live A- and HK-shares	Hierarchical, meeting-based	Manager; simulated-trading, risk, and news analysts	Manager decides; weekly strategy and market meetings; risk-alert meeting on composite risk score; dual reward from real and simulated trades	vs. prior re-run on NASDAQ-100: ARR 58.7%, Sharpe 3.11 vs. 49.3%, 2.41 HedgeAgents and 45.3%, 2.25 FinAgent; ablation: no market meeting, ARR to 48.6%; no strategy meeting, Sharpe to 2.51; no risk meeting, drawdown 16.9% to 21.9%
ContestTrade [239]	Chinese A-shares	Competitive (internal contest)	Several data-analysis and research agents	Capital to research agents by predicted Sharpe; factors chosen by knapsack over context budget	Ablation: no researcher contest, SR 3.12 to 1.78; no analyst contest, to 2.01; no contests and no deep research, to 0.07 (CR 52.8% to 3.0%)
AlphaAgents [240]	15 US technology stocks	Group chat, then debate to consensus	Fundamental, sentiment, valuation agents	Round-robin debate to unanimous buy or sell; risk profile set by prompt; no learning	vs. single (figure only, one run): multi-agent loses less than single-agent portfolios under risk-neutral prompts, all with negative cumulative return; all trail the equal-weight benchmark under risk-averse prompts

Continued on next page

Table 9 (continued)

System	Market	Organization	Roles	Decision aggregation and learning	Evidence for the multi-agent design
Macro Economists [63]	15 commodity ETFs	Debate between macro personas	Hawkish, dovish, and debate agents; rule-based agent with identical inputs	All emit tilt signals into one shared portfolio engine; no learning	vs. single and vs. rule, 5,000-resample bootstrap: debate Sharpe 0.57 vs. 0.57 hawkish ($p = 0.77$) and 0.001 above the hawkish-dovish average; hawkish and debate beat the 0.53 rule by 0.04 ($p = 0.07, 0.09$), dovish not significant ($p = 0.13$)
Expert Teams [147]	TOPIX 100, market-neutral	Hierarchical, three levels	Technical, quantitative, qualitative, news analysts; sector and macro agents; portfolio manager	Scores and short rationales flow upward; manager ranks stocks; no debate or learning	Leave-one-out ablation, 50 trials, fine-grained team: removing the technical agent cuts Sharpe by 0.40–0.66 at 20–50 stocks ($p < 0.0001$); removing the quantitative, qualitative, news, or macro agent raises it by 0.12–0.50; no single-agent baseline
MASS [74]	SSE 50, CSI 300, ChiNext 100	Population ensemble	Up to 512 independent agents of 16 styles, each scoring a random stock subset	Consensus minus disagreement across agent types; type mixture re-fitted daily to recent returns	Scaling from 64 to 512 agents raises rank IC, “approximately linear” per the authors (figure only); no single-agent baseline; without daily re-fitting of the type mixture rank IC collapses 8.2 to 0.8 on SSE 50; consensus without disagreement 8.2 to 6.3
Crypto portfolio MAS [141]	Top-15 layer-1 cryptocurrencies, weekly	Hierarchical, collaborative, or debate (all three compared)	Crypto (market-data) agent, news agent, trading agent	Trading agent integrates both reports with portfolio state; four-week rolling memory; zero-shot, CoT, RAG, or indicator-skill prompting; no learning	vs. single (GPT-4o, 2025, one run per configuration): best MAS 133.5% CR, Sharpe 1.50 vs. best single agent –8.8%, 0.01; mean CR of 12 MAS configurations 31.0% vs. –23.0% for 4 single-agent ones, gap holds under GPT-5 and Claude Sonnet 4.5; ablation of hierarchical zero-shot: no crypto agent, CR 52.2% to 9.6% (Sharpe 1.02 to 0.42), no memory to 40.7%, no news agent to 51.3% with higher volatility
TradExpert [42]	DOW 30	Mixture of experts	News, market, alpha, fundamental experts; general expert	General expert synthesizes reports; relaxed pairwise ranking for stock selection; LoRA-trained experts	Ablation: no market expert, AR 49.8% to 30.9%; no news expert, to 31.9% (Sharpe 5.01 to 1.74); no alpha expert, to 41.7%; no fundamental expert, to 44.3%
FS-ReasoningAgent [200]	BTC, ETH, SOL	Pipeline with fact/subjectivity split	Statistics, fact, subjectivity, two reasoning agents, trader, reflection	Trader synthesizes factual and subjective reasoning; periodic reflection	Ablation (o1-mini, BTC, one run): removing subjectivity reasoning cuts bull return 76.2% to 66.0%, removing statistics deepens bear loss –15.9% to –20.4%; with the split o1-mini leads on ETH and SOL bull, without it GPT-4o trails GPT-3.5 by 18% on SOL; buy-and-hold beats every variant in each bull window
CryptoTrade [119]	BTC, ETH, SOL	Modular pipeline	Market analyst, news analyst, trader, reflection	Trader synthesizes analyst reports; weekly reflection	Ablation (GPT-4o, bull window, labeled ETH though the full-prompt value matches the BTC row): no reflection 28.5% to 17.1%, no on-chain statistics to 12.7%, base prompt 8.4%
MarketSenseAI 2.0 [53]	S&P 100 and 500	Modular pipeline	News, fundamentals, dynamics, macroeconomic (RAG), signal agents	Signal agent synthesizes by chain-of-thought; monthly rebalancing; no learning	Index benchmarks only: cap-weighted portfolio 125.9% vs. 73.5% S&P 100 with deeper drawdown; attribution alpha 8.0% (equal-weighted; 10.6% cap-weighted), and Carhart/FF5 factors explain 85–88% of variance; no agent ablation

Continued on next page

Table 9 (continued)

System	Market	Organization	Roles	Decision aggregation and learning	Evidence for the multi-agent design
Meme-coin copy trading (Luo et al.) [140]	Solana meme coins	Modular pipeline (independent judges)	Wallet, coin, timing agents	Confidence-weighted average, weights tuned on validation set	Ablation: removing the wallet or coin agent turns copier return negative (−6.7%, −5.9% from +2.9%), removing timing leaves +1.2%; only method with positive copier return under bonding-curve impact, although XGBoost has higher AUC (0.727 vs. 0.716)

Table 10. Reported results of LLM trading agent systems and benchmarks, each against the passive baseline the paper reports over the same window. Rows are a purposive sample, not a census: a representative, highly cited agent-level system of each design family (single-agent memory, multimodal, hierarchical, debate, contest, population, mixture-of-experts, RL-trained, alpha mining), the multi-agent systems of Table 9 that report numerical returns (FS-ReasoningAgent and the crypto portfolio MAS are discussed in Appendix A4.3 instead; TradingGPT reports none), and every independent benchmark or re-evaluation in our corpus that runs LLM agents on market data and reports a return. CR = cumulative return, AR = annualized return, TR = total return, SR = Sharpe ratio, MDD = maximum drawdown, B&H = buy-and-hold. Return definitions differ: FinMem, FinCon, FLAG-Trader, and InvestorBench sum daily log returns; FinAgent annualizes linearly over seven months; Agent Market Arena reports cumulative returns but annualizes Sharpe over two months with a zero risk-free rate; AI-Trader reports the Sortino ratio; PolyBench reports per-trade Sharpe. Where a paper reports several assets, the best and worst are shown. Figures are gross of costs unless marked net. “Costs; LLM cost” gives the transaction-cost assumption, then reported API spend, call counts, or latency. n.r. = not reported; “figure only” = no number in text or tables. Protocol details are in Table 11.

System (LLM)	Market; test window	Return vs. passive, same window	SR; MDD (vs. passive)	Costs; LLM cost
<i>US equities, author-run</i>				
FinMem [231] (GPT-4; GPT-4-Turbo in preprint)	5 US stocks, one at a time; Oct 2022–Apr 2023	61.8% CR on TSLA (best) vs. B&H −18.6%; 4.9% on AMZN (worst) vs. −10.8%	2.68 vs. −0.54; MDD 10.8% vs. 55.3% (TSLA)	n.r.; n.r.
FinCon [232] (GPT-4-Turbo)	8 US stocks; two 3-stock portfolios; Oct 2022–Jun 2023	82.9% CR on TSLA (best of 8) vs. B&H 6.4%; 17.5% on NIO (worst) vs. −77.2%; 113.8% on portfolio 1 vs. 9.3% equal-weight	1.97 vs. 0.15 (TSLA); 3.27 vs. 0.49 (portfolio 1); MDD 29.7% vs. 58.2% (TSLA)	n.r.; n.r.
FinAgent [237] (GPT-4, GPT-4V)	5 US stocks and ETH; Jun 2023–Jan 2024	92.3% AR on TSLA (best of 6) vs. B&H 37.4%; 31.9% on AAPL (worst) vs. 13.0%	2.01 vs. 0.72; MDD 12.1% vs. 32.7% (TSLA)	n.r.; n.r.
TradingAgents [213] (GPT-4o, o1-preview)	3 US stocks; Jan–Mar 2024	26.6% CR on AAPL (best) vs. B&H −5.2%; 23.2% on AMZN (worst) vs. 17.1%	8.21 vs. −1.29, flagged by authors as anomalous; MDD 0.9% vs. 11.9% (AAPL)	n.r.; 11 LLM and 20+ tool calls per decision
MarketSenseAI 2.0 [53] (GPT-4o)	S&P 100; S&P 500 in 2024; Jan 2023–Dec 2024	125.9% TR gross, 123.0% net (cap-weighted, S&P 100) vs. index 73.5%; 55.7% equal-weighted vs. 42.3%	2.76 vs. 2.52; MDD 13.8% vs. 9.7%	10 bps per trade; n.r.
TradExpert [42] (LLaMA-2-7B, LoRA)	DOW 30, top-K buy-and-hold; 2023	49.8% AR vs. DJI 13.9%	5.01 vs. 1.22; MDD 6.6% vs. 9.7%	n.r.; 4.7 s per stock on one A5000 GPU
FLAG-Trader [217] (SmolLM2-135M, RL-trained)	5 US stocks, Oct 2020–May 2021; BTC, Apr–Nov 2023	50.4% CR on TSLA (best) vs. B&H 39.2%; 20.1% on MSFT (worst) vs. 15.3%; 45.5% on BTC vs. 21.8%	1.73 vs. 0.68 (BTC), below B&H on UVV; MDD 24.4% vs. 20.8% (BTC)	n.r.; n.r.
QuantAgents [118] (GPT-4o)	NASDAQ-100, 2021–2023; “live” A- and HK-shares, Q3 2024–Q1 2025	299.6% TR (58.7% AR) vs. NDX 32.5%; live 111.9% (A-shares), 97.7% (HK), no passive reported	3.11 vs. 0.64; MDD 16.9% vs. 35.6%; live SR 2.02, 1.76, MDD n.r.	n.r.; \$180 API over the 3-year backtest
<i>Chinese, Japanese, and other equities, author-run</i>				
Kou et al. [101] (GPT-4o)	SSE 50, Jan 2023–Jan 2024; CSI 300 and S&P 500 half-years	53.2% CR vs. SSE 50 −13.2%	0.29 vs. −0.06 (not annualized); MDD n.r.	Turnover-capped, no value; n.r.
ContestTrade [239] (DeepSeek-V3, R1)	A-shares; Jan–Jun 2025	52.8% CR net vs. CSI All Share 4.4%	3.12 vs. 0.46; MDD 12.4% vs. 13.8%	0.1% per trade; n.r.
MASS [74] (Qwen2.5-72B)	SSE 50, CSI 300, ChiNext 100; 2023	4.95% AR net on CSI 300 (best) vs. index −9.8%; 1.17% on ChiNext (worst) vs. −19.2%	2.23 vs. −2.92; MDD 14.0% vs. 21.4% (CSI 300)	0.1% round-trip; \$0.68–2.27 API per trading day

Continued on next page

Table 10 (continued)

System (LLM)	Market; test window	Return vs. passive, same window	SR; MDD (vs. passive)	Costs; LLM cost
R&D-Agent-Quant [121] (o3-mini; GPT-4o; o4-mini out of sample)	CSI 300, 2017–Aug 2020; CSI 500 and NASDAQ 100 out of sample, 2024–Jun 2025	14.2% AR net on CSI 300; no passive row, Alpha 158 library 5.7%; 19.8% and 28.4% out of sample	IR 1.74 vs. 0.85 (Alpha 158); 2.17 and 1.77 out of sample; MDD 7.4% vs. 7.7%	0.05% buy, 0.15% sell (CSI); 0.1% (NASDAQ); under \$10 per run
Expert Teams [147] (GPT-4o)	TOPIX 100 long-short, composite of six agent strategies; Sep 2023–Nov 2025	13.7% AR gross, 10.6% net vs. TOPIX 100 19.3%	1.22 gross, 0.95 net vs. 1.68; 50-50 blend 2.11; MDD n.r.	10 bps one-way; n.r.
AlphaAgents [240] (GPT-4o)	15 US technology stocks; Feb–May 2024	Figure only, vs. equal-weight 15-stock benchmark	Figure only; MDD n.r.	n.r.; n.r.
<i>Cryptocurrency, author-run</i>				
CryptoTrade [119] (GPT-4o)	BTC, ETH, SOL in bull, sideways, and bear windows; 2023	25.5% TR on ETH bull vs. B&H 22.6% (above B&H in 4 of 9 windows: BTC bear, ETH bull, SOL sideways and bear); 115.2% SOL bull vs. 176.7%; –16.3% SOL bear (worst) vs. –36.1%	0.18 vs. 0.14 (ETH bull); MDD n.r.	Proportional fee, rate n.r.; n.r.
Meme-coin copy trading [140] (GPT-4o)	Solana meme coins, 6,000 projects, held-out split; dates n.r.	3% mean net return per copied position (only positive method); 14% for the identified wallets; no passive	n.r.; n.r.	Bonding-curve impact; 1% platform fee; n.r.
<i>Cross-market, commodities, and prediction markets, author-run</i>				
HedgeAgents [116] (GPT-4-1106)	BTC, DJ30, USD/CNY; 2021–2023	71.6% AR (405% TR) vs. BTC 12.9%, DJIA 7.6%, FX 4.1%	2.41 vs. 0.54 (BTC); MDD 14.2% vs. 76.6% (BTC)	n.r.; \$15 API over 3 years
Macro Economists [63] (GPT-4o-mini)	15 commodity ETFs, weekly; Oct 2023–Feb 2026	7.74% AR hawkish (best), 7.61% dovish vs. inverse-volatility 6.68% and same-input rule 7.11%	0.57 vs. 0.52, 0.56 at 10 bps; MDD 9.4% vs. 9.5%	0–30 bps sweep; \$0.002 API per weekly call
Capponi et al. [29] (Gemini 3.1 Pro)	Kalshi cross-contract arbitrage, \$10,000; Feb–Mar 2026	14.1% net ROI over 46 trades; no passive	4.46 (short sample); MDD 33.6%	Kalshi taker fee 0.07 $N P(1 - P)$; n.r.
<i>Independent benchmarks and re-evaluations</i>				
StockBench [33] (13 LLMs)	20 DJIA stocks, \$100,000; Mar–Jun 2025	2.5% best (Qwen3-235B-Think), –2.8% worst (GPT-OSS-20B) vs. equal-weight B&H 0.4%; 8 of 13 above passive	Sortino 0.042 (Kimi-K2) vs. 0.016; MDD 11.2–14.9% vs. 15.2%	None; n.r.
InvestorBench [109] (13 LLMs)	7 US stocks, Oct 2020–May 2021; BTC, ETH, Apr–Nov 2023; one ETF, 2020	GPT-4, GPT-4o, o1-preview beat B&H on 4, 3, and 2 of 7 stocks (B&H 10.8–52.2%); BTC 34.1% CR (o1-preview) vs. 21.8%	1.61 vs. 0.99 (BTC, o1-preview, camera-ready values); MDD n.r.	n.r.; n.r.
FINSABER [113], re-run (GPT-4o-mini; GPT-4o)	63–91 historical S&P 500 members incl. delisted, four composites; 2004–2024, rolling	FinMem –1.7% to 4.1% AR and FinAgent 4.5–14.0% AR, net of commissions, vs. B&H 6.7–9.9%	FinMem –0.29 to 0.03, FinAgent –0.08 to 0.24 vs. B&H 0.32–0.70; MDD 10.3–28.1% vs. 14.1–35.1%	\$0.0049 per share; \$198 (FinAgent) and \$32 (FinMem) API per composite run
Profit Mirage [117], re-run (GPT-4o)	NASDAQ-100 matched windows, Q2–Q3 2021 vs. Q3–Q4 2024; 6 assets, Jul 2024–Jun 2025	Five agents, incl. Wang et al.’s QuantAgent [201], lose 50–72% of TR and 51–62% of SR post-cutoff; FinMem 52.3% TR on TSLA vs. B&H 57.9%	0.95 vs. 0.98 (FinMem, TSLA); MDD n.r.	Costs and slippage asserted, no values; n.r.
Alpha Illusion [228], re-run (LLM n.r.)	TSLA, NVDA, KO, XOM, MSTR, \$100,000; Jan 2025–Jan 2026	TradingAgents 6.4% gross, 2.3% net; QuantHarness (Xiong et al. [216]) –18.6% gross, –22.1% net; vs. equal-weight B&H 4.8%	0.43 to 0.22 and –0.96 to –1.15, gross to net; MDD per asset only	Commission, spread, impact, token cost; commissions \$105–1,632 per asset n.r.; n.r.
AI-Trader [52], paper trading (6 LLMs)	NASDAQ-100 hourly; SSE 50; 10 crypto pairs; Oct–Nov 2025, two weeks for crypto	US 9.56% CR best (MiniMax-M2), –0.06% worst vs. QQQ 1.87%; A-shares best 1.31% vs. SSE 50 1.65%; crypto best –12.2% vs. index –14.3%	Sortino 4.42 vs. 1.51; MDD 4.9% vs. 3.9% (US best)	n.r.; n.r.
LiveTradeBench [229], paper trading (21 LLMs)	15 US stocks, \$1,000; 10 Polymarket markets, \$500; Aug–Oct 2025, 50 days	Stocks 1.8–6.3% CR, all positive; Polymarket –57.6% to 20.5%, 4 of 21 positive; no passive	Stocks 0.60–2.64, Polymarket –5.26 to 2.38; MDD 1.8–3.5% and 7.5–72.1%	None; n.r.
Agent Market Arena [161], paper trading (5 backbones × 4 agents)	TSLA, BMRN, BTC, ETH; Aug–Sep 2025	InvestorAgent (GPT-4.1) 40.8% on TSLA vs. B&H 46.9%; TradeAgent (GPT-4.1) –38.7% (worst); DeepFundAgent (vote) 9.45% on BMRN vs. –6.9%	6.47 vs. 6.00; MDD 4.4% vs. 6.3% (TSLA)	n.r.; n.r.
Prediction Arena [236], real capital (6 frontier LLMs)	Kalshi, 29 series; Polymarket; \$10,000 each; Jan–Mar 2026, 57 days	Kalshi –16.0% (best) to –30.8%, net of fees; Polymarket –2.7% to 0.0%, mean –1.1% (Feb 9–Mar 9; one model did not trade); against market prices	n.r.; MDD 16.3–30.9%	Real exchange fees, value n.r.; token use tracked, n.r.
PolyBench [34] (7 LLMs)	38,666 Polymarket snapshots, \$10 lots; Feb 2026	Confidence-weighted return 17.6% (MiMo-V2-Flash) to –25.2% (MiniMax-M2.5), 2 of 7 positive; against market prices	Per-trade 0.02 to –0.22; MDD n.r.	Order-book slippage; \$1.34–11.50 API per model

Table 11. Audit of the evaluation protocols of the systems in Table 10, which inherit its row-selection rule (18 author-run and 10 independent rows); Capponi et al. is omitted (its backtest illustrates a relation-discovery method). Mode, coded from the backbone’s documented training cutoff and the test window: BT-IS = backtest overlapping the LLM’s training data; BT-PC = backtest after the training cutoff; Partial = window straddles the cutoff; Paper = simulated portfolio on real-time data; Real = real capital on an exchange; where a paper reports two backbone versions, both codes are given. Fill; costs: assumed execution price, then the cost or slippage assumption. Coding rules: a value or rule is *stated* when the paper gives it in text, tables, or code; *impl.* (implied) = inferred by us from the paper’s return formula, not stated; “asserted, no value” = the paper says costs or slippage are modeled but gives no rate; n.r. = not reported. No system reports fill rates or partial fills (as of September 2026).

System	Test window	Universe	Mode	Fill; costs	Runs	Baselines
<i>Author-run evaluations</i>						
FinMem [231]	Oct 2022–Apr 2023 (6 mo.)	5 US stocks	BT-IS (preprint GPT-4-Turbo); BT-PC with GPT-4	Close-to-close (impl.); n.r.	n.r.	Passive; DRL; LLM
FinCon [232]	Oct 2022–Jun 2023 (8 mo.)	8 US stocks; two 3-stock portfolios	BT-IS	Close-to-close (impl.); n.r.	5 (median)	Passive; DRL; mean-variance; LLM
FinAgent [237]	Jun 2023–Jan 2024 (7 mo.)	5 US stocks; ETH	BT-PC (assuming 2023 GPT-4/GPT-4V snapshots)	n.r.; commission, no value	n.r.	Rule; ML; DRL; LLM
TradingAgents [213]	Jan–Mar 2024 (3 mo.)	3 US stocks	BT-PC	n.r.; n.r.	n.r.	Passive; rule
MarketSenseAI 2.0 [53]	2023–2024 (24 mo.)	S&P 100; S&P 500 (2024)	Partial	n.r.; 10 bps	n.r.	Index; factor models
TradExpert [42]	2023 (12 mo.)	DOW 30	BT-PC	n.r.; n.r.	n.r.	ML; DRL; index
HedgeAgents [116]	2021–2023 (36 mo.); 2024 Q1–Q3 hold-out	BTC; DJ30; USD/CNY	Partial; hold-out BT-PC	n.r.; n.r.	n.r.	Passive; DRL; rule; LLM
QuantAgents [118]	2021–2023 (36 mo.); “live” Q3 2024–Q1 2025	NASDAQ-100; A- and HK-shares	Partial; live: n.r.	n.r.; costs, impact asserted, no values; capital, venue unstated	n.r.	Index; classical; DRL; LLM
ContestTrade [239]	Jan–Jun 2025 (6 mo.)	A-shares	BT-PC	n.r.; 0.1%	n.r.	Rule; ML; DRL; multi-agent
AlphaAgents [240]	Feb–May 2024 (4 mo.)	15 US tech stocks	BT-PC	n.r.; n.r.	n.r.	Equal-weight; single-agent
CryptoTrade [119]	2023 bull, sideways, bear	BTC; ETH; SOL	BT-IS (bear, sideways); Partial (bull)	Open (impl.); prop. fee, no value	n.r.	Passive; rule; deep time-series
Kou et al. [101]	Jan 2023–Jan 2024 (12 mo.); half-years 2021–2023	SSE 50; CSI 300; S&P 500	Partial	n.r.; turnover-capped, no value	n.r.	ML; DRL; LLM; index
FLAG-Trader [217]	Oct 2020–May 2021; Apr–Nov 2023 (BTC)	5 US stocks; BTC	BT-IS	Next-step price; n.r.	Median-SR run	Passive; LLM
Macro Economists [63]	Oct 2023–Feb 2026 (124 weeks)	15 commodity ETFs	BT-PC	Weekly, 1-wk lag; 0–30 bps sweep	Block bootstrap	Same-input rule; passive
Expert Teams [147]	Sep 2023–Nov 2025 (27 mo.)	TOPIX 100, market-neutral	BT-PC	Next-month open; 10 bps one-way (one section)	50 seeds (median, U tests)	Coarse-grained variant; index
MASS [74]	2023 (12 mo.); Q1 2025 leakage check	SSE 50; CSI 300; ChiNext 100	BT-IS; check BT-PC	First 15 min after open; 0.1% round-trip	n.r.	Rule; ML; DL; LLM
R&D-Agent-Quant [121]	2017–Aug 2020; OOS 2024–Jun 2025	CSI 300; CSI 500; NASDAQ 100	BT-IS; OOS Partial (o4-mini cutoff Jun 2024)	Close; 0.05/0.15% (CSI), 0.1% (NASDAQ)	n.r.	Factor libraries; 16 ML/DL models
Meme-coin copy trading [140]	Wallet histories, 6,000 projects	Solana meme coins	BT-PC	Bonding curve, impact modeled; 1% platform fee	n.r.	ML (XGBoost, LASSO, NN)
<i>Independent benchmarks and re-evaluations</i>						
FINSABER [113]	2004–2024 (20 yr), rolling	63–91 historical S&P 500 members incl. delisted	Mixed	n.r.; \$0.0049/share, \$0.99 min. per order	Paired <i>t</i> -tests; CAPM alpha	Rule; ARIMA; XGBoost; DRL; LLM
Profit Mirage [117]	Matched-return in- vs. post-cutoff windows	NASDAQ-100; 6 assets	BT-IS and BT-PC, by design	n.r.; costs and slippage asserted, no values	n.r.	FinLLMs; single- and multi-agent
Alpha Illusion [228]	Jan 2025–Jan 2026 (12 mo.)	TSLA; NVDA; KO; XOM; MSTR	n.r. (backbone)	n.r.; commission, spread, impact, token cost	n.r.	Equal-weight passive
StockBench [33]	Mar–Jun 2025 (4 mo.); regime windows 2025	20 DJIA stocks	BT-PC	Open; none (stated); infeasible orders revised	3 seeds (mean)	Passive

Continued on next page

Table 11 (continued)

System	Test window	Universe	Mode	Fill; costs	Runs	Baselines
InvestorBench [109]	2020–2021 (stocks); 2023 (crypto); 2020 (ETF)	7 US stocks; BTC; ETH; one ETF	BT-IS	Close-to-close (impl.); n.r.	5 (median)	Passive
PolyBench [34]	Feb 2026 snapshots	38,666 Polymarket markets	BT-PC	Order-book sweep, top 5 levels; fees n.r.	n.r.	None
AI-Trader [52]	Oct–Nov 2025 (5–6 wk; 2 wk crypto)	NASDAQ-100; SSE 50; 10 crypto pairs	Paper	n.r.; n.r.; infeasible orders rejected	n.r.	Index
LiveTradeBench [229]	Aug–Oct 2025 (50 days)	15 US assets; 10 Polymarket contracts	Paper	Close (stocks), live quote (Polymarket); none	n.r.	None
Agent Market Arena [161]	Aug–Sep 2025 (2 mo.)	TSLA; BMRN; BTC; ETH	Paper	Daily signal × daily return; fees n.r.	n.r.	Buy-and-hold
Prediction Arena [236]	Jan–Mar 2026 (57 d); 3-d paper cohort	29 Kalshi series; Polymarket (open)	Real (\$10k each); Paper (cohort 2)	Exchange fills at bid mark; fees; rejections reported	n.r.	Market prices

Table 12. Head-to-head evidence comparing LLM-based approaches with traditional baselines. “Traditional baselines” lists each paper’s non-LLM comparators (rule = technical rules or passive strategies; ML = classical machine learning; DL = deep time-series models; DRL = deep reinforcement learning; human = analysts or forecasters). All rows are author-run evaluations; independent benchmarks and re-evaluations, which hold costs, universes, and windows fixed, are consolidated with their baselines in Table 10. Each outcome opens with its code under the rules of Table 5 and its cost flag. Figures are as reported and not comparable across rows.

Study	Market and window	Traditional baselines	Outcome for the LLM approach
<i>Signal-level studies (LLM as perception or predictor)</i>			
Lopez-Lira and Tang [134]	4,123 US stocks, Oct 2021–May 2024, daily rebalanced	FinBERT [6], BERT-family; market	<i>Ahead (1/1); net (10 bps)</i> . GPT-4 overnight-news long-short Sharpe 2.97 at zero cost vs. −0.33 FinBERT; 1.29 at 10 bps round trip, unprofitable at 20 bps; decays 6.54 (2021Q4) to 1.22 (2024)
Kirtac and Germano [98]	US stocks, 965,375 articles 2010–2023; strategy Aug 2021–Jul 2023	Loughran–McDonald [136]; FinBERT [84]; BERT; market	<i>Ahead (1/1); net</i> . OPT long-short Sharpe 3.05 after 10 bps vs. 2.07 FinBERT, 1.23 dictionary, 1.45 market; dictionary scores lose significance ($t = 1.87$)
Guo and Hauptmann [73]	NA, EU, EM equities, 2003–2014 train, 2015–2019 out of sample, monthly	FinBERT [6], FinVader portfolios; universe	<i>Mixed (4/6); gross</i> . Averaged over the six LLM variants, beats the best sentiment or universe portfolio on Sharpe in 4 of 6 universe-portfolio cells, losing NA and EM long-only to FinBERT. NA long-only Sharpe: fine-tuned Llama 3 1.32 vs. 1.22 FinBERT, 0.68 universe; in EM, FinBERT long-only Sharpe 0.70 beats every LLM variant; no costs
Alpha-GPT [202]	WorldQuant IQC 2024 (interim, June 2024); a Joinquant high-frequency contest	Human quant teams	<i>Behind (0/2); ext.; costs n.r.</i> Top-10 worldwide of 41,000+ participants at stage 2; in the Joinquant contest 14% return, Sharpe 5.47 vs. 16%, 5.42 for the top-5% human and 21%, 6.88 for the winner, reached in one day vs. one month
Makri et al. [144]	ETH daily open, 2016–2023	PatchTST; LSTM; MLP; ANN	<i>Ahead (2/2), forecast only</i> . Frozen Llama 3 with tuned embeddings and norms: RMSE 0.052 vs. 0.067 LSTM, 0.073 PatchTST (normalized); few-shot 0.057 vs. 0.082 PatchTST; no trading evaluation
AIA Forecaster [3]	ForecastBench (498 questions) and 1,610 liquid market questions, 2025	Superforecasters; market prices	<i>Mixed (2/4), n.s., forecast only; ext.</i> Brier 0.108 vs. 0.111 superforecasters on FB-7-21 ($p = 0.15$) and 0.110 vs. 0.115 on FB-8-14; trails superforecasters on FB-Market (0.0753 vs. 0.0740) though beating its market price (0.0965), and trails the market on 1,610 liquid markets (0.126 vs. 0.111); ensemble with market 0.106
Izadyar [89]	Nine G10 rates, releases 1996–2024; strategy Jan 2001–Oct 2024, monthly	Dollar, carry, momentum, value factors	<i>Ahead (1/1); gross</i> . LLM-labeled macro releases: long-short Sharpe 0.58–0.60 at 36–48 month lookbacks (about 0.7 at 6 months, figure only) vs. 0.42 carry, 0.33 value; alpha 74% of return at 48 months; no actual-minus-consensus surprise-rule baseline

Agent-level studies, author-run (Trader; A: Alpha Miner; B: bounded component)

Continued on next page

Table 12 (continued)

Study	Market and window	Traditional baselines	Outcome for the LLM approach
FinMem [231]	5 US stocks, Oct 2022–Apr 2023	Buy-and-hold; PPO, DQN, A2C	<i>Ahead (5/5); gross.</i> Highest CR and Sharpe on all five stocks: TSLA 61.8%, 2.68 vs. 33.3%, 0.97 best DRL (DQN) and –18.6%, –0.54 B&H; AMZN 4.9% vs. –6.4% best DRL; risk setting chosen on the test window; 13.5 months of training data vs. ten years for DRL; no costs
FinCon [232]	8 US stocks and two 3-stock portfolios, Oct 2022–Jun 2023	Buy-and-hold; A2C, PPO, DQN; Markowitz; FinRL-A2C	<i>Ahead (10/10); gross.</i> Highest CR and Sharpe on all 8 stocks and both portfolios; portfolio 1 113.8%, 3.27 vs. 19.5%, 0.83 FinRL-A2C and 12.6%, 0.61 Markowitz; portfolio 2 32.9% vs. 15.1% equal-weight, with worse drawdown (21.5% vs. 14.7%); no costs
FinAgent [237]	5 US stocks and ETH, test Jun 2023–Jan 2024	Rule (MACD, KDJ, mean reversion); LGBM, LSTM, Transformer; DQN, SAC, PPO	<i>Mixed (3/6); gross.</i> 36% mean gain in annualized return over the best baseline per asset (TSLA 92.3% vs. 50.0% FinMem, the best baseline, 42.2% SAC, and 37.4% B&H; –4% on ETH vs. FinMem); best Sharpe on 3 of 6 assets; rule-based tools raise ETH drawdown and cut its return 54.8% to 43.1%; no costs
TradExpert [42]	DOW 30, 2023	SVM, XGBoost, LightGBM; A2C, PPO, EIIE, SARL, DeepTrader; index	<i>Ahead (1/1); gross.</i> 49.8% annualized, Sharpe 5.01, MDD 6.6% vs. 32.5%, 1.82, 15.3% DeepTrader (best DRL), 21.6%, 0.79 XGBoost, and 13.9%, 1.22, 9.7% index; no costs
CryptoTrade [119]	BTC, ETH, SOL; bull, sideways, bear windows of 2023	Buy-and-hold, SMA, SLMA, MACD, Bollinger; LSTM, Informer, Autoformer, TimesNet, PatchTST	<i>Mixed (1/9); unspec.</i> Highest Sharpe only in ETH bull; GPT-4o beats the best deep forecaster on return in 3 of 9 asset-regime cells, buy-and-hold in 4 of 9, and the best moving-average rule in 1 of 9 (ETH bull, 25.5% vs. 22.6% B&H, 22.1% LSTM, 10.2% SMA); proportional fee, rate n.r.
R&D-Agent-Quant (A) [121]	CSI 300, 2017–2020; CSI 500 and NASDAQ 100 out-of-sample 2024–2025	Alpha 101/158/360 and AutoAlpha libraries; 16 ML and deep stock models in Qlib with costs	<i>Ahead (3/3); net.</i> 14.2% annualized, IR 1.74 vs. 5.7%, 0.85 Alpha 158 and 9.0%, 1.34 MASTER (best baseline IR), net of 0.05/0.15% costs; out-of-sample IR 2.17 and 1.77 vs. 0.69 LSTM and 0.59 Alpha 360; no passive benchmark
MASS (A) [74]	SSE 50, CSI 300, ChiNext 100, 2023	Disagreement proxy rule; index; LightGBM; DTML, MASTER	<i>Ahead (3/3); net.</i> Highest rank IC on every universe (8.2 vs. 5.1 MASTER on SSE 50; 6.5 vs. 5.2 LightGBM on CSI 300); backtest 1.2–5.0% annualized net of 0.1% round trip vs. negative for every baseline except MASTER on CSI 300 (0.8%)
Kou et al. (A) [101]	SSE 50, Jan 2023–Jan 2024; CSI 300 and S&P 500 half-years	XGBoost, LightGBM, MLP; PPO; index	<i>Ahead (1/1), SSE 50 only; unspec.</i> 53.2% return vs. 9.5% XGBoost, 7.1% LightGBM, 2.9% PPO, –13.2% index; Sharpe 0.29 vs. 0.04 XGBoost (not annualized); CSI 300 and S&P 500 half-years compared with the index only; turnover-capped, cost rate n.r.
ContestTrade [239]	A-shares, Jan–Jun 2025	MACD, RSI+KDJ; LightGBM, LSTM; A2C, PPO; MASS	<i>Ahead (1/1); net.</i> 52.8% return, Sharpe 3.12, MDD 12.4% vs. 15.1%, 1.33, 17.1% PPO and 8.2%, 0.47, 8.3% RSI+KDJ; MASS loses 19.1% on the same window; net of 0.1% per trade
HedgeAgents [116]	BTC, DJ30, FX; 2021–2023	SAC, DeepTrader, AlphaMix+; time-series momentum; zero-mean reversion; buy-and-hold	<i>Ahead (1/1); gross.</i> 71.6% annualized, Sharpe 2.41, MDD 14.2% vs. 37.6%, 1.62, 25.6% AlphaMix+ (best DRL) and 19.1%, 0.78 momentum; 2024 hold-out not coded; no costs
QuantAgents [118]	NASDAQ-100; 2021–2023	SAC, DeepTrader, AlphaMix+; mean-variance; index	<i>Ahead (1/1); gross.</i> 58.7%, Sharpe 3.11, MDD 16.9% vs. 32.5%, 1.49 AlphaMix+, 11.3%, 0.72 mean-variance, and 9.8%, 0.64 NDX; live track has no passive baseline and is not coded; no costs
FLAG-Trader [217]	5 US stocks (Oct 2020–May 2021), BTC (Apr–Nov 2023)	Buy-and-hold; 13 prompted LLMs	<i>Mixed (5/6); gross.</i> RL-trained 135M model beats buy-and-hold on CR for all six assets (BTC 45.5% vs. 21.8%; TSLA 50.4% vs. 39.2%) and prompted GPT-4o on five of six (BTC 14.3%; not HON, 34.3% vs. 38.5%); Sharpe below B&H on UVV; no costs
FinRL-DeepSeek (B) [14]	NASDAQ-100; train 2019–2022, trade 2023; train 2013–2018, trade 2020–2023	PPO; CVaR-PPO; index	<i>Behind (0/1); gross.</i> The designated CPPO-DeepSeek’s information ratio (0.008) trails plain PPO (0.010); LLM signals raise CVaR-PPO information ratio (–0.015 to 0.008) and lower PPO’s (0.010 to –0.009); returns figure only, most variants below the index; two stochastic runs of the same setup reorder the agents
Macro Economists (Geng et al.; B) [63]	15 commodity ETFs, Oct 2023–Feb 2026, weekly	Rule-based z-score agent, identical inputs; inverse-volatility passive	<i>Ahead (1/1), n.s.; net.</i> Sharpe 0.57 hawkish, 0.57 debate, 0.56 dovish vs. 0.53 rule, 0.52 passive; hawkish and debate add 0.04 over rule ($p = 0.07$, 0.09, unadjusted), dovish not significant; debate adds nothing over hawkish ($p = 0.77$); 0.56 at 10 bps
QuantHarness [216]	8 instruments incl. ES, NQ, CL futures; 4h bars, 5,000 per asset; 1h figure only	Linear regression; XGBoost	<i>Behind (0/7); gross.</i> Highest 3-bar directional accuracy on all 8 (SPX 63.7% vs. 60.0% XGBoost); Sharpe below the better of the two baselines on all 7 instruments of its Sharpe table that have baselines (SPX 0.18 vs. 1.27 XGBoost; ES 0.34 vs. 0.60); no costs

Continued on next page

Table 12 (continued)

Study	Market and window	Traditional baselines	Outcome for the LLM approach
Emmanoulo- poulos et al. [47]	AAPL, NVDA, MSFT, F; Sep 2024–Apr 2025, daily	Buy-and-hold	<i>Mixed (2/4); net.</i> Averaged over backbones, beats buy-and-hold on MSFT and F but not AAPL or NVDA (the best backbone per asset wins all four). LLM-discovered SDE risk metrics raise seven LLM traders’ mean Sharpe 0.88 to 1.40 (as reported; the 0.88 news-only mean does not reproduce from the paper’s Table 2, which gives about 0.82–0.96); buy-and-hold Sharpe 3.53, 2.79, 0.09, −0.84 (mean 1.39), so the gain closes the gap rather than opening one; 8.8 bps per trade; no classical risk-model baseline
PACE (B) [174]	SZSE Level-1, Apr 2026, 1,680 parent orders (10 per day, 8 repeats)	TWAP; Almgren–Chriss; XGBoost, LSTM schedulers	<i>Ahead (2/2); exec.</i> Aggressive-side cost −2.26 bps vs. −3.28 TWAP, −2.93 Almgren–Chriss, −2.91 LSTM: 1.02 bps over TWAP (95% CI [0.15, 2.12]), 0.65 over the best scheduler (0.71 in the passive setting), untested for significance; no market impact; about \$30 API for all orders
<i>Not coded (abstract only)</i>			
Cheng et al. [35]	Chinese futures	Instrumented PCA factors	<i>Not coded.</i> 40 GPT-generated factors; significant alpha over IPCA in and out of sample, strongest after the GPT training cutoff; magnitudes n.r. (abstract only)