International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

LLM Agents in Finance: Architecture, Capability, Coordination, and Deployment Readiness — A Systematic Review and Integrative Framework

INPLASY202690083
doi: 10.37766/inplasy2026.9.0083
Received: 20 September 2026
Published: 20 September 2026

Wu, CK.

Corresponding author:
Chengkun Wu

838902979@qq.com

Author Affiliation:
HKU Business School, University of Hong Kong, Hong Kong SAR, China.

ADMINISTRATIVE INFORMATION

Support - No specific funding.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202690083

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 20 September 2026 and was last updated on 20 September 2026.

INTRODUCTION

Review question / Objective RQ1 (Architecture–Capability). What architectural primitives—LLM-only, tool-augmented, retrieval-augmented, and planning-and-execution—are used in financial LLM agent systems, and what capabilities and limitations do these primitives exhibit across financial task domains?
RQ2 (Coordination). How is multi-agent coordination operationalized in financial LLM systems across independent, competitive, cooperative-hierarchical, and protocol-based peer-to-peer paradigms, and what coordination mechanisms and failure modes—including communication overhead, incentive misalignment, consensus deadlock, and authorization/audit gaps—are documented?
RQ3 (Evaluation–Deployment–Governance). What evidence exists on the evaluation and deployment readiness of financial LLM agents across benchmark, simulation, live/production, and framework-demonstration settings, and what gaps remain in comparative effectiveness, benchmark-to-deployment validation, and governance/auditability?

Rationale The integration of Large Language Model (LLM) agents into financial services has accelerated rapidly, yet the empirical literature remains fragmented and lacks a systematic, finance-specific framework for understanding how these systems are architected, how they coordinate, and whether they are ready for deployment. This review addresses that gap.

Existing surveys have catalogued aspects of the LLM agent landscape but leave critical questions unanswered. Dong et al. (2025) mapped applications to financial tasks but focused primarily on single-agent capabilities without systematically treating multi-agent coordination. Saha et al. (2025) acknowledged multi-agent collaboration as an emerging direction but provided nascent treatment of coordination mechanisms. Aldridge et al. (2025) offered a broad

survey of agentic AI in finance without identifying coordination failure modes or linking them to system design. Ding et al. (2024) was limited to trading applications. Ferrag et al. (2026) covered multiple domains but kept multi-agent discussion generic rather than finance-specific. Raza et al. (2026) proposed a risk taxonomy for LLM-based multi-agent systems but did not address deployment readiness in financial contexts. Bulla et al. (2026) examined fraud detection in isolation. Across these seven reviews, no study systematically characterizes the full coordination spectrum—from independent parallel deployment through competitive debate to cooperative hierarchy and protocol-based peer-to-peer—using a recently and systematically selected body of empirical studies. No existing review identifies distinct coordination failure modes associated with each coordination level, assesses deployment readiness across evaluation settings, or proposes an integrative framework mapping the relationship between architectural primitives, task domains, and coordination mechanisms.

Three specific gaps motivate this review. First, the field has progressed from single-agent tools toward multi-agent coordinated systems, but the evidence base is not a longitudinal record of transition; it is a cross-section of a field that has largely skipped standalone single-agent deployment and built multi-agent systems on composable single-agent primitives. Twenty-five of the 29 included studies describe multi-agent systems, yet nearly all construct them from four recurring single-agent building blocks. Understanding these primitives—LLM-only reasoning, tool augmentation, retrieval augmentation, and planning-and-execution decomposition—and the coordination mechanisms that compose them is essential.

Second, while multi-agent coordination is widely discussed, the literature lacks systematic documentation of coordination failure modes. Communication overhead, incentive misalignment, consensus deadlock, and authorization/audit gaps recur across studies but have not been synthesized into a coherent taxonomy linked to coordination paradigms.

Third, evaluation remains concentrated in benchmark and simulation settings. Only three studies report live or production-pilot evidence, and no study validates the correlation between benchmark performance and deployment outcomes. This evidence gap is the single largest threat to the credibility of financial LLM agent research and limits the confidence with which practitioners and regulators can assess deployment readiness.

This review introduces a unified three-dimensional Architecture–Capability–Coordination framework grounded in Simon's theory of near-decomposable systems (1962) and Malone and Crowston's coordination theory (1990, 1994). The framework treats coordination not as an afterthought but as a first-class design variable shaping how architectural primitives are assembled into deployable financial systems. By systematically mapping the explored design space and identifying priority research needs—including head-to-head comparisons of coordination mechanisms, paired benchmark-to-deployment validation studies, finance-specific coordination primitives, and standardized governance and auditability reporting—this review provides both a descriptive account of current capabilities and a prescriptive agenda for closing the gap between technical capability and accountable deployment.

Condition being studied This is a non-health systematic review. The condition being studied is the design, coordination, evaluation, and deployment readiness of large language model (LLM) agents in financial services. It refers not to a disease or medical condition but to a socio-technical condition: the increasing use of autonomous or semi-autonomous LLM-based software agents for high-stakes financial decisions, and the architectural, coordination, governance, and evidentiary problems that accompany that use.

The population is financial LLM agents, including single-agent and multi-agent systems. The concept covers four architectural primitives—LLM-only, tool-augmented, retrieval-augmented, and planning-and-execution—and their capabilities and limitations across financial task domains. The context is finance, including algorithmic trading, investment management, risk management, financial analysis, regulatory compliance, customer service/advisory, fraud detection, payments, credit lending, climate risk, and decentralized finance.

The review examines four multi-agent coordination paradigms: independent ensembles, competitive debate, cooperative hierarchy, and protocol-based peer-to-peer. It documents four recurring coordination failure modes: communication overhead, incentive misalignment, consensus deadlock, and authorization/audit gaps. It also assesses evaluation and deployment readiness across benchmark, simulation, live/production, and

framework-demonstration settings, together with governance and auditability.

This condition is important because LLM agents are being proposed for financial decisions that carry immediate monetary, regulatory, and reputational consequences, yet the empirical literature has not been systematically mapped onto a finance-specific framework. Existing surveys catalogue application areas but do not characterize the coordination spectrum, identify finance-specific failure modes, or validate the relationship between benchmark performance and live deployment. The evidence base is early-phase, dominated by controlled evaluations and conference proceedings, and shows a highly asymmetric distribution of findings, with almost no null or failed-deployment results. The review therefore treats financial LLM agents as an emerging design space in which four composable architectural primitives are assembled into multi-agent systems under different coordination mechanisms. Studies are excluded if they concern non-financial applications, non-LLM agents, non-peer-reviewed sources, or lack empirical evaluation. The purpose is to describe how these agents are built, how they coordinate, where they fail, and how ready they are for accountable deployment.

METHODS

Search strategy This review searched two bibliographic databases: Web of Science Core Collection and Scopus. The final database search was conducted on 31 August 2026. The search was supplemented by forward and backward citation searching of all included studies and the seven background reviews; citation searching was conducted in Web of Science Core Collection and Scopus on 3 September 2026. Both databases were accessed through their standard web interfaces. No date restriction was placed on database coverage, but inclusion was limited to studies published between January 2020 and August 2026. Only English-language publications were included. Grey literature, preprints without subsequent peer-reviewed publication, white papers, blog posts, technical reports without peer review, and opinion pieces were excluded by design.

The search strategy combined three concept blocks: (1) LLM agents; (2) finance and financial domains; and (3) single-agent, multi-agent, or agent coordination terms. The strategy was developed from the review's PCC framework—Population = financial LLM agents; Concept = architecture, capability, coordination, and evaluation/deployment; Context = finance—and from an initial scoping scan of the literature.

For Web of Science Core Collection, the final search string was:

TS = ("LLM agent" OR "large language model agent" OR "agentic AI" OR "language model agent*") AND TS = ("finance" OR "trading" OR "investment" OR "risk" OR "compliance" OR "financial market") AND TS = ("single-agent" OR "multi-agent" OR "agent coordination").

For Scopus, the conceptual structure was implemented in the TITLE-ABS-KEY field as:

TITLE-ABS-KEY("LLM agent" OR "large language model agent" OR "agentic AI" OR "language model agent*") AND ("finance" OR "trading" OR "investment" OR "risk" OR "compliance" OR "financial market") AND ("single-agent" OR "multi-agent" OR "agent coordination").

The search string was validated by checking whether it retrieved ten studies already known to the reviewers. All ten were retrieved and subsequently assessed for eligibility. The third concept block—single-agent, multi-agent, or agent coordination—was the most restrictive element of the strategy. It may have excluded studies that implement coordination mechanisms without explicitly labelling them as single-agent, multi-agent, or agent coordination. No automation tools were used beyond Rayyan for collaborative screening management.

Records were exported to Rayyan. Two independent reviewers screened titles and abstracts against pre-defined eligibility criteria. Records were excluded if they clearly concerned non-financial applications of LLM agents, non-LLM agents in finance, review/commentary/opinion articles, or non-English publications. Disagreements were resolved through discussion with a third reviewer. Inter-rater agreement at title/abstract screening was Cohen's kappa = 0.84, indicating substantial agreement. Full texts of potentially eligible records were then retrieved and independently assessed by the same two reviewers. Inter-rater agreement at full-text screening was kappa = 0.91. Reasons for exclusion were documented and reported in the PRISMA flow diagram. No studies were excluded at the full-text stage. The final included sample comprised 29 peer-reviewed studies published between January 2025 and August 2026.

The full search yielded 447 records: 311 from Scopus and 136 from Web of Science Core Collection. After 96 duplicates were removed, 351 unique records were screened at title/abstract. Of these, 322 were excluded: 248 (77.0%) for non-financial application of LLM agents, 68 (21.1%) for being review, commentary, or opinion articles, and 6 (1.9%) for non-LLM agents in finance. The remaining 29 full-text reports were sought for retrieval and assessed for eligibility, and all 29 met the inclusion criteria.

In summary, the search strategy combines Web of Science Core Collection and Scopus, a three-block Boolean string, validation against ten known studies, forward and backward citation searching of included studies and background reviews, and two-stage independent screening in Rayyan. The strategy is reproducible from the terms and database syntax reported above.

Participant or population This is a non-health review; no human patients are involved. The population is financial LLM agents—single-agent and multi-agent systems—as the unit of analysis, rather than individual human users. Included agents perform financial tasks such as trading, investment management, risk management, financial analysis, compliance, customer service/advisory, fraud detection, payments, credit lending, climate risk, and DeFi. Studies on non-financial LLM agents, non-LLM agents, or without empirical evaluation are excluded.

Intervention This is a non-clinical review. The "interventions" evaluated are configurations of financial LLM agent systems, not medical interventions. They include four architectural primitives (LLM-only, tool-augmented, retrieval-augmented, planning-and-execution), four multi-agent coordination paradigms (independent ensembles, competitive debate, cooperative hierarchy, protocol-based peer-to-peer), and associated governance/authorization mechanisms. The review also examines documented failure modes (communication overhead, incentive misalignment, consensus deadlock, authorization/audit gaps) and evaluation/deployment settings (benchmark, simulation, live/production, framework demonstration). No head-to-head comparison of coordination mechanisms exists; the review maps and synthesizes these design choices.

Comparator This is a non-clinical review; no single comparator is pre-specified. Study-level comparators include non-LLM baselines (rule-based agents, traditional ML, reinforcement learning), single-agent versus multi-agent configurations, alternative architectural primitives, alternative coordination paradigms, and financial benchmarks such as buy-and-hold or human performance. No included study compares coordination mechanisms head-to-head on a matched task, so comparative effectiveness is described narratively rather than pooled.

Study designs to be included This non-clinical review includes peer-reviewed original research: empirical evaluations, benchmark tests, simulations/backtests, live or production-pilot deployments, and framework demonstrations of financial LLM agents. Excludes reviews, opinion/commentary, non-empirical papers, non-LLM agents, and non-financial applications.

Eligibility criteria Additional eligibility criteria beyond PICOS: peer-reviewed journal articles or conference proceedings only; English language; published January 2020–August 2026; original empirical research with extractable data. Excluded: non-peer-reviewed sources (preprints without subsequent publication, white papers, blogs, technical reports), review/opinion/commentary pieces, duplicate records, and studies lacking empirical evaluation or sufficient data. No restriction by geography or financial sub-domain.

Information sources Web of Science Core Collection and Scopus were searched on 31 August 2026. Forward and backward citation searching of all included studies and the seven background reviews was conducted in the same two databases on 3 September 2026. No trial registers, grey literature, preprint servers, or direct contact with authors were used. No date restriction was placed on database coverage; inclusion was limited to studies published January 2020–August 2026. Rayyan was used only for collaborative screening management.

Main outcome(s) This non-clinical review reports descriptive, design-oriented outcomes across three research questions. RQ1: identification of four architectural primitives (LLM-only, tool-augmented, retrieval-augmented, planning-and-execution) and their capabilities and limitations across financial task domains. RQ2: mapping of four coordination paradigms (independent ensembles, competitive debate, cooperative hierarchy, protocol-based peer-to-peer) and four documented failure modes (communication overhead, incentive misalignment, consensus deadlock, authorization/audit gaps). RQ3: evaluation and deployment readiness across benchmark, simulation, live/production, and

framework-demonstration settings, plus governance and auditability. Certainty of evidence (adapted GRADE): moderate for architectural primitives, very low for comparative coordination effectiveness, very low for deployed performance, low for governance/auditability. Reporting bias: 28 positive/mixed-positive, 1 null, 0 failed/withdrawn deployments. Outcomes are synthesized narratively; no pooled effect estimate is computed.

Quality assessment / Risk of bias analysis This non-clinical review adapted a quality assessment tool based on established criteria for systematic reviews of design knowledge. Each included study was assessed on five domains: (1) theoretical grounding (explicit research questions or hypotheses); (2) methodological transparency (sufficient detail for replication); (3) evaluation robustness (appropriate metrics, validation, statistical testing); (4) external validity (validation beyond controlled laboratory benchmarks); and (5) deployment status (reporting of production or live deployment results). Each domain was rated as low, unclear, or high risk of bias. Two reviewers independently assessed all studies, resolving disagreements by discussion with third-reviewer arbitration. An overall judgement followed a pre-specified rule: low risk when no domain was rated high and at most one was unclear; high risk when three or more domains were rated high; moderate risk otherwise. High-risk studies were flagged for sensitivity analysis.

Reporting bias was assessed at two levels. Within studies, reviewers compared outcomes and analyses specified in each study's method or system description with those reported in its results, categorizing them as reported, partially reported, or not reported. Since no included study had a publicly available protocol or registration, this assessment was limited to internal comparisons. Between studies, reviewers screened for null, negative, and failed-deployment results and recorded the direction of each study's principal finding. Because no pooled effect estimate was computed, formal small-study effect methods (funnel plots, asymmetry tests) were not applicable.

Certainty of evidence was assessed using an adapted GRADE approach for four outcome domains: architectural primitives, comparative coordination effectiveness, deployed financial performance, and governance/auditability. Each domain started at high certainty and was downgraded one level for serious concerns and two levels for very serious concerns across risk of bias, inconsistency, indirectness, and imprecision. Two reviewers rated independently, with a third arbiter when ratings differed by more than one level. Two pre-specified rules applied: evidence based primarily on framework demonstrations was rated no higher than low, and evidence from fewer than five contributing studies was rated no higher than low.

Strategy of data synthesis Given the heterogeneity of study designs, outcome measures, and evaluation settings, no meta-analysis will be performed and no pooled effect estimate will be computed. Data will be synthesized narratively, organized around the review's three research questions and the Architecture–Capability–Coordination framework. Studies will be grouped by (1) agent architecture (LLM-only, tool-augmented, retrieval-augmented, planning-and-execution); (2) financial task domain (trading, risk management, financial analysis, compliance, customer service/advisory, and related areas); and (3) coordination paradigm (independent ensemble, competitive debate, cooperative hierarchy, protocol-based peer-to-peer). A study may contribute to more than one synthesis where it addresses multiple dimensions.

Quantitative performance statistics will be extracted in the units reported by original authors —classification metrics (accuracy, F1, precision, recall, MCC), financial metrics (return, Sharpe ratio, maximum drawdown, Information Coefficient), and others—and summarized descriptively. Tables will map study characteristics, architectural primitives, coordination mechanisms, documented failure modes (communication overhead, incentive misalignment, consensus deadlock, authorization/audit gaps), evaluation settings (benchmark, simulation, live/production, framework demonstration), and governance/auditability features.

Subgroup analysis Subgroup analyses will be descriptive and narrative only; no inferential or pooled subgroup comparisons will be performed, given the absence of head-to-head coordination comparisons and the small number of studies in several categories. Pre-specified subgroups correspond to the review's framework dimensions:

Coordination paradigm: independent ensembles, competitive debate, cooperative hierarchy, protocol-based peer-to-peer.

Financial task domain: algorithmic trading, risk management, financial analysis, regulatory compliance, customer service/advisory, and other

domains (e.g., fraud detection, payments, credit lending, climate risk).

Evaluation setting: benchmark, simulation, live/production, framework demonstration.

Architectural primitive: LLM-only, tool-augmented, retrieval-augmented, planning-and-execution.

Risk of bias: low, moderate, high (sensitivity analysis excluding high-risk studies).

Additional subgroups may include publication year (2025 vs. 2026) and publication type (journal vs. conference proceedings). Findings will be summarized as counts and narrative patterns, with explicit note of categories containing fewer than five studies (e.g., competitive debate n=2; independent ensembles n=3). No formal statistical subgroup testing will be conducted.

Sensitivity analysis Because no meta-analysis or pooled effect estimate will be performed, sensitivity analyses will be descriptive and narrative only. Pre-specified analyses will test whether the review's central patterns depend on studies with the weakest evidence.

First, studies rated at high overall risk of bias will be excluded, and subgroup counts and narrative patterns will be re-tabulated. Second, framework-demonstration studies will be excluded to assess whether findings on architectural primitives and coordination paradigms remain when only benchmark, simulation, backtest, or live/production evidence is retained. Third, studies lacking statistical testing or comparator details will be excluded to test the robustness of evaluation-related conclusions.

Fourth, analyses will examine whether conclusions are driven by very small categories, including competitive debate (n = 2), independent ensembles (n = 3), live/production evidence (n = 3), and governance/auditability evidence (n = 6). Findings from categories with fewer than five studies will be reported as indicative only and will not be used for comparative claims.

Results will be reported narratively. If the central descriptive patterns remain unchanged after exclusions, confidence in those patterns will be strengthened within the limits of the evidence. If patterns change materially, the affected conclusions will be downgraded or reported as unstable. Sensitivity analysis cannot address reporting bias due to missing null or negative results, because those results are unavailable by definition; this limitation will be stated explicitly rather than treated as remediable through sensitivity testing. No statistical subgroup testing or pooled sensitivity estimate will be conducted.

Country(ies) involved China (single Chinese author; HKU Business School, University of Hong Kong).

Keywords LLM agents; financial AI; multi-agent systems; agent coordination; risk management; deployment readiness.

Contributions of each author

Author 1 - Chengkun Wu.
Email: 838902979@qq.com