Online Resource 2

Methods and supporting evidence for From Individual Decisions to Collective Outcomes: A Scoping Review of LLM Agents in Financial Markets

Digital Finance | John Garcia | School of Management, California Lutheran University, Thousand Oaks, CA, USA | jgarcia@callutheran.edu

This supplement contains the detailed search, coding, system, and comparative evidence, a summary of the evidence behind the main conclusions, and the completed reporting checklist for scoping reviews. Online Resource 1 provides the complete extraction tables and the component-level source audit.

Methods and coding

Eligibility and boundary rules

The review covers research posted or published between November 1, 2022, and September 27, 2026, in which an LLM acts as an agent in a financial setting. The LLM judges, chooses, forecasts, advises, trades, or participates in a market, either in place of a human subject or as a participant or decision system in its own right. Eligibility was set in a shared protocol before the search, following the scoping-review conventions of Arksey and O'Malley (2005) and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews (PRISMA-ScR) of Tricco et al. (2018). The protocol, reproduced in Online Resource 1 (sheet "Protocols") and in the protocols folder of Online Resource 3, was not registered. Five categories of work qualify: (a) LLMs as synthetic economic subjects in financial decision problems; (b) LLM-agent market simulations and agent-based models; (c) LLM agents performing financial tasks; (d) LLM financial judgments audited for bias, consistency, or misconduct; and (e) methodological studies of the validity of LLM agents as subjects or decision-makers. Work was excluded when the LLM served only as a text-feature extractor, sentiment scorer, or question-answering system without a decision role, or when it studied general economic or psychological behavior with neither a financial task nor a construct central to behavioral finance (such work is used as background), when humans made the decisions, or when the full record was not in English. Boundary rules were refined at consolidation and applied to the 155 records on which the arms disagreed. This refinement departs from the protocol as written before the search. Under those rules, a study without a financial task was kept when it tested a construct on which behavioral finance is built, such as risk and time preferences, loss aversion, anchoring, or rationality. An LLM was treated as a forecaster or decision system, rather than a feature extractor, when its judgment, forecast, proposed strategy, or advice enters an evaluated decision, even when a rule, optimizer, or learned policy executes it.

Independent search arms and screening counts

Three search arms, each operated by a separate LLM agent, ran independently under a shared protocol. The source-first arm swept the arXiv application programming interface (API) by subject category and the Crossref records of about one hundred finance, economics, and computer science journals. The concept-first arm ran academic keyword searches through a general web search engine (Google Scholar blocks automated queries), queried OpenAlex, searched SSRN through site-restricted web searches, and

chased citations; its Semantic Scholar keyword searches were largely blocked by rate limits. The seed-first arm expanded a set of foundational papers through their citation networks and through working-paper series (NBER, FEDS, and IDEAS-NEP listings). The arms screened 114,532, 675, and 45,802 records at title and 859, 396, and 1,897 at abstract, respectively, and together included 910 records. These collapsed to 550 unique papers after cross-arm deduplication by identifier and normalized title, including one late-detected duplicate merged by hand. A consolidation re-screen of inclusion conflicts removed 26 papers, and two more were removed as non-English or unlocatable at a primary source, leaving 522. The arms are independent discovery routes; they do not supply independent human screening judgments.

Supplementary discovery and omission checks

A citation-neighborhood pass pulled the 7,585 references and citations of 160 papers found by only one arm, screened the 3,265 unique in-window records that no arm had seen, and added 14 papers; the reviewers' own excluded lists yielded 7 more upon re-screening. A further discovery pass through the Consensus academic database ran 64 queries phrased for the taxonomy nodes and for records that earlier routes could not resolve; of 1,245 returned records, 619 were new candidates, and 40 papers were added. A verification of a reference collection assembled for this project found three in-scope papers that no route had returned, and a one-step backward citation chase from those three, together with the searches run to locate their public versions, added four more papers. One paper was added directly. A targeted omission verification through Consensus on September 26, 2026, ran ten additional queries, returning 200 hits representing 156 distinct records. It added two eligible papers and three general methods references that were used only as background. A final search on September 27, 2026, extended coverage to records public on that date. It screened the arXiv announcement of September 25, the last before the search (five papers added), 28 Crossref queries restricted to records created or published from September 24 to 27 (three papers), and nine Consensus queries restricted to September 2026 records (one paper); hit counts are in the search log. The arXiv API and OpenAlex were unavailable, and SSRN was reached only through Crossref records. Three retained decision-system records were excluded under the eligibility rules of Section 2.1 of the main text (see 'Scope boundary of the signal-only systems'). The corpus therefore holds 599 papers (Fig. 1), and the publication cutoff is September 27, 2026.

Bibliographic verification and ledger completeness

Every included record was verified against its primary metadata record: the Crossref entry of its digital object identifier (DOI), the arXiv API, or the publisher or repository page. Of the 599 records, 583 matched their primary record and 16 matched with a recorded discrepancy (most often a title or author order that differs between versions; one of the 16 has since been withdrawn by its authors). This verification establishes each record's bibliographic identity rather than its findings, which were coded separately and from the full text where one could be read (Section 2.3). Most abstract-stage exclusions are itemized with reasons in the review's ledger; the exceptions are 1,250 of the seed-first arm's 1,509, which were counted but not listed, and 11 of the source-first arm's 511, a residual between its screened and included counts that no list accounts for. Twenty-four records were excluded because their abstract (16) or primary record (8) could not be obtained by September 27, 2026, and four borderline records were excluded on scope; all are listed in the ledger. Title-stage exclusions are counted in the search log rather than listed, and the log records the date of every query and the hit count of all but 17 of its 444 entries, two of them final-search queries that could not be run. Because title-stage exclusions and part of the abstract-stage exclusions are not itemized, screening cannot be reconstructed record by record, and the review is a transparently bounded scoping review rather than an exhaustive systematic one.

Synthesis set and eight coding facets

The consolidated extraction produced a taxonomy of five categories and 31 nodes, developed inductively from the arms' coding and adjudicated before the synthesis began (Table 1). For the synthesis, 276 papers are coded on eight facets adapted from a codebook prepared for this review. The 276 are the 273 papers in categories a, b, d, and e, and three from category c: two system-level evaluation papers (Barton et al., 2026b; Duan et al., 2026) and one review of multi-agent trading evaluation (Nguyen & Pham, 2026). Five of the 276 are secondary syntheses, so the synthesis rests on 271 primary studies. Shares such as 30 of 271 describe the synthesis set, which is selected by category, rather than the 599-paper corpus. The first four facets are finance relevance and evidence stratum; the agent's intended role (human surrogate, financial decision tool, AI economic participant, or experimental instrument); demonstrated agency (from elicited judgment to sequential decision loops and tool-mediated action); and interaction structure (independent elicitation, coordination within a decision system, strategic economic interaction, network transmission, or human–AI co-participation). The remaining four are the environment and whether prices are exogenous or endogenous to the agents' decisions (orders, or forecasts that enter a price equation); the source of heterogeneity and whether it is imposed, emergent, or validated against human data; outcomes by level; and design and validation, including the human comparator. Of the 276 papers, 222 were read in full, 48 from their abstracts, and 6 from metadata alone.

Decision-system audit and review profiles

The 295 papers in category c, which design or benchmark decision systems for trading, portfolio choice, valuation and credit, forecasting, advice, and quantitative research, are audited based on a narrower set of fields. These fields are evaluation design, the stated test window, the models and version strings, whether look-ahead through the model's pretraining is discussed and mitigated, whether transaction costs are modeled, whether the LLM itself selects the action, whether the system is a team of LLM roles, and whether code is released. Of these, 220 were read in full. Three of the 295 are reviews filed under task nodes. The 29 papers in the review node and three surveys of LLM-driven simulation are profiled separately to position this review. The 295 exclude the two system evaluations coded in the synthesis. Trading costs are assessed only where they apply. A paper read in full is trade-relevant when it evaluates an implementable trading or investment strategy (140 of the 220), not applicable when it evaluates accuracy, classification, valuation, credit, advice, or benchmark answers (80), and unassessed when only the abstract or metadata were read (75).

Extraction and condition-coding rules

Numbers were copied as reported, with units, uncertainty, and a locator (table, section, or page); a field was coded "not reported" only after the relevant full text had been searched, "not extracted" when only the abstract was available, and "not applicable" when the design excludes the feature, such as trading costs in a valuation audit. For the 30 price-setting designs, the studies in which agents' decisions set prices, Online Resource 1 also codes twelve components of the five conditions in Section 8, derived from the coded facets. For each component, it records whether the source states it, whether the coder inferred it from displayed results, whether it is absent from the full text read, whether it is unassessable from an abstract, or whether it is inapplicable. The authors' interpretations and the coder's reading are kept in separate fields. The synthesis is narrative and organized by facet rather than by paper. Effect sizes are not pooled because studies with the same behavioral label usually measure different things: an anchoring index, a pass-through slope, and a share of anchored questions. Where a result in the synthesis rests on an abstract rather than a full text, the text says so. Coverage shares use assessable denominators: a positive code can come from any reading depth, but an absence counts only for a study

read in full, and a dated model snapshot can be assessed only in a full text. Online Resource 1 reports each such count with both its all-record and assessable denominators.

Detailed LLM-assisted review workflow

Large language models were used throughout this review; this section documents their use. LLM agents (Claude models, Anthropic, September 2026) carried out the searches, screening, extraction, verification, and facet coding, computed the counts reported here, wrote the code that draws the figures, and assisted with manuscript revisions, all under written protocols. The retained session logs record the model identifiers: claude-fable-5-1 for the three search arms, the consolidation, and the checkpoint check, and claude-opus-5-5 for facet coding, the system audit, the review profiles, verification, the final search, and manuscript revision. Sampling parameters were left at the platform defaults and not recorded. The task prompts for the retained agent runs are available from the author. Parallel coding agents produced one evidence file per paper, and separate agents that had not produced the evidence then checked the claims in every section against the cited sources. An LLM audit of the corpus and taxonomy preceded the synthesis. A model from another developer (GPT-6 Sol, OpenAI, September 2026; recorded by model name only) reviewed the manuscript and ran the targeted Consensus omission check, and GPT-6 Sol agents audited the 360 component codes of the 30 price-setting designs against their sources, revising 16 of them; LLM agents drafted revisions, which were checked against the sources in the same way. Codex (OpenAI, October 2026) assisted with journal-guideline checks, editorial assessment, and tracked revisions, recoded a stratified sample of studies for the cross-model check (Table S4), and adjudicated the count-relevant disagreements from that check against the sources (see “Source adjudication of count-relevant disagreements”). The author set the research question and scope of the review, ruled on the corpus and on cases referred to before the synthesis began, revised the manuscript, and takes responsibility for its content.

The author also manually validated the results reported in the main text before the October 9, 2026 publication update (the counts, figures, evidence locators, and each section’s claims) against the underlying sources and evidence files. Online Resource 1 itemizes the documented checks: 17 priority checks of the codes behind the headline counts and 24 stratified spot checks of synthesis judgments (seed 20260927), all of which confirmed the coded values, and 15 rulings on boundary and eligibility cases, of which one recoded a component and three excluded records. One of these rulings was later superseded by the source adjudication described below. The changes made after that validation, from the October 9 review of the LLMF-0415 publication and the October 11 source adjudication, were checked only by AI review passes, and the author's confirmation of them is not recorded. Later checks recoded five facet values and corrected one transcribed number (sheet “Recodes after check”). These checks are not blind independent double coding and yield no inter-rater agreement statistic (Section 10 of the main text). The cross-model check described below measures agreement between two AI coding processes rather than inter-rater reliability.

Table S1 lists each step of the workflow, who performed it, the checks applied, and where Online Resource 1 records them. Model identifiers are those recorded in the retained session logs.

Table S1 Steps of the LLM-assisted review workflow

Step	Performed by	Checks applied	Record in Online Resource 1
Three independent search arms	Claude agents (claude-fable-5-1), one per arm	Consolidation re-screened the 155 inclusion conflicts under the refined boundary rules	Search log; Exclusion ledger; Corpus
Consolidation, supplementary discovery, and final search of September 27, 2026	Claude agents (claude-fable-5-1 for the consolidation; claude-opus-5-5 for the final search); GPT-6 Sol for the targeted Consensus omission check	Author rulings on referred records	Search log; Literature additions; Exclusion ledger; Author rulings
Bibliographic verification	Claude agents	Author review of the 16 recorded discrepancies	Corpus
Facet coding of the 276 synthesis papers	Claude agents (claude-opus-5-5), one evidence file per paper	Claim checks by agents that had not coded the evidence; 24 author spot checks; recodes after later checks; cross-model recoding of a 54-study sample by Codex agents (Table S4)	Synthesis facets; Coded results; Author spot check; Recodes after check; Cross-model check
System audit, review profiles, and closest reviews	Claude agents (claude-opus-5-5)	Cost-applicability pass; uniform dated-snapshot rule applied after a later check; closest reviews reread in full at revision and their reference lists matched to the corpus record by record	System audit; Review profiles; Closest reviews; Review reference check; Recodes after check
Component codes of the 30 price-setting designs	Claude agents, from the coded facets	Original September 2026 GPT-6 Sol source audit of all 360 codes (293 verified, 32 uncertain, 35 inaccessible; 16 revised); 17 author priority checks and 15 rulings; October 9 Codex review of the LLMF-0415 publication; October 11 Codex source adjudication of seven count-relevant cross-model disagreements (two codes revised), rechecked by AI review passes on October 12	Conditions by study; source audit; Priority checks; Author rulings; Source adjudication; Coverage sensitivity
Counts, tables, and figures	Scripts written by Claude agents	Recomputed from the adopted codes by Online Resource 3 against expected outputs	Summary; Assessable denominators
Manuscript	Claude (revisions); GPT-6 Sol (review)	Claim checks of each section against its sources; author validation of the results reported before October 9, 2026	README entry “Review agents”

Category composition and growth

Table 1 summarizes the corpus. Task-performing agents dominate (326 papers, category c), followed by validity and methodology (79, e), audited judgments (78, d), synthetic economic subjects (69, a), and market simulations (47, b). The literature is young and mostly unrefereed: 350 papers are working papers, 133 appear in conference proceedings, and 116 in journals. Growth is uneven across categories (Table 1). From 2025 to 2026, the audit and validity strands roughly doubled (from 23 to 43 and from 24 to 50 papers), whereas the synthetic-subject strand fell from 28 to 22 papers and the market-simulation strand increased from 19 to 22. The 2026 counts cover only January 1 to September 27.

Cross-model coding check

Codex agents (OpenAI) recoded a stratified random sample of primary studies from the sources, using a coding workbook and codebook built from the protocol definitions. The sample was drawn on October

9, 2026, with seed 20261009 from a frame of 256 primary studies, which excludes secondary syntheses, metadata-only records, studies with an unknown interaction structure, and the author's four papers. It comprises all 30 price-setting designs, 8 of the 25 other multi-agent studies read in full, 14 of the 159 studies with independent elicitation only read in full, and 3 of the 42 studies coded from abstracts. Four Codex agents coded separate batches from freshly retrieved versions of the sources, and other Codex agents rechecked the codes against the sources, changing 13 of them. According to their working records, the agents opened neither the agents' codes, which were sealed at the draw, nor the manuscript, nor an earlier Codex file that had compared 300 coded fields with the agents' codes, five sampled studies among them. Excluding those five studies changes no agreement in Table S4 by more than 3 percentage points. Codex read 49 full texts and 4 abstracts; one assigned abstract was inaccessible, and its codes are "not determinable". One price-setting design (LLMF-0415) is excluded because a later Codex update, made with the author-supplied publication, also revised its codes in Online Resource 1.

Agreement is the share of studies with identical codes; Cohen's kappa corrects it for chance using each coder's distribution of codes, and Gwet's AC1 remains stable when one category dominates. Intervals are 95% percentile bootstrap intervals over studies (2,000 resamples, seed 20261010), and "not determinable" is kept as a category. Frame-weighted agreement weights each stratum by its frame size relative to its sample size. Codex coded 28 of the 29 sampled price-setting designs as endogenous and none of the other 25 studies; the remaining design was the inaccessible abstract. Of the 90 disagreements, 75 arose in multi-agent designs, where Codex also coded independent elicitation (18 studies) or network transmission (8) that the agents had not. Among the 27 designs compared on components, Codex coded more as having a same-design human benchmark (9 rather than 5, with one concurrent in both), as comparing dispersion with humans (5 rather than 3), and as varying the trading mechanism (4 rather than 3). The counts of dependence measures (5), non-LLM agents under the same rules (4), and familiarity checks (1) matched. Table S4 reports the comparison as scored, before adjudication; Online Resource 1 lists every pair of codes, from which Online Resource 3 recomputes the table, and the adjudication below changes adopted codes, not this comparison.

Table S4 Agreement between Codex's recoding and the agents' codes in 54 sampled primary studies, as scored before adjudication

Facet or component	Studies	Agreement, % (95% CI)	Cohen's kappa (95% CI)	Gwet's AC1	Disagreements	Frame-weighted agreement, %
Independent elicitation	54	64.8 (51.9, 77.8)	0.35 (0.19, 0.54)	0.54	19	88.4
Within-system coordination	54	90.7 (83.3, 98.1)	0.40 (−0.05, 0.79)	0.90	5	96.4
Strategic economic interaction	54	94.4 (87.0, 100.0)	0.89 (0.77, 1.00)	0.93	3	93.9
Network transmission	54	83.3 (72.2, 92.6)	0.34 (0.00, 0.63)	0.81	9	95.6
Human–AI co-participation	54	92.6 (85.2, 98.1)	0.18 (0.00, 0.49)	0.92	4	93.5
Price formation	54	88.9 (79.6, 96.3)	0.82 (0.69, 0.94)	0.86	6	79.5
Human-surrogate role	54	88.9 (79.6, 96.3)	0.79 (0.62, 0.93)	0.85	6	92.8
Human comparator	54	83.3 (74.1, 92.6)	0.74 (0.59, 0.88)	0.81	9	91.6
Dispersion compared with humans	54	85.2 (74.1, 94.4)	0.52 (0.24, 0.78)	0.83	8	80.7
Dated model snapshot	54	100.0 (100.0, 100.0)	1.00 (1.00, 1.00)	1.00	0	100.0

Facet or component	Studies	Agreement, % (95% CI)	Cohen's kappa (95% CI)	Gwet's AC1	Disagreements	Frame-weighted agreement, %
Out-of-time validation	54	77.8 (66.7, 87.0)	0.64 (0.46, 0.80)	0.72	12	78.2
C2b: dispersion compared with humans	27	88.9 (77.8, 100.0)	0.61 (0.00, 1.00)	0.87	3	–
C2c: dependence measured	27	100.0 (100.0, 100.0)	1.00 (1.00, 1.00)	1.00	0	–
C3b: familiarity with the replicated design checked	27	96.3 (88.9, 100.0)	0.92 (0.74, 1.00)	0.96	1	–
C4a: trading mechanism varied	27	96.3 (88.9, 100.0)	0.84 (0.00, 1.00)	0.95	1	–
C4b: non-LLM agents under the same rules	27	100.0 (100.0, 100.0)	1.00 (1.00, 1.00)	1.00	0	–
C5a: same-design human benchmark	27	85.2 (70.4, 96.3)	0.64 (0.30, 0.92)	0.81	4	–

Source adjudication of count-relevant disagreements

On October 11, 2026, Codex agents checked against the cited source versions the seven component disagreements in Table S4 in which Codex, but not the agents, had coded a component, so that a reported count could rise. Online Resource 1 (sheet "Source adjudication") records for each decision the frozen codes, the September audit reading, the adopted code, the source version and locator, and the rationale. Two codes changed. In the July 2025 version of Shah et al. (2025b), the version within the search window, Section 3.2.4 compares the mean bid–value gaps of LLM bidders in clock and sealed-bid auctions (1.83 and 6.26) with those reported for human bidders under the same affiliated-private-value design (2.48 and 6.52), so the same-design human benchmark is coded as a published comparison. This supersedes the author's ruling of October 1, 2026, which had kept the code negative on the agents' reading that no numerical comparison was reported. Hansen and Lee (2025) compare a market maker whose price updates with the order flow against a treatment that fixes the price (Sections 4 and 6.3), so the trading mechanism is coded as varied. The other five remain negative under the written definitions: a same-design human benchmark requires a quantitative comparison with observed human outcomes, which a qualitative reference to human behavior or a shared pricing mechanism does not supply (Galanis, 2026; Jia & Yuan, 2024; Tsuchihashi, 2023), and a dispersion comparison concerns dispersion between agents, which pooled frequencies of bid or decision categories do not measure (Hansen & Lee, 2025; Shah et al., 2025b). The adopted counts become six same-design human benchmarks (one concurrent and five published) and four mechanism variations; dispersion comparisons remain three, and the source-status totals become 305 verified, 32 uncertain, and 23 inaccessible. AI review passes (Codex and Claude) rechecked the two changed codes against the source texts on October 12, 2026. These are AI decisions; the author's confirmation of them is not recorded. The remaining disagreements, nearly all in facets, were not adjudicated, and Table S4 is not recomputed with the adopted codes.

Extended system evidence

Where the LLM selects the action

The audit shows that the LLM often does not choose the action. The LLM produces the final decision itself (a trade, allocation, forecast, rating, or recommendation) in 171 of the 295 papers, shares the decision with a reinforcement-learning policy, optimizer, or rule in 58, and supplies only signals or features in 54; the remaining 12 could not be classified. In none of the 20 classifiable quantitative-

research papers (one more is unclassified) does the LLM control trades: in 18 it only proposes factors, strategies, or code that machine-learning, linear, or genetic combiners and top-k rules turn into positions, and in two it shares the decision with such a rule. Among trading agents, the LLM selects the action in only 35 of the 69 papers that could be classified. Claims about "LLM trading agents" therefore often describe hybrid systems whose economic behavior depends on components that are not language models.

Scope boundary of the signal-only systems

The 54 systems in which the LLM supplies only signals or features stay within the eligibility rule of Section 2.1 of the main text, classified by coded decision role. In 19, the LLM proposes factors, strategies, or code that a rule, optimizer, or backtest turns into positions, which is why quantitative research automation qualifies even when an optimizer implements the strategy. In 12, a rule, optimizer, or learned policy turns the LLM's judgments about assets (forecasts, scores, classifications, directional views, or written analysis) into positions. In 10, the LLM advises or reports and a human or another agent decides; 12 are agent benchmarks scored against reference answers without a market action; and in one, LLM agents run the steps of a conventional optimization pipeline. Three records at this boundary are excluded under those rules. In the first, FinBERT, an encoder classifier, supplies sentiment and uncertainty features to an optimizer and a reinforcement-learning controller (LLMF-0522). In the second, according to its abstract, an attention-based long short-term memory (LSTM) network makes the forecast that other agents explain, and no evaluated LLM role is established (LLMF-0497; the full text could not be accessed). In the third, the pooled last-layer hidden states of the LLM agents are fused by a trained module that predicts direction (LLMF-0279). One further record, coded from its abstract with its action role not extracted (LLMF-0360), is counted as unclassified in Table 3 rather than among the signal-only systems.

Architecture, backbone, and target dependence of attacks

Where coordination has been studied as a channel, the evidence concerns how errors, manipulation, and anchors propagate inside one system. The aggregation rule governs how far an injected signal survives. In a trading system whose three analyst reports are aggregated in four ways, a sell-targeted attack through the market analyst succeeds on 0.3% of attackable days under decentralized majority voting and on 55.7% when a central node averages the reports, with GPT-4.1 as the backbone (6.8% and 11.5% with Qwen3-235B) (Na et al., 2026). For buy-targeted attacks the ordering weakens or reverses, because one compromised analyst can complete a two-of-three majority. In the same paper's four-role pipeline, a jailbroken terminal risk manager flips 95.5% and 98.9% of attackable days in both target directions.

Deployed fleets and configuration effects

Production records of two deployed fleets indicate that the machinery around the model sets behavior (Barton et al., 2026b; see also Barton et al., 2026a). In the fleet that traded perpetual futures, mostly on paper at live prices, each additional level of a user-facing risk slider is associated with a 0.425 increase in the leverage multiple, and agent fixed effects account for 60% of the variance in leverage (Barton et al., 2026b). The authors identify an interface effect through a regression discontinuity: selection jumps 1.75-fold at the cut between the third-ranked gainer, which the interface displays, and the fourth, which it does not, and median chosen leverage is 5.0 times in every volatility band. In the other fleet, which traded real capital, 1,544 of 3,454 active agents bought the same token within one hour. The bunching of displayed assets and a single token is consistent with shared inputs producing correlated order flow. Other live deployments in the corpus report mainly trading returns (Liu & Chen, 2026; Zhang et al., 2026b), and the peer-aware architecture of Liu and Ma (2026), read at metadata depth only, is not

synthesized. A market model that represents LLM traders as models rather than as configured systems misstates both the source of their heterogeneity and the channel through which they move together.

Review positioning

Extended comparison with existing reviews

Existing reviews organize the field by application (Jadhav & Mirza, 2025; Kong et al., 2024), institutional function (Dong et al., 2025), agent function (Batra et al., 2025), or design pattern (Nguyen & Pham, 2026), and also discuss market structure, primarily using evidence from reinforcement-learning agents (Aldridge et al., 2026). The closest in method maps 77 studies of LLM trading agents under a written protocol and audits a primary subset of 19 for protocol and reproducibility, finding that only a few report time-consistent test splits or cost models (Xia et al., 2026). Its unit, however, is the single trading pipeline rather than a market of interacting participants. Other reviews that treat validity systematically also confine it to the trading performance of decision tools (Fu, 2025; Zhu & Cai, 2026). A practitioner review that makes simulation one of four agent functions notes that validation of LLM simulations often relies on qualitative assessment and that no common measure of alignment with human behavior exists, but it summarizes market simulations study by study (Batra et al., 2025). One framework paper identifies the gap between agent design and market outcomes, but it opts for a reduced-form model instead of reviewing the LLM evidence (Gong, 2026). A cross-domain survey of LLM-driven simulation proposes validating micro and macro behavior against human data but does not cover asset markets (Gao et al., 2024a). The review closest in motivation, published shortly before the cutoff and read in full at revision, maps 769 OpenAlex records from 1990 to July 2026 and codes 96 studies, 24 in each of four mechanism families and 74 from their abstracts (Yang et al., 2026a). Its architecture–use–amplifier–outcome framework places systemic risk where common models, objectives, infrastructures, or narratives correlate actions and meet thin risk-bearing capacity, and it reports the weakest evidence for vendor commonality, agent autonomy, and equilibrium feedback. It proposes measuring commonality in inputs, representations, outputs, and actions and defines effective model diversity as the dispersion of actions conditional on the same shock; however, the evidence in its generative AI family comes from benchmarks, prompt comparisons, advice experiments, outages, and launch events. None of the studies it cites places LLM agents in a market, and calibrated agent-based stress tests appear among its ten proposed designs. Its outcome ladder records the furthest outcome each study observes, from model outputs through prices or market quality to system-level loss; it does not code price formation by agents, comparisons with human behavioral benchmarks, or population dispersion.

Two reviews of multi-agent systems were also read in full at revision. A survey from computer science organizes LLM-based multi-agent systems by enabling techniques, foundation models, and mechanisms of collective intelligence, under which it tabulates role-specialized trading frameworks alongside simulated agent societies such as TwinMarket (Liu et al., 2026b). It discusses market simulations such as StockAgent and the market of Gao et al. (2024b) one at a time, and the simulation workflow it outlines calibrates parameters against stylized facts. A PRISMA review that searched Scopus and Web of Science on June 20, 2026, retains 93 primary studies of financial multi-agent systems (Khoukhi & Slimani, 2026). Its 26 LLM-based multi-agent systems serve mainly trading, research, portfolio, and advisory tasks. Of its 20 studies of market simulation or microstructure, 12 are agent-based models and four use multi-agent reinforcement learning, and its rules class agent-based simulations with language-enabled agents as hybrid architectures. Its “human validation” denotes evaluation with financial experts or end users rather than comparison with human behavioral benchmarks, and a single reviewer coded the studies without an agreement statistic. Neither review codes price formation by agents, comparisons with human behavioral benchmarks, or dependence among agents.

Together, the three reviews cite 26 records in this corpus: three cited by Yang et al. (2026a), 12 by Liu et al. (2026b), and 11 by Khoukhi and Slimani (2026). Eight belong to the synthesis set, including four of the 30 price-setting designs; 15 belong to the system audit; and three are profiled reviews (Online Resource 1, sheet "Review reference check"). Their reference lists also include nine LLM-based financial systems that the search did not retrieve, eight of them cited by Khoukhi and Slimani (2026). None is a market simulation; if eligible, they would enter the system audit rather than the synthesis set. Three reviews or perspectives that Khoukhi and Slimani (2026) cite were also not retrieved and would, if eligible, join the review profiles; seven further works fall outside the eligibility rules. The check covers only the works each review cites, and Khoukhi and Slimani (2026) cite at most 38 of their 93 primary studies.

Evidence base for the four conclusions

Table S2 summarizes the evidence behind the four conclusions in Section 1 of the main text. The first and fourth rest on many studies and on the review's own audit counts. The second and third rest on a few price-setting designs, all of them working papers or conference papers, and are better read as hypotheses that these designs support than as established regularities. Publication status is that of the version recorded in Online Resource 1.

Table S2 Evidence behind the four conclusions in Section 1 of the main text

Conclusion	Main evidence	Designs and publication status	Main caveat
1. Individual behavior belongs to a configuration rather than to LLMs as a class	Sections 4.1–4.3 of the main text: consistency, risk, loss aversion, and anchoring across models, releases, prompts, and instruments	Many individual-level studies in categories a, d, and e, refereed and unrefereed	Instruments and statistics differ across studies, so configuration and instrument effects are not separated
2. Aggregate outcomes depend on the trading institution and population composition	Chupilkin (2026); Saxena et al. (2026); Henning et al. (2025); Shah et al. (2025b); Galanis (2026)	Four of the 30 price-setting designs vary the trading mechanism; the five studies named are working papers	Institution dependence rests on one study that varies institution and model family together; composition effects rest on a few mixed-population designs
3. Dependence is a plausible channel whose importance relative to individual bias is unmeasured	Begin et al. (2026); Saxena et al. (2026); Ross et al. (2026); Yang et al. (2025); Barton et al. (2026b)	Five of the 28 full-text price-setting designs measure some form of dependence; one conference paper and four working papers	The designs use different statistics (error correlation, common error component, action overlap, holdings overlap), and none compares the market consequences of dependence with those of individual bias
4. Direct aggregation evidence is scarce	Sections 3.2 and 8 of the main text; Fig. 3; Table 5	Counts over the 271 primary studies; of the 30 price-setting designs, 23 are working papers, five conference papers, and two journal articles	Single-pass LLM coding without a human reliability estimate (source adjudication of cross-model disagreements corrected two component codes; see "Source adjudication of count-relevant disagreements"), unestimated recall, and no price-formation coding in the system audit; counts of rare designs may understate the evidence

Extended individual-decision evidence

Shared priors and downstream evidence use

Across providers, models carry shared priors: on average, twelve models rank environmental and social performance as a more important driver of firm value than portfolio managers do, although two of the newest rank it below the managers' level and three others at about the same level (Li & Yu, 2025). Their expected returns for environmental, social, and governance (ESG) and market portfolios are less dispersed than retail investors', with standard deviations of 0.7 to 0.9 percentage points against 4 to 5 (Li & Yu, 2025). All six models in one audit score large firms above small ones when the evidence is balanced (Lee et al., 2025), and nine models rank money-like assets almost identically, with a Kendall's W of 0.92 (Wu, 2026). Default agents answer like young high-income investors in 83% of cases, although that group is less than 30% of the population, while seeding demographics restores human-like variation across groups (Fedyk et al., 2025). These priors travel into downstream decisions: a single steered neuron that sets a model's buy or sell leaning changes its search behavior, evidence selection, and memo content (Park et al., 2026).

Misconduct, incentives, and warnings

Misconduct mostly responds to incentives in the direction economic theory predicts, with model-specific magnitudes. Baseline propensity to misuse customer funds ranges from 0.10 to 1.00 across twelve models and falls with risk aversion, lower profit expectations, and regulation (Biancotti et al., 2025). A vague profit mandate raises risk-dismissing judgments by 6.8 percentage points and lowers recommendations to escalate issues to the board from 74.4% to 60.5%, with effects ranging from near zero in three models to about 20 points in another (So, 2026). Under pressure, GPT-4 trades on an insider tip and conceals the reason from its manager (Scheurer et al., 2024), and multi-turn concealment raises attack success on GPT-4o from 43.10% to 97.51% (Cheng et al., 2026). On clear-cut cases, by contrast, models outperform lay humans: no LLM's initial response endorses a fraudulent investment, against 13.3% to 14.1% of human participants (Powdthavee, 2026).

Extended validity evidence

Historical inputs in market simulations

Contamination also threatens market simulations and synthetic subjects. Five of the 30 price-setting designs feed real historical news to agents or replay a real episode, while a sixth uses real price histories only in separate backtests with exogenous prices (Ma et al., 2025). Two of the five mask dates and names, although one keeps month and day cues (Yang et al., 2025; Zhang et al., 2026a), and one withholds calendar dates but not asset names (Li et al., 2026b). Another argues that pre-training overlap is unlikely to distort outcomes that arise endogenously inside the simulation (Zou et al., 2026). The fifth, a simulation of the 2022 nickel squeeze, masks dates and identifying details yet concedes that its success may partly reflect the episode's presence in the pre-training data (Wang et al., 2025a). One study infers from a social-interaction ablation that leakage from historical news has a negligible impact (Yang et al., 2025), but none of the studies fully assess how much remembered outcomes influence simulated prices. Coding a design as using real news or episodes therefore records exposure to remembered outcomes, not a measured contamination effect.

Market-component coding

Definitions and study-level membership behind Table 5

Counts refer to the 30 price-setting designs in the synthesis set, in which prices respond to the LLM agents' own decisions (orders or forecasts), except where the cell states another base; unknown counts appear in parentheses. A component counted as absent was not found in the text read. A comparison of dispersion with human data is counted whatever its outcome; it must compare dispersion between individual agents, not pooled frequencies of responses. Dependence is counted when a study reports a correlation, overlap, or common-error statistic across the LLM agents' own errors, actions, or holdings; a study that correlates strategy archetypes, most of them rule-based, is not counted (Zou et al., 2026). A non-LLM baseline means rule-based, zero-intelligence, or other non-LLM agents trading under the same rules; automated market makers and non-LLM models run outside the market are not counted. Published same-design human evidence requires a quantitative comparison with observed human outcomes under the same core task, institution, and payoffs, which may use summary statistics from the original experiment; it need not entail reanalysis of microdata, and a qualitative reference to human behavior or a shared mechanism alone does not qualify. The evidence includes del Rio-Chanona et al. (2025), Hansen and Lee (2025), Saxena et al. (2026), Shah et al. (2025b), and Struski et al. (2026). Henning et al. (2025) separately run concurrent human markets under the same design.

Counts that do not follow from a single coded field rest on these studies: agent traits initialized from investor records (Yang et al., 2025); auction or trading mechanism varied as a treatment, including whether a market maker updates its price (Chupilkin, 2026; Hansen & Lee, 2025; Shah et al., 2025b; Tsuchihashi, 2023), which excludes a study that varies the contract traded (Maresca, 2026) and one that varies the sign of expectations feedback (del Rio-Chanona et al., 2025); non-LLM agents under the same rules (Hashimoto et al., 2025; Ross et al., 2026; Tsuchihashi, 2023; Zou et al., 2026); dependence measured (Barton et al., 2026b; Begin et al., 2026; Ross et al., 2026; Saxena et al., 2026; Yang et al., 2025); real news or episodes (Li et al., 2026b; Wang et al., 2025a; Yang et al., 2025; Zhang et al., 2026a; Zou et al., 2026); model familiarity with a replicated design checked (Saxena et al., 2026). Illustrative designs meet the condition in at least one respect named in the requirement column, not necessarily in every respect.

TwinMarket's weighted Jaccard similarity measures overlap in trade intensity, not common errors. Gao et al. (2024b) initialize their environment with five days of real prices; this historical initialization is coded separately from the five studies using real news or replayed episodes. TwinMarket reports three repeated robustness runs for GPT-4o and Gemini, so a single headline path is not evidence of an entirely single-run design. Where neither the text nor appendices establish an independent run count, the audit records unknown. The prompt supplements of Galanis (2026) and Struski et al. (2026) were read for this audit.

The publisher version of T. Gao et al. (2026), read in full on October 9, 2026, uses historical gold-market reports to calibrate investor types and capital rather than replaying an observed gold-price episode. The reported default scenario generates prices from the agents' orders and includes sampled agent-authored comments. Table 1 and Figure 1 report ten run-level results per architecture, supporting a positive repeated-run code by inference, but independent resets and seeds are not documented; the code remains uncertain. The publication does not supply a model name, a dated snapshot, or complete prompt text. Its architectural ablation uses two LLM populations under the same matching rules, so it is neither a non-LLM baseline nor an institutional treatment. No same-design human or behavioral-dispersion comparison is reported.

Claim-conditional component rules and unmeasured requirements

The necessary evidence depends on the claim a study makes (Table 6 in the main text). Online Resource 1 lists the necessary components and their values, including unknowns, for each of the 30 price-setting designs and the corresponding roles they claim, but it does not provide a summary score. Surrogate claims need a same-design human benchmark, independently repeated markets, a dispersion comparison with human data where the claim concerns distributional features, and, for replications, a verification of the models' familiarity with the original design. Participant claims need the configured sources of heterogeneity and, for simulations, independently repeated markets. Tool claims need repeated runs; time-consistent, out-of-sample, and net-of-friction tests are not coded here. Instrument claims need stability across meaning-preserving wordings and repeated runs. A non-LLM baseline under the same rules is necessary only for a claim that the LLM component explains outcomes beyond simpler agent rules. Because that claim is not coded, the baseline is reported for every design as a diagnostic, alongside dependence measures.

The production record in the set (Barton et al., 2026b) has no simulated market to repeat; its claims depend instead on the sampling frame of the accounts watched, the definition of exposure, the comparison group, and the observation window, so the main text limits its retail comparison to the trailing week it covers, and no retail sample size is inferred (see the qualifications below). These elements support a description of the fleet's actions; however, estimating its effects on prices or stability would also require a credible counterfactual, which is not recorded in the coding. Requirements such as a justified institution, modeled frictions, out-of-sample evaluation, stability across versions, and dependence among tools deployed in many copies are not coded or assessed. A study that claims several roles is assessed against each.

Scope and status of the source audit

The original September 2026 audit checked all 360 component decisions across the 30 price-setting designs. Online Resource 1 records the original code, adopted code, source, locator, supporting evidence, and verification status for each component. It distinguishes direct source statements, source-supported inferences, non-documentation in the material read, insufficient access, and non-applicability. In that original audit, of the 360 decisions, 293 carried the agent status verified, 32 uncertain, and 35 inaccessible. The original statuses are retained separately from subsequent publication access and from the later author review. A verified status means that an LLM agent found source support for the adopted code; subsequent author confirmations are recorded separately in the workbook's author-review sheets. The uncertain codes enter the sensitivity analysis below rather than being resolved by assumption. At the time of the original audit, full-text verification was attempted for all 30 studies; 27 were read in full, and three remained abstract-only. Reproducibility fields are documentation proxies: a dated identifier for one model does not document all models, and availability of prompts does not establish recovery of seeds, logs, model weights, or an execution environment.

On October 9, 2026, the author-supplied publisher publication for LLMF-0415 was reviewed by Codex AI agents, replacing its earlier abstract-only access. Eleven component decisions then received source-supported verified status; the repeated-run decision remains uncertain because independent resets and seeds are not reported. The source-status totals became 304 verified, 33 uncertain, and 23 inaccessible across the same 360 decisions; the October 11 source adjudication then verified one uncertain code, giving 305, 32, and 23. The price-setting subset now has 28 full-text readings and two abstract-only readings. Original September audit values and the existing dated author confirmations are retained; this publication update does not add a human coding confirmation or an inter-rater agreement estimate. The narrower interpretation distinguishes report-based parameter calibration from exposure to

replayed real outcomes, while treating the historical calibration as insufficient for out-of-time financial validation.

Coverage sensitivity and boundary cases

Table 5 reports the adopted codes, including the two adjudicated corrections. Online Resource 1 recomputes each component under three alternatives: treating as unknown every negative code whose source verification is marked uncertain; adopting the broader reading of four boundary cases; and combining the two. The broader readings add at most two to any count: mechanism variation rises from 4 to 6, and dependence measures and non-LLM baselines each rise by one, while same-design human benchmarks stay at six. Treating uncertain absences as unknown removes no positive code, as by construction it cannot; it mainly moves wording-stability tests (from 24 negative codes to 14) and dependence measures (from 23 to 19) into the unknown category. Under every treatment, no study that compares agent dispersion with human data also includes non-LLM agents under the same rules. These alternatives test sensitivity to uncertain codes, not the accuracy of the positive codes, the component definitions, or search recall.

The four boundary readings concern definitions rather than access; a fifth, which counted the human literature benchmarks of Shah et al. (2025b) as a same-design comparison (LLMF-0114), lapsed when the source adjudication found a numerical comparison under the same design. Two studies vary features other than the trading format: del Rio-Chanona et al. (2025) vary the sign of expectations feedback, and Maresca (2026) varies the market horizon and whether positions bear interest (LLMF-0094 and LLMF-0202); the broader reading counts these as institutional variation. Zou et al. (2026) build correlation networks across strategy archetypes, most of them rule-based, without a statistic that isolates the LLM agents (LLMF-0194); the broader reading counts them as a dependence measure. Barton et al. (2026b) compare the fleet with retrospective, position-level exit rules rather than interacting non-LLM traders (LLMF-0300); the broader reading counts these as a non-LLM comparison. The adopted codes keep the narrower definitions stated in the note on Table 5.

One stricter reading would lower a count by one: treating a single headline run with separate robustness runs as partial replication would remove Yang et al. (2025) from the repeated-market count (19 to 18). Requiring a numerical matched comparison would not remove Struski et al. (2026), whose Table 1 places the original human round-level trades, prices, and price dispersion beside the LLM results. Other uncertain codes rest on access rather than definition. For Wang et al. (2025a), the main text was read in full, but the cited supporting notes, which include further profiles and prompts, were not retrieved; the negative codes for dispersion comparison and dependence remain uncertain, while wording stability is now coded unknown following the author's confirmed ruling of October 1, 2026 (LLMF-0413).

Qualifications to central numerical comparisons

Begin et al. (2026), Appendix Table 5, report 68.0% ± 2.8% accuracy for the ten-agent scaling condition and 70.2% ± 1.7% for a standalone model across five independent trials of 50 questions. The table does not define these uncertainty terms; the main text therefore reports the means without assigning an interval interpretation. Section 4.1 instead calls 67.6% ± 5.2% an all-honest market estimate, while Table 4 assigns that estimate to the nine-honest, one-malicious composition. The main text uses the internally coherent scaling comparison. No formal equality test is inferred.

Saxena et al. (2026), Table 2 and Section 3.1, support similarity between Qwen3-14B and humans on several bubble measures, not a formal ranking of the closest model. Appendix B.5.1–B.5.2 report no speculative bubbles with reasoning disabled under the standard prompt in 40 runs; the nonlinear-extrapolation prompt restores bubbles in 52% of 40 runs. Chupilkin (2026), Tables 1 and 2, show

material changes across institutions and model families, but do not provide an effect decomposition establishing equal-sized effects.

Barton et al. (2026b), Table 2, report positive net ROI for 15% of week-active agent accounts versus 53% in the retail comparison during the trailing week ending August 15, 2026. This is not whole-deployment profitability. The text and table provide different retail denominator descriptions, so no retail sample size is inferred here. Ross et al. (2026) report the five-to-sevenfold tracking-error comparison at a 100% LLM share, relative to the noise-only baseline; the 25–30% claimed overvaluation is an imposed misleading narrative, not an independently established error in fair value. The 0.696 correlation in Begin et al. is a pooled honest–honest estimate across nine adversarial compositions; the ten-agent all-honest scaling condition has an empirical correlation of 0.655. These estimates are not treated as one cell. Henning et al. (2025) compare 20-instance LLM markets with 19 human markets averaging 18 participants; human participants also complete an intervening risk-elicitation task. Same-design here refers to the core trading institution and payoffs, not literal identity of every procedure.

Reporting checklist

Table S3 maps the items of the reporting checklist for scoping reviews (Tricco et al., 2018) to the main text and the Online Resources.

Table S3 PRISMA-ScR checklist items and where they are reported

Item	Topic	Where reported
1	Title	Title page and page 1: identified as a scoping review
2	Structured summary	Abstract (one paragraph)
3	Rationale	Section 1
4	Objectives	Section 1 (review question, second paragraph)
5	Protocol and registration	Section 2.1; this supplement (Eligibility and boundary rules); protocol text in Online Resource 1 (sheet “Protocols”) and Online Resource 3; not registered
6	Eligibility criteria	Section 2.1; this supplement (Eligibility and boundary rules)
7	Information sources	Section 2.2; this supplement (Independent search arms; Supplementary discovery and omission checks); last search September 27, 2026
8	Search	Online Resource 1 (sheet “Search log”: 444 dated entries with sources, available hit counts, and documented gaps)
9	Selection of sources of evidence	Section 2.2; Fig. 1; this supplement (Bibliographic verification and ledger completeness)
10	Data charting process	Section 2.3; this supplement (Detailed LLM-assisted review workflow; Table S1); Online Resource 1 (sheet “Protocols”, README entry “Review agents”, and the author worksheets)
11	Data items	Section 2.3; this supplement (Synthesis set and eight coding facets; Extraction and condition-coding rules); Online Resource 1 (sheet “Conditions key”)
12	Critical appraisal of individual sources of evidence	Sections 2.3 and 8 (component audit of the 30 price-setting designs); this supplement (Market-component coding)
13	Synthesis of results	Section 2.3 (narrative synthesis by facet; effect sizes not pooled)
14	Selection of sources of evidence (results)	Fig. 1; Section 2.2; Online Resource 1 (sheet “Exclusion ledger”)
15	Characteristics of sources of evidence	Section 2.4; Table 1; Online Resource 1 (sheet “Corpus”)

Item	Topic	Where reported
16	Critical appraisal within sources of evidence	Table 5; Online Resource 1 (sheets “Conditions by study”, “source audit”, and “Coverage sensitivity”)
17	Results of individual sources of evidence	Tables 2–4; Online Resource 1 (sheets “Coded results”, “Market studies”, and “System audit”)
18	Synthesis of results	Sections 3–8; Fig. 3; Tables 3, 5, and 6
19	Summary of evidence	Section 1 (four conclusions); Section 11; Table S2
20	Limitations	Section 10
21	Conclusions	Section 11
22	Funding	Statements and Declarations (no funding); the funding of included sources was not charted

References

Aldridge, I., An, J., Burke, R., et al. (2026). Agentic Artificial Intelligence in Finance: A Comprehensive Survey (arXiv:2604.21672). arXiv. https://arxiv.org/abs/2604.21672

Arksey, H., & O’Malley, L. (2005). Scoping studies: towards a methodological framework. International Journal of Social Research Methodology, 8, 19–32. https://doi.org/10.1080/1364557032000119616

Barton, T. J., Constantakis, C., Hauseman, P., et al. (2026a). Operating-Layer Controls for Onchain Language-Model Agents Under Real Capital (arXiv:2604.26091). arXiv. https://arxiv.org/abs/2604.26091

Barton, T. J., Constantakis, C., Hauseman, P., et al. (2026b). What LLM Trading Agents Actually Do in Production: A Six-Month, Population-Scale Record from Two Fleets (arXiv:2609.05663). arXiv. https://arxiv.org/abs/2609.05663

Batra, D., Hamill, C., Hartley, J., et al. (2025). A Review of LLM Agent Applications in Finance and Banking (SSRN Working Paper 5381584, version posted August 14, 2025). https://doi.org/10.2139/ssrn.5381584

Begin, J., Gho, B., Muppavarapu, S., et al. (2026). Preference Optimization Drives Monoculture in LLM Prediction Markets (arXiv:2606.26583). arXiv. https://arxiv.org/abs/2606.26583

Biancotti, C., Camassa, C., Coletta, A., et al. (2025). Chat Bankman-Fried: an Exploration of LLM Alignment in Finance (arXiv:2411.11853v3). arXiv. https://arxiv.org/abs/2411.11853v3

Cheng, G., Jin, H., Zhang, W., et al. (2026). Learning to Conceal Risk: Controllable Multi-turn Red Teaming for LLMs in the Financial Domain (arXiv:2509.10546v2). arXiv. https://arxiv.org/abs/2509.10546v2. Published version: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 41005–41020 (2026). https://doi.org/10.18653/v1/2026.acl-long.1903

Chupilkin, M. (2026). Artificial Institutions: How Institutional Design Shapes LLM Simulations (arXiv:2608.04020). arXiv. https://arxiv.org/abs/2608.04020

del Rio-Chanona, R. M., Pangallo, M., & Hommes, C. (2025). Can Generative AI agents behave like humans? Evidence from laboratory market experiments (arXiv:2505.07457). arXiv. https://arxiv.org/abs/2505.07457

Dong, Y., Wu, F., Zhang, K., et al. (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 (pp. 17889–17907). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.972

Duan, Q., Li, C., Wang, C., et al. (2026). Can Agentic Trading Systems Pay for Their Own Intelligence? (arXiv:2607.10286v2). arXiv. https://arxiv.org/abs/2607.10286v2

Fedyk, A., Kakhbod, A., Li, P., & Malmendier, U. (2025). AI and perception biases in investments: An experimental study (Working paper, draft of June 2025). https://eml.berkeley.edu/~ulrike/Papers/AI_PerceptionBias_aug2025.pdf. Published version: Journal of Financial Economics 185:104350 (2026). https://doi.org/10.1016/j.jfineco.2026.104350

Fu, W. (2025). The New Quant: A Survey of Large Language Models in Financial Prediction and Trading (arXiv:2510.05533). arXiv. https://arxiv.org/abs/2510.05533

Galanis, S. (2026). Information Aggregation with AI Agents (arXiv:2604.20050v3). arXiv. https://arxiv.org/abs/2604.20050v3

Gao, C., Lan, X., Li, N., et al. (2024a). Large language models empowered agent-based modeling and simulation: a survey and perspectives. Humanities and Social Sciences Communications, 11, 1259. https://doi.org/10.1057/s41599-024-03611-3

Gao, S., Wen, Y., Zhu, M., et al. (2024b). Simulating Financial Market via Large Language Model based Agents (arXiv:2406.19966). arXiv. https://arxiv.org/abs/2406.19966

Gao, T., Peng, L., Xu, Q., & Li, G. (2026). Heterogeneous-Agent Financial Market Simulation with Large Language Models. Procedia Computer Science, 287, 1016–1023. https://doi.org/10.1016/j.procs.2026.08.142

Gong, H. (2026). AI Agents in Financial Markets: Architecture, Applications, and Systemic Implications (arXiv:2603.13942v3). arXiv. https://arxiv.org/abs/2603.13942v3. Published version: FinTech 5(2):34 (2026). https://doi.org/10.3390/fintech5020034

Hansen, A. L., & Lee, S. J. (2025). Financial Stability Implications of Generative AI: Taming the Animal Spirits (arXiv:2510.01451). arXiv. https://arxiv.org/abs/2510.01451. Also issued as Finance and Economics Discussion Series 2025-090, Board of Governors of the Federal Reserve System. https://doi.org/10.17016/feds.2025.090

Hashimoto, R., Takayanagi, T., Suzuki, M., & Izumi, K. (2025). Agent-Based Simulation of a Financial Market with Large Language Models (arXiv:2510.12189). arXiv. https://arxiv.org/abs/2510.12189. Published version: LLM agents reveal how human bias shapes path-dependent market dynamics. Journal of Computational Social Science 9(2):32 (2026). https://doi.org/10.1007/s42001-026-00465-4

Henning, T., Ojha, S. M., Spoon, R., et al. (2025). LLM Agents Do Not Replicate Human Market Traders: Evidence From Experimental Finance (arXiv:2502.15800v3). arXiv. https://arxiv.org/abs/2502.15800v3

Jadhav, A., & Mirza, V. (2025). Large Language Models in equity markets: applications, techniques, and insights. Frontiers in Artificial Intelligence, 8, 1608365. https://doi.org/10.3389/frai.2025.1608365

Jia, J., & Yuan, Z. (2024). An Experimental Study of Competitive Market Behavior Through LLMs (arXiv:2409.08357v2). arXiv. https://arxiv.org/abs/2409.08357v2

Khoukhi, A., & Slimani, I. (2026). From Agent-Based Models to LLM-Based Financial Agents: A PRISMA-Based Systematic Review, Multidimensional Taxonomy, and Future Research Agenda (SSRN Working Paper 7493756). https://doi.org/10.2139/ssrn.7493756

Kong, Y., Nie, Y., Dong, X., et al. (2024). Large Language Models for Financial and Investment Management: Applications and Benchmarks. The Journal of Portfolio Management, 51, 162–210. https://doi.org/10.3905/jpm.2024.1.645

Lee, H., Seo, J., Park, S., et al. (2025). Your AI, Not Your View: The Bias of LLMs in Investment Analysis (arXiv:2507.20957v4). arXiv. https://arxiv.org/abs/2507.20957v4. Published version: Proceedings of the 6th ACM International Conference on AI in Finance, pp 150–158 (2025). https://doi.org/10.1145/3768292.3770375

Li, T., & Yu, L. (2025). ESG Beliefs of Large Language Models: Evidence and Impact (arXiv:2601.00836). arXiv. https://arxiv.org/abs/2601.00836

Li, Z., Su, Y., White, M., et al. (2026b). Your Agent Says Yes: Interpreting Adversarial Market Behavior Beyond Individual Transactions (arXiv:2609.07675v2). arXiv. https://arxiv.org/abs/2609.07675v2

Liu, J.-H., & Ma, Y. (2026). From Impulse to Introspection: Two-Round Peer-Aware Architectures for Large Language Model Trading Agents. In 2026 International Conference on Advances in Artificial Intelligence and Machine Learning (AAIML) (pp. 984–987). IEEE. https://doi.org/10.1109/AAIML67890.2026.11498104

Liu, K., & Chen, L. (2026). OOM-RL: Out-of-Money Reinforcement Learning Market-Driven Alignment for LLM-Based Multi-Agent Systems (arXiv:2604.11477). arXiv. https://arxiv.org/abs/2604.11477

Liu, Q., Yao, L., Wang, Z., et al. (2026b). Large language model-based multi-agent systems for financial markets simulation: a survey. Science China Information Sciences, 69, 171201. https://doi.org/10.1007/s11432-026-4986-x

Ma, T., Du, J., Huang, W., et al. (2025). Agent Trading Arena: A Study on Numerical Understanding in LLM-Based Agents (arXiv:2502.17967v2). arXiv. https://arxiv.org/abs/2502.17967v2. Published version: Findings of the Association for Computational Linguistics: EMNLP 2025, pp 5496–5514 (2025). https://doi.org/10.18653/v1/2025.findings-emnlp.294

Maresca, C. (2026). Can Interest-Bearing Positions Solve the Long-Horizon Problem in Prediction Markets? (arXiv:2602.21091). arXiv. https://arxiv.org/abs/2602.21091

Na, C., Ni, H., Szpruch, L., et al. (2026). Poisoning Agentic Alpha: Adversarial Vulnerabilities Across Roles and Architectures in Multi-Agent Trading Systems (arXiv:2608.24069). arXiv. https://arxiv.org/abs/2608.24069

Nguyen, P., & Pham, T. (2026). Toward Reliable Evaluation of LLM-Based Financial Multi-Agent Systems: Taxonomy, Coordination Primacy, and Cost Awareness (arXiv:2603.27539). arXiv. https://arxiv.org/abs/2603.27539. Published version: Trends and Applications in Knowledge Discovery and Data Mining, pp 324–336 (2026). https://doi.org/10.1007/978-981-92-2014-4_26

Park, S., Park, S., Lee, H., et al. (2026). Your AI, On a Dial: Controlling Investment Bias in LLMs with a Single Neuron (arXiv:2608.22852). arXiv. https://arxiv.org/abs/2608.22852

Powdthavee, N. (2026). Large Language Models Outperform Humans in Fraud Detection and Resistance to Motivated Investor Pressure (arXiv:2604.20652). arXiv. https://arxiv.org/abs/2604.20652

Ross, J., So, E., De Simone, Z., et al. (2026). Why Better Models Can Create Riskier Systems: Evidence from LLM Agents in Financial Markets (arXiv:2609.04373). arXiv. https://arxiv.org/abs/2609.04373

Saxena, M., Pangallo, M., Hommes, C., et al. (2026). Machine Spirits: Speculation and Adaptation of LLM Agents in Asset Markets (arXiv:2604.18602v2). arXiv. https://arxiv.org/abs/2604.18602v2

Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). Large Language Models can Strategically Deceive their Users when Put Under Pressure (arXiv:2311.07590v4). arXiv. https://arxiv.org/abs/2311.07590v4

Shah, A., Zhu, K., Jiang, Y., et al. (2025b). Learning from Synthetic Labs: Language Models as Auction Participants (arXiv:2507.09083v1). arXiv. https://arxiv.org/abs/2507.09083v1

So, E. (2026). The Profit Alignment Problem: How Profit Mandates Induce Alignment Failures in LLMs (arXiv:2609.07731). arXiv. https://arxiv.org/abs/2609.07731

Struski, P., Swistak, J., Okulska, I., & Biecek, P. (2026). Competitive Market Behavior of LLMs (arXiv:2609.02580). arXiv. https://arxiv.org/abs/2609.02580

Tricco, A. C., Lillie, E., Zarin, W., et al. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine, 169, 467–473. https://doi.org/10.7326/m18-0850

Tsuchihashi, T. (2023). How much do you bid? Answers from ChatGPT in first-price and second-price auctions. Journal of Digital Life, 3, 11. https://doi.org/10.51015/jdl.2023.3.11

Wang, C., Wang, C., Zeng, S., et al. (2025a). Advanced simulation paradigm of human behaviour unveils complex financial systemic projection (arXiv:2503.20787v2). arXiv. https://arxiv.org/abs/2503.20787v2

Wu, W. (2026). Auditing Asset-Specific Preferences in Financial Large Language Models: Evidence from Bitcoin Representations and Portfolio Allocation (arXiv:2606.02528v3). arXiv. https://arxiv.org/abs/2606.02528v3

Xia, Y., You, P., Wang, T., et al. (2026). Agentic Trading: When LLM Agents Meet Financial Markets (arXiv:2605.19337). arXiv. https://arxiv.org/abs/2605.19337

Yang, W., Liu, H., & Wang, Y. (2026a). From Algorithmic Trading to Generative AI: Artificial Intelligence, Systemic Risk, and the Transformation of Financial Markets. *Journal of Economic Surveys*. Advance online publication. https://doi.org/10.1111/joes.70167

Yang, Y., Zhang, Y., Wu, M., et al. (2025). TwinMarket: A Scalable Behavioral and Social Simulation for Financial Markets (arXiv:2502.01506v5). arXiv. https://arxiv.org/abs/2502.01506v5. Published version: Advances in Neural Information Processing Systems 38, pp 70830–70880 (2025). https://doi.org/10.52202/085713-2132

Zhang, C., Liu, X., Zhang, Z., et al. (2026a). When AI Meets Finance (StockAgent): Large Language Model-based Stock Trading in Simulated Real-world Environments (arXiv:2407.18957v5). arXiv. https://arxiv.org/abs/2407.18957v5. Published version: 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). https://doi.org/10.1145/3844605

Zhang, J., Liu, G., Johansson, O., et al. (2026b). Prediction Arena: Benchmarking AI Models on Real-World Prediction Markets (arXiv:2604.07355). arXiv. https://arxiv.org/abs/2604.07355

Zhu, L., & Cai, M. (2026). Artificial Intelligence in Equity and Crypto Markets: Progress, Profitability Evidence, and the Limits of Automated Investing (arXiv:2609.04917). arXiv. https://arxiv.org/abs/2609.04917

Zou, M., Chen, J., Luo, A., et al. (2026). FinEvo: From Isolated Backtests to Ecological Market Games for Multi-Agent Financial Strategy Evolution (arXiv:2602.00948). arXiv. https://arxiv.org/abs/2602.00948