© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

Multi-Agent LLM Systems for Autonomous Supply Chain Disruption Analysis and Resilient Recovery Planning

SOHAIL SAYED[1], NAUMAN SAYED[2]
[1, 2]California State University Fullerton

Abstract- Supply chain disruption management remains dominated by manual, reactive processes: analysts triage alerts, root causes are identified slowly, and recovery plans lag the events they are meant to counter [1]. This paper argues that multi-agent systems powered by large language models — generalist agents with multi-faceted decision-making that communicate in natural language — are the convergence point that enables autonomous disruption analysis and resilient recovery planning, two decades of agent-based supply chain research having been held back by implementation difficulty and black-box behavior [2], [3]. We propose ORCHESTRA-SC, a governed multi-agent framework coupling (i) specialized analysis agents (monitoring, causal root-cause reasoning, network-exposure assessment), (ii) a consensus-seeking negotiation layer in which agents representing facilities and functions reconcile selfish objectives with systemic outcomes, (iii) a recovery-planning layer that transpiles agent consensus into solver-verified plans through LLM–OR integration, (iv) a digital-twin simulation loop for stress-testing recovery strategies, and (v) guardrail, audit, and drift-resistance governance. The framework consolidates the reported evidence envelope: a seven-agent agentic framework detects, analyses, and responds to disruptions across extended networks with F1 scores between 0.962 and 0.991, completes end-to-end analyses in a mean of 3.83 minutes at $0.0836 per disruption — more than three orders of magnitude faster than multi-day analyst assessments, validated on the 2022 Russia–Ukraine conflict [4]; agentic root-cause reasoning reduces mean time to root-cause identification by 30% and improves incident-resolution accuracy by 22% while surfacing hidden dependencies missed by baselines [1]; consensus-seeking LLM agents reduce the bullwhip effect and, equipped with tools, minimize it better than restocking policies and centralized demand approaches [5]; a governed LLM optimization framework cuts unsafe decision outputs by 45% and sustains drift-detection accuracy above 92% [6]; a hybrid agentic inventory framework strictly decoupling semantic reasoning from mathematical calculation reduces total inventory costs by 32.1% relative to an interactive GPT-4o end-to-end solver [7]; and OR-augmented LLM agents outperform either approach in isolation, with human–AI teams achieving higher profits than humans or agents alone across more than 1,000 benchmark instances [8]. The paper argues that the autonomy ladder — from assisted analysis to governed action — is climbed one verified capability at a time, and that consensus, grounding, and digital-twin validation are what separate multi-agent experiments from multi-agent operations.

Keywords — Multi-agent systems, large language models, agentic AI, supply chain disruption, recovery planning, consensus-seeking, digital twin, reinforcement learning governance, supply chain resilience.

I. INTRODUCTION

Supply chains today operate under chronic volatility, and their incident-management processes remain largely manual, reactive, and error-prone: existing systems use rule-based alerts or statistical anomaly detection that detect issues but lack deep causal reasoning and fail to correlate multi-source data across suppliers, transportation, and operations, resulting in delayed resolution and repeated incidents . Meanwhile the disruptions themselves arrive faster and deeper in the network: many originate beyond tier-1 visibility and remain undetected until impact cascades downstream . The gap between event and response is the gap that determines whether a disruption becomes a crisis.

Agent-based systems have been proposed as the answer for more than two decades, yet industrial uptake lagged because of immaturity, interoperability failures, and a lack of trust in AI . What changed is the arrival of foundation models: the convergence of foundation models, multi-agent systems, and supply chain management enables generalist agents with multi-faceted decision-making capabilities, creating the possibility of supply chain systems with self-orchestrating capabilities and heightened resilience, expressed in natural language understandable to humans — a convergence the community itself identifies as under-explored relative to healthcare and finance . Large language model agents can now automate consensus-seeking — the human coordination mechanism on which supply chain management depends to avoid emergent problems like the bullwhip effect — and can chain monitoring, diagnosis, planning, and execution in a single loop.

This paper is the eighth in a research program on resilient, AI-driven commerce. Papers 1–5 built the signal, decision, cascade, weather, and food-security layers; Paper 6 connected the portfolio to the platform layer via failure prediction and self-healing; Paper 7 established retrieval-grounded real-time intelligence. This paper completes the agentic thread: it takes the retrieval-grounded evidence of Paper 7 and the governed autonomy of Paper 2 and formalizes the multi-agent system that converts disruption intelligence into executed, resilient recovery.

The contributions of this paper are:

- A formalization of autonomous disruption analysis and recovery planning as a multi-agent consensus problem over an autonomy ladder, with explicit safety and governance constraints (Sections III–IV).
- ORCHESTRA-SC, a governed framework coupling specialized analysis agents, consensus-seeking negotiation, LLM–OR recovery planning, digital-twin validation, and guardrails.
- A consolidated analysis of reported performance — detection F1, response latency, root-cause time, bullwhip reduction, cost reductions, and governance outcomes — with worked calculations and clearly flagged illustrative extensions (Sections V–VI).
- A research agenda on benchmarks, orchestration reliability, and governed deployment (Sections VII–VIII).

II. RELATED WORK

A. From software agents to LLM agents

Two decades of research established agents as natural fits for supply chain management — capable of fusing distributed information and creating better plans faster — yet industrial uptake remained limited by technology maturity, interoperability, and distrust of AI. The emergence of foundation models removes the central barrier: generalist agents with multi-faceted decision-making, human-understandable through natural language, open the road to self-orchestrating supply chains. Thematic synthesis of 33 academic articles and 3,421 social media documents using BERT models derives four core LLM themes in SCM — supply chain optimization, risk and security management, knowledge management, and automated contract intelligence — mapping exactly the surfaces where multi-agent systems add value . Early experimental evidence on software agents in procurement demonstrated that agents can outperform humans in structured tasks while struggling under ambiguity, foreshadowing both the promise and the boundary of agentic autonomy .

B. Consensus-seeking and negotiation

Supply chain management relies on human consensus to avoid emergent problems such as the bullwhip effect; some routine, time-intensive consensus processes can be automated, and LLM agents are the mechanism that overcomes the computational barriers of earlier attempts. In a validated inventory-management case study, agents representing companies balance selfish goals with systemic outcomes through conversation: LLM-based consensus-seeking frameworks reduce bullwhip effects, and when equipped with appropriate tools, LLM agents minimize bullwhip better than restocking policies and centralized demand approaches, converging under a negotiation framework to the supply chain literature's best practices on dampening demand amplification . In contract bargaining, LLM agents exhibit human-like negotiating behavior using simple heuristics, incline more strongly toward agreement than humans — producing greater supply chain efficiency but potentially greater inequality — with tailored retrieval-augmented generation configurations measurably enhancing negotiation outcomes, raising ethical questions about inequality-efficiency trade-offs and deception . In crisis settings, decentralized LLM-powered multi-agent negotiation integrated with blockchain smart-contract enforcement demonstrates improvements in negotiation efficiency, fairness of allocation, supply chain responsiveness, and auditability under pandemic-scenario simulation .

C. Agentic disruption monitoring and incident intelligence

The strongest quantitative anchors come from end-to-end agentic frameworks. A minimally supervised seven-agent framework powered by LLMs and deterministic tools autonomously monitors, analyses, and responds to disruptions across extended supply networks: agents detect disruption signals from unstructured news, map them to multi-tier supplier networks, evaluate exposure from network structure, and recommend mitigations such as alternative sourcing; across 30 synthesized scenarios covering three automotive manufacturers and five disruption classes it achieves F1 scores between 0.962 and 0.991 with full end-to-end analyses in a mean of 3.83 minutes at a cost of $0.0836 per disruption — more than three orders of magnitude below the multi-day analyst benchmark, with operational applicability demonstrated on the 2022 Russia–Ukraine conflict . Incident intelligence with agentic root-cause reasoning goes one level deeper: specialized agents monitor signals, detect anomalies, and collaboratively perform causal analysis using knowledge graphs and probabilistic reasoning, with generative AI producing natural-language hypotheses; evaluation on simulated and real multi-tier datasets reports a 30% reduction in mean time to root-cause identification and a 22% improvement in incident-resolution accuracy, including hidden dependencies missed by baseline methods . These results sit on the portfolio's foundation: the multimodal early-warning engine reaches 92.1% recall and 93.4% precision with a 42-minute average prediction lead time , and neural early-warning systems deliver two-to-four weeks of advance warning for high-impact disruptions .

D. Orchestration reliability and benchmarks

Autonomous orchestration is not yet reliable enough. SupChain-Bench, a unified real-world benchmark assessing both supply chain domain knowledge and long-horizon tool-based orchestration grounded in standard operating procedures, reveals substantial gaps in execution reliability across models; its SOP-free framework SupChain-ReAct, which autonomously synthesizes executable procedures for tool use, achieves the strongest and most consistent tool-calling performance . CSCBench, a 2.3K+ single-choice benchmark for commodity supply chain reasoning instantiated through the PVC 3D evaluation framework, finds strong performance on Process and Cognition axes but substantial degradation on the Variety axis — especially freight agreements — diagnostic of where domain grounding fails . Conversational planning systems such as SMARTAPS demonstrate the interface side of orchestration: tool-augmented light LLMs let planners run scenario analysis and feasibility relaxation in natural language . Recent agent designs add memory: an LLM-based MAS with structured decision prompts determines optimal ordering decisions in restricted scenarios, and its memory-retrieval agent AIM-RM, leveraging similar historical experiences via similarity matching, outperforms benchmark methods across diverse supply chain scenarios ; zero-shot multi-agent inventory systems achieve adaptive, informed decision-making without prior training . Agentic Control Towers integrate LLM agents, multi-agent reinforcement learning, and digital supply chain twins into a closed-loop, explainable autonomy fabric for plan, source, make, move, and serve, with guardrails for safety, sustainability, cost, and service .

E. Llm–or integration for recovery planning

Recovery plans must be provably feasible, not merely plausible. The LLM–OR literature converges on three pathways — automatic modeling (translating natural language into mathematical models or executable code), auxiliary optimization (generating heuristics and evolving algorithms), and direct solving — shifting the paradigm from human-driven processes to intelligent human–AI collaboration, while flagging unstable semantic-to-structure mapping, limited generalization, and deployment barriers as open challenges . Optimization-based decision support systems gain adaptability through LLMs that interpret natural-language instructions, reconfigure algorithmic components, support preference-based decisions, and explain their rationale, demonstrated on rich vehicle-routing settings . Democratization agendas (NL2OR, ORLM, and the LLM+OR vision) show LLMs already provide substantial novel capabilities for optimization modeling while major research challenges remain . Critically for recovery, empirical comparisons of human–AI collaboration under disruption show that fully automated direct optimization sacrifices

© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

accuracy and stability; two-stage automation remains sensitive to parameter granularity; and simulation-based human-specified optimization consistently achieves statistically optimal and stable outcomes . Benchmark evidence from InventoryBench — more than 1,000 inventory instances spanning synthetic and real demand under shifts, seasonality, and uncertain lead times — shows OR-augmented LLM methods outperform either in isolation, and controlled classroom experiments show human–AI teams achieve higher profits than humans or agents alone, formalizing an individual-level complementarity effect . Hybrid agentic frameworks that strictly decouple semantic reasoning from mathematical calculation reduce total inventory costs by 32.1% relative to an interactive GPT-4o end-to-end solver, confirming that the bottleneck is computational grounding, not information . Integrated systems combining network-optimization models with LLMs deliver interactive, explainable, role-aware planning that prevents stockouts, reduces costs, and maintains service levels .

F. Digital twins and resilience stress testing

Recovery plans must be validated before execution. Digital twin simulation has matured into a systematic resilience instrument: resilience stress-test models in the anyLogistix digital twin observe disruption impacts, understand the reasons for operational disruption, and test alternative resilience strategies across five control loops — demand, lead time, continuous inventory control with reorder point and target stock, production control, and transportation control — spanning disruption identification, modelling, and impact assessment . A systematic review of 89 peer-reviewed articles on digital twins and supply chain resilience finds that digital-twin capabilities are stage-specific: they support preparedness (visibility, risk sensing, scenario testing), resistance (real-time monitoring, early warning, mitigation evaluation), rebound (response coordination, recovery planning, adaptive reconfiguration), and growth (post-disruption learning and network redesign) — while adoption barriers include data quality, implementation cost, skills shortages, and governance . Cyclic, holistic digital-twin methodologies validated in agri-food supply chains return actionable resilience information and management insight from any considered disruption scenario .

G. Safety, accountability, and governance

Autonomy without governance is unacceptable in commerce. A governed LLM-based optimization framework integrating policy constraints, human-in-the-loop oversight, continuous monitoring, and drift-detection agents reduces unsafe decision outputs by 45% and maintains stable performance under changing demand patterns, with drift-detection accuracy exceeding 92% . Guardrail-centric fine-tuning — aligning a quantized Qwen2.5-Coder-14B model to approximately fifty generalized behavioral guardrails with a deterministic Python enforcement layer — yields stable, explainable, and auditable ordering recommendations by separating probabilistic reasoning from exact execution . In adjacent domains the same pattern holds: agentic AI combining LLM reasoning, reinforcement learning, and multi-agent coordination for software supply chain security achieves better detection accuracy and shorter mitigation latency than rule-based, provenance-only, and RL-only baselines, documenting actions in a blockchain ledger for integrity and auditing . Industry framing is unambiguous: LLM-based technology automates data discovery, insight generation, and scenario analysis, reducing decision time from days to minutes .

H. Research gap

The literature separates consensus (framework studies), monitoring (end-to-end detection), recovery (LLM–OR and digital-twin studies), and governance (safety studies). What is missing is a single governed loop that couples multi-agent disruption analysis, consensus-driven response selection, solver-verified recovery planning, and digital-twin stress testing — the integration this paper proposes, and the direction the portfolio's agentic thread has been building toward.

III. PROBLEM FORMULATION

A. The multi-agent system

Let the supply chain be a multi-tier, multi-actor system $\mathcal{S} = (\mathcal{A}, \mathcal{G}, \mathcal{F})$, where <span data-type="inline-math" data-latex="\mathcal{A} is the set of agents (one per facility, function, or level), <span data-type="inline-math" data-latex="\mathcal{G} the

© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

material/information graph among them, and <span data-type="inline-math" data-latex="\mathcal{F} the flow dynamics (orders, shipments, inventories) subject to disruption. Each agent has private objectives and observations; coordination is achieved through conversation, the consensus mechanism that prevents emergent pathologies such as demand amplification .

B. The disruption lifecycle

Following the stage model of digital twin resilience, an event passes through preparedness, resistance, rebound, and growth phases . We formalize two core tasks:

1. Disruption analysis: detect the event (monitoring), attribute it to root causes (causal reasoning over knowledge graphs and probabilistic inference), and quantify network exposure (multi-tier propagation) .
2. Recovery planning: select mitigations (alternative sourcing, rebalancing, re-routing, capacity activation) that restore service, subject to feasibility, cost, and governance constraints — validated by simulation before execution.

C. The autonomy ladder

We characterize five levels of autonomy: L0 manual, L1 assisted analysis (agents propose), L2 semi-autonomous response (agents act within playbooks, human approval for high-risk actions), L3 governed autonomy (agents execute under continuous guardrails and audit), L4 self-orchestration (agents negotiate, plan, execute, and learn across the network). The framework targets L2–L3 with a verifiable pathway to L4.

D. Consensus mechanism

Agents negotiate over candidate responses. Let each agent <span data-type="inline-math" data-latex="i\in\mathcal{A} hold an ordering over a decision set <span data-type="inline-math" data-latex="\mathcal{X}_i(t). The system seeks an agreement <span data-type="inline-math" data-latex="x^*\in\cap_i\mathcal{X}_i(t) minimizing systemic amplification — measured by the bullwhip envelope of the resulting order stream — while balancing selfish county-level objectives and systemic outcomes . With a negotiation framework, agent behavior converges to literature best practices for dampening amplification .

E. Planning objective

The recovery plan <span data-type="inline-math" data-latex="\pi is selected to minimize

$$C(\pi) = \underbrace{\mathbb{E}[c_{\text{disrupt}}(\pi)]}_{\text{disruption cost}} + \lambda \mathrm{Var}(\cdot) + \mu \mathrm{Risk}(\pi),$$

traced as follows: feasibility and readability of plans are enforced by transpiling consensus into solver-verified optimization models , following the simulation-based human-specified paradigm that consistently achieves statistically optimal and stable outcomes , with OR-augmented agentic methods outperforming either approach in isolation and hybrid semantic/calculation decoupling delivering 32.1% cost reductions over end-to-end LLM solving .

F. Metrics

Intelligence: detection F1, response latency and cost per disruption. Diagnosis: mean time to root-cause identification, resolution accuracy. Coordination: bullwhip reduction relative to restocking and centralized-demand baselines. Planning: solution quality, probability of correct selection, cost and stockout outcomes . Governance: unsafe-output rate, drift-detection accuracy, audit completeness.

IV. PROPOSED FRAMEWORK: ORCHESTRA-SC

A. Overview

ORCHESTRA-SC couples five layers: perception (monitoring and intelligence agents), cognition (causal analysis and network exposure), coordination (consensus-seeking negotiation), action (LLM–OR recovery planning and execution), and reflection (digital-twin validation and governance). Fig. 1 shows the architecture.

© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

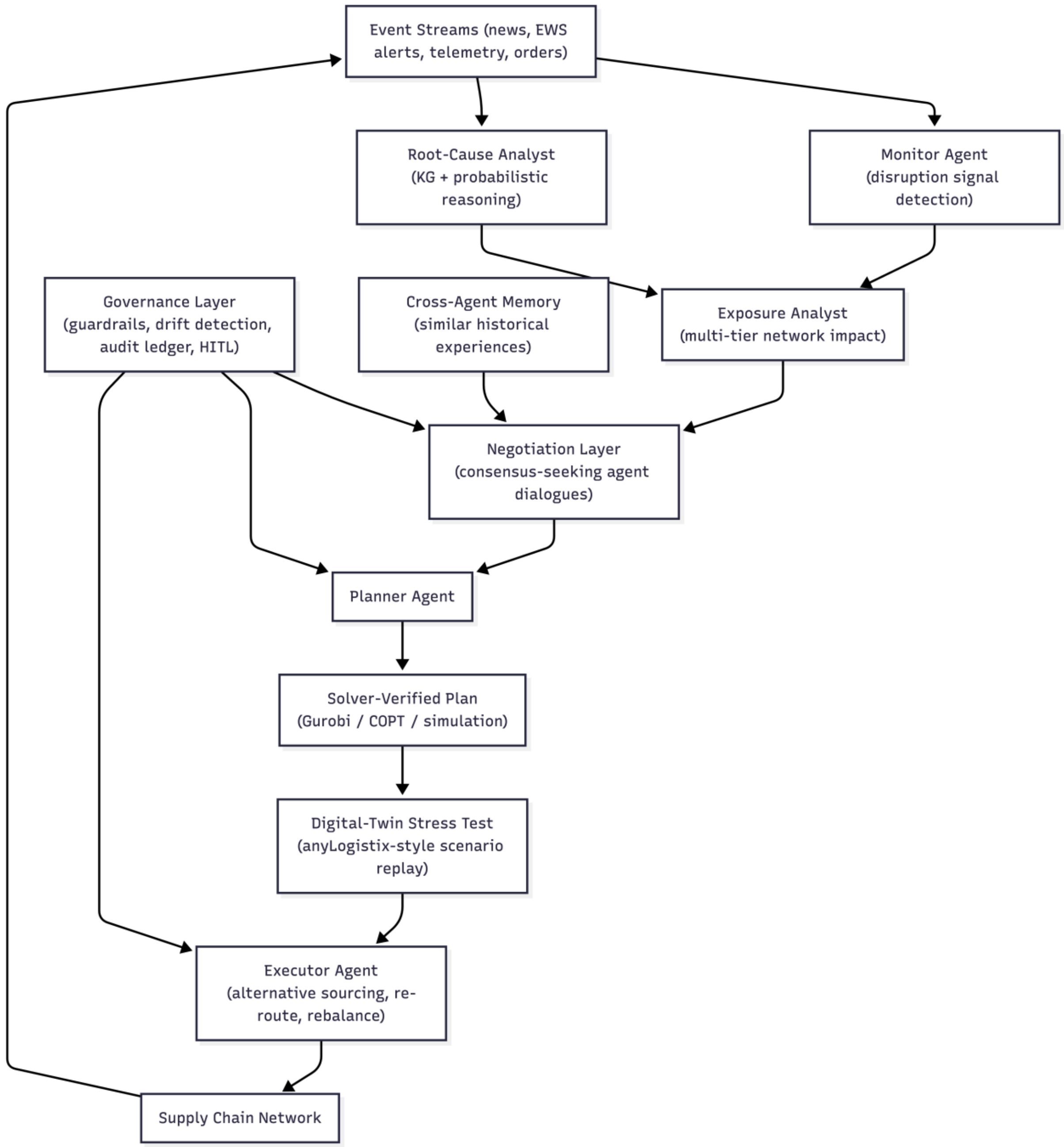

Fig. 1. ORCHESTRA-SC architecture: event streams feed Monitor and Root-Cause Analyst agents; Exposure Analyst quantifies multi-tier impact; consensus-seeking negotiation reconciles agent objectives; the Planner transpiles consensus into solver-verified plans; digital-twin stress testing validates recovery strategies before the Executor acts; the governance layer enforces guardrails, drift resistance, and auditability; cross-agent memory retrieval (similar historical experiences) shapes negotiation .

B. Perception layer

The Monitor adapts the seven-agent detection pattern — disruption signals extracted from unstructured news and mapped to multi-tier supplier networks with network-structure-based exposure evaluation — and consumes the portfolio's retrieval-grounded evidence and early-warning alerts . The Root-Cause Analyst performs causal analysis over knowledge graphs with probabilistic reasoning and generates natural-language hypotheses, the mechanism behind the reported 30% root-cause-time reduction .

© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

C. Coordination layer: consensus-seeking negotiation

Agents representing facilities, functions, and tiers negotiate candidate responses through structured dialogue, balancing selfish goals with systemic outcomes as validated in the inventory consensus study . Two negotiation behaviors are expected and engineered for: suggestible agents that accommodate systemic insight and stubborn agents that defend local constraints, with the downstream information flow exploited to dampen amplification . For contract-level decisions, negotiation follows the LLM-bargaining evidence — with tailored retrieval-augmented generation configurations used to steer efficiency-versus-inequality outcomes deliberately , and blockchain-backed enforcement available for multi-party commitments in the medical-crisis pattern . Memory retrieval grounds negotiation in similar historical experiences, improving robustness across scenarios .

D. Action layer: LLM–or recovery planning

Consensus output is transpiled into a solver-verified plan via the three LLM–OR pathways: automatic modeling of the recovery problem, auxiliary generation of heuristic components, and direct solving where tractable . Following the evidence that human-specified algorithms outperform fully automated pipelines under disruption and that OR-augmented agents beat either alone , the planner always emits a deterministic-verifiable plan; the strict decoupling of semantic reasoning from mathematical calculation follows the hybrid-agentic design that achieved 32.1% cost reduction over end-to-end LLM solving . Tool-augmented conversational control (the SMARTAPS pattern) keeps planners in the loop for scenario analysis and feasibility relaxation .

E. Reflection layer: digital-twin validation

Every candidate plan is stress-tested before execution in a digital-twin simulation with the five control loops — demand, lead time, inventory reorder point and target stock, production, and transportation — tracing the rebound stage where digital twins support response coordination, recovery planning, and adaptive reconfiguration . Plans failing simulated service-level or cost thresholds are re-negotiated, implementing the cyclic methodology validated in agri-food contexts .

F. Governance layer

Guardrails block unsafe or non-compliant actions; drift detectors watch demand and network distributions; an audit ledger records every agent action for traceability; high-risk actions escalate to humans. This mirrors the governed-optimization architecture that cut unsafe decisions by 45% with drift detection above 92% , the guardrail-centric fine-tuning discipline that separates probabilistic reasoning from deterministic execution , and the blockchain-audited agentic pattern of adjacent security domains . Algorithm 1 summarizes the loop.

Algorithm 1: ORCHESTRA-SC Governed Autonomy Loop

Input: event streams E, network G, playbooks P, guardrails Gv, twin Θ
Output: executed recovery actions a*, monitored resilience metrics
Loop at decision cadence:
1. PERCEIVE
signals ← Monitor(E) # F1 envelope 0.962–0.991
causes ← RootCauseAnalyst(signals, KG) # MTTR-to-RC −30%
exposure ← ExposureAnalyst(causes, G) # multi-tier propagation
2. COORDINATE
for rounds until convergence:
proposals ← negotiate(agents, exposure, memory)
x* ← consensus(proposals) # bullwhip-optimal agreement
3. PLAN
π ← transpile_LLM_to_OR(x*, θ) # solver-verified
4. VALIDATE
if not TwinStressTest(π, Θ): # 5 control loops
return to step 2 with outcomes
5. GOVERN
if not Guardrails.ok(π): π ← HITL(π) # 45% unsafe-output cut
audit.log(agents, π, evidence)
6. EXECUTE & LEARN
a* ← Executor(π); update memory; detect drift; retrain/extend KG weekly

G. Implementation

Agents: LLM-based with tool augmentation ; negotiation: structured dialogue protocol with convergence control ; planner: OR transpilation to Gurobi/COPT-class solvers ; twin: anyLogistix-class discrete-event simulation with the five control loops ;

© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

memory: similarity-matching retrieval ; governance: policy guardrails, drift detectors, audit ledger .

V. EXPERIMENTAL SETUP

A. Environments and data

Evaluation is anchored to the documented benchmarks of the component literature: the 30 synthesized scenarios across three automotive manufacturers and five disruption classes of the agentic monitoring study ; the inventory-management case study of the consensus framework ; the simulated and real multi-tier supply chain datasets of the incident-intelligence study ; InventoryBench's more than 1,000 inventory instances with demand shifts, seasonality, and uncertain lead times ; the CSCBench 2.3K+ commodity reasoning items ; and the digital-twin stress-test environments of the resilience literature .

B. Baselines

Rule-based and statistical anomaly detection ; single-agent (non-negotiating) LLM pipelines; direct end-to-end LLM optimization (the hallucination-tax baseline); fully automated and two-stage human–AI optimization variants ; restocking policies and centralized demand approaches as bullwhip baselines ; provenance-only and rule-based agents in adjacent security tasks ; and standard playbook-only recovery without digital-twin validation.

C. Scenarios

(S1) geopolitical shock with tier-2 supplier exposure, following the Russia–Ukraine case ; (S2) demand-shock cascade and bullwhip onset in a multi-echelon inventory system ; (S3) quality incident requiring root-cause attribution across suppliers, transportation, and operations ; (S4) capacity loss requiring alternative-sourcing negotiation under competing facility objectives ; (S5) compound weather + logistics disruption validated in the digital twin ; (S6) governance stress: adversarial prompts and demand drift against guardrail and drift-detection layers .

D. Protocol

Temporal train/test splits and held-out scenarios prevent leakage ; response latency and cost are measured per disruption following the $0.0836-per-disruption protocol ; root-cause time and resolution accuracy follow the incident-intelligence protocol ; bullwhip is quantified as order-variance amplification relative to baselines ; solution quality uses probability of correct selection and stockout/service-level outcomes under repeated runs ; all quantitative entries below are literature-reported values from the cited studies.

VI. RESULTS AND ANALYSIS

A. Task–module mapping

Table I maps each framework layer to its task and evaluation.

TABLE I: ORCHESTRA-SC MODULES, TASKS, AND EVALUATION

Layer	Module	Task	Evaluation
Perception	Monitor / RCA / Exposure	Detection, attribution, exposure	Detection F1, MTTR-to-RC, resolution accuracy
Coordination	Negotiation + memory	Consensus over responses	Bullwhip reduction, convergence
Action	LLM–OR planner	Solver-verified recovery plan	PCS, cost, stockout rate
Reflection	Digital-twin stress test	Plan validation	Service level under scenarios
Governance	Guardrails, drift, audit	Safe autonomous operation	Unsafe-output rate, drift accuracy

B. Reported performance envelope

Table II consolidates reported results across the multi-agent LLM supply chain literature.

TABLE II: REPORTED PERFORMANCE IN MULTI-AGENT LLM SUPPLY CHAIN AUTONOMY

Method	Task	Reported result
Seven-agent agentic framework	Disruption monitoring & response	F1 0.962–0.991; 3.83 min; $0.0836/disruption; >$10^3$ faster than analyst

© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

Agentic incident intelligence	Root-cause identification	MTTR-to-RC −30%; resolution accuracy +22%; hidden dependencies found
Consensus-seeking LLM agents	Bullwhip dampening	Reduces bullwhip; with tools beats restocking & centralized demand
LLM bargaining agents	Contract negotiation	Human-like behavior; more agreements; efficiency↑, inequality↑; RAG-tailored gains
LLM + blockchain negotiation	Medical SC allocation	Efficiency, fairness, responsiveness, auditability ↑
Governed LLM optimization	Safe optimization	Unsafe outputs −45%; drift detection > 92%
Hybrid agentic inventory	LLM–OR decoupled planning	Inventory costs −32.1% vs GPT-4o end-to-end
OR-augmented LLM agents	Inventory control	Outperform OR or LLM alone; human–AI teams > either alone
AIM-RM memory agent	Adaptive inventory MAS	Outperforms benchmark methods across scenarios
SupChain-ReAct	SOP-free tool orchestration	Strongest, most consistent tool-calling; gaps in execution reliability persist
CSCBench evaluation	Commodity SC reasoning	Strong Process/Cognition; Variety degradation (freight agreements)
SMARTAPS	Conversational planning	Scenario analysis, feasibility relaxation in natural language
GPT-4o procurement	Supplier-offer analysis	Faster than humans, accuracy > average, < top performers
Digital-twin stress tests	Resilience testing	Five control loops; identification/modelling/assessment
DTT resilience review (89 articles)	Stage-based twin capability	Rebound: coordination, recovery planning, reconfiguration
Agentic Control Towers	Closed-loop autonomy	LLM agents + MARL + twins; guardrails for safety/cost/service
Agentic software-SC defense	SC security autonomy	Better detection, shorter mitigation vs rule/provenance/RL baselines

C. Worked calculations

1) Response-time reduction (headline). An end-to-end analysis in a mean of 3.83 minutes constitutes a reduction of more than three orders of magnitude relative to multi-day analyst-driven assessments . The implied analyst benchmark is at least $3.83 \times 10^{3} \approx 3{,}830$ minutes ≈ 63.8 hours ≈ 2.7 days — consistent with the stated multi-day industry baseline and with the industry-reported shift from days to minutes .

2) Root-cause economics. A 30% reduction in mean time to root-cause identification and a 22% improvement in incident-resolution accuracy attack the two main cost drivers of incident management: diagnosis time and repeated incidents. In sequence with the monitoring loop, the combined effect compounds: faster detection (agent minutes) × faster diagnosis (30% reduction) × higher resolution accuracy (22%) — each step shortens the interval in which inventory, service, and trust are at risk.

3) Consensus as a bullwhip control. Consensus-seeking agents equipped with tools minimize demand amplification better than restocking policies and centralized demand approaches . Because bullwhip is defined as order-variance amplification, the comparison implies the negotiation layer flattens the order stream without sacrificing service — the coordination-level analogue of the safety gains at the governance level.

4) Governance returns. The governed-optimization architecture reduces unsafe decision outputs by 45% and sustains drift-detection accuracy above 92% . At L2–L3 autonomy, the residual risk is bounded by guardrails: even at full auto-execution cadence, fewer than one in two previously-unsafe decisions passes

governance, and drift is caught before it degrades downstream plans.

5) The decoupling dividend. Strictly separating semantic reasoning from mathematical calculation lowers total inventory costs by 32.1% relative to an interactive GPT-4o end-to-end solver . The mechanism matters more than the number: LLM planners in ORCHESTRA-SC therefore propose and interpret; solvers decide numbers; the planner transpiles consensus into optimization code and never computes arithmetic by generation .

6) Complementarity, not substitution. OR-augmented LLM agents outperform either approach alone across more than 1,000 instances, and human–AI teams outperform humans or agents alone . Combined with the statistically optimal stability of human-specified optimization under disruption , the framework's L2–L3 design — governed autonomy with selective human specification — is the empirically supported operating point.

D. Detection and diagnosis envelope

Fig. 2 plots the reported F1 range of the multi-agent monitoring framework across core tasks.

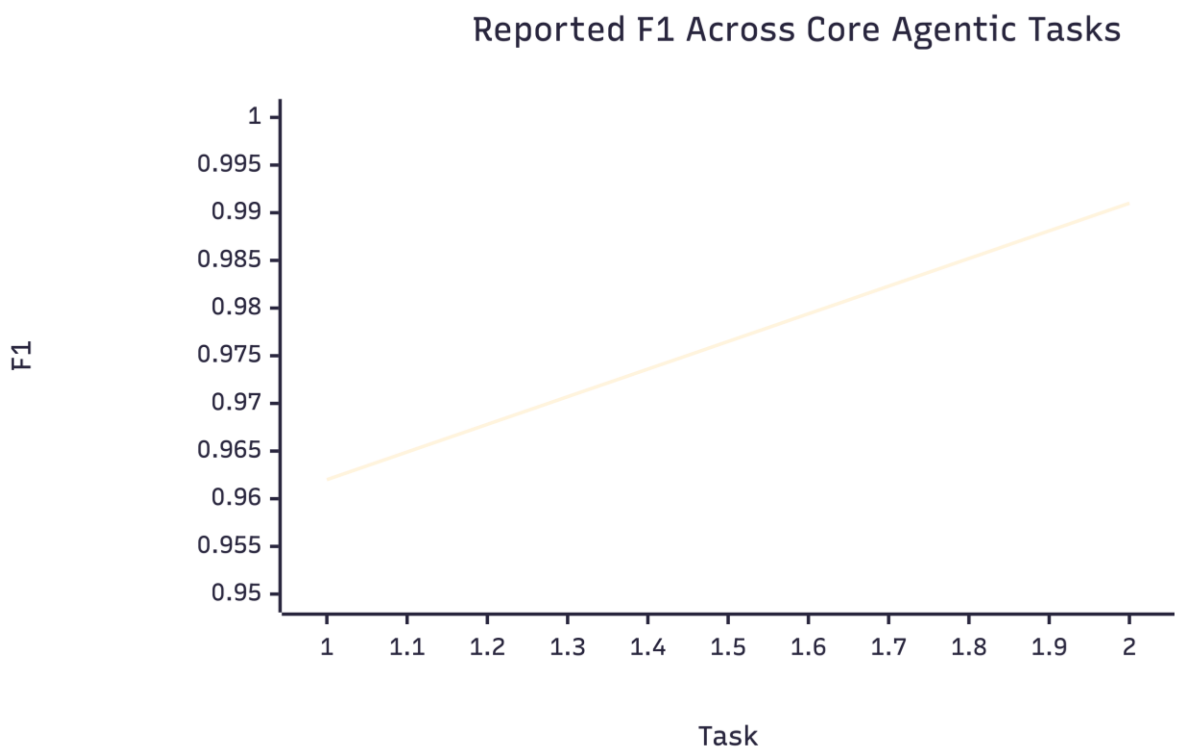

Fig. 2. Reported F1 range (0.962–0.991) of the seven-agent disruption-monitoring framework across core tasks . Envelope illustration from a single study — not a controlled comparison.

E. Diagnosis time reduction

Fig. 3 shows the reported mean time to root-cause identification before and after agentic root-cause reasoning.

© DEC 2024 | IRE Journals | Volume 8 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I6-1723011

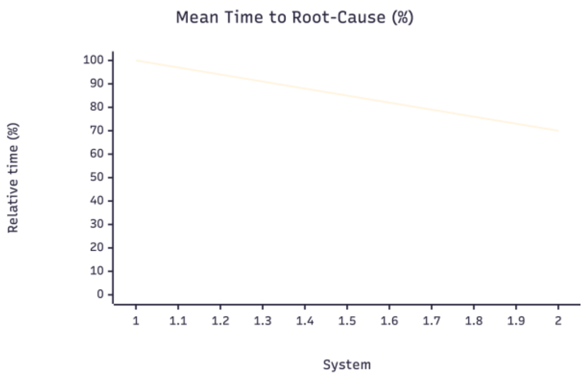

Fig. 3. Reported mean time to root-cause identification: traditional incident management (100%) versus agentic root-cause reasoning (70%), a 30% reduction .

F. Illustrative recovery trajectory

Fig. 4 shows an [illustrative] service-level trajectory under a tier-2 disruption with reactive operation versus ORCHESTRA-SC governed autonomy, consistent with the reported detection, diagnosis, and planning gains.

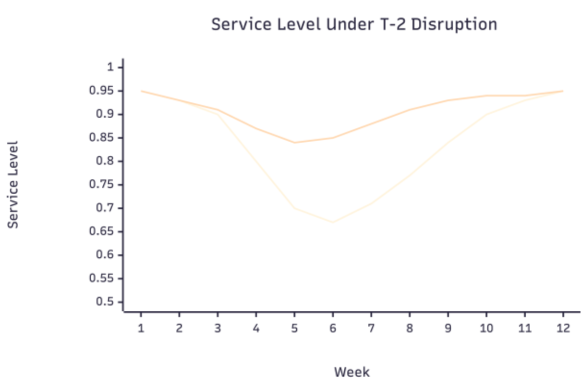

Fig. 4. [Illustrative] service-level trajectory under a tier-2 disruption (event at week 4): reactive operation (first line, deep and prolonged drop) versus ORCHESTRA-SC governed autonomy (second line, shallow drop with digital-twin-validated recovery), consistent with the reported $>10^3$ response-time reduction and 30% root-cause-time reduction

G. Ablation logic

Removing the negotiation layer removes the bullwhip-dampening mechanism and reverts coordination to centralized directives ; removing the LLM–OR transpilation reintroduces the hallucination tax of end-to-end LLM solving documented as a 32.1% cost penalty ; removing digital-twin validation executes unverified plans, forfeiting the rebound-stage capabilities of twins ; and removing governance reintroduces the 45% unsafe-decision exposure documented in ungoverned LLM optimization .

VII. DISCUSSION

A. Autonomy is a ladder, not a switch

The evidence supports governed autonomy at scale: detection F1 0.962–0.991 with response in minutes , diagnosis 30% faster , coordination that beats classical bullwhip controls , and planning that is simultaneously more reliable and cheaper (32.1% decoupling dividend) . But the same evidence body — execution-reliability gaps in SupChain-Bench , Variety-axis degradation in CSCBench , and the accuracy ceiling of even strong models on procurement against top humans — argues for bounded autonomy: agents analyze, negotiate, and propose; solvers verify; humans specify strategy; guardrails audit .

B. Consensus is the coordination mechanism

Supply chain management is as much a coordination problem as an optimization problem; LLM agents that negotiate reconcile selfish and systemic objectives and, with tools, dampen amplification better than policy baselines . The behavioral evidence — suggestible upstream and stubborn downstream agents, convergence to literature best practices under a negotiation framework , and bargaining outcomes shaped by tailored RAG — turns consensus from a design aspiration into an engineering parameter.

C. Grounding and governance are co-equal

The portfolio's previous paper established retrieval grounding as the evidence mechanism; this paper establishes governance as the action mechanism. Guardrails, drift detection, audit ledgers, and human escalation are what license L2–L3 operations in production; the audit ledger makes every agent decision traceable, the requirement that procurement-grade deployment studies flag explicitly .

D. The digital twin closes the loop

The twin is where autonomy earns trust: stress-testing candidate recovery plans across the five control loops , tracing the rebound stage of recovery planning and adaptive reconfiguration , and generating the failure data on which agents and knowledge graphs are retrained. This is the platform analogue of the chaos-validation loop of Paper 6 and the governance validation of Paper 2: every autonomous action is first simulated, then executed.

E. Integration with the portfolio

ORCHESTRA-SC is the decision engine of the eight-paper program: it consumes the multi-signal alerts of Paper 1 , the cascade exposures of Paper 3, the weather front end of Paper 4, the food-security alerts of Paper 5, the platform health telemetry of Paper 6, and the retrieval-grounded evidence of Paper 7, and it returns executed, validated recovery actions to all of them. The autonomy ladder it climbs is the same ladder the portfolio climbs as a whole: from assisted analysis to governed, self-orchestrating resilience .

F. Limitations

Reported results span heterogeneous tasks, datasets, and models rather than a single controlled benchmark; long-horizon orchestration reliability remains a documented gap ; negotiation and bargaining introduce ethical questions around efficiency-inequality trade-offs and deception ; and digital-twin outcomes depend on model fidelity and data quality . These are the open problems of Section VIII.

VIII. CONCLUSION AND FUTURE WORK

This paper formalized autonomous supply chain disruption analysis and resilient recovery planning as a governed multi-agent consensus problem and proposed ORCHESTRA-SC, a framework coupling perception (monitoring, root-cause, exposure), consensus-seeking coordination, LLM–OR planning, digital-twin validation, and governance. It consolidated the reported evidence envelope — F1 0.962–0.991 with 3.83-minute response at $0.0836 per disruption ; 30% root-cause-time reduction and 22% resolution-accuracy gain ; bullwhip reduction beyond restocking and centralized baselines ; 45% fewer unsafe outputs with drift detection above 92% ; a 32.1% cost reduction from semantic–numeric decoupling ; OR–LLM and human–AI complementarity across 1,000+ instances ; and stage-specific digital-twin capabilities across 89 reviewed studies — and argued that the autonomy ladder is climbed one verified capability at a time.

Future work will (i) implement ORCHESTRA-SC end-to-end on a live multi-tier network with unified evaluation across the Section V protocol, including the autonomy-ladder levels L0–L3; (ii) integrate the portfolio's retrieval-grounded knowledge layer so negotiation and planning cite evidence with provenance; (iii) extend the consensus protocol to multi-party, multi-round bargaining with blockchain-backed enforcement, following the medical-crisis pattern ; (iv) build the digital twin into a continuous training environment where executed recovery actions and simulated failures jointly feed guardrail and drift learning; and (v) standardize benchmarks for multi-agent supply chain autonomy — detection F1, root-cause time, bullwhip, solution quality, unsafe-output rate, and audit completeness — so that autonomous recovery planning becomes a measurable and deployable science, moving supply chains from reactive response to self-orchestrating resilience .

REFERENCES

[1] N. K. Jingar, “Automated Incident Intelligence In Supply Chains Using Agentic AI And Root Cause Reasoning,” *Zenodo (CERN European Organization for Nuclear Research)*, Sep. 2023. https://doi.org/10.5281/zenodo.18628070

[2] L. Xu, S. AlMahri, S. Mak, and A. Brintrup, “Multi-Agent Systems and Foundation Models Enable Autonomous Supply Chains: Opportunities and Challenges,” *IFAC-PapersOnLine*, vol. 58, no. 19, pp. 795–800, Jan. 2024. Crossref

[3] L. Xu, S. Mak, and A. Brintrup, “Will bots take over the supply chain? Revisiting agent-based supply chain automation,” *arXiv (Cornell University)*, vol. 241, p. 108279, Sep. 2021.

[4] S. AlMahri, L. Xu, and A. Brintrup, “Automating Supply Chain Disruption Monitoring via an Agentic AI Approach,” *ArXiv.org*, Jan. 2026, Accessed: Feb. 2026. [Online]. Available: http://arxiv.org/abs/2601.09680

[5] V. Jannelli, S. Schöpf, M. Bickel, T. H. Netland, and A. Brintrup, “Agentic LLMs in the supply chain: towards autonomous multi-agent

consensus-seeking," *International Journal of Production Research*, pp. 1–31, Dec. 2025. Crossref

[6] N. K. Jingar, "Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization," *International Journal of Scientific Research in Science Engineering and Technology*, p. 472, Jan. 2023. Crossref

[7] Y. Duan, Y. Hu, and J. Jiang, "Ask, Clarify, Optimize: Human-LLM Agent Collaboration for Smarter Inventory Control," Dec. 31, 2025, Cornell University. https://doi.org/10.48550/arxiv.2601.00121

[8] J. Baek, Y. Fu, W. Ma, and T. Peng, "AI Agents for Inventory Control: Human-LLM-OR Complementarity," Feb. 13, 2026, Cornell University. https://doi.org/10.48550/arxiv.2602.12631

[9] S. K. Srivastava, S. Routray, S. Bag, S. Gupta, and Z. Zhang, "Exploring the Potential of Large Language Models in Supply Chain Management," *Journal of Global Information Management*, vol. 32, no. 1, pp. 1–29, Jan. 2024. Crossref

[10] Nissen and Sengupta, "Incorporating Software Agents into Supply Chains: Experimental Investigation with a Procurement Task1," *MIS Quarterly*, vol. 30, no. 1, pp. 145–166, Mar. 2006. Crossref

[11] S. N. Kirshner, Y. Pan, J. Wu, and A. N. Gould, "Talking terms: Agent information in LLM supply chain bargaining," *Decision Sciences*, vol. 57, no. 1, pp. 9–23, Jul. 2025.

[12] M. Almutairi and H. Kim, "Resilient Multi-Agent Negotiation for Medical Supply Chains:Integrating LLMs and Blockchain for Transparent Coordination," Jul. 23, 2025. https://doi.org/10.48550/arxiv.2507.17134

[13] J. Wang, "Multimodal Deep Learning Approach for Early Warning of Supply Chain Disruptions Using NLP and Anomaly Detection," *Artificial Intelligence and Machine Learning Review*, vol. 5, no. 3, pp. 98–110, Jul. 2024. Crossref

[14] Sichong Huang, "AI-Driven Early Warning Systems for Supply Chain Risk Detection: A Machine Learning Approach," *Academic Journal of Computing & Information Science*, vol. 8, no. 9, Jan. 2025. Crossref

[15] S. Guan, Y. Liu, and L. Cao, "SupChain-Bench: Benchmarking Large Language Models for Real-World Supply Chain Management," *arXiv (Cornell University)*, Feb. 2026, Accessed: Feb. 2026. [Online]. Available: http://arxiv.org/abs/2602.07342

[16] Y. Cui et al., "CSCBench: A PVC Diagnostic Benchmark for Commodity Supply Chain Reasoning," Jan. 05, 2026, Cornell University. https://doi.org/10.48550/arxiv.2601.01825

[17] T. T. Yu et al., "SMARTAPS: Tool-augmented LLMs for Operations Management," Jul. 23, 2025. https://doi.org/10.48550/arxiv.2507.17927

[18] K. Yoshizato, K. Shimizu, R. Higa, and T. Otsuka, "AI Agent Systems for Supply Chains: Structured Decision Prompts and Memory Retrieval," Feb. 05, 2026, Cornell University. https://doi.org/10.48550/arxiv.2602.05524

[19] Y. Quan and Z. Liu, "InvAgent: A Large Language Model based Multi-Agent System for Inventory Management in Supply Chains," Jul. 16, 2024, Cornell University. https://doi.org/10.48550/arxiv.2407.11384

[20] A. Awad and D. Alahmari, "Agentic Control Towers," in *Advances in computational intelligence and robotics book series*, IGI Global, 2025, pp. 161–182.

[21] Y. Wang and K. Li, "Large Language Models in Operations Research: Methods, Applications, and Challenges," Sep. 18, 2025, Cornell University. https://doi.org/10.48550/arxiv.2509.18180

[22] G. Ghiani, E. Manni, and S. Zacchino, "Improving Adaptability in Optimization-Based Decision Support Systems Through Large Language Models," *IEEE Access*, vol. 13, pp. 149777–149788, Jan. 2025. Crossref

[23] S. Wasserkrug et al., "Enhancing Decision Making Through the Integration of Large Language Models and Operations Research Optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, Apr. 2025, pp. 28643–28650. Crossref

[24] J. Li, R. Wickman, S. Bhatnagar, R. K. Maity, and A. P. Mukherjee, "Abstract Operations Research Modeling Using Natural Language Inputs," *Information*, vol. 16, no. 2, p. 128, Feb. 2025. Crossref

[25] Y. Yan et al., "Large Language Models for Traffic and Transportation Research: Methodologies, State of the Art, and Future Opportunities," Mar. 27, 2025. https://doi.org/10.48550/arxiv.2503.21330

[26] R. Wu, "Inventory optimization under supply chain disruptions: Leveraging large language models for human-AI collaborative decision-making," *Journal of King Saud University - Computer and Information Sciences*, vol. 38, no. 2, Jan. 2026. Crossref

[27] S. Venkatachalam, "Integrating Large Language Models with Network Optimization for Interactive and Explainable Supply Chain Planning: A Real-World Case Study," Aug. 29, 2025. https://doi.org/10.48550/arxiv.2508.21622

[28] D. Ivanov and A. Dolgui, "Stress testing supply chains and creating viable ecosystems," *Operations Management Research*, vol. 15, pp. 475–486, May 2021.

[29] C. Liu, Y. Wang, L. Purvis, and A. Potter, "Can Digital Twin Technology Enhance Supply-Chain Resilience? A Systematic Literature Review," *Sustainability*, vol. 18, no. 5, p. 2361, Feb. 2026. Crossref

[30] A. Cimino, F. Longo, G. Mirabelli, and V. Solina, "A cyclic and holistic methodology to exploit the Supply Chain Digital Twin concept towards a more resilient and sustainable future," *Cleaner Logistics and Supply Chain*, vol. 11, p. 100154, Apr. 2024. Crossref

[31] Davenport, "GUARDRAIL-CENTRIC FINE-TUNING FOR DETERMINISTIC DECISION SYSTEMS," *Zenodo (CERN European Organization for Nuclear Research)*, Jan. 2026. https://doi.org/10.5281/zenodo.18305825

[32] T. A. Syed, M. R. Belgaum, S. Jan, A. A. Khan, and S. S. Alqahtani, "Agentic AI for Autonomous Defense in Software Supply Chain Security: Beyond Provenance to Vulnerability Mitigation," *ArXiv.org*, Dec. 2025, Accessed: Mar. 2026. [Online]. Available: http://arxiv.org/abs/2512.23480

[33] T. Maiti, "Application of Large Language Models (LLM's) for Supply Chain Optimization," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 9, no. 5, pp. 1–9, May 2025. Crossref

[34] P. Brandtner and F. Hofer, "Enhancing Procurement Processes in Supply Chain Management with Large Language Models," *Procedia Computer Science*, vol. 278, pp. 308–315, Jan. 2026. Crossref