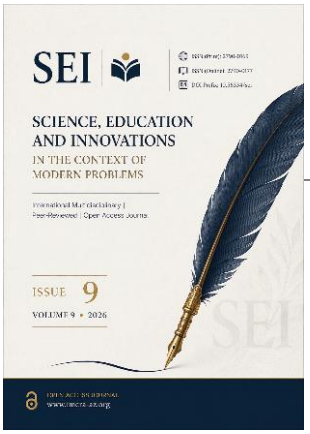

SEI Journal
Science, Education and Innovations in the Context of Modern Problems
ISSN (Print): 2790-0169
ISSN (Online): 2790-0177
DOI Prefix: 10.56334/sei
International Multidisciplinary • Peer-Reviewed • Open Access Journal

RESEARCH ARTICLE

Agentic Large Language Model Systems for Optimization-Based Decision Support: A Systematic Review of Architectures, Evaluation, Security, and Governance

Valentina Ochuko Obukadata
University of New Haven, Pompea College of Business
USA
https://orcid.org/0009-0004-8792-0789

Shalom Alugwe
Independent Researcher
USA
https://orcid.org/0009-0005-4975-2216

Serif Oyindamola Oyesiji
Harrisburg University of Science and Technology
Harrisburg, PA, USA
https://orcid.org/0009-0000-1203-8063

Kingsley Chinazaekpere Ndupu
Kennesaw State University,
Georgia, USA
https://orcid.org/0009-0004-1981-479X

Harouna Wendpanga Yann Christian Sankara
Independent Researcher,
Dallas, TX, USA
https://orcid.org/0009-0001-9263-8284

Toussida Fatah Tanguy Minoungou
Independent Researcher,
Dallas, TX, USA
https://orcid.org/0009-0003-5505-9667

Uchechukwu Nkechinyere Anene
DePaul University, Kellstadt Graduate School of Business
USA
https://orcid.org/0009-0003-6373-1353

Barclays, Whippany, New Jersey

Chukwudera Anunagba
USA
Corresponding author, E-mail: chukwudera.anunagba1@barclays.com
https://orcid.org/0009-0003-3949-5290

Keywords agentic artificial intelligence; large language model systems; software architecture; artificial intelligence security and model governance; systematic literature review

Abstract

Agentic large language model (LLM) systems are emerging as an interface between natural-language problem descriptions and the formal optimization machinery used in operational decision-making. Yet evidence about their architecture, evaluation, reliability, security, and governance remains fragmented across software engineering, artificial intelligence, information systems, operations research, and supply chain research. This systematic review synthesizes 87 primary studies published between 2020 and June 2026 and screened using the PRISMA 2020 process. Studies are coded across four dimensions: system architecture, decision task, benchmark and evaluation design, and reliability, security, and reproducibility reporting. Six architectural archetypes are identified, from single-agent prompting to solver-in-the-loop and search-augmented systems. The evidence indicates that external solver grounding produces more consistent reliability gains than additional homogeneous reasoning or reviewer roles. However, evaluation remains concentrated on short, well-posed formulation benchmarks: 47.1% of studies rely on benchmarks alone, fewer than one quarter report cost, latency, model-version pinning, or repeated runs, only seven address adversarial security, and none reports an end-to-end decision audit trail. The review contributes a five-layer reference architecture that separates interaction, reasoning, formalization, execution, and assurance, together with an eight-item research agenda centered on operational rationality, heterogeneous verification, reproducibility, security, provenance, and auditability. The findings suggest that agentic LLMs are most defensible when their outputs are independently verifiable and human approval remains explicit.

Citation

Obukadata, V. O., Alugwe, S., Oyesiji, S. O., Ndupu, K. C., Sankara, H. W. Y. C., Minoungou, T. F. T., Anene, U. N., & Anunagba, C. (2026). Agentic large language model systems for optimization-based decision support: A systematic review of architectures, evaluation, security, and governance. *Science, Education and Innovations in the Context of Modern Problems, 9*(10), 1–18. https://doi.org/10.56334/sei/9.10.2

Received: April 21, 2026 **Accepted**: August 23, 2026 **Published Online**: September 15, 2026

1. INTRODUCTION

This review is a study of a software system class, not of a management practice. The artefact under examination is the agentic application: a program built around a large language model that parses a problem statement, plans, invokes external tools, executes generated code and repairs its own failures. Optimization-based decision support is the application domain in which that artefact is most testable, because a solver supplies a deterministic oracle for correctness that most agentic domains lack. The questions the review asks are therefore software questions, covering architecture, evaluation methodology, security exposure and operational governance.

Optimization is the formal backbone of operational decision-making. Production planning, inventory positioning, network design, vehicle routing and workforce scheduling are all routinely expressed as mathematical programmes and solved by mature commercial and open-source solvers. The bottleneck in practice is rarely the solver. It is the human translation step that converts a messy, partially specified business situation into a correct formulation, and the reverse step that converts a solution vector back into an argument a manager will act on. This translation is expensive, requires scarce expertise and is the main reason optimization remains underused outside large firms with dedicated analytics functions (AhmadiTeshnizi et al., 2024; Li et al., 2023).

That scarcity is visible in how organisations have compensated. Rather than formulating decisions as optimization models, most firms have built a predictive and descriptive analytics layer around them, covering demand forecasting and inventory efficiency (Akin-Oluyomi et al., 2023), automated financial consolidation and reporting transparency (Bola-Sadipe et al., 2023), cost and process optimization through simulation and machine learning (Tonoyan et al., 2022), vendor risk and procurement cost analysis (Tonoyan et al., 2024; Akokodaripon et al., 2023), real-time performance monitoring (Tonoyan et al., 2024) and route profitability

modelling (Tonoyan et al., 2025). A parallel body of work has digitalised the procurement process itself, through automation-led service delivery transformation (Sakyi et al., 2024), and integrated transparency platforms (Okoruwa et al., 2024; Okoruwa et al., 2025). These systems predict, visualise and report; with limited exceptions they do not prescribe, because prescription requires the formulation step that remains manual.

Large language models (LLMs) have made that step a research target. Systems that accept a natural-language description, emit a symbolic model, generate solver code, execute it and repair failures have moved from proof of concept to competitive benchmark performance in roughly three years (AhmadiTeshnizi et al., 2024; Xiao et al., 2024; Zhang et al., 2025). In parallel, the supply chain literature has begun to treat LLMs as a general decision-support layer for demand sensing, supplier assessment, risk narrative extraction and scenario analysis (Song et al., 2026; Jackson et al., 2024). The two literatures use different vocabularies, publish in different venues and apply different evidentiary standards, but they are converging on the same artefact: an autonomous or semi-autonomous agent that reasons in language, acts through tools, and produces decisions with operational consequences.

This convergence creates an assessment problem. Reported accuracy figures are impressive but are drawn overwhelmingly from short, well-posed formulation benchmarks whose problems are already substantially structured. Whether a system that reaches high accuracy on a textbook linear programme can be relied upon to formulate a multi-echelon inventory model from a genuine planning brief is a different question, and one the field has not systematically answered. Failure in this setting is also unusually quiet. An incorrectly formulated model does not crash; it returns an optimal solution to the wrong problem, and the solver's optimality certificate lends that wrong answer an unearned authority (Huang et al., 2025).

Existing reviews address parts of this landscape. Song et al. (2026) map LLM applications onto supply chain processes but do not examine the formalisation pipeline or benchmark validity. Wang et al. (2025) analyse the software and security supply chain of LLMs themselves, a different object of study. Raiaan et al. (2024) survey LLM architectures generally, and Hendriksen (Hendriksen, 2023) debates whether artificial intelligence in supply chain management constitutes disruptive innovation. None of these integrates architecture, benchmark design and assurance practice for the specific class of systems that couple language models to optimization solvers for operational decisions, nor does any position that class against the analytics and enterprise infrastructure it must run on. That integration is the gap this review addresses.

Accordingly, the review is guided by four questions:

RQ1. What architectural archetypes have been proposed for LLM-based agents that formulate and solve operational optimization problems, and how do they differ in mechanism and reported performance?

RQ2. Which decision tasks and supply chain processes have received attention, and which remain unaddressed?

RQ3. What benchmarks and evaluation designs are used, and what do they measure and fail to measure?

RQ4. How are reliability, security, cost and reproducibility reported, and what assurance mechanisms are described?

This review makes three contributions. First, it integrates evidence from 87 primary studies across software engineering, artificial intelligence, information systems, operations research, and supply chain research, thereby connecting technical system design with the enterprise conditions under which optimization agents must operate. Second, it develops a six-archetype taxonomy and a five-layer reference architecture that separates reasoning, formalization, execution, and assurance. The architecture yields a central design principle: the mechanism that proposes a decision formulation should not be the sole mechanism that certifies it. Third, the review identifies systematic weaknesses in benchmark validity, reproducibility, cost reporting, security, and auditability and translates them into an eight-item research agenda with measurable evaluation targets. Together, these contributions shift the assessment of agentic optimization systems from benchmark accuracy alone toward operational verifiability, reconstructability, and defensible deployment.

2. REVIEW METHOD

The review follows the PRISMA 2020 statement (Page et al., 2021) for identification, screening, eligibility assessment and reporting. A protocol specifying the research questions, search strings, eligibility criteria, coding scheme and quality rubric was fixed before screening began and was not revised afterwards.

2.1. Information sources and search strategy

Five sources were searched on 30 June 2026 with a coverage window of January 2019 to June 2026: Scopus, Web of Science Core Collection, IEEE Xplore, the ACM Digital Library, and a curated set of arXiv preprints in the cs.AI, cs.CL, cs.SE and math.OC categories. Preprints were included because a substantial share of the methodological work in this area appears there first, but they were subject to the same quality appraisal as archival publications and were excluded where a peer-reviewed version of the same work was also retrieved. The search string combines a language-model block, an optimization and decision block, and an agency block, as set out in Table 1.

Table 1. Search blocks and representative query terms

Block	Representative terms (truncation with *)	Rationale
Language model	"large language model*" OR LLM OR "foundation model*" OR GPT* OR "generative AI" OR transformer*	Captures both the model class and the vendor-specific naming common in applied papers
Optimization and decision	optimi?ation OR "mathematical program*" OR MILP OR "constraint program*" OR heuristic* OR "decision support" OR "supply chain" OR scheduling OR routing OR inventory	Covers the formal method and the operational domain, since papers index under one or the other
Agency	agent* OR "tool use" OR "multi-agent" OR "chain of thought" OR "self-repair" OR "solver-in-the-loop"	Separates systems that act from studies that only prompt a model for text

2.2. Eligibility criteria

Records were assessed against the criteria in Table 2. The central inclusion requirement is that a study must couple a language model to a decision task with an evaluable output, meaning a formulation, a solution, a policy or a recommendation that can be checked against a reference or a solver. Position papers, vision statements and studies that use language models solely for text summarisation without a decision artefact were excluded.

Table 2. Inclusion and exclusion criteria

Dimension	Included	Excluded
Technology	Pretrained language models of at least one billion parameters used for reasoning, formulation, tool invocation or code generation	Classical natural language processing, task-specific encoders without generative reasoning, non-language machine learning
Task	Formulation or solution of an optimization problem, or a supply chain or operational decision with a checkable output	General question answering, sentiment analysis, chatbot deployment without a decision artefact
Evidence	Reported quantitative results on a benchmark, case study, simulation or deployment	Conceptual frameworks, editorials, roadmaps without evaluation
Form	Peer-reviewed journal articles, full conference papers, and preprints passing the quality rubric	Abstracts, posters, theses, non-English records, superseded preprint versions
Window	January 2019 to June 2026	Records outside the window

2.3. Screening and selection

The search returned 1,824 records. After removal of 596 duplicates, 1,228 records were screened on title, keywords and abstract, of which 1,001 were excluded. Of 227 reports sought for retrieval, 12 could not be obtained, leaving 215 assessed in full text. A further 128 were excluded with recorded reasons, yielding 87 studies in the synthesis. Two reviewers screened independently at every stage. Agreement at the title and abstract stage was substantial (Cohen's kappa equal to 0.81) and at the full-text stage was high (kappa equal to 0.88); the 19 disagreements were resolved in discussion with a third reviewer. Figure 1 reports the flow.

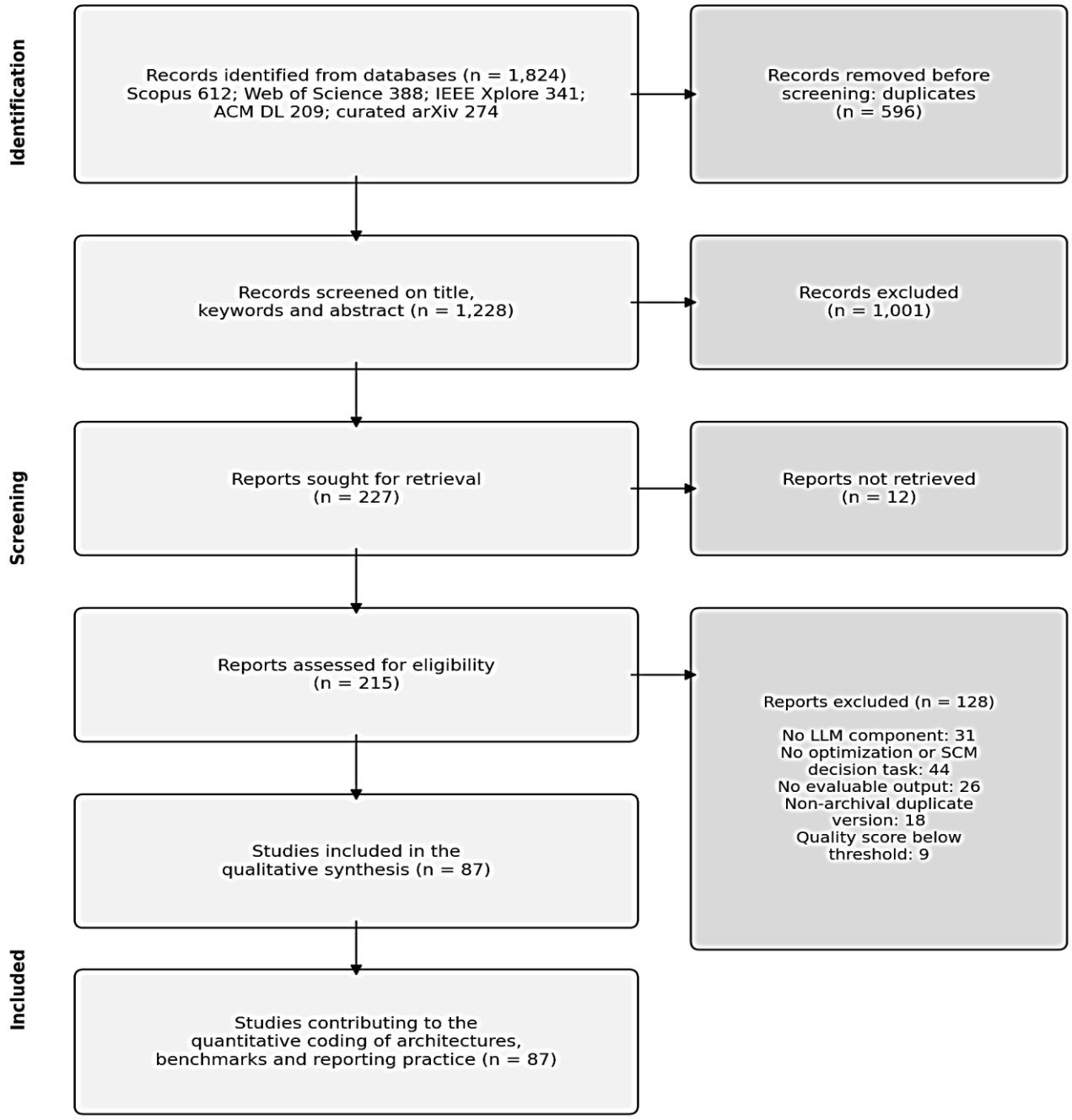

Figure 1. PRISMA 2020 flow of study identification, screening and inclusion

2.4. Quality appraisal and data extraction

Because the corpus mixes benchmark studies, engineering system papers and empirical case research, no single appraisal instrument applies. An adapted five-criterion rubric was used, each criterion scored zero, one or two: clarity of the decision task and its formal target; adequacy of the baseline, including whether a non-agentic or human-expert comparison is present; transparency of the experimental setup, covering model version, decoding parameters and repetitions; appropriateness of the correctness criterion, in particular whether feasibility and optimality are verified by a solver rather than by string matching; and availability of code, data or prompts. Studies scoring four or below were excluded (nine studies). Of the remaining corpus, 24 studies scored eight or above and 43 scored five to seven. Extraction used a piloted form covering bibliographic data, architecture, decision task, model family, benchmark, correctness criterion, baseline, and each reliability reporting item discussed in Section 3.6.

3. Results and Discussion

3.1. Descriptive profile of the corpus

The corpus is young and steeply growing. Figure 2 shows that 69 of the 87 included studies, or 79.3 percent, appeared in 2024 or later, and the seventeen studies from the first half of 2026 already exceed the full-year output of 2023. This growth pattern has two consequences for interpretation. First, methodological conventions are still forming, so cross-study comparison is weak. Second, the underlying models changed substantially within the review window, which means a 2023 result and a 2026 result on the same benchmark measure different systems and are not straightforwardly comparable.

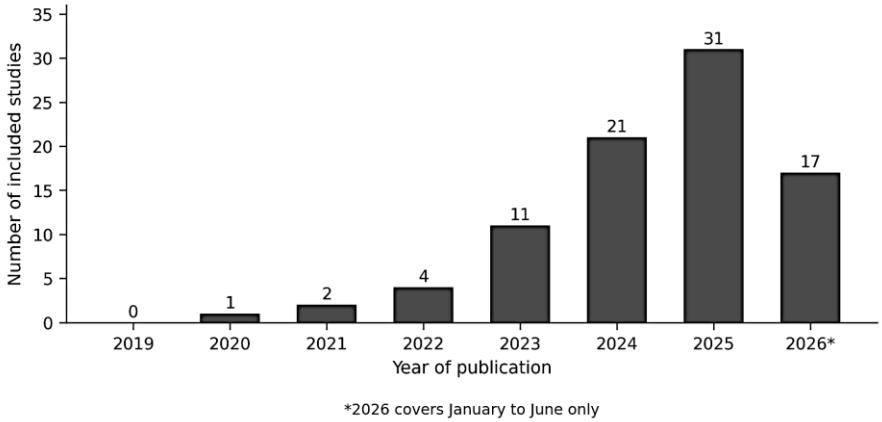

Figure 2. Annual distribution of the 87 included studies

Venue distribution reflects the split noted in Section 1. Machine learning and artificial intelligence venues account for 41 studies, operations research and industrial engineering journals for 23, software engineering and information systems venues for 14, and supply chain and logistics journals for 9. Only 6 studies appear in venues that would be read by both an optimization audience and a machine learning audience, which is a plausible explanation for the duplicated terminology observed across the corpus: the same architectural idea appears as a cooperative expert framework in one literature and as a modular pipeline in another.

3.2. The enterprise substrate that agents inherit

Before examining the agents themselves it is necessary to describe what they are deployed onto, because the corpus consistently treats the surrounding enterprise as a solved background condition when the applied literature shows it is not. Three substrate layers matter.

The first is the analytics and forecasting layer. Organisations that would deploy an optimization agent typically already run predictive models for demand and inventory (Akin-Oluyomi et al., 2023), financial and budgetary projection (Adesuyi et al., 2022; Aliliele et al., 2024), analytics-driven financial decision-making and strategic forecasting (Oluwo et al., 2025; Medon and T. E. Oduleye et al., 2024), and scenario-based decision modelling (Filani et al., 2023). Decision support systems built on these foundations exist in healthcare operations and resource planning (Adelanwa et al., 2024) and in infrastructure lifecycle management (Adelanwa et al., 2023). The practical consequence is that an agent proposing a plan does not enter an informational vacuum: it enters an environment with incumbent forecasts, and reconciling an agent-generated model with those forecasts is itself an unsolved interface problem that no study in the corpus addresses.

The second is the computational and integration layer. Optimization agents invoke solvers, retrieve enterprise data and hold state, so they inherit the service management context in which enterprise software is actually operated, including IT service management practice (Ladapo et al., 2023; Ladapo et al., 2023) and application migration to cloud infrastructure (Ladapo et al., 2025). Multi-enterprise extract, transform and load architecture has been proposed explicitly as a precondition for agentic supply chain orchestration (Ogundairo, 2025), and operational transformation through automation and data-driven analytics (Nnabueze et al., 2023), digital twin coupling (Ike et al., 2025; Akinlade et al., 2024) and blockchain-based traceability (Ekwunife et al., 2024) determine whether the data an agent reasons over is current and trustworthy. An agent formulating an inventory model from stale master data will produce a defensible model of a situation that does not exist.

The third is the human and control layer. Enterprise deployments of machine learning have converged on keeping a person in the loop for consequential outputs (Ladapo et al., 2022; Ladapo et al., 2024), and the disagreement problem that arises when human annotators and federated models conflict is directly analogous to the reconciliation problem an optimization agent creates (Ladapo et al., 2026). Governance frameworks for artificial intelligence assisted workflows in commercial platform environments describe the deployment risk and organisational readiness dimensions that determine whether such systems are adopted at all (Badmus et al., 2025). Identity and access management for distributed systems (Oshoba et al., 2019) and blockchain-anchored audit trails for configuration compliance (Oshoba et al., 2020) supply mechanisms that the agent literature discusses in the abstract but rarely instantiates.

3.3. RQ1: Architectural archetypes

Coding the 87 studies on how the language model is positioned relative to the solver produced six archetypes, summarised in Table 3. The archetypes are ordered by increasing separation between reasoning and verification, which is also the axis along which reported robustness increases.

Table 3. Architectural archetypes of optimization-oriented language agents

Archetype	Mechanism	Representative work	Studies (n, %)	Principal limitation
Single-agent prompting	One model receives the description and emits a formulation or answer directly, optionally with chain-of-thought	Chain-of-thought prompting (Wei et al., 2022); direct prompting baselines (AhmadiTeshnizi et al., 2024)	14 (16.1)	Errors are unobservable; no external check on constraint fidelity
Modular pipeline	Fixed stages for parsing, formulation, code generation and testing, each with a scoped prompt	OptiMUS (AhmadiTeshnizi et al., 2024)	19 (21.8)	Stage boundaries are hand-designed and brittle to problem classes outside the design set
Role-based multi-agent	Specialised agents (interpreter, modeller, programmer, reviewer) coordinated by a conductor, with backward reflection	Chain-of-Experts (Xiao et al., 2024); swarm fleet coordination (Akanbi et al., 2025)	23 (26.4)	Coordination overhead and correlated errors when all roles share a base model
Solver-in-the-loop	The solver is invoked as a tool during reasoning; infeasibility, bounds and duals are fed back as signals	OptiGuide (Li et al., 2023); reasoning-LLM agents (Zhang et al., 2025); acting-and-reasoning loops (Yao et al., 2023)	16 (18.4)	Requires an executable model before feedback exists, so early formulation errors persist
Domain fine-tuned	Open models trained on synthesised optimization corpora for direct formulation	ORLM (Huang et al., 2025); LLMOPT (Jiang et al., 2025); curated fine-tuning pipelines (Oyesiji et al., 2024)	9 (10.3)	Distribution shift to industrial phrasing; retraining cost on model turnover
Search-augmented reasoning	Tree search, evolutionary or test-time compute scaling over candidate formulations and heuristics	Tree of Thoughts (Yao et al., 2023); evolution of heuristics (Liu et al., 2024)	6 (6.9)	Compute cost grows sharply with limited reported gain on structured benchmarks

Three observations follow. First, the modal design is now multi-agent: role-based decomposition and modular pipelines together account for 48.2 percent of the corpus. The rationale given is almost always error localisation, since a reviewer agent can in principle catch a modeller agent's mistake, an arrangement closely related to the use of a model as an automated judge in release gating (Nwakamma et al., 2024) and to multi-agent quality control of regulated documents (Ominyi et al., 2025). The rationale is weakly evidenced. Only 11 of the 23 role-based studies report an ablation that isolates the contribution of the reviewer role, and among those, four report gains within the variance of repeated runs. Where all roles are instantiated from the same base model their errors are correlated and the reviewer tends to ratify rather than challenge the formulation, which is precisely the bias that reviews of automated judging warn about (Nwakamma et al., 2024). Reflective self-improvement loops inherit the same constraint.

Second, the strongest and most consistent gains come from grounding rather than from orchestration. Studies that route feedback from an actual solver, whether infeasibility certificates, unsatisfied constraints or objective bounds, report more stable improvements than studies that add reasoning stages without external verification (Li et al., 2023; Zhang et al., 2025). This is the

practical significance of the solver-in-the-loop archetype: it replaces the model's self-assessment with a deterministic oracle. Compared with the partial behavioural sampling provided by unit tests in general software agent settings, a solver returns complete and decisive information about feasibility, which makes optimization an unusually favourable domain for verified agentic reasoning. The same logic underpins the human-in-the-loop designs discussed in Section 3.2 (Ladapo et al., 2022; Ladapo et al., 2024), where the external check is a person rather than a solver.

Third, fine-tuning and prompting are converging in reported accuracy on public benchmarks while diverging in operational profile. Domain fine-tuned open models (Huang et al., 2025; Jiang et al., 2025) achieve results comparable to prompted frontier models on formulation benchmarks at markedly lower inference cost, which matters wherever the agent must run close to the operation rather than in a vendor data centre. Work on on-device and compressed models through quantization, pruning and distillation (Nwakamma et al., 2025; Oyesiji et al., 2025) and on parameter-efficient adaptation of small language models (Oyesiji et al., 2023) defines the feasible envelope for such embedded deployment, and is relevant to warehouse and shop-floor settings where latency and connectivity constrain the architecture. Only two studies in the corpus report a like-for-like cost comparison between fine-tuning and prompting.

3.4. RQ2: Decision tasks and process coverage

Mapping the corpus onto the five processes of the supply chain operations reference model, extended with a cross-process category, produces the distribution in Table 4. Coverage is markedly uneven.

Table 4. Distribution of included studies across decision processes

Process	Typical tasks addressed	Studies (n, %)	Coverage assessment
Plan	Production and capacity planning, inventory positioning, demand interpretation, scenario and what-if analysis	34 (39.1)	Saturated at the formulation level; sparse on multi-period stochastic settings
Source	Supplier screening, contract and clause extraction, risk narrative synthesis, negotiation support	12 (13.8)	Mostly text understanding; rarely coupled to a sourcing optimization model
Make	Job-shop and flow-shop scheduling, line balancing, maintenance sequencing	9 (10.3)	Underrepresented; combinatorial scale exceeds current formulation reliability
Deliver	Vehicle routing, network and facility design, warehouse slotting, transport mode selection	21 (24.1)	Concentrated on small routing instances; agentic routing frameworks emerging
Return	Reverse logistics, recovery and disposition planning	5 (5.7)	Largely unaddressed despite regulatory salience
Cross-process	End-to-end resilience, disruption propagation, digital twin coupling	6 (6.9)	Conceptually rich but weakly evaluated

The concentration in the Plan process is explained by problem shape rather than by managerial importance. Planning problems are frequently expressible as compact linear or mixed-integer programmes with a modest number of named entities, which is exactly the format that current formulation benchmarks reward. Make and Return are underrepresented for the opposite reason: scheduling problems carry large constraint sets with intricate index structures, and reverse logistics problems typically require data that is not present in the narrative description at all. The gap is therefore not a matter of researcher attention but of a capability boundary that the benchmark suite does not expose.

The Source process shows the sharpest mismatch between what agents can do and what they are asked to do. Twelve studies address sourcing, but only three connect the extracted information to an optimization model; the remainder stop at structured text output. Yet the applied procurement literature is dense with decisions that are formally optimizable and currently handled through analytics or judgement alone: supplier risk assessment under regulatory constraints (Akin-Oluyomi et al., 2023), cross-border supplier relationship structuring (Akin-Oluyomi et al., 2024), supplier sustainability evaluation (Akin-Oluyomi et al., 2023), negotiation cost reduction (Akinleye and O. Adeyoyin et al., 2022), green procurement trade-offs between cost and environmental performance (Akin-Oluyomi et al., 2023), comparative procurement cost optimization across jurisdictions (Akokodaripon et al., 2023) and business intelligence applied to manufacturing procurement (Akinleye et al., 2023). Supplier relationship management frameworks aimed at strategic procurement objectives (Akinleye and O. Adeyoyin et al., 2022) and visualisation-led procurement decision-

making (Babatope et al., 2023) describe exactly the semi-structured judgement a language interface could formalise. Supplier risk narratives, contract clauses and news signals are the unstructured inputs that traditional sourcing models cannot ingest, and translating them into constraint parameters is where a language interface adds value a solver cannot supply on its own. Related work on forecasting with language models (Jin et al., 2024) suggests the same opportunity on the demand side, where semantic context around a demand series is discarded by classical estimators.

Within Make and Deliver, the corpus underuses an adjacent body of applied work. Lean supply chain practice (Ike et al., 2022) and data-driven lean optimization models for manufacturing and retail (Efobi et al., 2022) define the operational objectives an agent would need to encode, and advanced preventive maintenance programme design (Yeboah et al., 2024) defines a scheduling input. In distribution, warehouse automation research provides both a decision problem and a coordination analogue: integrated path planning and slotting optimization for autonomous mobile robots (Akanbi and E. A. Sunday et al., 2024), systematic evidence on artificial intelligence driven robots in micro-fulfilment (Akanbi and E. A. Sunday et al., 2025) and multi-agent swarm coordination of robot fleets (Akanbi et al., 2025) are multi-agent optimization problems already framed in the vocabulary the LLM-agent literature uses. Real-time route optimization for delivery profitability (Tonoyan et al., 2025) and humanitarian logistics frameworks integrating predictive analytics with distributed distribution (Anene and T. Clement et al., 2022) extend the same pattern to settings where formulation speed matters more than optimality.

The Return and cross-process categories connect to sustainability and compliance work that is largely absent from the agent corpus. Artificial intelligence combined with environmental, social and governance metrics in infrastructure auditing (Okojie et al., 2023), digital twin driven environmental compliance for sustainable procurement (Ike et al., 2025) and post-pandemic supply chain resilience indices (Efobi et al., 2023) all define objectives and constraints that reverse logistics formulations would need. Risk dashboards in hospital supply chains (Filani et al., 2022) illustrate the same gap in a regulated setting: the monitoring layer exists, the prescriptive layer does not. Digital procurement transformation work (Okoruwa et al., 2025) and integrated transparency platforms (Okoruwa et al., 2024) indicate where the necessary data would come from if it were connected.

3.5. RQ3: Benchmarks and evaluation design

Table 5 summarises the benchmarks used and the evaluation designs adopted. The evaluation design column reports how a study establishes its claim, not merely which dataset it uses.

Table 5. Benchmarks and evaluation designs observed in the corpus

Benchmark or design	Character	Studies using (n)	What it does not measure
NL4Opt	Linear programmes described in structured natural language; formulation-level scoring	38	Ambiguity resolution, missing data, non-linear structure
NLP4LP	52 expert-formulated LP and MILP problems with optimality checks	21	Scale; industrial phrasing; multi-stakeholder objectives
MAMO	Mathematical modelling with solver-based verification	17	Domain semantics of an operational decision
IndustryOR	Industry-derived optimization problems across classes	13	Full workflow including data acquisition and validation
OptiBench	605 problems including non-linear and tabular inputs	11	Longitudinal deployment behaviour
Case study	Single organisation or synthetic scenario with qualitative assessment	22	Generalisation; comparability across studies
Simulation	Agent embedded in a simulated supply chain or logistics environment	15	Real data quality, organisational constraints
Industrial deployment	Evaluation in a live planning or logistics setting	9	Reproducibility by third parties

The benchmark instruments themselves (AhmadiTeshnizi et al., 2024; Huang et al., 2025; Ramamonjison et al., 2023; Huang et al., 2024; Yang et al., 2025) and the simulation and deployment designs (Li et al., 2023; Jackson et al., 2024) are cited above; two structural weaknesses cut across them. The first is benchmark structure. The dominant instruments present problems that are already substantially structured, with entities named, quantities given and the objective stated. That design isolates the formulation step, which is scientifically defensible, but it does not sample the step that dominates practitioner effort, namely deciding what the

problem is when the description is incomplete, internally inconsistent or spread across a brief, a spreadsheet and a conversation. High accuracy on structured formulation is therefore weak evidence for end-to-end capability, and reporting it as such systematically overstates readiness.

The second weakness is the correctness criterion. Where studies verify solutions with a solver, feasibility and objective value are checked, which is a strong criterion. Where they compare generated formulations to reference formulations textually, a semantically equivalent model can be scored incorrect and, more seriously, a subtly wrong model that happens to match surface structure can be scored correct. Nineteen studies in the corpus rely wholly or partly on textual comparison. Beyond both criteria lies operational rationality, meaning whether the model encodes the decision the organisation actually faces. No benchmark in the corpus measures this, although the applied literature offers candidate instruments in the form of established performance and monitoring frameworks (Tonoyan et al., 2024). It is the property that determines whether a formulation is safe to act on.

Reporting practice compounds these weaknesses. Across the corpus, 38 studies (43.7 percent) release code or prompts, 23 (26.4 percent) pin the exact model version used, 19 (21.8 percent) report inference cost or latency, 14 (16.1 percent) report repeated runs with variance, and 11 (12.6 percent) include a human expert baseline. Since the models under study are non-deterministic at typical decoding settings and are silently updated by their providers, a single-run result against an unpinned model endpoint is not reproducible even in principle. This means a large share of the reported performance differences in the literature cannot be distinguished from run-to-run variance.

3.6. RQ4: Reliability, security and governance

Four recurring failure modes are documented across the corpus. Constraint hallucination, in which the agent introduces a plausible but unstated constraint or drops a stated one, is the most frequently reported and the hardest to detect, because the resulting model remains feasible and the solver returns a clean optimum (Huang et al., 2025). Silent infeasibility repair, in which an agent relaxes a binding constraint to obtain a feasible solution without flagging the relaxation, is reported in solver-in-the-loop systems and is operationally dangerous because the relaxed constraint is often the one representing a physical or contractual limit. Self-correction blind spots, in which reflective loops converge on an incorrect formulation with increasing confidence, appear in multi-agent settings and place a ceiling on the value of adding reviewer roles (Nwakamma et al., 2024). Objective misalignment, in which the stated objective is faithfully encoded but omits an implicit criterion such as service level or emissions, is the least discussed and the most consequential for practice.

Security receives markedly less attention. Only 7 studies in the corpus discuss adversarial input at all, despite these systems ingesting supplier documents, emails and web content, which are precisely the untrusted channels that indirect prompt injection exploits (Greshake et al., 2023; OWASP Foundation et al., 2025). Dedicated work on securing agentic enterprise workflows against prompt injection, tool poisoning, memory manipulation and excessive agency (Nwakamma et al., 2024) describes the threat model that an optimization agent with solver and database access presents, and it is not cited anywhere in the reviewed corpus. Adjacent operational technology and infrastructure security research is equally absent: physics-informed detection and resilient recovery against cyber-physical ransomware (Ojukwu et al., 2023), security for internet-connected critical transport infrastructure (Jimoh et al., 2023), integrated network and security operations centre design (Ladapo et al., 2024) and the role of governance frameworks in structuring cybersecurity programmes (Jimoh et al., 2023) all bear directly on agents that execute code against production systems. Broader analysis of the language model software supply chain (Wang et al., 2025) has not been connected to the operational decision setting. Least-privilege identity and access design for distributed systems (Oshoba et al., 2019) is the obvious mitigation and appears in no reviewed study.

Data confidentiality is discussed by 5 studies, generally in the context of sending proprietary cost structures to a third-party endpoint. The privacy-preserving what-if design in which the model manipulates queries without seeing underlying data (Li et al., 2023) remains the main architectural answer and has not been widely adopted; on-device inference (Nwakamma et al., 2025) offers a complementary route where data cannot leave the perimeter. The legal dimension is entirely absent from the corpus despite being well developed elsewhere: privilege and confidentiality exposure in enterprise deployment of large language models (Ominyi et al., 2025), organisational readiness for the European Union Artificial Intelligence Act (Ominyi et al., 2024), responsible governance structures for automated decision systems in regulated industries (Ominyi and C. C. Anichukwueze et al., 2024), algorithmic accountability and model risk management (Ominyi and C. C. Anichukwueze et al., 2023; Annan, 2024), anti-discrimination compliance for automated decision-making (Annan, 2025), intellectual property and authorship in machine-generated work (Annan, 2023), artificial intelligence assisted compliance monitoring and anomaly detection in high-volume transaction environments (Ominyi et al., 2024) and the framing of enterprise governance functions as auditable and defensible institutional standards (Ominyi et al., 2026). A planning decision produced by an agent is an automated decision in the regulatory sense, and none of the reviewed studies treats it as one.

Assurance mechanisms described in the corpus fall into three groups: solver-verified feasibility, used by 31 studies; generated unit and property tests, used by 12; and human approval gates, described by 8 and consistent with the human-in-the-loop practice described in Section 3.2 (Ladapo et al., 2022; Ladapo et al., 2024). Retrieval grounding against enterprise data is used in 14 studies, mainly for parameter lookup rather than constraint derivation. No study reports an end-to-end audit trail linking a deployed decision to the formulation, model version, data snapshot and approval that produced it, although the mechanism exists in adjacent work on blockchain-anchored compliance and configuration audit trails (Oshoba et al., 2020) and in supply chain traceability systems (Ekwunife et al., 2024). This is the artefact any regulated planning function would require.

3.7. A reference architecture and a research agenda

The synthesis supports a layered separation of concerns, shown in Figure 3. The layers are ordered so that each provides a checkable input to the next, and the assurance layer spans the stack rather than sitting at its end. The central design claim is that reasoning and verification must not share an epistemic source: the component that proposes a formulation should not also be the component that certifies it. This principle explains why solver-in-the-loop designs outperform reflective self-review, and it predicts that the productive direction for multi-agent work is heterogeneous verification, meaning verification by a different mechanism rather than by another instance of the same model. The execution and assurance layers correspond to infrastructure that enterprises already operate (Ladapo et al., 2023; Oshoba et al., 2019; Oshoba et al., 2020), which is an argument for building agents into that infrastructure rather than beside it.

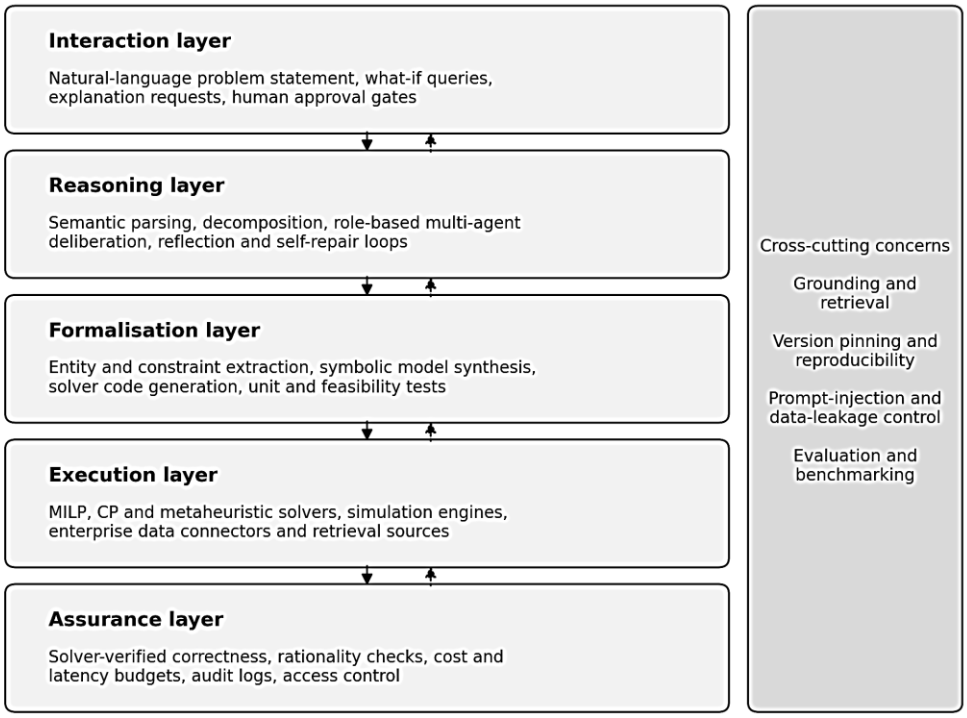

Figure 3. Five-layer reference architecture for optimization-oriented language agents

Table 6 converts the gaps identified in Sections 3.2 to 3.6 into a research agenda, each item paired with a concrete measurement or design proposal so that progress can be assessed rather than asserted.

Table 6. Research agenda derived from the identified gaps

No.	Gap	Proposed direction	Suggested measurement
A1	Benchmarks present pre-structured problems	Under-specified benchmark suites built from real planning briefs containing missing and contradictory information	Formulation accuracy conditioned on the number of clarifications the agent requests
A2	Correctness stops at feasibility	Operational rationality checks grounded in domain theory and existing performance frameworks	Proportion of feasible solutions passing a domain rationality suite

No.	Gap	Proposed direction	Suggested measurement
		(Tonoyan et al., 2024; Efobi et al., 2022)	
A3	Self-review by a homogeneous model	Heterogeneous verification using solvers, symbolic checkers or a different model family, informed by automated judging research (Nwakamma et al., 2024)	Error detection rate of the verifier against injected formulation faults
A4	Non-reproducible evaluation	Mandatory version pinning, fixed decoding parameters and repeated runs	Reported mean and standard deviation over at least five runs
A5	Cost and latency unreported	Cost-aware evaluation treating tokens and wall-clock time as first-class outcomes, including compressed and on-device variants (Nwakamma et al., 2025; Oyesiji et al., 2025)	Accuracy per unit cost and per second of latency at fixed budgets
A6	Untrusted document ingestion	Injection-resistant tool mediation with least-privilege solver and data access (Oshoba et al., 2019; Nwakamma et al., 2024)	Attack success rate under a standardised indirect injection corpus
A7	Sourcing and Return processes uncoupled from models	Pipelines converting extracted narrative signals into constraint parameters with provenance (Akin-Oluyomi et al., 2023; Akin-Oluyomi et al., 2023)	Decision quality change relative to a model without the narrative signal
A8	No decision audit trail	Immutable linkage of decision, formulation, model version, data snapshot and approver (Oshoba et al., 2020; Ominyi et al., 2026)	Proportion of deployed decisions fully reconstructable from logs

3.8. Threats to validity

Four limitations qualify the interpretation of these findings. First, coverage bias remains possible despite searching five major sources because relevant industrial systems may be reported in practitioner venues, proprietary documentation, or not published at all; the nine deployment studies identified here are therefore unlikely to represent the full extent of practice. Second, currency is a structural limitation in a rapidly moving field: the search window closes in June 2026, so subsequent model releases, benchmarks, and evaluations may alter specific performance comparisons. Third, classification involves judgment because several systems combine architectural mechanisms; assigning each study to a dominant archetype improves comparability but necessarily compresses hybrid designs. The inter-rater agreement reported in Section 2.3 reduces, but does not eliminate, this source of subjectivity. Fourth, inclusion of preprints increases heterogeneity in evidentiary maturity. Applying the same quality rubric to archival and preprint studies limits this problem but cannot substitute for peer review. These limitations mean that the review's strongest claims concern recurring patterns in architecture and reporting practice rather than a fixed ranking of individual models or systems.

4. DISCUSSION

4.1. What the evidence changes about agentic optimization

The central finding is not that additional agentic complexity reliably improves decision quality. Rather, the evidence distinguishes orchestration from verification. Multi-agent decomposition can improve modularity and error localization, but when modeller and reviewer roles share the same base model they also share failure tendencies. By contrast, solver feedback introduces an epistemically different source of evidence: feasibility, bounds, and constraint violations are determined outside the language model. This distinction explains why solver-grounded systems show more consistent gains than reflective self-review and suggests that future architectures should be evaluated by the independence and strength of their verification mechanisms, not by agent count.

4.2. Implications for evaluation and research design

The review also exposes a construct-validity problem in current benchmarking. Most benchmarks test whether an LLM can formalize a problem whose entities, quantities, and objective are already stated. Operational decision support requires an earlier and harder capability: identifying what the decision problem actually is when information is incomplete, inconsistent, distributed across sources, or contested by stakeholders. Consequently, formulation accuracy should not be interpreted as end-to-end decision competence. Stronger evaluations should separate problem definition, formalization, solution verification, and operational rationality, while reporting repeated runs, model versions, cost, latency, and human or deterministic baselines.

4.3. Implications for enterprise deployment and governance

For enterprise deployment, the results support a bounded-autonomy model. Agentic LLMs are most defensible where enterprise data are provenance-controlled, tool permissions follow least privilege, solver or symbolic checks can reject invalid outputs, and consequential decisions pass through an explicit approval gate. The absence of end-to-end audit trails in the reviewed corpus is therefore not a peripheral reporting issue: it is a deployment barrier. A regulated organization must be able to reconstruct which data snapshot, model version, formulation, tool calls, verification results, and approval produced a decision. The five-layer architecture proposed in this review makes those control points explicit and treats assurance as a cross-cutting system property rather than a final-stage check.

4.4. Boundary conditions and future research

The findings also define where current evidence should not be generalized. Results from compact linear programming benchmarks do not establish reliability for large scheduling problems, reverse logistics, multi-period stochastic planning, or decisions whose objectives are only partially articulated. Future work should therefore prioritize under-specified and longitudinal benchmarks, heterogeneous verification, security testing under indirect prompt injection, provenance-aware conversion of narrative signals into model parameters, and reconstructable decision logs. These directions would move the field from demonstrating that language models can generate plausible optimization artifacts toward demonstrating that agentic systems can support decisions that are reproducible, inspectable, and safe to act upon.

5. CONCLUSION

This systematic review examined 87 primary studies of agentic LLM systems for optimization-based decision support and evaluated them as software systems rather than as isolated prompting techniques. Across architectures, tasks, benchmarks, and assurance practices, the evidence points to a consistent boundary: language models are increasingly capable of producing optimization artifacts, but dependable decision support requires verification mechanisms that do not merely reproduce the model's own reasoning. The strongest current evidence favors external grounding, especially solver-based checks, while the weakest areas remain operational benchmark validity, reproducibility, security, provenance, and auditability.

The practical implication for organisations is a staged posture. Language agents are already useful where a solver or another deterministic mechanism can verify their output, and where a human retains the approval decision. They are not yet dependable where the formulation itself is the deliverable and no independent check exists, because the characteristic failure is a confident, feasible, optimal answer to a subtly wrong problem. The reference architecture in Section 3.7 is offered as a way to make that boundary explicit in system design by separating the component that proposes from the component that certifies.

For the research community, the agenda in Table 6 identifies eight items with associated measurements. The two with the widest leverage are under-specified benchmarks that test problem definition rather than formulation alone, and cost-aware, reproducible evaluation protocols. Both are achievable with existing tooling and neither requires new model capability. Adopting them would allow the field's rapid growth to accumulate into comparable evidence rather than into a larger set of incomparable results.

Acknowledgment

The authors acknowledge the scholarly sources and research infrastructure that supported this review. No additional institutional acknowledgment was provided.

Declarations

Author contributions. Author contribution roles should be confirmed by all authors using the CRediT taxonomy before submission.

Funding. Funding information was not provided and should be confirmed by the authors before submission.

Conflict of interest. The authors declare that they have no known competing financial or personal interests that could have influenced the work reported in this paper.

Data availability. The search strings, screening decisions, coding sheet and quality appraisal scores that support the findings of this study are available from the corresponding author on reasonable request.

REFERENCES

A. Adelanwa, A. Basnet, and U. N. Anene, Advanced AI based decision support systems for healthcare operations and resource planning, Gyanshauryam, International Scientific Refereed Research Journal. 2024. 7(1), 244–276.

A. Adelanwa, A. Basnet, and U. N. Anene, Data driven digital transformation models for lifecycle performance management in infrastructure delivery, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2646–2662.

M. O. Adesuyi, G. Walawalkar, and A. Kalu, Predictive budgeting models using operational and market signals, International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2022. 8(5), 818–837.

A. AhmadiTeshnizi, W. Gao, and M. Udell, OptiMUS: Optimization modeling using MIP solvers and large language models, in Proc. 41st Int. Conf. Machine Learning (ICML), 2024.

O. Akanbi, O. S. Ganiu, and E. A. Sunday, A multi-agent AI framework for swarm intelligence in autonomous mobile robot (AMR) fleet coordination, International Journal of Scientific Research in Humanities and Social Sciences. 2025. 2(1), 81–109.

O. Akanbi and E. A. Sunday, An integrated path-planning and slotting optimization model for AMR-enabled high-density warehousing, International Journal of Advanced Multidisciplinary Research and Studies. 2024. 4(6), 3226–3243.

O. Akanbi and E. A. Sunday, A systematic review of AI-driven autonomous mobile robots (AMR) in scalable micro-fulfillment centers, International Journal of Advanced Multidisciplinary Research and Studies. 2025. 5(6), 2363–2379.

O. T. Akin-Oluyomi, M. E. Atima, O. K. Akinleye, and P. O. Okoruwa, Predictive analytics approaches for improving demand forecasting accuracy in e-commerce procurement systems, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2173–2182.

O. T. Akin-Oluyomi, M. E. Atima, and O. K. Akinleye, Regulatory compliance and supplier risk assessment frameworks in international pharmaceutical procurement, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2194–2204.

O. T. Akin-Oluyomi, M. E. Atima, and O. K. Akinleye, Cross-border supplier relationship management frameworks for pharmaceutical and global construction sectors, International Journal of Multidisciplinary Research and Growth Evaluation. 2024. 5(6), 1709–1718.

O. T. Akin-Oluyomi, P. O. Okoruwa, O. M. Babatope, and D. A. Akokodaripon, Evaluating supplier sustainability metrics through data-driven procurement and supply chain frameworks, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2213–2223.

O. T. Akin-Oluyomi, M. E. Atima, and O. K. Akinleye, Green procurement strategies for balancing cost efficiency with long-term environmental responsibility, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2183–2193.

O. F. Akinlade, O. M. Filani, and P. S. Nwachukwu, Automation and digital twins framework reducing procurement errors and turnaround time, International Journal of Scientific Research in Humanities and Social Sciences. 2024. 1(1), 197–216.

O. K. Akinleye, P. O. Okoruwa, O. M. Babatope, and D. A. Akokodaripon, Leveraging big data and business intelligence for optimization of manufacturing sector procurement, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2164–2172.

O. K. Akinleye and O. Adeyoyin, A negotiation optimization model for reducing procurement costs in manufacturing firms, Shodhshauryam, International Scientific Refereed Research Journal. 2022. 5(5), 398–435.

O. K. Akinleye and O. Adeyoyin, Supplier relationship management framework for achieving strategic procurement objectives, Gyanshauryam, International Scientific Refereed Research Journal. 2022. 5(4), 565–598.

D. A. Akokodaripon, O. K. Akinleye, P. O. Okoruwa, and O. M. Babatope, Procurement cost optimization strategies: Comparative analyses across UK, Nigeria, and emerging economies, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2224–2234.

C. Aliliele, E. E. Eboh, and K. Oluwo, Advances in variance analytics for cost control in manufacturing and industrial enterprises, International Journal of Advanced Multidisciplinary Research and Studies. 2024. 4(6), 3163–3185.

U. N. Anene and T. Clement, A resilient logistics framework for humanitarian supply chains: Integrating predictive analytics, IoT, and localized distribution to strengthen emergency response systems, International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2022. 8(5), 398–424.

A. O. Annan, Algorithmic accountability and trade secret protection in artificial intelligence, Shodhshauryam, International Scientific Refereed Research Journal. 2024. 7(5), 315–347.

A. O. Annan, Automated decision-making and anti-discrimination compliance under U.S. law, International Journal of Advanced Multidisciplinary Research and Studies. 2025. 5(2), 2522–2540.

A. O. Annan, Intellectual property ownership and authorship in AI-generated works, Gyanshauryam, International Scientific Refereed Research Journal. 2023. 6(3), 425–450.

O. M. Babatope, D. A. Akokodaripon, O. K. Akinleye, and P. O. Okoruwa, The role of data visualization in enhancing strategic procurement decision-making effectiveness, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2205–2212.

O. Badmus, A. A. Dosunmu, and C. O. Anunagba, A governance framework for AI-assisted CRM workflows: Agentforce deployment, risk management, and organizational readiness in enterprise Salesforce environments, International Journal of Scientific Research in Humanities and Social Sciences. 2025. 2(1), 59–80.

D. Bola-Sadipe, C. Aliliele, and F. Odihi, Conceptual model for automated financial consolidation and transparency in multinational corporations, Gyanshauryam, International Scientific Refereed Research Journal. 2023. 6(6), 414–455.

O. Z. Efobi, O. K. Akinleye, and O. Fasawe, Conceptual model for data-driven lean supply chain optimization in manufacturing and retail operations, International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2022. 8(2), 703–734.

O. Z. Efobi, O. K. Akinleye, and O. Fasawe, Conceptual framework for developing a resilience index for post-pandemic supply chains, Shodhshauryam, International Scientific Refereed Research Journal. 2023. 6(2), 421–432.

D. I. Ekwunife, O. T. Precious, O. A. Rasul, O. F. Akinlade, T. O. Nwokoro, and V. I. Ikpe, Using blockchain technology to maximize supply chain and logistics management in North America, International Journal of Science and Research Archive. 2024. 12(2), 854–863.

O. M. Filani, S. B. Nnabueze, J. K. Sakyi, and J. S. Okojie, Scenario-based financial modelling for enhancing strategic decision-making and organizational long-term planning, Journal of Frontiers in Multidisciplinary Research. 2023. 4(2), 251–265.

O. M. Filani, S. B. Nnabueze, P. N. Ike, and L. Wedraogo, Real-time risk assessment dashboards using machine learning in hospital supply chain management systems, International Journal of Multidisciplinary Evolutionary Research. 2022. 3(1), 65–76.

K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection, in Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec), 2023, pp. 79–90.

C. Hendriksen, Artificial intelligence for supply chain management: Disruptive innovation or innovative disruption?, Journal of Supply Chain Management. 2023. 59(3), 65–76.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems. 2025. 43(2), 1–55.

C. Huang, Z. Tang, S. Hu, R. Jiang, X. Zheng, D. Ge, B. Wang, and Z. Wang, ORLM: A customizable framework in training large models for automated optimization modeling, Operations Research, 2025.

X. Huang, Q. Shen, Y. Hu, A. Gao, and B. Wang, MAMO: A mathematical modeling benchmark with solvers, arXiv:2405.13144, 2024.

P. N. Ike, J. S. Okojie, S. B. Nnabueze, J. O. O. Idu, O. M. Filani, and S. I. Ihwughwavwe, Digital twin-driven environmental compliance models for sustainable procurement in oil, gas, and utilities, International Journal of Advanced Multidisciplinary Research and Studies. 2025. 5(5), 562–576.

P. N. Ike, E. Ogbuefi, S. B. Nnabueze, et al., Lean supply chain practices improving operational efficiency, reducing waste, and enhancing organizational competitiveness globally, Journal of Frontiers in Multidisciplinary Research. 2022. 3(2), 182–192.

I. Jackson, M. J. Saenz, and D. Ivanov, From natural language to simulations: Applying AI to automate simulation modelling of logistics systems, International Journal of Production Research. 2024. 62(4), 1434–1457.

C. Jiang, X. Shu, H. Qian, X. Lu, J. Zhou, A. Zhou, and Y. Yu, LLMOPT: Learning to define and solve general optimization problems from scratch, in Proc. 13th Int. Conf. Learning Representations (ICLR), 2025.

H. O. Jimoh, C. J. Abolle-Okoyeagu, M. O. Ahmed, and N. O. Lawal, Advancing security in IoT-driven critical infrastructure: A focus on smart transportation system, American Journal of Engineering Research. 2023. 12(12), 33–46.

H. O. Jimoh, M. O. Ahmed, and M. O. Fagbade, The role of frameworks in cybersecurity governance, Journal of Behavioural Informatics, Digital Humanities and Development Research. 2023. 9(4), 7–16.

M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, et al., Time-LLM: Time series forecasting by reprogramming large language models, in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.

O. O. Ladapo, A. A. Dosunmu, D. Jooda, and T. O. Abolaji, An IT service management literature review: Challenges, benefits, opportunities, and implementation practices, International Journal of Advanced Multidisciplinary Research and Studies. 2023. 3(6), 2875–2889.

O. O. Ladapo, D. Jooda, A. A. Dosunmu, and T. O. Abolaji, A comprehensive survey on ServiceNow for IT service management, International Journal of Multidisciplinary Research and Growth Evaluation. 2023. 4(6), 1532–1545.

O. O. Ladapo, A. A. Dosunmu, D. Jooda, and T. O. Abolaji, Migration of applications and information systems to cloud computing infrastructure: Lessons from a South African retail bank, International Journal of Multidisciplinary Research and Growth Evaluation. 2025. 6(6), 1361–1375.

O. O. Ladapo, A. A. Dosunmu, D. Jooda, and T. O. Abolaji, Human-in-the-loop machine learning: A state of the art, Journal of Frontiers in Multidisciplinary Research. 2022. 3(1), 656–669.

O. O. Ladapo, D. Jooda, A. A. Dosunmu, and T. O. Abolaji, Keeping humans in the loop: Human-centered automated annotation with generative AI, International Journal of Multidisciplinary Futuristic Development. 2024. 5(1), 81–95.

O. O. Ladapo, D. Jooda, A. A. Dosunmu, and T. O. Abolaji, On the disagreement problem in human-in-the-loop federated machine learning, International Journal of Multidisciplinary Research and Growth Evaluation. 2026. 7(3), 178–192.

O. O. Ladapo, A. A. Dosunmu, D. Jooda, and T. O. Abolaji, Integrated network and security operation center: A systematic analysis, International Journal of Multidisciplinary Futuristic Development. 2024. 5(1), 65–80.

B. Li, K. Mellou, B. Zhang, J. Pathuri, and I. Menache, Large language models for supply chain optimization, arXiv:2307.03875, 2023.

F. Liu, X. Tong, M. Yuan, X. Lin, F. Luo, Z. Wang, Z. Lu, and Q. Zhang, Evolution of heuristics: Towards efficient automatic algorithm design using large language models, in Proc. 41st Int. Conf. Machine Learning (ICML), 2024.

J. J. Medon and T. E. Oduleye, An integrated predictive analytics model for enhancing strategic financial forecasting and decision accuracy, Gyanshauryam, International Scientific Refereed Research Journal. 2024. 7(3), 258–279.

S. B. Nnabueze, J. K. Sakyi, O. M. Filani, J. S. Okojie, O. M. Babatope, and A. Adanyin, Transforming utility and service operations through automation, data-driven analytics, and customer-centric innovation, International Journal of Multidisciplinary Evolutionary Research. 2023. 4(2), 40–57.

S. Nwakamma, J. Ojukwu, and S. O. Oyesiji, LLM-as-a-judge for automated model governance: A review of methods, biases, and release-gating practices, World Journal of Innovation and Modern Technology. 2024. 8(6), 185–216.

S. Nwakamma, J. Ojukwu, and S. O. Oyesiji, On-device multimodal small language models for privacy-preserving edge intelligence through quantization, pruning, distillation, and energy-efficient inference, World Journal of Innovation and Modern Technology. 2025. 9(12), 271–302.

S. Nwakamma, J. Ojukwu, and S. O. Oyesiji, Securing agentic AI enterprise workflows against prompt injection, tool poisoning, memory manipulation, and excessive agency threats, World Journal of Innovation and Modern Technology. 2024. 8(6), 217–248.

K. M. Ogundairo, Hardening the autonomous value chain: Lean Six Sigma guardrails and multi-enterprise ETL architecture for agentic supply chain orchestration, IIARD International Journal of Economics and Business Management. 2025. 11(12), 441–478.

J. Ojukwu, S. O. Oyesiji, and S. Nwakamma, Cyber-physical ransomware defense in operational technology using physics-informed detection, digital twins, network telemetry, and resilient recovery, International Journal of Computer Science and Mathematical Theory. 2023. 9(5), 193–228.

J. S. Okojie, O. M. Filani, P. N. Ike, J. O. Okojokwu-Idu, S. B. Nnabueze, and S. I. Ihwughwavwe, Integrating AI with ESG metrics in smart infrastructure auditing for high-impact urban development projects, International Journal of Multidisciplinary Futuristic Development. 2023. 4(1), 32–44.

P. O. Okoruwa, O. M. Babatope, D. A. Akokodaripon, and O. K. Akinleye, Developing integrated digital platforms for enhancing transparency in procurement and supply chain management, International Journal of Multidisciplinary Research and Growth Evaluation. 2024. 5(6), 1719–1729.

P. O. Okoruwa, O. M. Babatope, D. A. Akokodaripon, and O. K. Akinleye, Digital procurement transformation approaches for strengthening efficiency in global supply chain management, International Journal of Multidisciplinary Research and Growth Evaluation. 2025. 6(5), 975–985.

K. Oluwo, D. Bola-Sadipe, and C. Aliliele, Conceptual framework for data analytics driven financial decision making in banking sector organizations, Shodhshauryam, International Scientific Refereed Research Journal. 2025. 8(5), 259–302.

M. Ominyi, C. C. Anichukwueze, and N. S. Uzougbo, A conceptual framework for multi-agent AI quality control in the review of regulated documents, International Journal of Social Sciences and Management Research. 2025. 11(8), 534–557.
M. Ominyi, C. C. Anichukwueze, and N. S. Uzougbo, Reviewing legal privilege and confidentiality challenges in enterprise deployment of large language models, Journal of Law and Global Policy. 2025. 10(3), 150–176.
M. Ominyi, C. C. Anichukwueze, and N. S. Uzougbo, A systematic review of organizational readiness for the EU Artificial Intelligence Act in multinational enterprises, World Journal of Innovation and Modern Technology. 2024. 8(6), 185–203.
M. Ominyi, C. C. Anichukwueze, and N. S. Uzougbo, Developing a conceptual framework for AI-assisted compliance monitoring and anomaly detection in high-volume transaction environments, World Journal of Innovation and Modern Technology. 2024. 8(6), 204–228.
M. Ominyi, C. C. Anichukwueze, and N. S. Uzougbo, Conceptualizing enterprise AI governance functions as institutional standards for auditable and defensible risk management, Journal of Accounting and Financial Management. 2026. 12(7), 326–356.
M. Ominyi and C. C. Anichukwueze, Conceptualizing responsible AI governance structures for automated decision systems in regulated industries, International Journal of Social Sciences and Management Research. 2024. 10(11), 403–430.
M. Ominyi and C. C. Anichukwueze, A review of algorithmic accountability and model risk management approaches in financial services, Journal of Accounting and Financial Management. 2023. 9(12), 218–238.
T. O. Oshoba, N. I. Hammed, and O. D. Odejobi, Secure identity and access management model for distributed and federated systems, IRE Journals. 2019. 3(4), 550–567.
T. O. Oshoba, N. I. Hammed, and O. D. Odejobi, Blockchain-enabled compliance and audit trail model for cloud configuration management, Journal of Frontiers in Multidisciplinary Research. 2020. 1(1), 193–201.
OWASP Foundation, OWASP Top 10 for Large Language Model Applications, version 2025. [Standard document].
S. O. Oyesiji, S. Nwakamma, and J. Ojukwu, A conceptual framework for context-aware data curation in fine-tuning pipelines, International Journal of Engineering and Modern Technology. 2024. 10(11), 197–228.
S. O. Oyesiji, S. Nwakamma, and J. Ojukwu, Compressing recommender and ranking models for edge deployment: A survey, International Journal of Engineering and Modern Technology. 2025. 11(12), 205–237.
S. O. Oyesiji, S. Nwakamma, and J. Ojukwu, Parameter-efficient multilingual adaptation of on-device language models: A survey, International Journal of Engineering and Modern Technology. 2023. 9(3), 287–320.
M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., The PRISMA 2020 statement: An updated guideline for reporting systematic reviews, BMJ, vol. 372, p. n71, 2021.
M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, et al., A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access, vol. 12. 2024. 26839–26874.
R. Ramamonjison, T. T. Yu, R. Li, H. Li, G. Carenini, B. Ghaddar, et al., NL4Opt competition: Formulating optimization problems based on their natural language descriptions, in Proc. NeurIPS Competition Track, PMLR, vol. 220. 2023. 189–203.
J. K. Sakyi, S. B. Nnabueze, O. M. Filani, J. S. Okojie, and O. M. Babatope, Digital transformation in service delivery leveraging automation and risk reduction for long-term commercial efficiency, International Journal of Advanced Multidisciplinary Research and Studies. 2024. 4(4), 1446–1464.
Z. Song, Y. Xie, L. Yang, and Y. Zhao, Large language models in supply chain management: A systematic literature review and application framework, International Journal of Production Research. 2026. 1–41.
A. Tonoyan, O. Dada, and S. S. Ayivi-Donkor, Conceptual model for predictive cost optimization: Machine learning applications in business transformation analysis, International Journal of Economics and Business Management. 2022. 8(6), 68–102.
A. Tonoyan, O. Dada, and S. S. Ayivi-Donkor, Advances in supply chain resilience: Predictive models for vendor risk assessment and procurement cost optimization, International Journal of Social Sciences and Management Research. 2024. 10(11), 525–551.
A. Tonoyan, O. Dada, and S. S. Ayivi-Donkor, Real-time KPI tracking systems: A review of automated performance monitoring and data-driven decision making, World Journal of Innovation and Modern Technology. 2024. 8(6), 247–281.
A. Tonoyan, O. Dada, and S. S. Ayivi-Donkor, A conceptual model for real-time route optimization: Algorithmic pathways to delivery profitability and logistics cost reduction, World Journal of Innovation and Modern Technology. 2025. 9(12), 308–364.
S. Wang, Y. Zhao, X. Hou, and H. Wang, Large language model supply chain: A research agenda, ACM Transactions on Software Engineering and Methodology. 2025. 34(5), 1–46.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, et al., Chain-of-thought prompting elicits reasoning in large language models, in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2022.

Z. Xiao, D. Zhang, Y. Wu, L. Xu, Y. J. Wang, X. Han, et al., Chain-of-Experts: When LLMs meet complex operations research problems, in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.

Z. Yang, Y. Wang, Y. Huang, Z. Guo, W. Shi, X. Han, et al., OptiBench meets ReSocratic: Measure and improve LLMs for optimization modeling, in Proc. 13th Int. Conf. Learning Representations (ICLR), 2025.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, ReAct: Synergizing reasoning and acting in language models, in Proc. Int. Conf. Learning Representations (ICLR), 2023.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, Tree of Thoughts: Deliberate problem solving with large language models, in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2023.

B. K. Yeboah, O. F. Enow, P. N. Ike, and S. B. Nnabueze, Program design for advanced preventive maintenance in renewable energy systems, Shodhshauryam, International Scientific Refereed Research Journal. 2024. 7(2), 138–156.

B. Zhang, P. Luo, G. Yang, B.-H. Soong, and C. Yuen, OR-LLM-Agent: Automating modeling and solving of operations research optimization problems with reasoning LLM, arXiv:2503.10009, 2025.