systems

MDPI

Article

The Role of Large Language Models in Supply-Chain Decisions: A Systematic Review and Research Agenda

Guanghui Cheng and Yong He *

School of Economics and Management, Southeast University, Nanjing 211189, China; guanghui@seu.edu.cn
* Correspondence: hy@seu.edu.cn

Highlights

Please indicate how your work links to systems science via your contributions to systems practice, theory, and/or methodology.

- Frames each LLM application as a component of a supply-chain decision system, defined by the decision it supports and the procedure that receives its output.
- Uses a TCCM synthesis of 81 studies to compare applications by decision, LLM task, method integration, and evaluation benchmark, linking systems practice with evaluation methodology.

What are the main findings and/or the implications of the main findings?

- LLM output enters supply-chain decisions in three ways: as evidence for human or analytical review, as the formulation of the model a solver then optimizes, or directly as negotiation offers and replenishment orders.
- Evaluation should test whether generated models preserve operating constraints and how LLM-based negotiation distributes gains between partners.

Abstract

Large language models (LLMs) are increasingly applied to supply-chain problems that depend on business documents and natural-language interaction. This review examines how these models support different supply-chain decisions and how they work alongside established analytical methods. Combining bibliometric analysis with a theory, context, characteristics, and methodology (TCCM) synthesis, we examine 81 articles identified through searches covering 40 AJG-ranked operations research and operations management journals, completed in November 2026. In supplier assessment and risk monitoring, LLMs mainly interpret documentary evidence and prepare structured inputs for analysis. In planning, they translate operating requirements into optimization or simulation models and explain the solutions those models compute. In negotiation and replenishment, the generated responses enter the operating choice itself. Most applications rely on prompting coupled with external solvers or evaluators. Alongside these implementations, organisational and analytical studies examine how expected model capabilities shape firm adoption and supply-chain contracts. LLMs therefore tend to complement established decision methods rather than replace them, and both their value and their risks depend on the specific choice they support. We then set five directions: semantic errors in operating decisions, prompts and retrieved histories, explanations and human judgement, information available at the decision date, and shared model development and maintenance.

Check for updates

Academic Editor: Polinpapilinho Katina

Received: 23 October 2026
Revised: 11 November 2026
Accepted: 13 November 2026
Published: 17 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: large language models; supply-chain management; decision support; operations research; systematic literature review

1. Introduction

Supply-chain decisions routinely depend on unstructured text that must be interpreted before any formal analysis can begin. Commercial contracts set out the obligations between trading partners, and a disruption report only acquires operational meaning once it is read against a firm's open orders and capacity commitments. Operations research connects demand estimates and operating parameters to replenishment or capacity decisions [1,2], and machine learning supports demand prediction and risk detection [3]. Neither, however, removes the need to translate prose statements into structured inputs for the decision at hand. Large language models (LLMs) can carry out much of this translation through natural-language interaction, and studies of knowledge work show that they substantially reduce the time and effort required for text-intensive tasks [4,5].

In supply-chain operations, however, the effect of such task-level gains on actual decisions depends on what the model's output is asked to do. Generating code from a process description helps build a logistics simulation [6]. Generating pairwise supplier comparisons, by contrast, feeds directly into a selection score [7]. The same split appears even within a single supply-chain function. In procurement, an LLM may summarize bid documents for a buyer to review, or it may make price offers directly to a supplier [8]. The first output prepares evidence for human judgement; the second acts in the transaction itself. Evaluating an application therefore means tracing the model's output into the downstream decision procedure. Agreement with an expert rater and profit earned at the bargaining table measure fundamentally different outcomes.

Prior reviews offer an important starting point for this analysis. Earlier AI reviews map computational techniques to supply-chain domains and outline research gaps [9–11]. Implementation roadmaps and capability frameworks instead examine how generative models fit into organisational workflows [12,13]. More recently, several reviews have focused directly on LLMs and generative AI in supply chains. Song et al. [14] link LLM capabilities to supply-chain processes through a theory-informed framework, Panja et al. [15] compare generative-AI technologies across logistics and supply-chain functions, and Bahroun et al. [16] survey technical implementations and reported outcomes by process area, noting that most studies remain simulation-based prototypes with limited field validation. Building on these process- and capability-level mappings, the present review examines how LLMs operate inside specific supply-chain decision problems. It connects the language task assigned to the model, the analytical method that processes its output, and the operational or strategic outcomes against which it is judged. Table 1 summarizes these related reviews and frameworks.

To conduct this comparison, we review 81 articles retrieved from 40 AJG-ranked operations research and operations management journals. Our full-text synthesis tracks how each implemented application feeds the model's output into a concrete evaluation or optimization procedure, and pairs this technical evidence with empirical studies of organisational adoption and game-theoretic models of expected model capabilities. Organising the review through the theory, context, characteristics and methodology (TCCM) framework [17] anchors these comparisons in the specific research question and operating setting of each study. Three research questions guide the analysis:

RQ1: Which supply-chain decisions are studied, and what tasks are assigned to LLMs?

RQ2: How are LLMs combined with existing decision methods, and how do these arrangements vary across decision problems?

RQ3: What improvements and unresolved problems are reported when LLMs are used in these decisions?

Table 1. Related reviews and conceptual frameworks.

Study	Literature Period	Research Question	Main Finding
AI reviews			
Min (2010) [9]	Not reported	AI for which decisions?	Decision-aid uses and research needs.
Toorajipour et al. (2021) [10]	2008–2018	Which techniques fit SCM?	Techniques mapped to supply-chain fields.
Sharma et al. (2022) [11]	1994–2021	Where is AI research concentrated?	Application clusters and research gaps.
Frameworks and research roadmaps			
Richey et al. (2023) [12]	Not reported	What enables AI adoption?	Roadmap for organisational implementation.
Jackson et al. (2024) [13]	Not reported	Which GenAI capabilities fit operations?	Capabilities mapped to operating functions.
LLM and generative-AI reviews			
Song et al. (2026) [14]	To April 2025	How do LLMs fit SCOR?	Theory-informed application pathways.
Panja et al. (2026) [15]	To December 2024	Where are GenAI models used?	Forecasting focus; scarce field validation.
Bahroun et al. (2026) [16]	To June 2025	How is GenAI implemented?	Mostly prototypes; limited system-wide KPIs.
Present review	To November 2026	How do LLMs support supply-chain decision-making?	LLMs complement established decision methods; their roles and value vary across decision problems.

Note. SCOR: Supply-Chain Operations Reference model.

By linking language-model tasks to the decision methods and outcomes that evaluate them, this review makes three main contributions. First, it shows that LLMs generally complement established analytical procedures rather than replace them. They convert unstructured records into decision inputs, or they help planners formulate and interpret mathematical models. A direct-action role appears mainly in negotiation and ordering. Second, by bringing technical implementations together with organisational adoption and channel-contract studies, it connects what an LLM workflow can demonstrably achieve in experiments with the capability assumptions that drive firm investment and contracting decisions. Third, it develops a decision-oriented research agenda. The agenda asks how semantic errors propagate into operating commitments, how shared prompt and retrieval histories shift bargaining surplus, and how firms can govern and maintain shared models over time.

The remainder of the paper is organised as follows. Section 2 sets out the conceptual background, and Section 3 details the search, screening and coding protocol. Sections 4 and 5 present the bibliometric profile and the TCCM synthesis. Section 6 sets out the future research agenda, and Section 7 concludes.

2. Conceptual Background

2.1. Supply-Chain Decisions

Because supply-chain decisions coordinate activities across firm boundaries, their outcomes depend on commitments made elsewhere in the network [18]. A plant's production schedule relies on upstream component deliveries as well as downstream demand. Sharing demand information changes replenishment rules and safety stock across tiers [19], and contract terms determine whether a buyer's order quantity aligns with its supplier's incentives [20]. These decisions also span very different time horizons and degrees of reversibility (Figure 1). Qualifying a supplier may lock in a multi-year partnership, but replenishment orders under that contract can be revised daily.

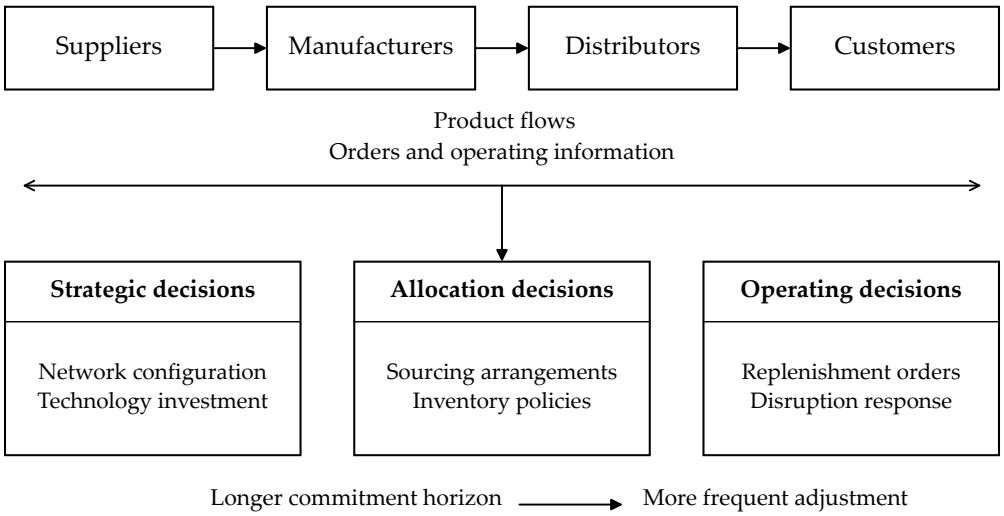


Figure 1. Supply-chain relationships and decision horizons.

In this review, a decision problem is defined by the choice facing a manager and the operating constraints under which that choice is made. Those conditions also determine the benchmark against which an LLM application should be judged. Agreement with expert rankings tests a supplier-scoring tool; on-time delivery tests the sourcing arrangement that follows. Consistent with prescriptive analytics, which evaluates information through the quality of the action it supports [2], an LLM's role depends on whether it prepares evidence for a human decision maker, formulates the analytical problem solved by another method, or proposes the operating action directly.

2.2. Large Language Models and Their Use in Business Tasks

Generative language models produce text or executable code in response to user instructions and supplied context. In-context examples can steer task behavior without updating model weights [21], and instruction tuning aligns responses more closely with user intent [22]. Alongside autoregressive generators, pretrained encoders such as BERT build contextual representations for text classification or variable construction [23]. These two uses assign language modeling very different roles. An encoder that extracts an econometric risk index from corporate filings does something quite different from a generative model prompted to negotiate a contract or recommend an order quantity.

The surrounding system architecture determines what information a model can access and how its output is used. Retrieval-augmented generation (RAG) grounds model responses in passages retrieved from an external document store [24], so that an extracted supply link can be traced back to the corporate disclosure that reported it [25], and operational staff can query internal training manuals [26]. Where numerical precision is

needed, tool integration hands calculations or optimization routines off to an external solver, and agentic workflows iterate this process by selecting subsequent actions based on feedback from the environment [27]. In practice, these arrangements are frequently combined (Table 2), as when retrieved domain examples guide a model in writing database queries or formulating solver code.

Table 2. LLM arrangements in supply-chain applications.

Arrangement	How the Model Is Used	Implication for a Supply-Chain Task
Prompting [7,21]	Instructions and examples specify the requested response without changing model parameters.	A supplier assessment can depend on how requirements and scoring instructions are expressed.
Fine-tuning [28,29]	Additional training adapts model parameters to a task or domain.	Performance depends on whether the training material represents the documents and cases encountered in use.
Retrieval-augmented generation [24,25]	Retrieved material supplies external context for the generated response.	Supplier or disruption information can be connected to a source document and updated without retraining the model.
Tool integration [30]	The model produces inputs for an external program and interprets its results.	A solver calculates the plan; the LLM translates the user's requirements into the problem to be solved.
LLM agents [8,27]	Model outputs guide successive actions in an interactive setting.	An offer made in negotiation changes the information and choices available in subsequent exchanges.

2.3. LLMs in Relation to Established Decision Methods

Most supply-chain LLM applications plug into analytical methods already established in digital supply-chain modeling and profit-oriented optimization [31,32]. Depending on the workflow, the language model may write simulation code from a text specification [6], provide pairwise criteria weights within a structured supplier evaluation [7], explain inventory analytics to managers in natural language [33], or translate a planner's verbal request into solver constraints and summarize the resulting schedule [30]. Table 3 separates the language task handled by the LLM from the calculation performed by the host method.

Table 3. LLM tasks within established decision methods.

Decision Method	Purpose of the Method	LLM Contribution to Examine
Optimization [30,34]	Select an action satisfying operating constraints under a specified objective.	Formulating the problem, supplying inputs or interpreting a solution.
Simulation [6,35]	Estimate system behavior under alternative operating policies or scenarios.	Constructing executable models or translating scenario descriptions into model changes.
Multicriteria evaluation [7,36]	Compare alternatives using explicit evaluation requirements.	Extracting supplier attributes, proposing criteria or supplying comparative judgements.

Table 3. *Cont.*

Decision Method	Purpose of the Method	LLM Contribution to Examine
Strategic interaction [8,37]	Examine choices whose consequences depend on another participant's response.	Making offers or modeling how expected LLM benefits affect adoption and contractual incentives.
Predictive or empirical analysis [38,39]	Relate observed information to an outcome of interest.	Producing a text representation or measure that enters a subsequent analytical model.

How an application should be evaluated follows directly from this division of labor. When a model extracts facts or generates code, verification can check the extracted statements against source documents and test whether the code compiles and runs. A separate validation step is still needed to confirm that the code captures the intended operating rules. When the model proposes an action directly, evaluation must also account for the wider trading relationship. Achieving a lower purchase price in a single negotiation is distinct from maintaining a supplier's willingness to trade over time [40], and the historical examples retrieved into a prompt can shift how total surplus is divided between buyer and supplier [8]. In game-theoretic adoption models, by contrast, equilibrium investment and contract choices follow from the capability gains assumed at the outset [41].

3. Review Methodology

3.1. Review Design and Search Strategy

Following systematic study selection, we combine bibliometric mapping with a TCCM-based full-text synthesis [17,42–44], reported in accordance with PRISMA 2020 and PRISMA-S guidelines [45,46]. The bibliometric stage tracks publication growth, citation links, keyword co-occurrence, and title-and-abstract topic clusters. The TCCM synthesis then compares how individual studies design, integrate, and evaluate LLMs across specific supply-chain decisions (Figure 2).

We searched Web of Science Core Collection and Scopus on 5 November 2026 across all 40 journals rated 3, 4, or 4* in the OR&MANSCI (27 journals) and OPS&TECH (13 journals) fields of the AJG 2024 guide. Search strings combined language-model terms with supply-chain terms and journal ISSNs for articles published between 2019 and 2026 (Table 4; full queries and ISSNs are listed in Section S1 of the Supplementary Materials). Functional terms such as procurement, inventory, and scheduling were included so that the search would also capture domain studies that did not use the phrase "supply chain" in their title or abstract. Broader generative-AI terms were included for the same reason, since organisational and analytical papers often frame their inquiry at that level; final inclusion was determined by full-text inspection of the technology and decision problem.

To complement the database search, we conducted forward and backward citation tracking in OpenAlex on 5–6 November 2026, starting from the DOIs of the 76 provisionally eligible database articles (78 at final reassessment) and nine related reviews. Eligible primary studies were traced in both directions. Reviews were used for backward reference tracking only. After removing seed records and duplicates, this process yielded 4589 backward-traced records and 1270 unique forward-traced records (with 80 records appearing in both directions, for a combined pool of 5779 unique candidates). Applying the same journal, year, article-type, and topic filters retained 37 candidates for screening, and a second snowball iteration on newly identified papers returned no further eligible records.

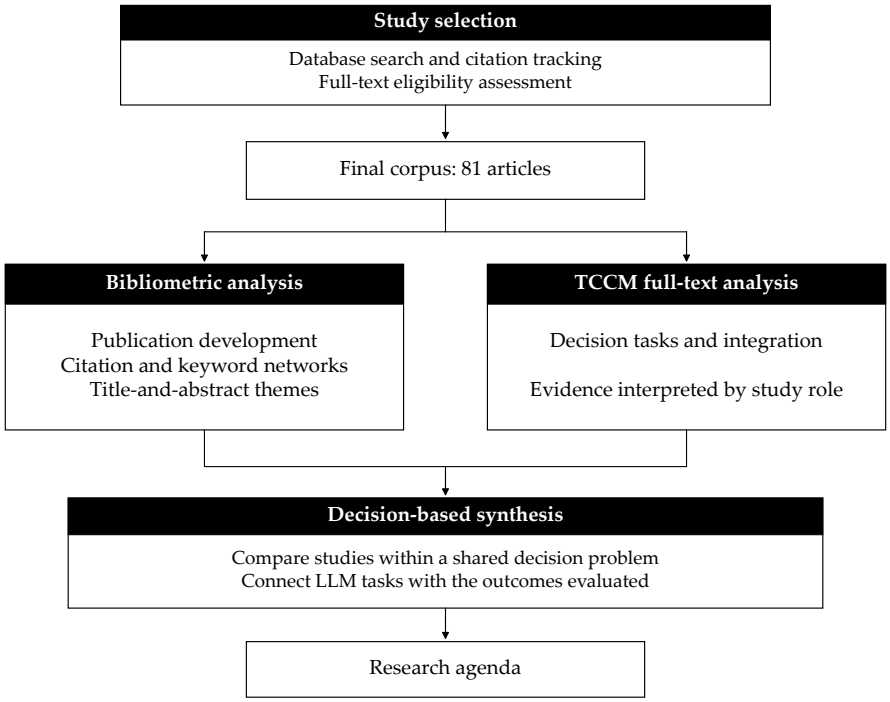


Figure 2. Study selection and analysis.

Table 4. Search terms and database implementation.

Component	Specification
Language models	"large language model"; "large language models"; LLM; LLMs; "LLM-based"; "LLM-driven"; "LLM-powered"; "LLM-assisted"; ChatGPT; "GPT-4*"; "GPT-3*"; "generative pre-trained transformer"; "generative pretrained transformer"; "foundation model"; "foundation models"; "generative AI"; "generative artificial intelligence"; "retrieval-augmented generation"; "retrieval augmented generation"; "prompt engineering"; "LLM agent"; "LLM agents"; "language model*"; GPT; LLaMA; Gemini; DeepSeek; Qwen; Claude; "agentic AI"; "agentic artificial intelligence"; "AI agent*"; chatbot*; "conversational agent*"; "conversational AI"; GenAI; "Gen-AI".
Supply chain	"supply chain"; "supply chains"; "supply network"; "supply networks"; logistic; logistics; procurement; sourcing; supplier; suppliers; inventory; production; manufactur*; scheduling; transport*; forecast*; warehous*; "operations management"; purchas*.
Web of Science	TS = (language-model block) AND TS = (supply-chain block) AND IS = (40 journal ISSNs) AND PY = (2019–2026) AND DT = (Article).
Scopus	TITLE-ABS-KEY(language-model block) AND TITLE-ABS-KEY(supply-chain block) AND ISSN(40 journal ISSNs) AND PUBYEAR greater than 2018 AND PUBYEAR less than 2027 AND LIMIT-TO(DOCTYPE, "ar").
Journal frame	Forty AJG 2024 journals rated 3, 4 or 4* in OR&MANSCI (27 journals) and OPS&TECH (13 journals).

Note. In search terms, * denotes a wildcard matching zero or more characters.

3.2. Eligibility and Study Selection

To be included, an article had to investigate a supply-chain problem in which language modeling played a substantive role in the core analysis (Table 5), whether as an implemented operational tool, a text-measurement instrument for empirical research, or the focus of an

adoption or contracting study. Papers that mentioned LLMs only in passing, used them merely for copy-editing, or suggested supply chains only as a future application area were excluded. Studies framed around general generative AI were retained only where the substantive analysis encompassed language-model capabilities.

Database queries returned 228 records from Web of Science and 195 from Scopus (423 in total). Removing 174 duplicates left 249 unique records for title-and-abstract screening, of which 122 were excluded and 127 advanced to full-text review. Among the 37 citation-traced candidates, 31 were excluded at the title-and-abstract stage, and six proceeded to full-text review. In total, we screened 286 records, excluded 153 based on titles and abstracts, and assessed 133 full texts. Full-text review excluded 52 articles. Twenty fell outside the supply-chain decision scope (FT1), 13 did not involve language-model technology (FT2), eight used an LLM only incidentally (FT3), and 11 were review papers or commentaries without an original investigation (FT4). The final corpus comprises 81 articles, 78 from the database search and three from citation tracking (Figure 3).

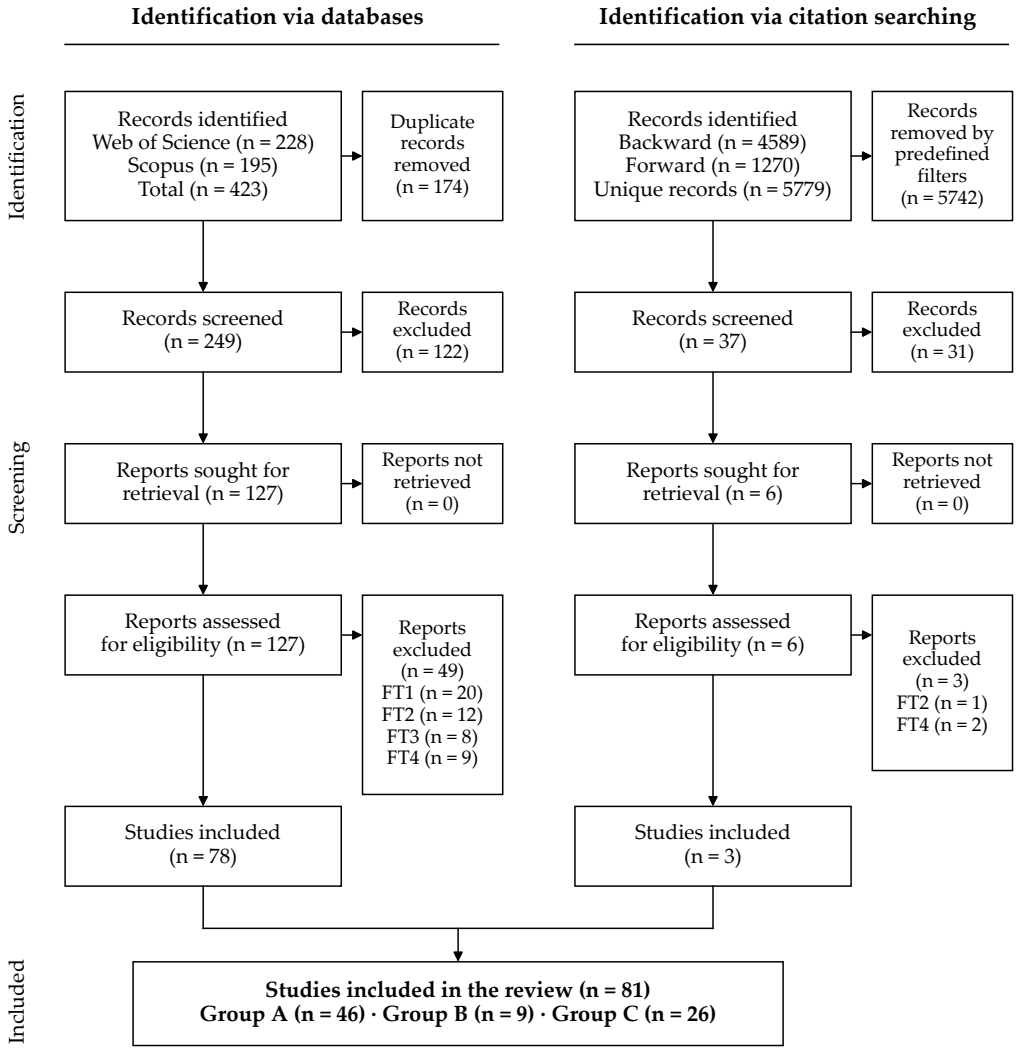


Figure 3. Study identification and selection following PRISMA 2020.

We grouped the 81 articles into three categories according to how the language model entered the paper's main contribution. Group A (operational applications, 46 articles: 45 generative-LLM workflows and one embedding-only model) comprises studies that

execute a language model whose output feeds into a supply-chain task or decision-support procedure. Group B (research-measurement studies, nine articles) uses model-generated text codings or indices as empirical variables to study a supply-chain phenomenon. Group C (adoption, analytical and design studies, 26 articles) examines organisational adoption, models expected LLM capabilities in game-theoretic equilibria, or proposes conceptual system architectures without executing an operational workflow. Where an article combined several of these elements, we assigned it according to its primary contribution.

Table 5. Study eligibility criteria.

Criterion	Included	Excluded
Publication	English-language articles in the specified journal frame, retrieved within the search window and Article filter.	Items outside the frame or indexed as reviews, editorials, chapters, or conference contributions.
Technology	The language model contributes through one of the three study roles defined in this section.	Background mentions or ancillary use unrelated to the central contribution; conventional AI or generative systems without a substantive connection to language modeling.
Decision setting	A supply-chain problem is central to the application or research question.	Supply chains appear only as background or a suggested future application.
Research content	An implemented application, analytical model, empirical study, or developed conceptual argument.	General commentary lacking a substantive investigation or framework.

3.3. Full-Text Coding and Comparative Synthesis

Each included article was read in full and coded across the TCCM dimensions defined in Table 6 [17], capturing the focal decision problem, model inputs and outputs, integration architecture and evaluation design (see Table S1 for the complete study-level coding). Every paper received a single primary decision area and a single primary research methodology. Language tasks and technical integration mechanisms were coded as multi-label fields, since a workflow often combined several operations.

Table 6. TCCM dimensions and data extraction.

Dimension	Information Extracted	Analytical Purpose
Bibliographic profile	Year and journal; authors and affiliations; DOI, keywords, and database citation fields.	Describe publication development and intellectual connections.
Theory	Named explanatory perspective and its role in the argument.	Explain the expected relationship between a task and its outcome.
Context	Industry and location; unit of analysis; evidence sources and operating conditions.	Establish the setting in which findings apply.
Characteristics	Decision problem; model and task; input and output; technical integration and associated decision method.	Compare the allocation of work within similar decisions.
Methodology	Research design; execution or assumed capability; benchmark, evaluation and reported outcome.	Interpret the result through the comparison producing it.

After the first author coded all 81 articles, a second coder independently double-coded a stratified random subsample of 20 studies (24.7%, sampled across study groups with random seed 20261106) while blinded to the initial codes. Prior to reconciliation, inter-rater agreement (Cohen's κ) reached 0.815 for primary decision area, 0.921 for research methodology and 0.920 for model role; across the multi-label fields, pooled label-level κ was 0.825 for realized tasks and 0.933 for technical integration, with mean Jaccard similarities of 0.867 and 0.925, respectively. Thirteen field-level discrepancies across seven articles were resolved by an independent domain expert against the full texts and codebook, and the reconciled dataset is used throughout our analysis.

Within each decision area, we traced how model outputs entered the downstream procedure and what benchmark was used to judge performance. This let us separate linguistic accuracy from end-to-end operational gains. Empirical adoption findings were read in light of their sampling and research design, and analytical models in light of their structural capability assumptions. For the 46 operational applications in Group A (including the embedding-only study), we also conducted a structured appraisal of evaluation rigor and reporting transparency (Table S2). For each study, we recorded whether it benchmarked against a baseline or alternative treatment, conducted component ablations or parameter sensitivity checks, repeated runs to account for stochastic output variation (as distinct from internal iterations within a single solver run), identified the underlying model family or API snapshot, and provided verifiable prompt templates or reusable protocols.

3.4. Bibliometric Methods

We analyzed publication trends and journal outlets across all 81 articles, using first-author affiliations to map geographical distribution and author keywords (available for 80 papers) to map thematic links. Citation counts were compiled separately from Web of Science (73 indexed records; 1232 citations) and Scopus (72 indexed records; 1514 citations). Within-corpus citation links were reconstructed by matching DOIs in Web of Science reference lists against our 81 included studies. To build the keyword co-occurrence network, we standardized spelling variants, removed generic technology and supply chain terms, and retained keywords appearing in at least two papers (counting each term once per article). Edges required at least one co-occurrence, weighted by association strength (co-occurrence divided by the product of the two terms' frequencies), and were partitioned into communities using the Louvain algorithm with resolution 1 and random seed 42 [47].

To complement the keyword analysis, we modeled title-and-abstract topics with BERTopic [48] using normalized all-MiniLM-L6-v2 embeddings, UMAP dimensionality reduction (15 neighbors, five dimensions, minimum distance 0, seed 42) and HDBSCAN clustering (minimum cluster size eight, minimum samples two), with English stop words stripped when extracting unigram and bigram topic labels. These BERTopic clusters capture broad lexical similarity in titles and abstracts. The full-text TCCM coding, reported in Section 5, captures the specific operational tasks and evaluation methods. Cluster stability was verified across ten random seeds via pairwise adjusted Rand indices and topic-level Jaccard similarities, alongside a parameter grid sweep reported in Table S3.

4. Bibliometric Analysis

4.1. Publication Development and Research Outlets

The 81 reviewed articles span 19 journals and 263 distinct authors. Measured by journal issue year, output rose sharply from two articles in 2023 and three in 2024 to 14 in 2025 and 59 in 2026, plus three early-access articles already assigned 2027 issue dates by our November 2026 search cutoff. As Figure 4 shows, just two outlets, the *International Journal of Production Economics* (23 articles) and the *International Journal of Production Research* (19 articles),

account for 52% of the corpus. The remainder is spread across journals in engineering management, computational optimization and inter-firm supply-chain management. First-author affiliations span 17 countries or regions, led by China (35 articles) and the United States (12), followed by Australia and Germany (five each), and 18 scholars author or co-author two or more papers in the sample (see Tables S4 and S5 for complete journal, regional and author breakdowns).

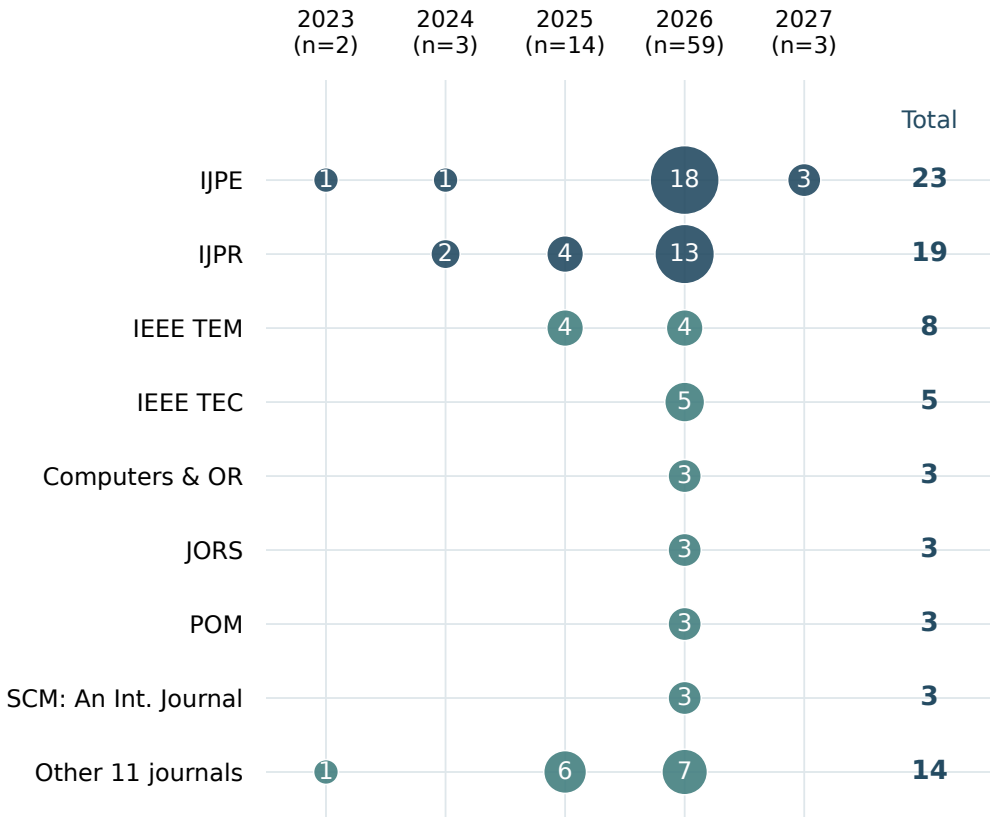


Figure 4. Articles by publication outlet and issue year. *Note*. Bubble size and labels indicate article counts.

4.2. Citation Structure of the Literature

Matching reference lists within the sample reveals 142 direct citation links among 65 connected articles, leaving 16 recent papers with no within-corpus links yet (Figure 5). The most heavily cited anchor papers inside the network connect early empirical and conceptual discussions of generative-AI adoption [49–51], natural-language code generation for logistics simulation [6], and inter-firm supply-chain coordination [52]. Across the wider literature, the indexed papers have accumulated 1232 citations in Web of Science (across 73 records) and 1514 in Scopus (across 72 records), with study-level counts reflecting both external reach and time since publication (Table S6).

4.3. Research Themes

After generic technology and supply-chain terms are filtered from the 80 articles with author keywords, the co-occurrence network retains 40 recurring terms and 106 links across eight Louvain communities (Figure 6; term frequencies by year appear in Table S7). One set of clusters centers on optimization, simulation and digital twins, tying language models to mathematical decision support. Other clusters group around supplier selection and negotiation in inter-firm trading, or around supply-chain visibility, network mapping and natural-language processing.

BERTopic modeling of titles and abstracts partitions the corpus into four overarching themes (Figure 7): risk, models and decision frameworks (29 articles); generative AI and organisational adoption (22); supplier/platform interaction and collaboration (15); and scheduling, optimization and heuristics (15). All ten random seeds recover this four-topic structure, at a mean adjusted Rand index of 0.83 and topic Jaccard similarities of 0.84–1.00 (Table S3). Cross-tabulating these themes against full-text research designs reveals a sharp methodological divide (Figure 7). All 15 scheduling and heuristics papers rely on computational or task experiments, 11 of the 15 supplier/platform interaction papers develop game-theoretic models, and the organisational adoption cluster splits evenly between survey studies (seven articles) and computational experiments (seven) alongside qualitative and conceptual work. Section 5 looks beneath these abstract-level clusters, comparing how specific language tasks, operating contexts and evaluation methods interact within each supply-chain decision.

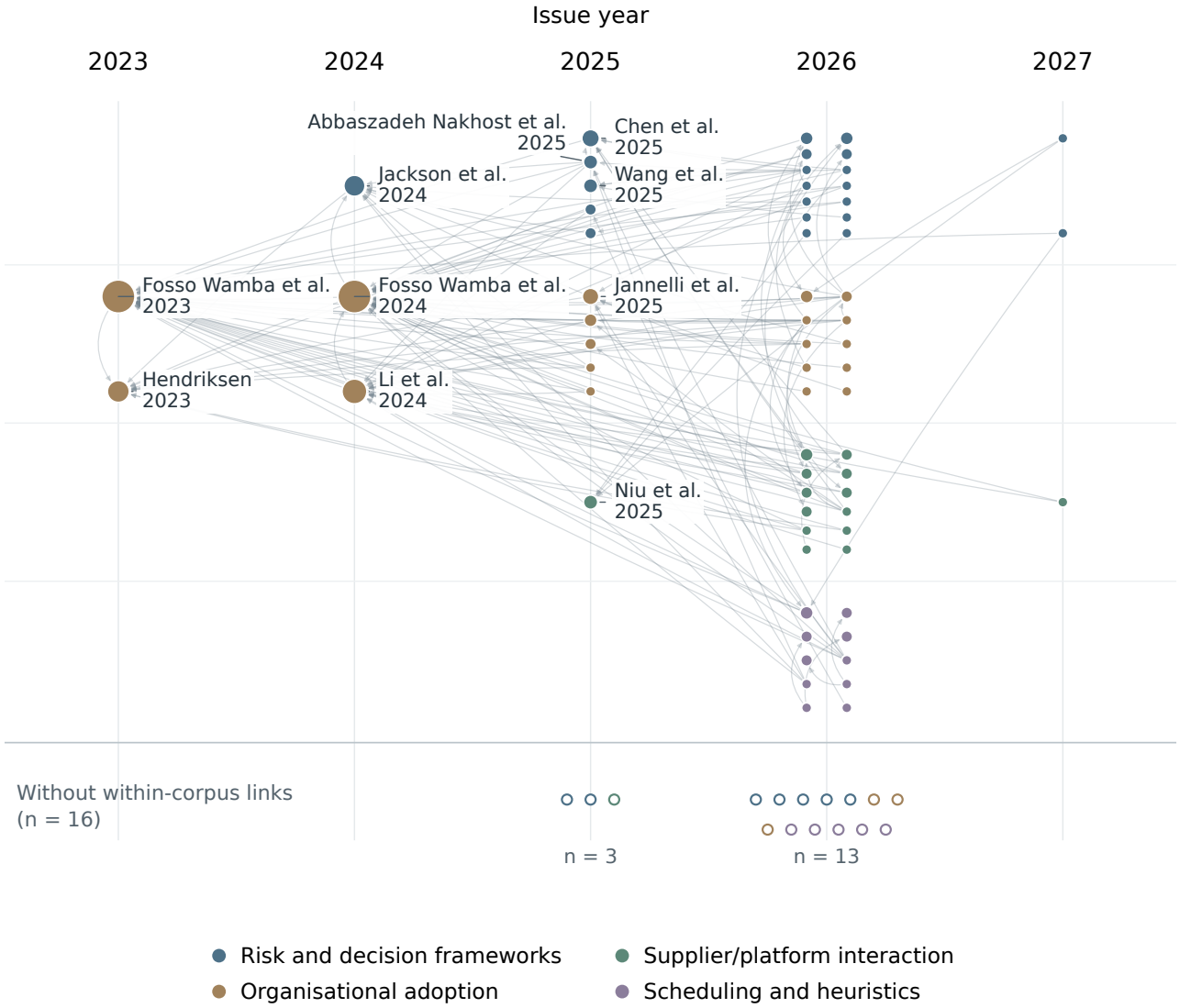


Figure 5. Within-corpus citation network by issue year. *Note*. Node size indicates within-corpus citations received; labels show authors and publication years.

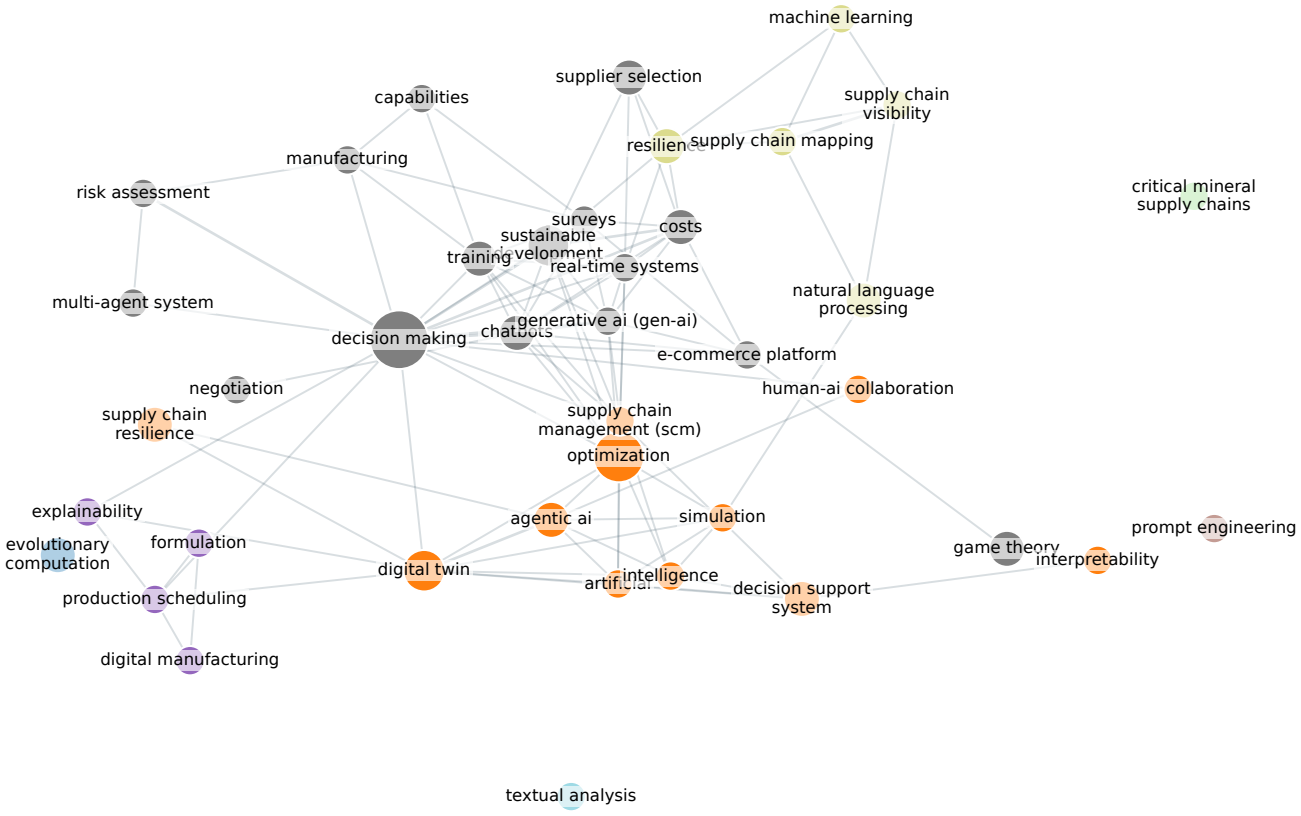


Figure 6. Author-keyword co-occurrence network. *Note.* Terms occur in at least two articles. Node size indicates frequency; colors identify communities.

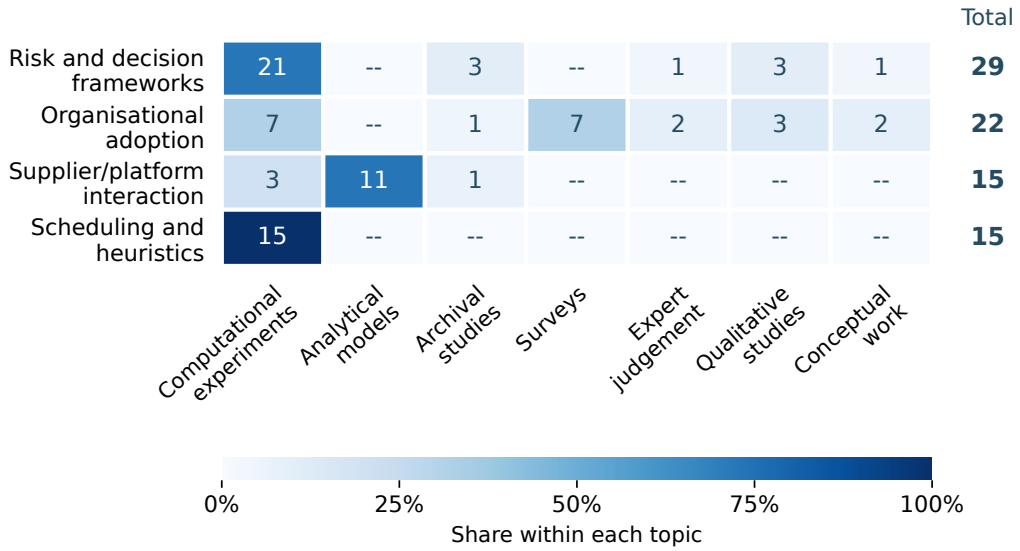


Figure 7. Research methods within the four title-and-abstract themes. *Note.* Cells show article counts; shading indicates within-theme shares. Dashes indicate zero articles.

5. Content Analysis Based on the TCCM Framework

5.1. Theoretical Perspectives

Theoretical framing across the corpus depends on whether a study examines workflow integration inside a firm or strategic interaction across firm boundaries. Internally oriented studies frequently draw on information-processing theory, which links an organisation's

need for data-processing capacity to the environmental uncertainty it faces [53]. In a closed-loop steel digital twin, the LLM acts as a natural-language interface that translates user questions into database queries for an operational allocation model [28]. In a circular hotel-linen study, domain experts first establish the relevant operational barriers, and LLM-generated priority rankings are then benchmarked against expert judgements [36].

To explain why managers adopt and rely on these tools, empirical studies draw on task–technology fit [54], with survey evidence on ChatGPT showing that perceived fit with daily tasks drives user satisfaction and adoption intention [55]. Dynamic-capability theory [56] then extends this view beyond initial acceptance, framing both organisational barriers to LLM adoption [57] and how LLM augmentation supports managerial improvisation when routine procedures break down [58].

When applications cross firm boundaries, theoretical attention shifts from internal efficiency to asymmetric information, external dependence and rent division. Signaling theory helps explain why LLM evaluations of public procurement bids score baseline vendor qualifications more consistently than strategic differentiation claims [59], and why a major customer's generative-AI announcement triggers measurable abnormal returns in its suppliers' share prices [60]. Resource-dependence theory is used to read how defense manufacturers balance buffering and bridging strategies in their annual reports [61]. Bargaining and co-opetition models treat the strategic tension more directly. Retrieving past contract records alters how negotiation surplus is split between buyer and supplier [8], and spillover effects from joint model fine-tuning can support co-investment or leave competing manufacturers worse off [37].

5.2. Research Contexts

The empirical contexts in our sample range from text-heavy commercial negotiations to physical production facilities governed by hard capacity constraints (Table 7). In document-centered settings, the relevant detail sits in firm-specific records. Evaluating Ohio public IT tenders requires interpreting how diverse vendors frame their qualifications against standard solicitation rules [59]. Querying a Vietnamese steel digital twin requires mapping local delivery and scrap-recovery logs [28]. Assessing circular hotel-linen operations in China depends on the collection and laundering arrangements linking hotels and recyclers [36]. In production and logistics scheduling, the environment is defined by explicit physical rules, such as berth and bay limits in a Japanese wholesale market [62] or shop-floor routing constraints in an Italian manufacturing digital twin validated with plant experts [63].

Studies also operate at different units of analysis, which shapes what outcomes they can observe. Individual- and firm-level surveys capture employee willingness to use ChatGPT [55], buyer–supplier coordination across Chinese firms, and managerial responsiveness in UK manufacturers [52,58]. Dyadic and network-level studies instead track bilateral bargaining offers [8], game-theoretic contracting across competing channels [37,41,64], and macro-level trade decoupling and commodity price movements extracted from financial news [38,39]. Standardized ordering vignettes isolate behavioral biases under controlled laboratory conditions [65]. User acceptance, firm-level operating performance, and inter-firm profit division are different outcomes, and the unit of analysis decides which of them a study can observe.

Table 7. TCCM comparison of representative studies.

Study	Theory and Context	Characteristic and Method	Main Contribution
[65]	Behavioral operations; inventory and other decision vignettes	GPT choices across 18 biases; repeated experiments	Model improvement differs between calculable tasks and preference judgements.
[59]	Signaling; actual Ohio procurement bids	Retrieved documents and repeated LLM scores; human comparison	Qualification and competitive differentiation involve different judgement ambiguity.
[8]	Bargaining and information asymmetry; bilateral contracts	Retrieved agreement histories; agent negotiation	Context choice changes both efficiency and profit allocation.
[35]	Resilience and inventory buffering; critical minerals	Document extraction linked to agent-based simulation	Narrative evidence can become stock-policy inputs.
[28]	Information processing; closed-loop steel	Query generation linked to a digital twin and allocation procedure	Language connects user requests to operational data.
[36]	Information processing; circular hotel linen	Barrier identification and ranking; expert comparison	Supplied criteria support judgement more reliably than unaided barrier discovery.
[58]	Dynamic capabilities; UK manufacturers	Augmentation and improvisation; organisational survey	Managerial use connects language capability to response capacity.
[61]	Resource dependence; defense-company annual reports	LLM-assisted content coding; longitudinal interpretation	Buffering and bridging explain resilience development.
[37]	Co-opetition; competing manufacturers	Shared adaptation investment; analytical equilibrium	Spillovers change the incentives to collaborate in model development.
[66]	Information asymmetry; bus-financing documents	Language-enhanced text workflow; manual comparison	Faster evidence preparation involves a quality trade-off.

5.3. Characteristics of LLM Applications

We organise our synthesis of application characteristics around nine primary supply-chain decision areas, tracing what task the LLM performs and how its output enters the decision procedure. The first eight subsections examine specific operational and strategic decision problems. The final subsection covers cross-functional decision support and general organisational adoption. Where an article spans multiple processes, it is assigned to its primary decision problem in the descriptive counts: inventory and production planning accounts for 17 articles, cross-functional adoption and decision tools for 14, risk monitoring and resilience for 11, procurement and supplier management for nine, channel contracts and technology collaboration for nine, logistics and network design for eight, circularity and sustainability for six, demand and price forecasting for four, and supply-chain finance for three (Figure 8).

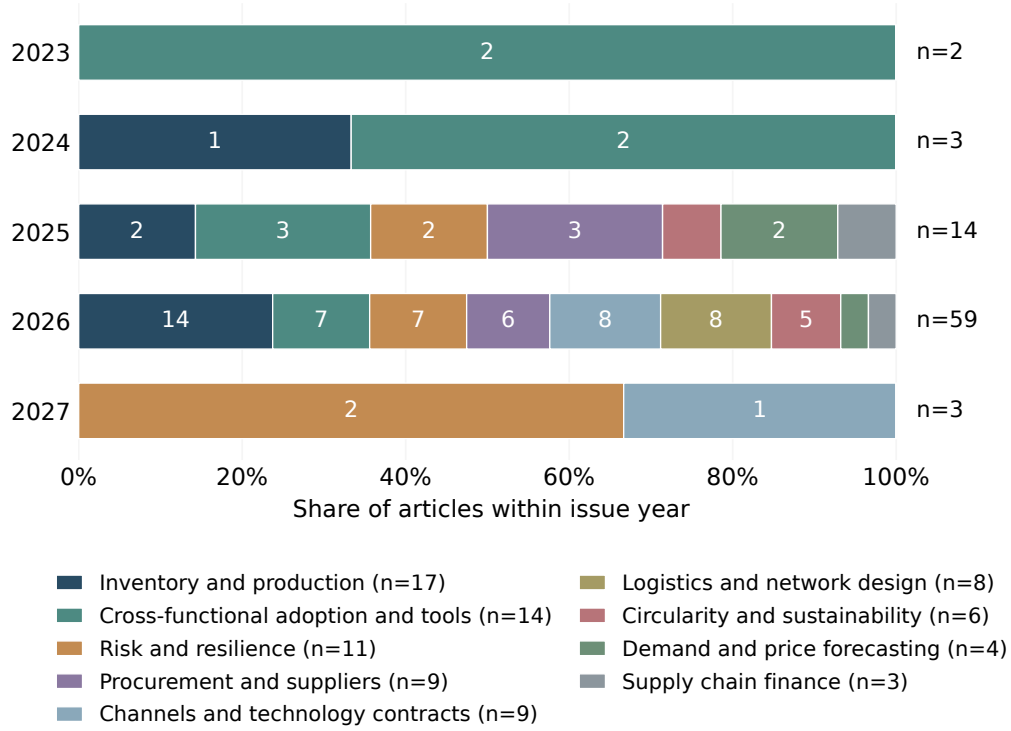


Figure 8. Decision-area composition by issue year. *Note*. Segment labels show article counts.

Across the 81 studies, 46 implement operational applications (Group A: 45 generative-LLM workflows and one embedding-only predictor), nine use language models as text-measurement instruments for empirical research (Group B), and 26 examine adoption empirically (13 studies), model capabilities analytically (11), or propose conceptual system architectures (two; Figure 9). Within the 46 Group-A workflows, the most common realized tasks are generating solver code, heuristics, or database queries (18 studies), explaining analytical results or supporting interactive dialogue (15), and synthesizing or drafting text (12). Eleven applications propose operating actions or policies directly, including three multi-round negotiation agents. The nine Group-B measurement studies focus on extracting empirical indicators from corporate or news archives (eight studies), assigning risk or sentiment scores (three), and summarizing disclosures (three). Multi-step workflows often span more than one task category.

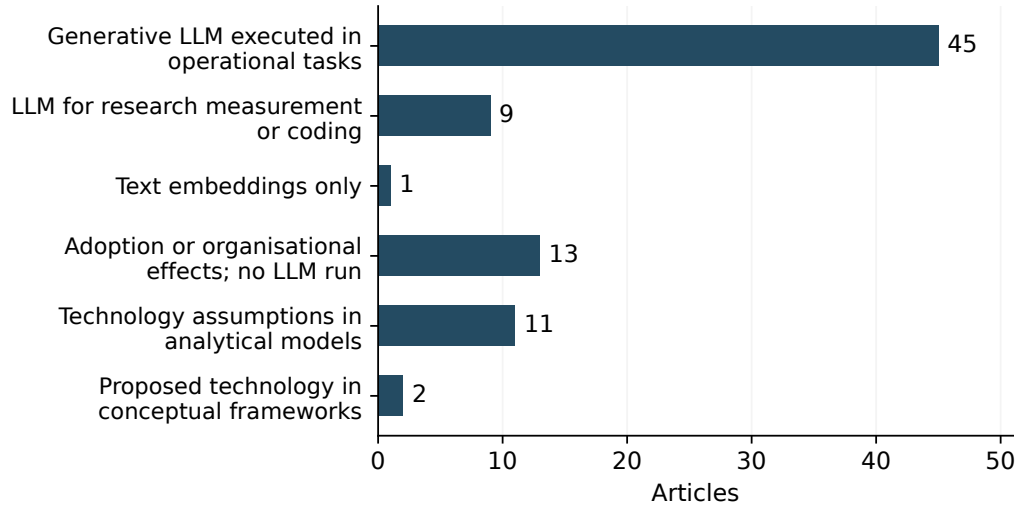


Figure 9. Roles assigned to language models.

Technically, 45 of the 46 Group-A applications rely on prompting or in-context design, supplemented by retrieval augmentation in eight studies and domain fine-tuning in six. All nine Group-B measurement studies also use prompting, one with retrieval and one with fine-tuning. Most operational workflows pair the language model with an external analytical engine: 29 link the LLM to mathematical optimization or simulation, nine integrate it with predictive machine-learning models, and four deploy interacting multi-agent LLMs. This hybrid pattern is especially clear in automated heuristic design and scheduling [34,67–77], where the LLM proposes candidate dispatch rules or solver code but an external evaluator checks feasibility and computes objective values (Figures 10 and 11).

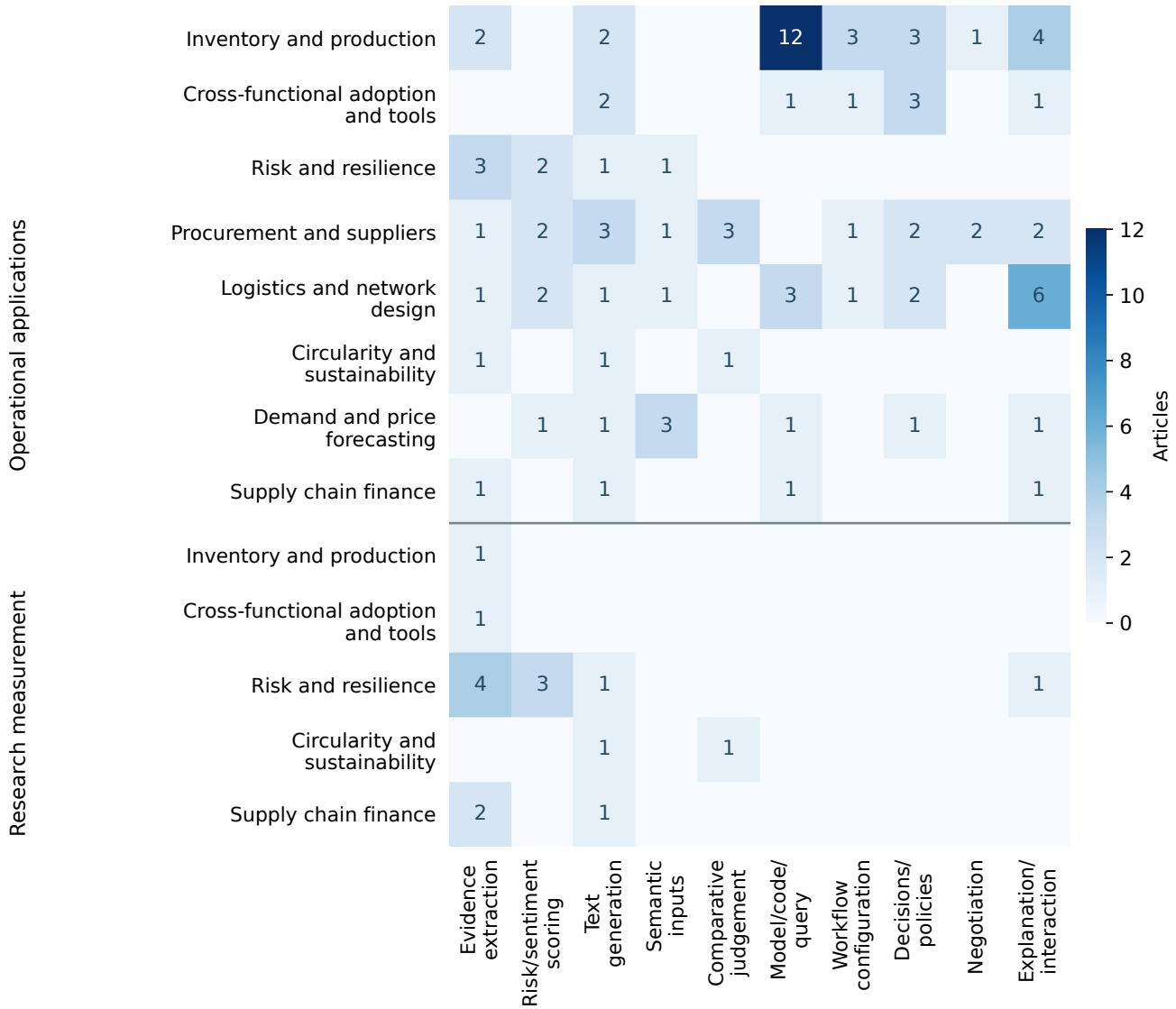


Figure 10. Implemented LLM tasks across decision areas. *Note*. Studies may report multiple tasks.

5.3.1. Procurement and Supplier Management

Supplier selection requires translating broad purchasing goals into structured evaluation criteria [78,79], and procurement studies test LLMs at different stages of this workflow. Sadeghi et al. [7] prompt ChatGPT to propose selection criteria and pairwise comparisons within an analytic hierarchy process, benchmarking the resulting weights against those of 14 heavy-machinery purchasing professionals. Moving from criteria weighting to actual bid evaluation, McKinley et al. [59] compare repeated LLM scores across 123 vendor proposals from 31 Ohio IT tenders against 102 human evaluator records, finding that model scores

align much more closely with human judgements on objective vendor qualifications than on open-ended claims of competitive differentiation. Other procurement workflows use the LLM as an upstream evidence filter: extracting and human-verifying green supply-chain practices before generating advisory recommendations [80], configuring rare-event supplier-risk detection pipelines [81], or feeding text-derived external risk scores into a Bayesian reinforcement-learning and multiobjective supplier-selection model tested through baseline comparisons and component ablations [82].

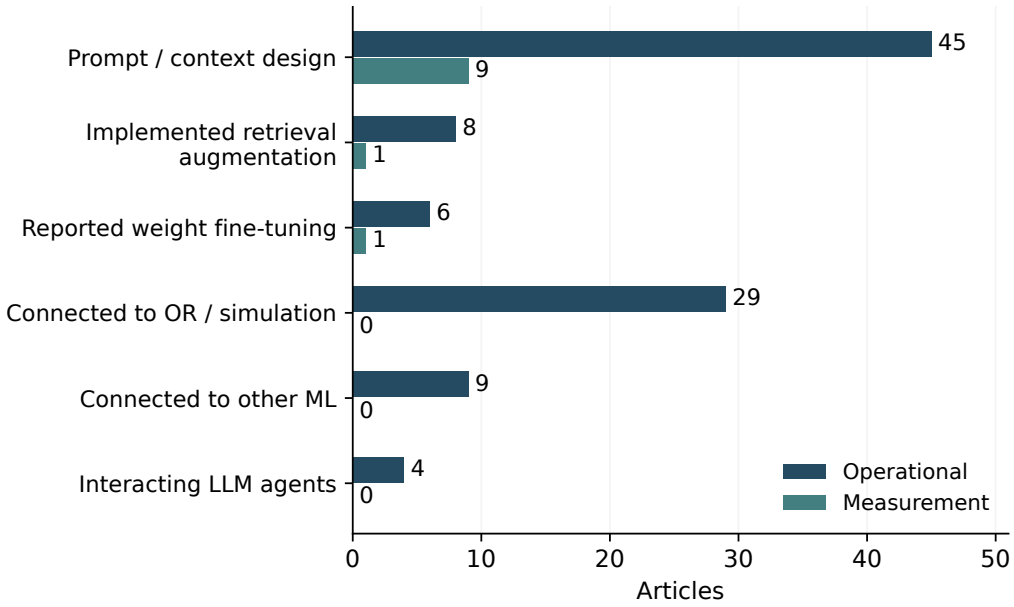


Figure 11. Technical integration of LLM applications. *Note*. Studies may use multiple integration methods.

When the task shifts from evaluating bids to negotiating terms, the model's output becomes a direct commercial offer to a trading partner. In simulated bilateral bargaining benchmarked against prior human experiments, Kirshner et al. [8] show that supplying an LLM agent with retrieved past agreements curated for joint efficiency raises total channel efficiency from 95.05% to 98.01%, but compresses the supplier's share of profit. Curating the retrieval library around supplier-favorable deals reverses that rent division, so the retrieved contracts carry a bargaining posture as well as a factual record. Herold et al. [40] find a similar trade-off across three behavioral experiments with students and professional buyers. An LLM buyer prompted to bargain aggressively secures deeper price discounts. Cooperative prompting preserves the human supplier's trust and willingness to continue trading. Prompt instructions and retrieved contract histories both change how trading surplus is split [8,40], so the price saved on one deal is an incomplete measure of purchasing performance. Game-theoretic work derives when expected LLM capabilities sustain emergency sourcing collaboration [83]. Event-study evidence shows that a major customer's generative-AI adoption announcement triggers significant stock-market reactions among its upstream suppliers [60].

5.3.2. Inventory and Production Planning

In inventory and production planning, better information creates value only when it improves replenishment or scheduling choices [19,84–86], and LLM studies enter this process either by building the mathematical models or by generating order quantities directly. On the model-building side, Jackson et al. [6] prompt an LLM to write Python 3 simulation code from verbal descriptions of two inventory systems and a queueing problem,

noting that explicit variable definitions or starter code are often needed to produce runnable scripts. Domain fine-tuning pushes this further by translating natural-language scheduling problems into executable mixed-integer programs verified through solver execution [29]. A critical-mineral stockpiling study uses retrieval-augmented extraction across 972 policy and market documents to populate parameters and decision rules for an agent-based simulation, testing 24 stockpiling policy combinations across four disruption scenarios (100 runs per cell) [35].

When LLMs are asked to place replenishment orders directly across multi-period horizons, they must adapt to evolving inventory states and demand shocks. In a simulated three-echelon supply chain, a consensus-seeking study compares independent LLM agents against information-sharing, tool-assisted, and negotiated agent configurations under a shared demand path, finding that the simpler independent Flash setup generates lower order variance than multi-agent negotiation [87]. In a decentralized dual-channel inventory game driven by retail demand from three stores, direct LLM ordering policies earn consistently lower profits than specialized reinforcement learning across repeated simulation runs [88]. The two uses need different checks. Formulating a simulation or optimization model requires verifying that physical constraints were not dropped in translation. A live ordering policy needs its multi-period stability benchmarked against established control and reinforcement-learning heuristics.

A third group of studies uses the LLM as an interpretive bridge between complex planning models and human managers. Inside an Italian manufacturing digital twin, the language model translates decision-tree representations of candidate scheduling policies into natural-language explanations, evaluated across four synthetic datasets and one plant dataset by five production experts [63]. A cross-task experiment tests how LLM temperature settings trade off operational accuracy against solution novelty [89]. An econometric study constructs a servitization measure using bag-of-words text analysis and validates it against GPT-4-based service identification in a bullwhip-effect analysis [90]. A conceptual paper outlines how agentic digital twins could pair conversational interfaces with simulation-based planning [91].

5.3.3. Logistics and Supply Network Design

In logistics and network design, language models extend machine-learning-assisted combinatorial optimization [92] by translating verbal scenario changes into mathematical formulations or database queries. LLM4SCM [30] modifies a supply network optimization model from natural-language prompts and summarizes the resulting Gurobi solution across 30 test scenarios, benchmarking five LLM backbones against prompt-only baselines and expert formulations. The narrative summaries can still misstate numerical outputs. In one documented case, the summary reports an objective value of 187,500 when the solver returned 208,500. In a closed-loop green-steel digital twin [28], the LLM instead writes structured database queries that feed delivery and scrap-recovery records into a downstream allocation heuristic, evaluated across 160 ground-truth queries and 800 linguistic paraphrases to compare base models, RAG and fine-tuning. Language models also generate the risk parameters or disruption scenarios under which logistics networks are stress-tested: a commercial-property study scores weather narratives (reaching a 0.81 correlation with author ratings across 200 events, versus 0.51 for a rule-based baseline) before running multicriteria site evaluation [93]. A semiconductor hub-location study uses Flan-T5 to generate disruption scenarios for a hybrid simulation–genetic-algorithm solver [94], extending classic reliability-based facility location [95].

Interactive logistics prototypes keep conversational dialogue separate from numerical optimization. In some, the LLM explains a solved schedule. In others, it proposes physical design changes. In a wholesale produce market testbed [62], a locally hosted Gemma model

answers user questions about truck-unloading schedules, and conventional optimization and simulation engines compute the actual bay assignments. An agentic control-tower proof of concept pairs tool-assisted user dialogue with a separate multi-period simulation of information sharing [96]. CORTEX-V [97] uses a vision-language model to propose candidate facility-layout modifications directly, passing each proposed layout to executable verification functions and benchmarking performance across 30 runs against optimization baselines and component ablations.

5.3.4. Risk Monitoring and Supply-Chain Resilience

In supply-chain risk management, early warning and network visibility matter only insofar as they inform mitigation choices such as dual sourcing, inventory buffering, or capacity reconfiguration [98–101]. Several studies test whether LLMs can reconstruct multi-tier supply networks from unstructured public disclosures. Company names are recovered far more often than the supply links between firms. When a RAG pipeline maps electronics supply chains from corporate filings [25], checking 347 extracted firms against S&P Capital IQ confirms high precision (0.827) but low recall (0.199). A supply-chain knowledge-graph study achieves 0.973 precision@1 on entity disambiguation but only 0.328 F1 on relation extraction [102]. Both pipelines miss many of the inter-firm dependencies needed to gauge disruption propagation. An automotive study fuses text embeddings with graph topology across 1031 suppliers to predict firm-level resilience against conventional baselines and feature ablations [103]. A critical-mineral study shows that cobalt news sentiment shifts substantially with the domain background injected into the prompt [104]. UNISONE feeds text-mined disruption alerts directly into a capacity-reconfiguration model evaluated across repeated disruption simulations [105].

A second group of resilience papers uses LLM-coded text as an empirical lens to study how firms respond to macro-level shocks. Text indicators extracted from investor-relations disclosures enter panel regressions and inventory simulations [106]. Human-validated news classifications track firm-level supply-chain decoupling and partner reconfiguration across borders [38]. Longitudinal content coding of defense contractors' annual reports traces shifts between buffering and bridging strategies [61], and a comparative study of 20 European multinationals codes corporate reports (reaching 0.85 human–model agreement) to map geopolitical-risk responses [107]. UK survey evidence links LLM augmentation to managerial improvisation [58]. Structured expert evaluations assess generative-AI risks in sustainable manufacturing [108].

5.3.5. Demand and Commodity Price Forecasting

Two forecasting studies use LLMs to supply text features to downstream time-series predictors. The other two test direct forecasting or LLM-generated forecasting programs. In agricultural commodity markets, DeepSeek sentiment scores extracted from news articles feed into a transformer-based wheat-price predictor [39]. Ablation and baseline checks show that text signals do not help at every horizon. At five days, a standard GRU baseline still has lower error. An energy-material forecasting study encodes LLM-generated summaries of historical market conditions alongside numerical price series [109]. The summaries outperform raw scraped text in repeated tests. They are generated retrospectively with a pretrained model, so it is unclear whether they contain knowledge from after the decision date.

In direct demand forecasting, model generation matters. One laboratory comparison sets 123 graduate students against five LLMs on 24 retail sales series [110]. Earlier models (ChatGPT-3.5 and Llama 2) produce significantly larger forecasting errors than human participants. Newer frontier models (ChatGPT-4, Bard and Bing) perform comparably to humans. For spare-parts forecasting, an LLM instead selects forecasting methods and gen-

erates Python scripts, which are executed locally and debugged through feedback. Across four simulated and two empirical datasets (30 replications per dataset) [111], averaging predictions across repeated runs stabilizes accuracy, detailed context prompts lower mean error, and a machine-learning meta-tree that selects among candidate forecasting methods outperforms any single static prompt.

5.3.6. Circularity and Sustainable Supply-Chain Decisions

In circular and sustainable supply chains, operational studies use LLMs to structure evaluation criteria or extract physical flow quantities from field reports. A circular hotel-linen study in China [36] finds that prompting GPT to discover implementation barriers from scratch recovers only 15 of the 19 barriers rated as significant by the expert panel. Once the full barrier set and definitions are supplied in a two-stage prompt, its priority rankings align closely with expert weights. The alignment holds across sensitivity checks on prompt phrasing, expert weighting, and model backbone. In humanitarian relief operations, an LLM extracts quantities of reusable and recyclable materials from unstructured situation reports, which are verified before setting the capacity bounds of a multiobjective allocation model [112]. Extraction accuracy therefore determines the feasible region of the downstream optimization.

Beyond immediate operational allocation, a palm-oil study combines multiple LLMs with manual source verification to compile a 249-indicator biodiversity taxonomy for upstream supplier auditing [113]. Other studies treat sustainability as an adoption and contracting problem. A game-theoretic carbon-abatement model analyses how an assumed AI-driven reduction in abatement failure risk alters buyer sourcing and hedging strategies [114]. Survey evidence links generative-AI adoption to international green supply-chain performance via dynamic capabilities [115]. Another survey traces how resource orchestration translates decarbonization analytics into environmental performance [116].

5.3.7. Channel Contracts and Collaborative Technology Investment

The nine channel-contracting papers develop game-theoretic models to examine how anticipated LLM capabilities alter technology adoption, pricing and profit division across supply-chain tiers. In manufacturing and service settings, models represent the technology through shorter production lead times and lower coordination costs [117]. In e-commerce channels, by contrast, the focal tools interact directly with consumers. Platform explanation interfaces reshape buyer valuation biases [64]. Retail language models shift end-market demand [118], including when competing retailers deploy AI-generated review summaries [119]. AI shopping assistants interact with supplier adaptation effort, consumer scepticism [120] and heterogeneous privacy preferences [41]. Across these equilibria, a capability that expands total channel demand can still redistribute margins between platforms, manufacturers and retailers, so channel members disagree over wholesale versus agency contracts. The conflict is sharper when competing manufacturers co-invest in fine-tuning a shared industry model [37]. As in horizontal information-sharing dilemmas [121], knowledge spillovers between rivals determine whether joint model development yields mutual gains or leaves both firms worse off. Two related analytical models extend these contracting questions to community group buying and AI service pricing [122,123].

5.3.8. Supply-Chain Finance

In supply-chain finance, studies use LLMs both to speed up credit due diligence and to extract off-balance-sheet financing records from corporate filings. An operational due-diligence study benchmarks manual document review against four LLM-assisted workflows across 2527 bus-procurement web pages and 84 financial documents [66]. The

GPT-4 workflow cuts processing time from 691 min to 212 min but its expert-rated output quality edges down from 5.00 to 4.75. In an illustrative application of its own framework, an operations–finance study prompts an LLM, with human auditing and multi-model cross-checks, to mine 10-K filings across 5455 US corporations and identifies 257 buyers that use reverse factoring or supplier-finance programs [124]. Language models also support qualitative coding in a multiple-case study of blockchain-enabled supply-chain finance [125], and they generate text-based self-attribution variables from earnings-call transcripts to test how firms explain operating performance [126].

5.3.9. Cross-Functional Decision Support and Organisational Adoption

Studies in the final category examine organisational adoption and cross-functional decision support across the supply chain as a whole. Early survey and multi-method papers map initial adoption stages and perceived operational benefits [49,50]. Later firm-level surveys link generative-AI use to operational performance through buyer–supplier coordination [52]. In a survey of 382 supply-chain practitioners, task–technology fit drives adoption intention through user satisfaction. The hypothesized trust pathways are statistically non-significant [55]. Structured expert evaluations rank implementation barriers across supply-chain processes [57,127]. Qualitative interviews explore day-to-day human–AI collaboration in managerial decision-making [128] and organisational readiness for agentic AI [129]. Conceptual work links adoption breadth to the decision authority delegated to AI systems [13,51].

A parallel set of cross-functional studies tests model behavior and system architectures that cut across multiple operating domains. Controlled behavioral experiments test whether GPT models exhibit classic human decision biases across 18 operational vignettes [65]. Parameter experiments examine how adjusting decoding temperature shifts the balance between factual reliability and creative problem-solving [89]. At the architectural level, conceptual and prototype studies integrate conversational LLM interfaces with simulation-based digital twins and supply-chain control towers [91,96]. Empirical measurement work uses LLM-coded corporate disclosures to analyze how executives attribute operational outcomes [126].

5.4. Research Methodologies

Across the 81 articles, computational and task experiments represent the dominant methodology (46 studies), followed by analytical game-theoretic models (11), survey research (seven), qualitative case or content studies (six), econometric or archival analyses (five), expert-judgement and structural-modeling studies (three), and conceptual frameworks (three). All 11 primary analytical models appear in 2026 or 2027 issues, in the same period as a rise in experimental implementations (Figure 12). Within the 46 Group-A operational applications, 25 evaluate their workflows on real enterprise or industry records, 19 rely on synthetic or standardized benchmark problems, one reports a live production deployment, and one presents an unbenchmarked proof of concept. Real procurement tenders preserve the incomplete, strategic wording of commercial bids [59]. Synthetic multi-echelon simulations allow controlled stress-testing of ordering policies [87,88]. Thirty-eight of the 46 operational studies evaluate a downstream decision-support output or operational metric, such as forecast error, supplier rank correlation, or solver feasibility. Eight stop at intermediate linguistic metrics.

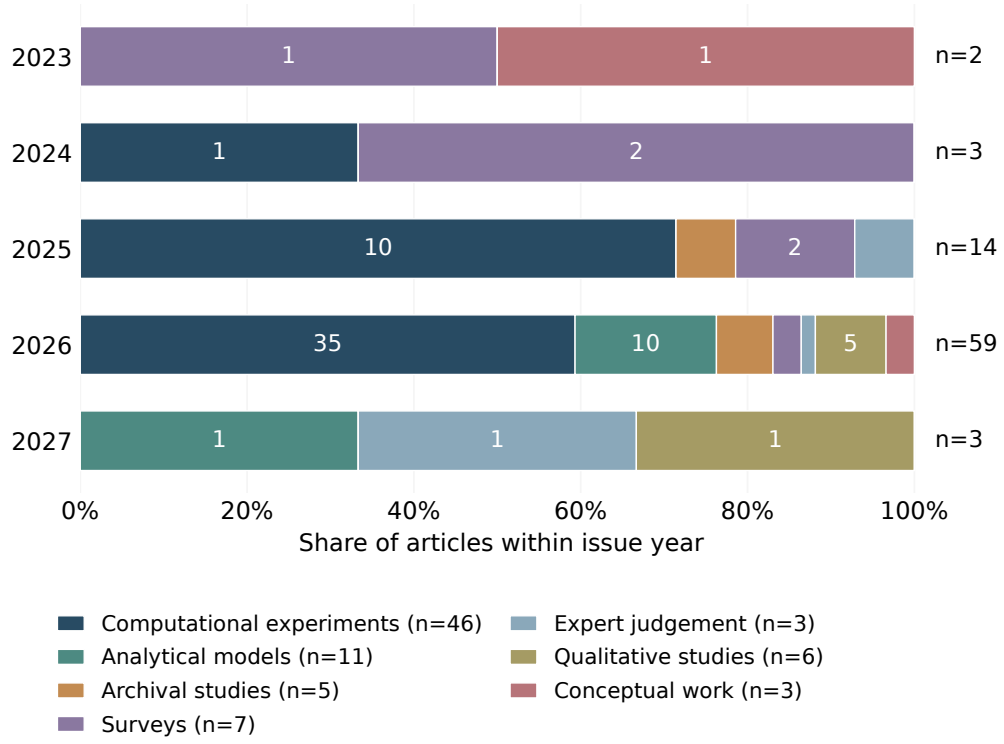


Figure 12. Research-method composition by issue year. *Note*. Segment labels show article counts.

Our study-level methodological appraisal (Table S2) shows how unevenly these experimental designs isolate the language model's contribution. Of the 46 operational studies, 45 include a baseline or treatment comparison and 35 test component ablations or parameter variations. Only 26 repeat model or system runs to account for stochastic generation variance. Thirty-nine studies identify the specific LLM family or version tested, and 27 provide verifiable prompt templates or reproducible workflow protocols in the paper or supplementary files. An end-to-end benchmark shows whether the complete pipeline works. A component ablation isolates how much the LLM adds over the external solver or retrieval database. Repeated runs show whether a stochastic model gives stable recommendations.

Non-experimental methods address questions that laboratory prototypes cannot observe. Archival and qualitative studies use validated LLM text coding to track cross-border supply-chain decoupling [38] and long-run shifts in defense firms' resilience strategies [61]. Practitioner surveys link perceived task fit and organisational routines to actual adoption behavior [52,55,58]. Game-theoretic models formalize when competitive spillovers or channel power make joint model development and adoption incentive-compatible [37,41].

6. Research Agenda

Improving an LLM's score correlation with human evaluators does not automatically change which supplier wins a contract. Tweaking a negotiation prompt, by contrast, can immediately alter the price offered to a trading partner. The five directions below follow language-model outputs into operating commitments and ask how inter-firm incentives govern these systems over time (Figure 13).

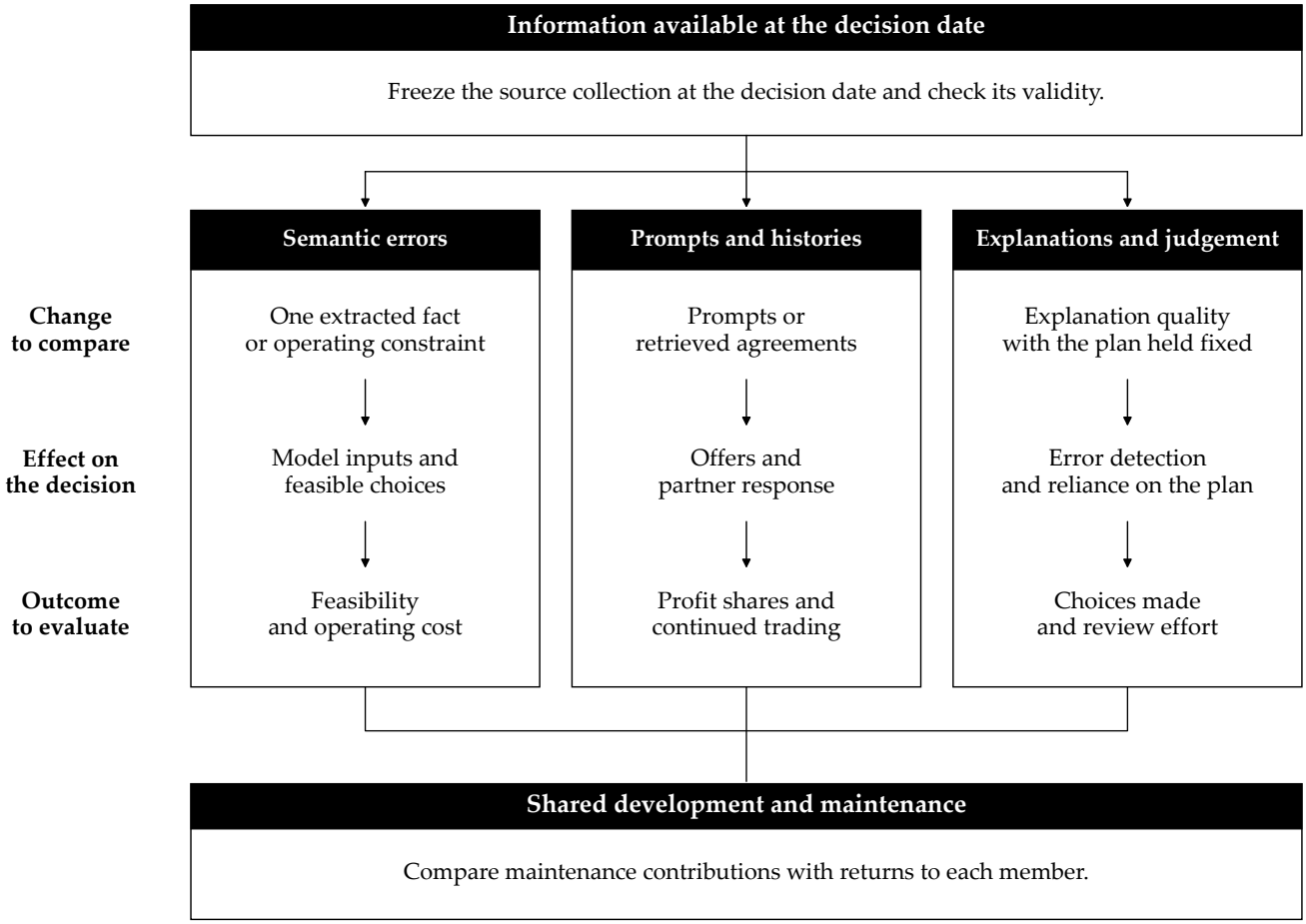


Figure 13. Research agenda for LLM-supported supply-chain decisions.

6.1. Semantic Errors and Operating Decisions

Network-mapping studies [25,102] and optimization-formulation tools [30,34] show that language models can misidentify supply links or silently drop operating constraints and still return fluent, executable outputs. How far these errors travel into later decisions, and what they cost, is not yet known. Missing a second-tier dependency can leave a backup sourcing plan exposed to the same regional disruption. Misreading a capacity bound can make a solver's schedule infeasible on the shop floor. Future research should follow an error from the source text through to the operating commitment, and ask where an automated check or a short human review would change the decision. General hallucination research focuses on factual faithfulness [130]. Supply-chain workflows add an operational trade-off between acting on ambiguous text and pausing to ask for clarification. Robust and distributionally robust optimization already provide tools for hedging against parameter uncertainty [131–133]. Researchers can build on them to model structural ambiguity, where the planner is unsure whether a contractual clause or capacity restriction applies at all. The open question is when an LLM agent should hedge mathematically and when it should query a human manager before locking in an order.

6.2. Prompts and Retrieved Histories in Trading Relationships

Experimental bargaining studies show that both the contract precedents retrieved into a prompt and the behavioral instructions given to an LLM buyer shift how profit is divided and whether human suppliers remain willing to cooperate [8,40]. Supply-chain relationships run on repeated transactions, so an aggressive prompt or a buyer-skewed

retrieval library that captures extra margin today may erode supplier trust and participation in later rounds. The same problem appears when trading partners or industry platforms keep a shared RAG repository of historical contracts. Even if every contributed record is factually accurate, a firm can selectively disclose agreements that anchor the model toward its preferred pricing or penalty clauses. Classic information-sharing theory [121] and recent co-opetition models of LLM fine-tuning [37] leave open who curates shared retrieval libraries, when selective disclosure distorts bargaining outcomes, and when firms prefer to keep their prompt histories private.

6.3. Explanations and Human Judgement

Several reviewed systems use LLMs to explain candidate policies or solver outputs to managers [30,62,63]. A high rating for fluency does not, by itself, mean the explanation produced a better operating choice. Theories of appropriate reliance [134] and human–AI complementarity [135] suggest holding the underlying optimization plan fixed and testing whether natural-language explanations help practitioners catch binding-constraint errors or numerical inconsistencies, such as the objective-value mismatch in [30], relative to solver tables alone. Expertise and checking cost belong in the same test. A veteran warehouse supervisor may immediately spot an infeasible truck-unloading sequence that a novice planner accepts at face value, and cross-checking source passages against solver outputs takes managerial time. Such studies should record both error-detection rates and review time before commitment, and ask when natural-language explanations improve human override decisions.

6.4. Information Available at the Decision Date

Backtesting LLM-based forecasts or risk indicators on historical periods introduces a methodological hazard whenever pretrained weights contain knowledge of events that occurred after the simulated decision date. It appears in our sample when a model generates retrospective narratives of past commodity markets [109], interprets dated cobalt news under varying prompt contexts [104], or codes historical corporate disclosures to predict inventory and trade shifts [38,106]. If a prompt dated in 2021 elicits representations shaped by post-2021 training data, the backtest overstates the signal a manager actually had at the time. Future forecasting and risk studies should keep signals drawn from contemporaneous documents with time-bounded models separate from retrospective reconstructions. The test of those signals is the inventory or procurement decision they support, not point-forecast error alone [2,136].

6.5. Shared Model Development and Maintenance

Analytical co-opetition models examine when competing firms should co-invest in fine-tuning a shared LLM [37]. The empirical applications in our review show that different adaptation assets carry very different competitive risks. Sharing domain-specific query templates [28] or scheduling code syntax [29] standardizes technical translation without exposing proprietary strategy. Pooling historical contract clauses or pricing prompts leaks sensitive bargaining posture. Future collaboration models can separate these technical and strategic layers and link the cost of domain adaptation to the operational decisions each partner makes.

Maintaining an application after deployment alters these investment incentives further. Current analytical models typically treat AI capability and development cost as one-off parameters [41,64,117,118,123], yet real supply-chain documents and foundation models drift continuously: supplier qualifications expire, disruption alerts lose relevance once capacity recovers, and vendor model updates can silently shift the recommendations or offers generated by an unchanged prompt. Just as physical digital twins require live state synchronization [137], text-driven decision systems require ongoing curation of retrieval

libraries and prompt guardrails. When a buyer and supplier share an application, maintenance costs are often asymmetric. One party may supply proprietary operational data while the other funds model updates. That split creates free-riding and hold-up risks, and dynamic contracting research should model them explicitly.

7. Discussion and Conclusions

Across the 81 reviewed studies, LLMs enter supply-chain decision-making in three distinct ways: interpreting unstructured documents for supplier evaluation and risk monitoring, formulating or explaining mathematical models in planning and logistics, and generating direct commercial offers or replenishment orders (RQ1). Most applications do not replace operations research methods. They pair prompting or retrieval with multicriteria scoring, mathematical programming, or simulation. Negotiation agents are the exception, since they embed the model directly in bilateral bargaining (RQ2). These workflows speed up evidence preparation and model formulation, and they also bring operational risks: dropped physical constraints, a slight quality loss against manual auditing [66], and shifts in profit division across trading partners [8] (RQ3). An LLM in supply-chain management therefore needs to be judged on more than linguistic fluency. Evaluations should check which operational constraints are preserved and how the resulting gains are distributed across the network.

7.1. Theoretical and Managerial Implications

For operations management theory, our synthesis extends prescriptive analytics and decision-focused learning, which evaluate predictions by the quality of the decisions they induce [2,138–140], to an upstream specification stage: when an LLM translates verbal requests or field reports into solver code, language interpretation defines the mathematical problem itself [30,34]. A solver can return a mathematically optimal solution that is operationally invalid if the translation omitted a shop-floor or capacity constraint. Information-processing theory and task–technology fit [53,54] set the same condition. An LLM adds operational value only when its compression of unstructured text preserves the structural rules governing the decision.

Across firm boundaries, technical design choices are also governance choices. System prompts and retrieved contract examples steer bargaining behavior, so an LLM workflow that raises joint channel surplus can still leave one partner with a smaller share [8,40]. Models of technology adoption and supply-chain contracting therefore need to specify who controls the prompt and retrieval context and how surplus is divided, in addition to an aggregate productivity parameter. Figure 14 brings these mechanisms together: in supplier assessment [7,59] and planning [6,30] (upper path), the LLM prepares inputs or formulations that an analytical procedure evaluates before a manager commits; in automated bargaining [8,40] (lower path), the model's output goes directly to a trading partner whose strategic response co-determines the outcome.

For supply-chain managers, this distinction governs how much decision authority to delegate to an LLM and where to place human verification gates. In procurement, a team can use an LLM to extract bid attributes and keep the final vendor trade-offs in human hands. In production planning, natural-language model modifications can be restricted to sandboxed simulation until constraint fidelity is verified. Balancing augmentation and automation [141,142] requires calibrating human reliance to actual system reliability [134,143]. Review effort should target errors that would alter an operating commitment, and auditing time should be weighed against the hours saved in document preparation [66]. Managers often over-rely on fluent advice or discount algorithmic recommendations unevenly [135,144,145]. Conversational explanations in digital twins or planning tools [63]

should therefore expose the underlying constraints and source records instead of only summarizing outputs. Organisations can then track which errors human reviewers actually catch before they expand the model's autonomy.

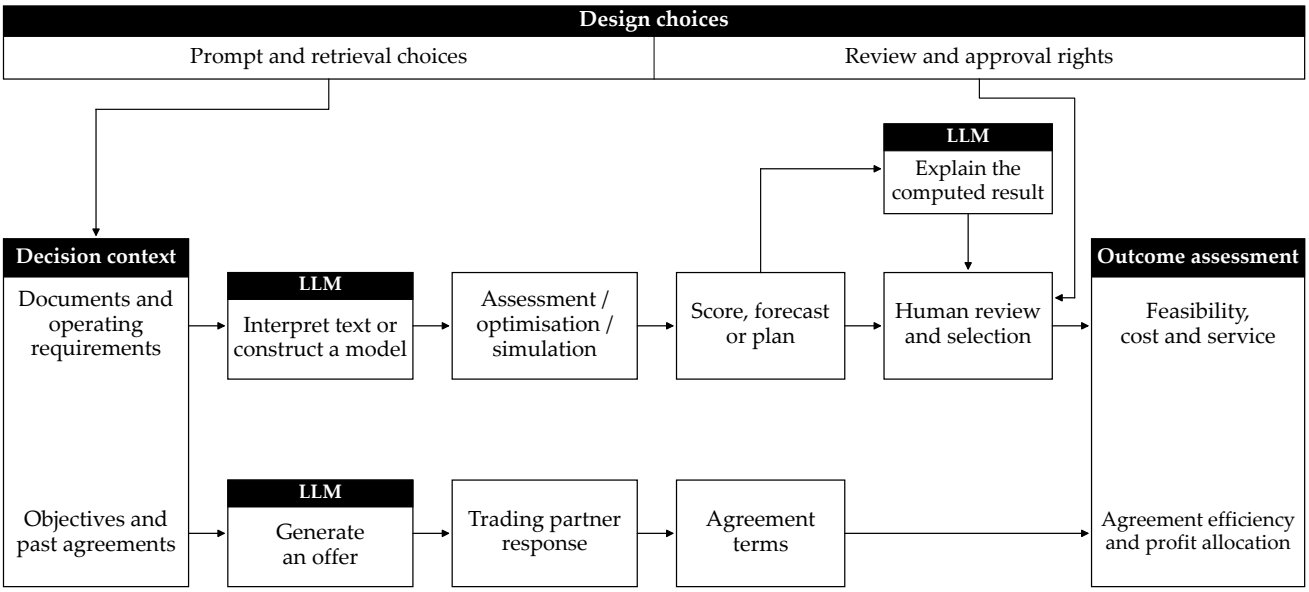


Figure 14. LLM contributions within supply-chain decision processes.

7.2. Scope and Limitations

Several limitations of our review scope should be noted. By restricting our search to English-language articles in 40 AJG-ranked operations research and operations management journals, we focused on peer-reviewed studies anchored in supply-chain decision problems; technical computer-science proceedings and industry white papers may document newer model architectures before they reach journal publication. Most operational studies in our sample also evaluate prototypes on offline enterprise datasets or synthetic simulations instead of longitudinal field deployments. Their reported gains therefore reflect specific dataset conditions and model snapshots. Longitudinal field studies are needed to test how these workflows hold up as operating environments shift and commercial foundation models are updated.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/systems14111478/s1, Table S1: Study-level classifications (81 studies); Table S1 (continued): Theory and research context; Table S2: Evidence appraisal of operational-application studies (46 studies); Table S3: Topic-model stability and parameter checks; Table S4: Publication profile and first-author affiliation; Table S5: Authors contributing at least two included articles; Table S6: Citations to the included articles; Table S7: Author-keyword frequency and occurrence by issue year; Section S1: Search strategies.

Author Contributions: Conceptualization, G.C. and Y.H.; methodology, G.C. and Y.H.; software, G.C.; formal analysis, G.C.; data curation, G.C.; writing—original draft preparation, G.C.; writing—review and editing, G.C. and Y.H.; visualization, G.C.; supervision, Y.H.; project administration, Y.H.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (Nos. 72571062 and 72171047) and HKUST-SEU Joint Research Seed Fund (No. 5014002601).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Choi, T.M.; Wallace, S.W.; Wang, Y. Big Data Analytics in Operations Management. *Prod. Oper. Manag.* **2018**, *27*, 1868–1883. [CrossRef]
2. Bertsimas, D.; Kallus, N. From Predictive to Prescriptive Analytics. *Manag. Sci.* **2020**, *66*, 1025–1044. [CrossRef]
3. Baryannis, G.; Validi, S.; Dani, S.; Antoniou, G. Supply chain risk management and artificial intelligence: State of the art and future research directions. *Int. J. Prod. Res.* **2019**, *57*, 2179–2202. [CrossRef]
4. Noy, S.; Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **2023**, *381*, 187–192. [CrossRef]
5. Brynjolfsson, E.; Li, D.; Raymond, L. Generative AI at Work. *Q. J. Econ.* **2025**, *140*, 889–942. [CrossRef]
6. Jackson, I.; Saénz, M.J.; Ivanov, D. From natural language to simulations: Applying AI to automate simulation modelling of logistics systems. *Int. J. Prod. Res.* **2024**, *62*, 1434–1457. [CrossRef]
7. Sadeghi, N.; Dehghani, M.; Kabadayi, N. The AI-driven Decision-Making (AIDM) Framework: Integrating AHP and ChatGPT for Supplier Selection. *Ann. Oper. Res.* **2026**, *359*, 2593–2615. [CrossRef]
8. Kirshner, S.N.; Pan, Y.; Wu, J.X.; Gould, A. Talking terms: Agent information in LLM supply chain bargaining. *Decis. Sci.* **2026**, *57*, 9–23. [CrossRef]
9. Min, H. Artificial intelligence in supply chain management: Theory and applications. *Int. J. Logist. Res. Appl.* **2010**, *13*, 13–39. [CrossRef]
10. Toorajipour, R.; Sohrabpour, V.; Nazarpour, A.; Oghazi, P.; Fischl, M. Artificial intelligence in supply chain management: A systematic literature review. *J. Bus. Res.* **2021**, *122*, 502–517. [CrossRef]
11. Sharma, R.; Shishodia, A.; Gunasekaran, A.; Min, H.; Munim, Z.H. The role of artificial intelligence in supply chain management: Mapping the territory. *Int. J. Prod. Res.* **2022**, *60*, 7527–7550. [CrossRef]
12. Richey, R.G., Jr.; Chowdhury, S.; Davis-Sramek, B.; Giannakis, M.; Dwivedi, Y.K. Artificial intelligence in logistics and supply chain management: A primer and roadmap for research. *J. Bus. Logist.* **2023**, *44*, 532–549. [CrossRef]
13. Jackson, I.; Ivanov, D.; Dolgui, A.; Namdar, J. Generative artificial intelligence in supply chain and operations management: A capability-based framework for analysis and implementation. *Int. J. Prod. Res.* **2024**, *62*, 6120–6145. [CrossRef]
14. Song, Z.; Xie, Y.; Yang, L.; Zhao, Y. Large language models in supply chain management: A systematic literature review and application framework. *Int. J. Prod. Res.* **2026**, *64*, 7144–7184. [CrossRef]
15. Panja, S.; Zheng, T.; Glock, C.H. Potentials of generative AI in logistics and supply chain management: A guiding review on applications and future research directions. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
16. Bahroun, Z.; Saihi, A.; As'ad, R.; Tanash, M. A systematic analysis of generative artificial intelligence for supply chain transformation. *Supply Chain Anal.* **2026**, *13*, 100188. [CrossRef]
17. Paul, J.; Rosado-Serrano, A. Gradual Internationalization vs Born-Global/International new venture models: A review and research agenda. *Int. Mark. Rev.* **2019**, *36*, 830–858. [CrossRef]
18. Mentzer, J.T.; DeWitt, W.; Keebler, J.S.; Min, S.; Nix, N.W.; Smith, C.D.; Zacharia, Z.G. Defining supply chain management. *J. Bus. Logist.* **2001**, *22*, 1–25. [CrossRef]
19. Cachon, G.P.; Fisher, M. Supply Chain Inventory Management and the Value of Shared Information. *Manag. Sci.* **2000**, *46*, 1032–1048. [CrossRef]
20. Cachon, G.P.; Lariviere, M.A. Supply Chain Coordination with Revenue-Sharing Contracts: Strengths and Limitations. *Manag. Sci.* **2005**, *51*, 30–44. [CrossRef]
21. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
22. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. [CrossRef]
23. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
24. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
25. Jackson, I.; Saénz, M.J.; Ivanov, D.; Ma, B.J. Supply chain mapping through retrieval-augmented generation: Applications to the electronics industry. *J. Oper. Res. Soc.* **2026**, *77*, 1798–1818. Correction in *J. Oper. Res. Soc.* **2026**, *77*, i. [CrossRef]

26. Gezdur, A.; Bhattacharjya, J. Innovators and transformers: Enhancing supply chain employee training with an innovative application of a large language model. *Int. J. Phys. Distrib. Logist. Manag.* **2025**, *55*, 394–408. [CrossRef]
27. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations*; OpenReview.net: Amherst, MA, USA, 2023. Available online: https://arxiv.org/abs/2210.03629 (accessed on 12 November 2026).
28. Le, C.A.; Kannan, D.; Govindan, K.; Nguyen, T.T. Integrated large language models and digital twin-based decision support in closed-loop supply chain operations: Green steel transition case study. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
29. Amarasinghe, P.T.; Nguyen, S.; Sun, Y.; Arisian, S.; Alahakoon, D. Business optimization for digital manufacturing: A fine-tuned large language model approach. *Int. J. Prod. Econ.* **2026**, *295*, 109934. [CrossRef]
30. Sukharamwala, U.C.; Won, D. LLM4SCM: A framework for intelligent supply chain decisions. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
31. Chen, X.; Cheng, G.; He, Y. Mathematical Modeling and Optimization of Platform Supply Chain in the Digital Era: A Systematic Review. *Mathematics* **2025**, *13*, 2863. [CrossRef]
32. Jarquin-Segovia, R.; Marmolejo-Saucedo, J.A. Optimization Techniques for Improving Economic Profitability Through Supply Chain Processes: A Systematic Literature Review. *Mathematics* **2026**, *14*, 185. [CrossRef]
33. Zabraoui, O.; Hmamou, Y.; Chafi, A.; Alami, S.K. Conversational AI for inventory analytics: Leveraging large language models in supply chain management. *Comput. Ind. Eng.* **2026**, *220*, 112298. [CrossRef]
34. Li, Q.; Zhang, L.; Mak-Hau, V. An LLM-powered MILP modelling engine for workforce scheduling guided by expert knowledge. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
35. Nie, W.; Li, F.; Tsolakis, N.; Kumar, M. Integrating digitally enhanced data extraction and simulation modelling for AI-driven supply chain resilience: An operational research framework for strategic stockpiling of critical minerals. *J. Oper. Res. Soc.* **2026**, *77*, 1770–1797. [CrossRef]
36. Du, B.; Chen, G.; Xie, S.; Geng, D. Assessing Large Language Models for Decision Support in Novel Circular Supply Chains: A Case of Hotel Linen. *IEEE Trans. Eng. Manag.* **2026**, *73*, 2397–2425. [CrossRef]
37. Niu, B.; Lin, C.; Xie, F.; Zhang, J. Fine-tuning collaboration in a Co-opetitive manufacturing supply chain. *Int. J. Prod. Econ.* **2027**, *303*, 110207. [CrossRef]
38. Lei, G.; Wei, J.; Ma, J. Firm-level trade decoupling risk and cross-border supply chain reconfiguration: Evidence from Chinese firms. *Int. J. Oper. Prod. Manag.* **2026**, *46*, 1330–1354. [CrossRef]
39. Yi, Z.; Liang, Z.; Xie, T.; Fu, Y.; Liu, Y. Unveiling dynamics in agricultural supply chain: A transformer-enhanced framework for commodity price modeling. *Int. J. Prod. Econ.* **2026**, *294*, 109820. [CrossRef]
40. Herold, S.; Heller, J.; Rozemeijer, F.; Mahr, D. Brave new procurement deals: An experimental study of how generative artificial intelligence reshapes buyer–supplier negotiations. *J. Purch. Supply Manag.* **2025**, *31*, 101012. [CrossRef]
41. Dou, R.; Nan, G.; Pan, Y.; Liu, X. Balancing privacy considerations and customization preferences for consumer: LLMs adoption and coordination strategies in supply chains. *Int. J. Prod. Econ.* **2026**, *292*, 109827. [CrossRef]
42. Donthu, N.; Kumar, S.; Mukherjee, D.; Pandey, N.; Lim, W.M. How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res.* **2021**, *133*, 285–296. [CrossRef]
43. Tranfield, D.; Denyer, D.; Smart, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *Br. J. Manag.* **2003**, *14*, 207–222. [CrossRef]
44. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
45. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
46. Rethlefsen, M.L.; Kirtley, S.; Waffenschmidt, S.; Ayala, A.P.; Moher, D.; Page, M.J.; Koffel, J.B.; PRISMA-S Group. PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Syst. Rev.* **2021**, *10*, 39. [CrossRef]
47. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
48. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794. [CrossRef]
49. Fosso Wamba, S.; Queiroz, M.M.; Chiappetta Jabbour, C.J.; Shi, C. Are both generative AI and ChatGPT game changers for 21st-Century operations and supply chain excellence? *Int. J. Prod. Econ.* **2023**, *265*, 109015. [CrossRef]
50. Fosso Wamba, S.; Guthrie, C.; Queiroz, M.M.; Minner, S. ChatGPT and generative artificial intelligence: An exploratory study of key benefits and challenges in operations and supply chain management. *Int. J. Prod. Res.* **2024**, *62*, 5676–5696. [CrossRef]
51. Hendriksen, C. Artificial intelligence for supply chain management: Disruptive innovation or innovative disruption? *J. Supply Chain Manag.* **2023**, *59*, 65–76. [CrossRef]
52. Li, L.; Liu, Y.; Jin, Y.; Cheng, T.C.E.; Zhang, Q. Generative AI-enabled supply chain management: The critical role of coordination and dynamism. *Int. J. Prod. Econ.* **2024**, *277*, 109388. [CrossRef]

53. Galbraith, J.R. Organization Design: An Information Processing View. *Interfaces* **1974**, *4*, 28–36. [CrossRef]
54. Goodhue, D.L.; Thompson, R.L. Task-Technology Fit and Individual Performance. *MIS Q.* **1995**, *19*, 213–236. [CrossRef]
55. Aslam, J.; Saleem, A.; Lai, K.H. Supply Chain Management in the Era of Generative AI (ChatGPT): Technology Fit and Psychological Drivers of Adoption. *IEEE Trans. Eng. Manag.* **2025**, *72*, 3817–3831. [CrossRef]
56. Teece, D.J.; Pisano, G.; Shuen, A. Dynamic capabilities and strategic management. *Strateg. Manag. J.* **1997**, *18*, 509–533. [CrossRef]
57. Sonar, H.; Ghag, N.; Sharma, I. Bridging theory and practice in AI-driven supply chains: Prioritizing LLM adoption challenges and SCOR-based applications. *Int. J. Prod. Econ.* **2026**, *296*, 110008. [CrossRef]
58. Chowdhury, S.; Mangla, S.K.; Li, G. From augmentation to action: AI-LLM–enabled improvisation and reactive capability in digital manufacturing supply chains. *Int. J. Prod. Econ.* **2026**, *302*, 110163. [CrossRef]
59. McKinley, F.A.; Dohmen, A.E.; Castillo, V.E. Do Humans and GAI See Eye to Eye? Implications of LLM Scoring Volatility in Supplier Evaluations. *J. Bus. Logist.* **2026**, *47*, e70072. [CrossRef]
60. Qian, Z.; Peng, J.; Li, J. The Impact of Generative AI Announcements on Suppliers: Evidence From the Stock Market. *Prod. Oper. Manag.* **2026**, *35*, 2037–2053. [CrossRef]
61. Aspeteg, J.; Holtström, J.; Rehme, J.; Havemo, E. Supply chain resilience development in the defence industry: An AI-assisted content analysis of annual reports. *Supply Chain Manag. Int. J.* **2026**, *31*, 38–52. [CrossRef]
62. Matsuno, K.; Iizuka, H.; Chou, C.W.; Ohno, T. Intelligent scheduling in agricultural wholesale markets: An optimisation approach with simulation-driven and GenAI-based interaction support. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
63. Genetti, S.; Scarton, G.; Formentini, M.; Iacca, G. An intelligent Digital Twin based on machine learning for interpretable decision-making in manufacturing. *Int. J. Prod. Econ.* **2026**, *291*, 109841. [CrossRef]
64. Du, G.; Ding, Y.; Wang, Z.; Yue, X. Partial understanding or full comprehension: Should platform deploy large language model to bridge consumer perception biases? *Int. J. Prod. Econ.* **2026**, *301*, 110139. [CrossRef]
65. Chen, Y.; Kirshner, S.N.; Ovchinnikov, A.; Andiappan, M.; Jenkin, T. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? *Manuf. Serv. Oper. Manag.* **2025**, *27*, 354–368. [CrossRef]
66. Wang, J.; Liu, G.; Cheng, Y.; Xu, X.; Li, Z. Leveraging Internet-Sourced Text Data for Financial Analytics in Supply Chain Finance: A Large Language Model-Enhanced Text Mining Workflow. *IEEE Trans. Eng. Manag.* **2025**, *72*, 1924–1938. [CrossRef]
67. Chen, S.; Liu, Y. Illuminating Heuristic Design: Language-Model-Driven Quality-Diversity Evolution for Combinatorial Optimization. *IEEE Trans. Evol. Comput.* **2026**, *advance online publication*. [CrossRef]
68. Simchi-Levi, D.; Dai, T.; Menache, I.; Wu, M.X. Democratizing Optimization with Generative AI. *Oper. Res.* **2026**, *advance online publication*. [CrossRef]
69. Chen, S. Coupling-aware destruction with large-language-model guidance for energy-efficient blocking hybrid flow shop scheduling. *Comput. Oper. Res.* **2026**, *196*, 107645. [CrossRef]
70. Hamdan, S.; Brahimi, N. Evaluating large language models for complex supply chain optimisation: A focus on routing problems. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
71. Çetinkaya, İ.O.; Büyüktahtakın, İ.E.; Shojaee, P.; Reddy, C.K. Discovering heuristics with Large Language Models (LLMs) for mixed-integer programs: Single-machine scheduling. *Comput. Oper. Res.* **2026**, *186*, 107325. [CrossRef]
72. Xiang, W.; Zhang, J.; Zhang, P.; Wang, M. Dynamic scheduling of wafer batch processing machines via reinforcement learning enhanced by expert-guided lightweight LLM. *Int. J. Prod. Res.* **2026**, *64*, 2731–2752. [CrossRef]
73. Huang, J.; Zhang, H.; Liu, Q.; Li, X.; Gao, L. Automated design of scheduling heuristics for dynamic flexible job shop scheduling with evolutionary textual gradients. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
74. Li, R.; Wang, L.; Sang, H.; Yao, L.; Pan, L. LLM-Assisted Automatic Memetic Algorithm for Lot-Streaming Hybrid Job Shop Scheduling With Variable Sublots. *IEEE Trans. Evol. Comput.* **2026**, *30*, 1377–1389. [CrossRef]
75. Yu, F.; Gao, L.; Li, X.; Lu, C.; Liu, Q. Automated Scheduling Heuristic Generation and Evaluation via Large Language Model. *IEEE Trans. Evol. Comput.* **2026**, *advance online publication*. [CrossRef]
76. Xu, M.; Liu, J.; Ong, Y.S. EvoSpeak: Large Language Models for Interpretable Genetic Programming-Evolved Heuristics. *IEEE Trans. Evol. Comput.* **2026**, *advance online publication*. [CrossRef]
77. Zhao, F.; Wu, B.; Wang, L.; Sang, H. A Large Language Model-Assisted Reinforcement Learning Framework with Evolutionary Algorithm for Hybrid Flow-Shop Scheduling. *IEEE Trans. Evol. Comput.* **2026**, *advance online publication*. [CrossRef]
78. Weber, C.A.; Current, J.R.; Benton, W.C. Vendor selection criteria and methods. *Eur. J. Oper. Res.* **1991**, *50*, 2–18. [CrossRef]
79. Saaty, T.L. A scaling method for priorities in hierarchical structures. *J. Math. Psychol.* **1977**, *15*, 234–281. [CrossRef]
80. Zhou, L.; Shi, X.; Wang, Z.; Ma, C.; Gao, L. Exploration of applications with ChatGPT for green supply chain management. *Ann. Oper. Res.* **2025**, *355*, 1503–1536. [CrossRef]
81. Dang, J.; Della Vedova, J.; Yang, Y. The 1% problem: LLM orchestration for rare event detection in manufacturing supply chains. *Int. J. Prod. Econ.* **2026**, *300*, 110119. [CrossRef]
82. Lin, C.Y.; Tseng, T.L.; Xu, H. GPT-Augmented Bayesian Reinforcement Learning Framework for Multiobjective Supplier Selection. *IEEE Trans. Eng. Manag.* **2025**, *72*, 3779–3804. [CrossRef]

83. Liu, M.; Huang, R.; Liang, K.; Wu, H. Recovering store brand supply from disruption: AI-LLMs-enabled procurement collaboration. *Int. J. Prod. Econ.* **2026**, *297*, 110014. [CrossRef]
84. Lee, H.L.; Padmanabhan, V.; Whang, S. Information Distortion in a Supply Chain: The Bullwhip Effect. *Manag. Sci.* **1997**, *43*, 546–558. [CrossRef]
85. Lee, H.L.; So, K.C.; Tang, C.S. The Value of Information Sharing in a Two-Level Supply Chain. *Manag. Sci.* **2000**, *46*, 626–643. [CrossRef]
86. Oroojlooyjadid, A.; Snyder, L.V.; Takáč, M. Applying deep learning to the newsvendor problem. *IISE Trans.* **2020**, *52*, 444–463. [CrossRef]
87. Jannelli, V.; Schöpf, S.; Bickel, M.; Netland, T.; Brintrup, A. Agentic LLMs in the supply chain: Towards autonomous multi-agent consensus-seeking. *Int. J. Prod. Res.* **2025**, *advance online publication*. [CrossRef]
88. Zhou, Q.; Yang, Y.; Ma, F.; Cheng, T.C.E. Multi-agent deep reinforcement learning for ordering and inventory allocation in a decentralized two-echelon dual-channel supply chain. *Int. J. Prod. Econ.* **2026**, *299*, 110067. [CrossRef]
89. Zouaghi, I.; Beldjoudi, S.; Gunasekaran, A.; Fosso Wamba, S.; Sahrane, S. Balancing performance and innovation in AI-driven supply chains through temperature-scaled hallucination control. *Int. J. Prod. Res.* **2026**, *64*, 4341–4378. [CrossRef]
90. Niu, Y.; Wu, J.; Jiang, S.; Jiang, Z. The Bullwhip Effect in Servitized Manufacturers. *Manag. Sci.* **2025**, *71*, 1–20. [CrossRef]
91. Ivanov, D. Agentic digital twins: Bridging model-based and AI-driven decision-making support for a new era of supply chain and operations management. *Int. J. Prod. Res.* **2026**, *64*, 6574–6590. [CrossRef]
92. Bengio, Y.; Lodi, A.; Prouvost, A. Machine learning for combinatorial optimization: A methodological tour d'horizon. *Eur. J. Oper. Res.* **2021**, *290*, 405–421. [CrossRef]
93. Quan, Y.; Liu, Z.; Benaben, F.; Montreuil, B. Leveraging large language models to enhance multi-agent risk assessment in supply chain networks. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
94. De, A.; Kundu, T. Strategic Hub-Location Decisions for Semiconductor Supply Chains under War, Cyber, Geopolitical, and Network Disruptions. *IEEE Trans. Eng. Manag.* **2026**, *advance online publication*. [CrossRef]
95. Snyder, L.V.; Daskin, M.S. Reliability Models for Facility Location: The Expected Failure Cost Case. *Transp. Sci.* **2005**, *39*, 400–416. [CrossRef]
96. Khojasteh, Y.; Mousavi Jahan Abadi, S.M. Agentic Supply Chain Control Towers: A framework for autonomous and resilient operations. *Int. J. Prod. Econ.* **2027**, *303*, 110251. [CrossRef]
97. Aslam, M.W.; Zhang, Z.; Qasim, K.U. CORTEX-V: A cognitive reasoning toolkit for vision-based, cost-efficient layout optimization. *Comput. Oper. Res.* **2026**, *190*, 107413. [CrossRef]
98. Tang, C.S. Perspectives in supply chain risk management. *Int. J. Prod. Econ.* **2006**, *103*, 451–488. [CrossRef]
99. Tomlin, B. On the Value of Mitigation and Contingency Strategies for Managing Supply Chain Disruption Risks. *Manag. Sci.* **2006**, *52*, 639–657. [CrossRef]
100. Snyder, L.V.; Atan, Z.; Peng, P.; Rong, Y.; Schmitt, A.J.; Sinsoysal, B. OR/MS models for supply chain disruptions: A review. *IIE Trans.* **2016**, *48*, 89–109. [CrossRef]
101. Ivanov, D.; Dolgui, A. Viability of intertwined supply networks: Extending the supply chain resilience angles towards survivability. A position paper motivated by COVID-19 outbreak. *Int. J. Prod. Res.* **2020**, *58*, 2904–2915. [CrossRef]
102. AlMahri, S.; Xu, L.; Brintrup, A. Enhancing supply chain visibility with knowledge graphs and large language models. *Int. J. Prod. Res.* **2026**, *64*, 2178–2209. [CrossRef]
103. Li, X.; Ding, L.; Zhao, Y. An integrated approach to predict supply network resilience using LLM embeddings and multilayer perceptrons. *J. Oper. Res. Soc.* **2026**, *advance online publication*. [CrossRef]
104. Joglekar, N.; Nie, W.; Alem Fonseca, M.; Tsolakis, N.; Kumar, M. An AI-driven approach to assess sentiments and interpret context in a critical mineral supply chain. *Int. J. Prod. Res.* **2025**, *advance online publication*. [CrossRef]
105. Lin, K.Y.; Wu, S.Y.; Matsuno, K. Generative AI-Driven resilience in supply chain management: UNISONE framework for disruption modelling and capacity optimisation. *Int. J. Prod. Res.* **2025**, *advance online publication*. [CrossRef]
106. Fang, J.; Su, B.; Wang, S.; Wang, B. Uncovering the risks of digital supply chains: A large language model framework for semantic identification and validation. *Int. J. Prod. Econ.* **2026**, *291*, 109858. [CrossRef]
107. Huchzermeier, A.; Kuppek, M. Managing geopolitical risk in global supply chains: A qualitative study of European multinational enterprises. *Int. J. Prod. Econ.* **2027**, *303*, 110226. [CrossRef]
108. Lo, H.W. A novel risk assessment model for assessing generative AI risks in sustainable manufacturing: Insights from SME machinery firms. *Int. J. Prod. Econ.* **2027**, *303*, 110237. [CrossRef]
109. Zhu, X.; Li, M.; Xia, P.; Lei, H.; Peng, Z. A Novel Hybrid Price Prediction Method Using Multimodal Deep Learning and LLM for New Energy Power Raw Materials. *IEEE Trans. Eng. Manag.* **2026**, *73*, 1348–1362. [CrossRef]
110. Abolghasemi, M.; Ganbold, O.; Rotaru, K. Humans vs. large language models: Judgmental forecasting in an era of advanced AI. *Int. J. Forecast.* **2025**, *41*, 631–648. [CrossRef]

111. Abbaszadeh Nakhost, M.; Bloemer, A.; Minner, S. Full automation or just a copilot? A study of large language models on spare parts demand forecasting with OpenAI o1-mini. *Int. J. Prod. Res.* **2025**, *advance online publication*. [CrossRef]
112. Baharmand, H.; Behdani, B. Generative artificial intelligence and mathematical modelling for eco-friendly humanitarian supply chains. *Int. J. Prod. Res.* **2026**, *advance online publication*. [CrossRef]
113. Sorooshian, S.; Sarkis, J.; Mina, H. A taxonomy of biodiversity indicators for evaluating supply chain sustainability. *Supply Chain. Manag. Int. J.* **2026**, *advance online publication*. [CrossRef]
114. Tong, J.; Zhai, Y.; Cheng, T.C.E. LLM-enabled carbon emissions abatement in green production: Single or double hedging? *Int. J. Prod. Econ.* **2026**, *296*, 109933. [CrossRef]
115. Wang, S.; Zhang, H. Generative artificial intelligence and international green supply chain performance: Dynamic capabilities as strategic enablers. *Supply Chain. Manag. Int. J.* **2026**, *advance online publication*. [CrossRef]
116. Hou, Y.; Yang, S.; Li, L.; Chen, L. Unlocking Environmental Sustainability With Generative Artificial Intelligence: Insights From Resource Orchestration Theory. *IEEE Trans. Eng. Manag.* **2025**, *72*, 3080–3093. [CrossRef]
117. Tan, B.Q.; Kang, K.; Zhong, R.Y. When to use large language models for digital manufacturing in supply chains? *Int. J. Prod. Econ.* **2026**, *295*, 109932. [CrossRef]
118. Yu, S.; Zhang, M.; Wu, X.; Zhao, Z.; Huang, G.Q. Do large language models enlarge demand? Optimal channel structures and multi-agency mode in online retailing. *Int. J. Prod. Econ.* **2026**, *298*, 110048. [CrossRef]
119. Li, Z.; Li, Y.; Zhou, J.; Ye, F.; Zhan, Y. How AI-Generated Summaries of Reviews are Reshaping E-Commerce With Competing Retailers? *IEEE Trans. Eng. Manag.* **2026**, *73*, 1116–1131. [CrossRef]
120. Wang, Q.; Dai, X.; Li, M. Generative AI in retail platform operations: Considering supplier effort exertion and consumer skepticism. *Int. J. Prod. Econ.* **2026**, *292*, 109870. [CrossRef]
121. Li, L. Information Sharing in a Supply Chain with Horizontal Competition. *Manag. Sci.* **2002**, *48*, 1196–1212. [CrossRef]
122. Zhou, F.; Zhang, C.; Tiwari, S.; Huang, X.; Basu, P. Can Gen-AI promote community group buying? A tripartite evolutionary game analysis. *Int. J. Prod. Econ.* **2026**, *291*, 109721. [CrossRef]
123. Ning, Y.; Shi, Z.; Tong, Y. Value-based or one-time? Optimal pricing modes for generative AI services in e-commerce platforms. *Int. J. Prod. Econ.* **2026**, *292*, 109828. [CrossRef]
124. Zhang, Y.; Peng, B.; Wu, J. Integrating operations and finance for sustainable development: Theory, practice, and opportunities. *Prod. Oper. Manag.* **2026**, *advance online publication*. [CrossRef]
125. Sriraman, S.; Wuttke, D.; Babich, V.; Rosenzweig, E. Blockchain-Enabled Supply Chain Financing (BCF). *Prod. Oper. Manag.* **2026**, *advance online publication*. [CrossRef]
126. Niu, Y.; Qiao, X.; Wu, J.; Yang, X. Internal Credit and External Blame: Self-Attribution in Operations and Supply Chain Performance. *Manuf. Serv. Oper. Manag.* **2026**, *advance online publication*. [CrossRef]
127. Maghroor, H.R.; Madanchi, F.; O'Neal, T. Leveraging Generative AI for sustainable supply chain: Adoption challenges and strategic insights. *Prod. Plan. Control* **2025**, *advance online publication*. [CrossRef]
128. Qarqash, Z.; Putintseva, A.; Gruchmann, T.; Ivanov, D. Generative AI in supply chains: A qualitative study on decision-making and human-AI collaboration. *Prod. Plan. Control* **2026**, *advance online publication*. [CrossRef]
129. Trifone, N.; Dallari, F.; Brintrup, A. Drivers and constraints of agentic AI in supply chain management: An exploratory study. *Int. J. Prod. Res.* **2026**, *64*, 9080–9119. [CrossRef]
130. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
131. Bertsimas, D.; Sim, M. The Price of Robustness. *Oper. Res.* **2004**, *52*, 35–53. [CrossRef]
132. Delage, E.; Ye, Y. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Oper. Res.* **2010**, *58*, 595–612. [CrossRef]
133. Mohajerin Esfahani, P.; Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **2018**, *171*, 115–166. [CrossRef]
134. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef]
135. Vaccaro, M.; Almaatouq, A.; Malone, T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat. Hum. Behav.* **2024**, *8*, 2293–2303. [CrossRef]
136. Petropoulos, F.; Apiletti, D.; Assimakopoulos, V.; Babai, M.Z.; Barrow, D.K.; Ben Taieb, S.; Bergmeir, C.; Bessa, R.J.; Bijak, J.; Boylan, J.E.; et al. Forecasting: Theory and practice. *Int. J. Forecast.* **2022**, *38*, 705–871. [CrossRef]
137. Ivanov, D.; Dolgui, A. A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. *Prod. Plan. Control* **2021**, *32*, 775–788. [CrossRef]
138. Ban, G.Y.; Rudin, C. The Big Data Newsvendor: Practical Insights from Machine Learning. *Oper. Res.* **2019**, *67*, 90–108. [CrossRef]
139. Elmachtoub, A.N.; Grigas, P. Smart "Predict, then Optimize". *Manag. Sci.* **2022**, *68*, 9–26. [CrossRef]

140. Donti, P.; Amos, B.; Kolter, J.Z. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5484–5494. Available online: https://proceedings.neurips.cc/paper/2017/hash/3fc2c60b5782f641f76bcefc39fb2392-Abstract.html (accessed on 12 November 2026).
141. Jarrahi, M.H. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Bus. Horiz.* **2018**, *61*, 577–586. [CrossRef]
142. Raisch, S.; Krakowski, S. Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Acad. Manag. Rev.* **2021**, *46*, 192–210. [CrossRef]
143. Parasuraman, R.; Riley, V. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum. Factors* **1997**, *39*, 230–253. [CrossRef]
144. Dietvorst, B.J.; Simmons, J.P.; Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **2015**, *144*, 114–126. [CrossRef]
145. Logg, J.M.; Minson, J.A.; Moore, D.A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* **2019**, *151*, 90–103. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.