



Large language models in supply chain management: a systematic literature review and application framework

Zhe Song^a, Ying Xie^b, Lichao Yang^a and Yifan Zhao^a

^aFaculty of Engineering and Applied Sciences, Cranfield University, Cranfield, UK; ^bFaculty of Business and Management, Cranfield University, Cranfield, UK

ABSTRACT

Modern supply chains face unprecedented complexity, volatility, and data heterogeneity, which challenge the effectiveness of traditional decision-making tools. Large language models (LLMs) offer a promising new approach to intelligent supply chain transformation with their contextual reasoning and semantic generalisation capabilities. However, existing research on LLMs in supply chain management (SCM) remains fragmented and exploratory, lacking a unified framework to guide theoretical development and practical deployment. Through a systematic literature review, the research identifies state-of-the-art applications of LLMs across SCM activities and proposes a structured framework for LLM-SCM applications. Guided by theoretical support and the Context–Mechanism–Outcome framework, the proposed framework maps LLM capabilities to the five core processes of the supply chain operation reference (SCOR) model, highlighting specific intervention points and application pathways. It demonstrates how LLMs can support resilient and insight-driven planning, ethical and sustainable sourcing, collaborative and traceable smart making, resilient and experience-driven delivering, and transparent and empathetic return. The framework not only enhances conceptual clarity on the role of LLMs in SCM but also provides methodological guidance for future research and practice. By aligning with the values of Industry 5.0, including resilience, human-centricity, and sustainability, the framework contributes to advancing intelligent and adaptive supply chain systems.

KEYWORDS

Large language models; supply chain management; SCOR model; industry 5.0; systematic literature review; application framework

1. Introduction

Supply chain management (SCM) is a critical component of organisational strategy and has a direct impact on the competitive advantage of the enterprise (Ketchen and Giunipero 2004). However, SCM is inherently complex, characterised by non-linear dynamics, multi-stakeholder coordination, and high uncertainty, particularly under the compounding effects of geopolitical disruptions, demand volatility, and sustainability imperatives (Aylak 2025; Ivanov et al. 2023). Traditional supply chain analytics and automation systems often fall short in flexibility and semantic reasoning, especially when dealing with unstructured, cross-domain, or real-time information. In this context, large language models (LLMs) offer a new computational paradigm, combining deep contextual understanding with broad generalisation capabilities, that may transform how supply chains are designed, planned, and governed (Hu et al. 2025; Qu et al. 2024).

As Industry 5.0 advocates for systems that are human-centric, sustainable, and resilient (Nahavandi 2019), there is a growing demand for intelligent technologies that

can support such values. LLMs, built upon the Transformer architecture (Vaswani et al. 2017), have rapidly emerged as general-purpose AI systems with strong capabilities in understanding, reasoning, and generation (Brown et al. 2020). Their integration into enterprise applications creates new opportunities for augmenting decision-making, automating knowledge-intensive tasks, and enabling adaptive intelligence at scale (Bommasani et al. 2022). Unlike Industry 4.0, Industry 5.0 emphasises collaborative intelligence, socio-environmental responsibility, and AI systems that augment rather than replace human capabilities (European Commission 2021; Nahavandi 2019). Within this paradigm, the integration of LLMs into SCM is not merely a technological upgrade but also a strategic response to Industry 5.0's guiding principles.

An increasing number of studies have begun to apply LLMs to various activities in SCM, such as product quality inspection (Knott, Perez-Cruz, and Defraeye 2023), supplier identification (Li, Ko, and Ameri 2025), and risk assessment (Fan, Wu, and Yang 2025). These studies

have demonstrated the potential of LLMs in processing unstructured data, enhancing semantic reasoning, and improving intelligent decision-making. However, most existing research remains exploratory, focusing on isolated use cases, small-scale experiments, and prototype systems. While some technical progress has been made, these studies often fail to align LLM applications with end-to-end supply chain processes, and the connection between model outputs and strategic objectives in SCM remains unclear. Besides, current research has yet to establish a theoretically structured framework to help researchers systematically and effectively integrate LLM competencies with specific management challenges and their desired target outcomes. Therefore, this study conducted a systematic literature review with the aim of establishing a framework of LLM application paths in SCM, which can be used to explore the potential application opportunities, implementation pathways, and development trends of LLM's competencies in various SCM processes. This review addresses two questions:

- *What motivates the adoption of LLMs in SCM?*
- *What are the impacts of LLM applications on SCM*

To answer the above research questions, this study first constructs a novel structured competency matrix of LLMs, with a particular focus on how key architectural innovations, especially five Transformer-based variants, are translated into context-specific functional capabilities. This innovative conceptual mapping provides the technical foundation for understanding the alignment between model design and practical task requirements, which is an area that has been rarely addressed in prior SCM-focused LLM research. Building on this foundation, a systematic literature review is conducted, guided by Dynamic Capabilities Theory (DCT) and Context–Mechanism–Outcome (CMO) framework, to identify the motivations, applications, and impacts of LLMs within current SCM research. The resulting framework presents a new, systematised, process-oriented pathway that, grounded in the principles of Industry 5.0, supports the ethical, sustainable, resilient, and human-centric adoption of LLMs across supply chain processes. This pathway is expected to assist both academic and industrial communities in advancing the responsible integration of LLMs into future-oriented SCM.

The main contributions of this study are as follow:

- **Mapping the competency matrix of LLMs:** This study constructs a Transformer-based LLM competency matrix that systematically links architectural innovations, particularly five representative variants, with their functional potential in specific scenarios. This mapping offers a conceptual foundation for

understanding how technical design translates into application feasibility.

- **Proposing a unified framework to guide responsible LLM integration:** Based on a systematic literature review across SCOR processes and informed by DCT and CMO framework, the LLM-SCM framework is developed to provide a process-oriented and theoretically grounded structure for deploying LLMs in SCM. This framework focuses on the research motivation, application methods and impact of LLM applications in SCM.
- **Identifying future application pathways and research challenges:** Grounded in the proposed framework and the principles of Industry 5.0, this study outlines practical pathways for applying LLMs within ethical, sustainable, resilient, and human-centric supply chains. It also identifies current technical and managerial barriers, offering strategic directions for future research.

The rest of this paper is structured as follows. Section 2 outlines the evolution of LLMs, their architectural variants, and a competency matrix. Section 3 presents a systematic review of LLM applications in SCM, covering motivations, methods, and impacts. Section 4 introduces the LLM-SCM application framework, mapping application pathways across SCOR processes in line with Industry 5.0. Key findings and limitations are presented in Section 5.

2. Large language models

This subsection first introduces the development history and architectural foundations of LLMs. It then categorises their Transformer-based variants in accordance with performance demands and downstream use cases. Finally, it contextualises these models through a structured capability mapping to illustrate how they address SCM challenges.

2.1. Model development & basic architecture

2.1.1. Model development

Since the late 2000s, large language models (LLMs) have progressed through several critical milestones that have gradually expanded their applicability to complex decision-oriented domains such as supply chain management. In the early phase (~1990–2017), neural models such as Word2Vec (Mikolov et al. 2013) and Long Short-Term Memory (LSTMs) (Hochreiter and Schmidhuber 1997) improved textual representation but remained limited in modelling long-range dependencies, restricting their usefulness for end-to-end supply chain analysis involving interconnected processes. A major turning

point came in 2017 with the introduction of the Transformer architecture (Vaswani et al. 2017), which enabled scalable modelling of long-range dependencies and laid the technical foundation for applying language models to complex, multi-stage supply chain tasks. Between 2020 and 2021, large-scale pretraining further enhanced LLMs' generalisation ability (Brown et al. 2020; Devlin et al. 2019; Radford et al. 2019), enabling zero- and few-shot learning that reduced the need for task-specific data in downstream SCM applications such as forecasting and planning. Subsequent alignment techniques, such as instruction tuning (Ouyang et al. 2022), improved the reliability and controllability of LLM outputs, which is particularly important in SCM contexts where erroneous recommendations may lead to operational or contractual risks. Since 2024, frontier models such as GPT-4 (OpenAI 2023) have begun to demonstrate structured reasoning, multimodality, and tool-use, marking the shift toward autonomous agents capable of planning and decision-making.

2.1.2. LLM architecture

Contemporary LLMs predominantly rely on Transformer-based architectures (Vaswani et al. 2017), which enable scalable reasoning over long and heterogeneous textual inputs commonly found in supply chain operations. The encoder and decoder architecture of Transformers enables them to integrate information across documents, temporal dimensions, and organisational boundaries. A detailed structural diagram is provided in Appendix A along with Figure A1. The flexibility of the Transformer design enables modifications that support specialised functions, making it the basis for nearly all contemporary language models used in enterprise settings, including those deployed for procurement, logistics, and resource planning.

2.2. Transformer variants and their relevance to SCM problems

To adapt to specific operational constraints and enterprise tasks, Transformer models have undergone targeted architectural or functional improvements. These variants support a range of enterprise functions such as long-text processing, computational efficiency, and multimodal integration. Table 1 maps five representative Transformer variants to the language model capabilities they enable and the corresponding applications in supply chain management. This mapping provides practical insight into how model design supports real-world decision-making and automation tasks in logistics, sourcing, and production operations. Additionally, recent advances from the computer science community have also explored multi-agent LLM architectures that

enable coordinated reasoning and task decomposition (Wu et al. 2024), providing important methodological foundations for complex decision-making scenarios relevant to supply chain coordination. Appendix B summarises five popular Transformer variants tailored for distinct requirements, along with representative models and their key optimisations for each variant.

Each variant offers enhancements that extend the standard Transformer architecture to new management contexts. The following sections describe the technical rationale for each variant while highlighting its practical role in supporting SCM workflows.

Figure 1 illustrates the progressive role of transformer-based LLMs in SCM, from basic text comprehension to advanced multimodal integration. Each stage is annotated with the corresponding transformer types, reflecting their technological evolution and suitability for increasingly complex SCM tasks. Moving from basic to advanced applications. At the foundation, text-focused transformers enable simple text comprehension tasks such as contract analysis and invoice classification. As capability develops, transformers for reasoning and optimisation and task adaptation models extend this to forecasting, analysis, and decision generation, supporting functions like demand prediction, stock adjustment and replenishment planning. With greater complexity, long-context transformers enable auto-simulation, allowing warehouse modelling and route planning. At a more integrated level, long-context and optimisation transformers support intelligent systems that handle negotiation and act as supply chain assistants. Finally, multimodal transformers provide a linkage between language and visual data, combining text and graphics for warehouse operations. Representative models of the five variants can be found in the Appendix C.

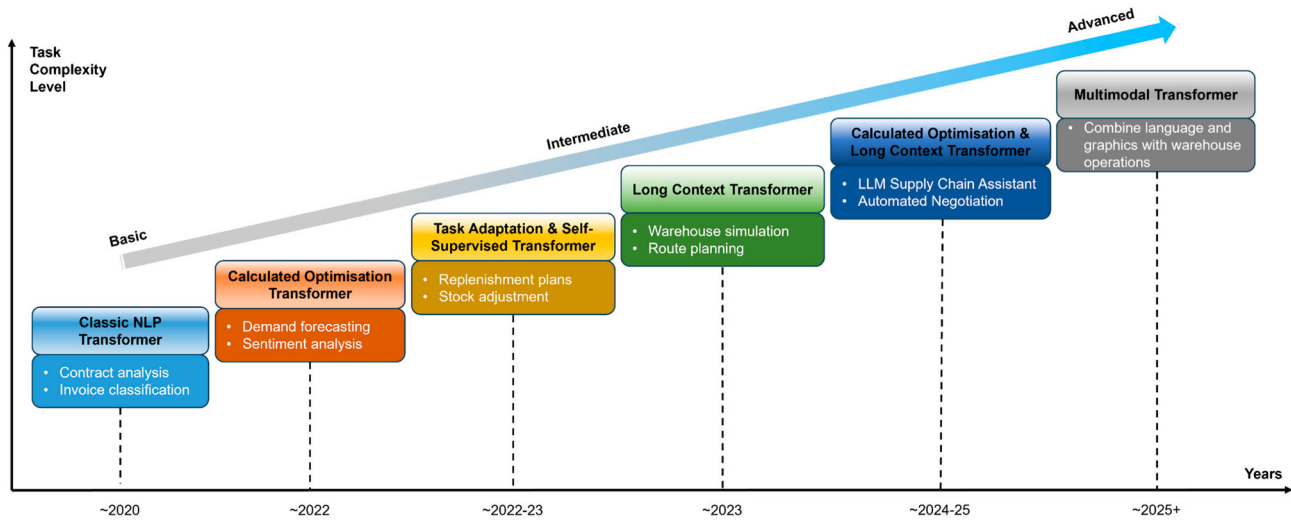
2.3. LLM competency matrix

To provide a unified understanding of LLM competencies in supply chain contexts, this study maps five representative Transformer variants to a structured five-dimensional competency framework. The text-focused Transformer lays the architectural foundation and corresponds to the Model Architecture and Technical Features dimension, relevant for basic SCM tasks such as contract parsing and invoice classification. Transformers for Reasoning and Optimisation improve training efficiency and data processing by enabling faster and more cost-effective model fine-tuning. This technical advantage lowers the resource threshold for adapting LLMs to downstream SCM applications, such as demand forecasting or cost-sensitive sourcing, where rapid iteration and task-specific customisation are essential. Task Adaptation Transformers contribute to core

Table 1. Transformer in SCM.

LLM architecture	Stages	LLM Architecture capability	Supply chain application	Reference
Text-focused Transformers	Basic	<ul style="list-style-type: none"> Natural Language Understanding and Generation 	<ul style="list-style-type: none"> Extracting insights from contracts, invoices, emails, or shipping documents Automated reporting, customer communication, or supplier queries 	Chiu, Tai, and Chu (2024); Shishehgharkhaneh et al. (2024)
Transformers for Reasoning and Optimisation	Intermediate	<ul style="list-style-type: none"> Logical and Mathematical Reasoning Capability Multilingual Capability 	<ul style="list-style-type: none"> Demand forecasting, cost estimation, and inventory optimisation using logic models Cross-border logistics support through multilingual processing 	Hu et al. (2025); Taiwo, Yussif, and Zayed (2025)
Task adaptation Transformer	Intermediate	<ul style="list-style-type: none"> Code Generation and Debugging Capability Personalisation and User Adaptability Tool Invocation and Plugin Integration Capability 	<ul style="list-style-type: none"> Tailored decision support for procurement, planning, or inventory control Automating workflows in supply chain platforms (e.g., writing SQL or scripts) 	Caetano, Oliveira, and Ramos (2025); Mishra and Senapati (2025)
Long Context & Transformer	Advanced	<ul style="list-style-type: none"> Multi-turn Dialogue and Context Memory Retrieval-Augmented Generation (RAG) Capability 	<ul style="list-style-type: none"> Analysing risks in multi-step logistics or production planning 	Wang et al. (2025b)
Multimodal Transformer	Advanced	<ul style="list-style-type: none"> Multimodal Support Multilingual Capability 	<ul style="list-style-type: none"> Visual inspection of goods, damage detection from images OCR and analysis of scanned supply chain documents in multiple languages 	Fan and Chiu (2024); Satirapiwong and Siriborvornratanakul (2021)

Note: The categories presented in this table are conceptual labels developed in this study to synthesise different transformer-based architectural tendencies in the SCM literature.

**Figure 1.** Evolution of Transformer Variants in SCM Applications.

language understanding and task generalisation, forming the basis of the Competency Dimensions, and support diverse applications like dynamic replanning or multilingual supplier matching. Multimodal Transformers, through their ability to process heterogeneous inputs (e.g., images and text), raise considerations in the Safety and Ethical Dimensions, especially for visual inspection in manufacturing or reverse logistics classification. Finally, Long Context Transformers enable scalability and extended reasoning, mapping to the Application and Deployment Dimensions, such as multi-echelon planning or end-to-end logistics simulation. The full matrix is provided in Appendix D.

The five dimensions of information related to LLM competencies, and their influencing factors, are shown in Figure 2:

Based on the analysis of the five Transformer-based variants and the competency matrix presented earlier, it is evident that:

- Model architecture and technical features, along with training and data characteristics, determine LLM competencies and their relative strengths. For example, advanced architectures such as Long Context Transformers enable LLMs to process extended procurement contracts, supplier agreements, or production planning documents, thereby improving information traceability and coordination across supply chain functions.
- Training and data characteristics, as well as LLM competencies, influence the model's security and ethical

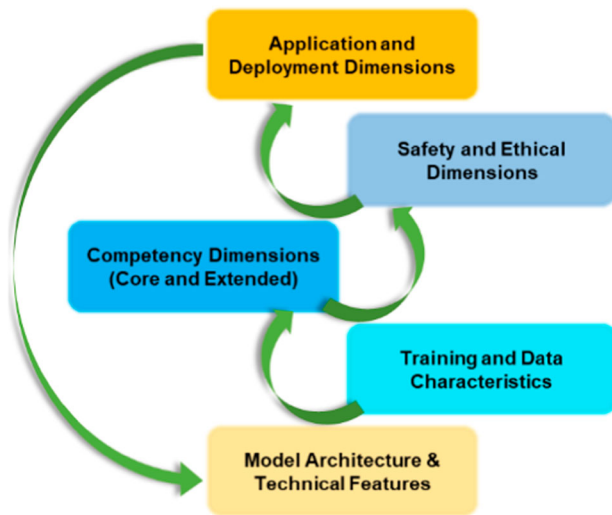


Figure 2. Five feature dimensions of LLM with associated relationships.

performance. In sourcing and reverse logistics processes, this affects the model's ability to mitigate biased supplier evaluations or prevent hallucinated classifications, which could otherwise lead to regulatory, contractual or compliance risks.

- LLM competencies, security, and ethical considerations collectively impact LLM deployment and application. During logistics execution, for instance, explainability, robustness and human oversight become critical when deploying LLMs to support autonomous delivery, disruption management, or multilingual negotiation with logistics partners.
- Feedback from deployment and use in operational settings inform ongoing improvements to LLM architecture and training strategies. In supply chain contexts, user interaction logs from planning assistants, digital twins, or warehouse simulations, provide evidence that can guide iterative fine-tuning, performance optimisation and context specific adaptation of LLMs.

The interdependence of these dimensions implies that decisions about LLM deployment in supply chain settings must be grounded not only in model performance but also in the specific operational context, regulatory environment, and integration demands of the application.

While Figure 2 does not directly represent the task-specific competencies of large language models, it outlines the progression of key technical traits that influence design, safety, and deployment strategies. These evolving features have important implications for practical SCM use cases. Section 3 further explores how different LLM characteristics translate into tangible outcomes across supply chain processes.

3. Systematic literature review on LLMs in SCM

By synthesising DCT with LLM applications in SCM via the SCOR model and operationalising this linkage through a CMO logic, this literature review positions LLMs as key technological enablers of dynamic capabilities, which underpin the explanation of the motivation for adopting LLMs. The review illustrates how AI reshapes traditional supply chain processes by embedding intelligence, agility, and responsiveness.

3.1. Theoretical support

3.1.1. Dynamic capabilities theory (DCT)

DCT provides a valuable perspective for understanding how organisations develop, deploy, and renew capabilities in response to rapidly changing environments. Originally introduced by Teece et al. (1997), the theory outlines how firms can gain a sustainable advantage through continuous adaptation, particularly under conditions of uncertainty. At the core of DCT are three interrelated capacities: sensing, seizing, and transforming. These represent an organisation's ability to detect environmental changes, respond with timely and effective actions, and reconfigure internal assets to maintain competitiveness.

The relevance of DCT to SCM has been widely acknowledged in recent literature. Supply chains today operate in contexts marked by demand volatility, geopolitical disruptions, technological innovation, and sustainability pressures (Ivanov 2022). In such settings, firms are expected to exhibit not only operational efficiency but also strategic agility. DCT has thus been used to explain how firms build resilience, recover from disruptions, and engage in continuous innovation (Blome, Schoenherr, and Rexhausen 2013; Wieland 2021). For example, sensing capabilities allow supply chains to anticipate risks by monitoring supplier performance and customer sentiment (Amellal et al. 2024; Wan et al. 2021). Seizing enables firms to make informed decisions in real time, such as rerouting logistics or adjusting sourcing policies (Painuly and Sharma 2024). Transforming refers to the ability to reconfigure supply chain structures, workflows, and partnerships, often through digital or data-driven innovation (Singh and El-Kassar 2019).

While DCT has been applied to explain a range of supply chain strategies, its potential to interpret technology-enabled transformation remains underexplored. This study proposes that the adoption of LLMs represents a class of digital innovation that can directly enhance an organisation's dynamic capabilities. LLMs support sensing by processing unstructured market data, identifying patterns, and detecting early signals of disruption. They enable seizing by facilitating decision-making through

Table 2. Mapping DCT to SCM processes and LLM contributions.

DCT component	SCM process	Enabling mechanism	LLM contributions	Reference
Sensing	Plan, Source	Sensing market changes, scenario planning	LLMs can analyse vast market and demand data to generate forecasts, simulate scenarios, and support strategic planning.	Wan et al. (2021); Zhang, Xia, and Xie (2024)
		Supplier sensing and selection, contract adaptation	LLMs can extract insights from supplier performance data and automate compliance review of contracts.	
Seizing	Make, Deliver	Resource reconfiguration, process innovation	LLMs support adaptive scheduling, real-time troubleshooting, and operational documentation.	Hu et al. (2025); Mishra and Senapati (2025)
		Customer engagement, logistics agility	LLMs improve customer interaction (e.g., chatbots), optimise routing via natural language analysis of delivery notes, and support decision-making in disruptions.	
Transforming	Plan, Make, Return, cross-process	Learning from failures, reverse logistics innovation, introduce text information, replaces manual testing.	LLMs facilitate root cause analysis and the development of return policies through the synthesis of unstructured data. Enhance enterprise risk awareness. Restructure product development capabilities and speed.	Mishra and Senapati (2025); Sliz (2024); Wang et al. (2025a)

scenario modelling, demand prediction, and adaptive control. They also contribute to transforming by allowing for the automation of knowledge workflows, semantic coordination across teams, and reconfiguration of operational routines. In this study, these DCT capabilities are not treated as abstract organisational attributes; instead, they are activated through specific LLM-enabled mechanisms, which are subsequently analysed using a CMO lens.

3.1.2. SCOR framework in supply chain management

The SCOR model, developed by the Supply Chain Council in the United States (now part of APICS), is a framework for modelling SCM processes. It is designed to help organisations analyse, improve, and optimise their supply chain operations (Stewart 1997). The SCOR model divides supply chain processes into five main management processes: Plan, Source, Make, Deliver, and Return (Stewart 1997). Analysing supply chain operations through the SCOR model allows companies to systematically understand the connection between technological value and business needs (Thilakarathna, Dharmawardana, and Rupasinghe 2015). With this model, the application of LLMs can be integrated into each process step, clarifying the motivation, specific uses, and practical impacts. This approach helps firms align advanced AI technologies with their operational processes in a standardised way, supporting the transformation towards an intelligent supply chain (Prakash 2013).

3.1.3. Mapping DCT to LLM-enabled SCM processes

To better illustrate the theoretical fit between DCT and the application of LLMs in SCM, Table 2 presents a mapping across the five core supply chain processes defined in the SCOR model. Each cell indicates how

LLMs activate specific dynamic routines to enhance process agility and responsiveness.

3.2. Systematic literature review methodology

This study conducted a systematic literature review (SLR) in accordance with the PRISMA 2020 guidelines (Page et al. 2021) to ensure transparency and reproducibility. The literature search was carried out in April 2025 across two major academic databases: Scopus and Web of Science (WoS). The keyword selection was guided by the research aim of exploring how LLMs are applied across SCM activities. Core terms such as “transformer”, “large language model”, “generative artificial intelligence”, and “ChatGPT” were selected to capture the breadth of LLM-related technologies. Synonyms and variations,– including acronyms like “GPT”, plural forms like “models”, and broader terms like “NLP*”,– were included to enhance retrieval sensitivity and ensure comprehensive coverage of the LLM landscape.

In parallel, supply chain-related terms such as “supplier”, “demand forecasting”, “inventory management”, and “supply chain*” were chosen to represent key functional areas of SCM aligned with the SCOR model. These terms reflect the scope of the review, which focuses on core operational processes where LLMs may provide value. The combination of these two keyword domains allows the search strategy to identify relevant interdisciplinary literature at the intersection of AI and SCM.

A total of 1,490 records were initially retrieved,– 1,081 from Scopus and 409 from WoS. Records were limited to peer-reviewed articles, published in English, between 2020 and 2025, resulting in 541 records. After removing 243 duplicates, 298 unique records were retained for title, keyword, and abstract screening. Following this

screening, 280 articles were sought for full-text retrieval. Eighteen records could not be retrieved, leaving 262 articles for eligibility assessment. After removing 187 studies that were irrelevant (e.g., focused on unrelated topics such as nonlinear programming or high-voltage transformers), 93 records remained.

A full-text assessment was then performed, during which 57 studies were excluded due to a lack of focus on LLMs (e.g., papers focusing purely on transformer architectures without discussing language models). Ultimately, 36 studies were included in the full text assessment. In addition to the initial database search, a snowballing strategy was applied to ensure broader coverage of the relevant literature. Both backward snowballing (screening the reference lists of included studies) and forward snowballing (examining papers citing the included studies) were conducted. Through this process, an additional 25 studies were identified and included. These studies met the same predefined inclusion and exclusion criteria as those used in the main search, ensuring that they were relevant and of comparable quality. The inclusion of these records was therefore not intended to increase the number of articles artificially but to strengthen the comprehensiveness and robustness of the review. The literature search process is illustrated in Figure 3.

3.3. Descriptive statistics

Figure 4 shows the distribution of selected studies across the five core processes in the SCOR model. The results indicate that existing research is more concentrated in the Plan and Make stages, more specifically, focusing on demand forecasting and intelligent manufacturing. In contrast, the Source and Return processes remain underexplored, highlighting potential opportunities for future work.

Figure 5 illustrates the distribution of LLM competency types covered in the literature. Most studies focus on core competencies such as natural language understanding and logical or mathematical reasoning. Some have started to explore more complex competencies, including multi-turn dialogue, retrieval-augmented generation, and multimodal integration. This indicates that the role of LLMs in supply chain tasks is gradually shifting from a basic tool-based function to a more cognitive one.

Figure 6 presents the annual distribution of included studies from 2020 to 2025. The results reveal a sharp increase in publications beginning in 2023, with a notable peak in 2024. This upward trend reflects the growing scholarly attention to the intersection of large language

models and supply chain management, particularly following the public release and rapid evolution of generative AI models. Although 2025 shows a slight decline, this is likely due to the timing of the literature search, conducted in April 2025. The trend indicates the timeliness of this review and indicates the emergence of a rapidly developing research domain.

The reviewed studies are published across a diverse set of journals, highlighting the interdisciplinary nature of LLM-enabled supply chain management research. As summarised in Table 3, the *International Journal of Production Research* represents the most frequent outlet within the reviewed corpus, followed by *IEEE Access* and a small number of journals focusing on sustainability and intelligent decision-support. Beyond these primary outlets, publications are widely dispersed, suggesting that this research stream is still emerging and spreading across multiple academic communities.

The methodological profile of the reviewed literature is dominated by analytical modelling and optimisation-based research. These studies typically develop forecasting, inventory, routing, or scheduling models enhanced by LLM or Transformer-based techniques and evaluate performance through numerical experiments or simulation.

Simulation and computational modelling form the second largest group, using agent-based models, digital twins, or LLM-enabled simulation generation to analyse dynamic supply chain behaviour. Empirical studies based on real-world data remain limited but are increasing, particularly in risk measurement, forecasting, and text-based intelligence applications. Conceptual and framework-oriented research has begun to emerge, while survey-based and mixed-method studies remain scarce.

From a geographical perspective, the reviewed studies are primarily authored by researchers based in China, Europe, and North America. This concentration indicates that LLM-enabled SCM research is currently driven by regions with strong artificial intelligence research capacity, advanced digital infrastructure, and manufacturing- or logistics-intensive supply chain environments. Contributions from other regions remain relatively limited, pointing to opportunities for future research across diverse institutional and regulatory contexts.

3.4. Thematic analysis

3.4.1. Context–mechanism–outcome framework

Originally developed within Realist Evaluation (Pawson and Tilley 1997), the CMO approach emphasises that an outcome only occurs when a particular mechanism is triggered within a specific context. Compared to

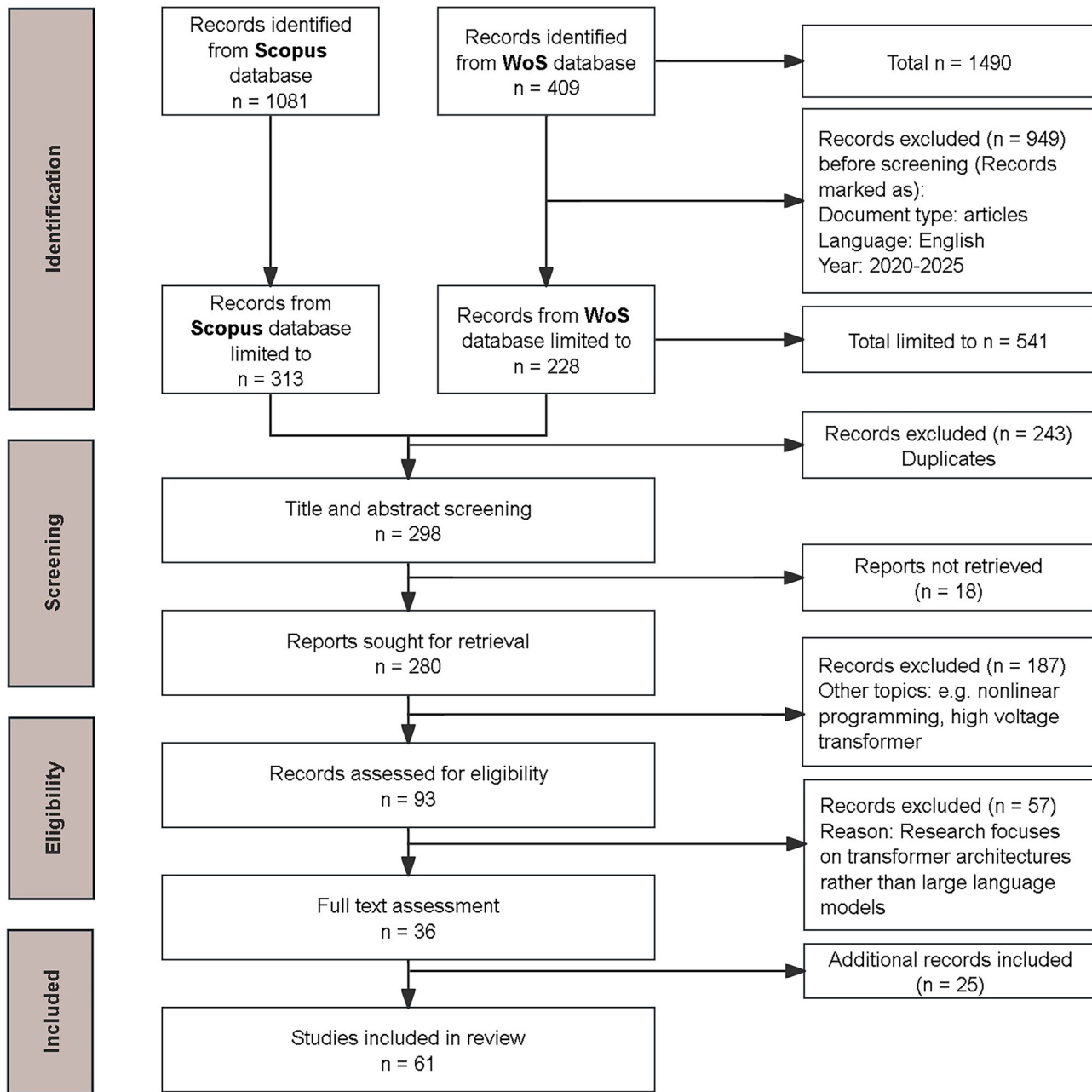


Figure 3. Literature search process.

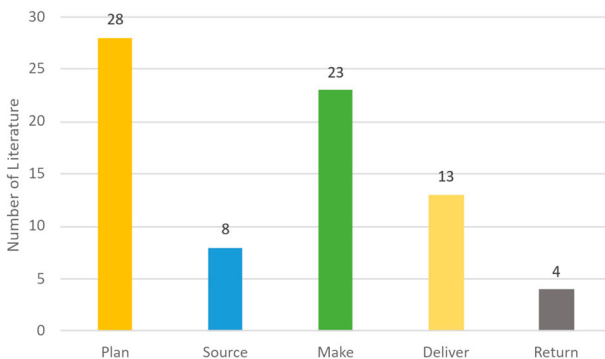


Figure 4. Literature distribution of the application of LLMs in the five SCM processes.

traditional input–output models, CMO is better suited to analysing the non-linear interactions between technology and social systems, especially in cases involving general-purpose AI technologies like LLMs operating in dynamic and uncertain supply chain environments.

In this review, the CMO framework is employed not as a substitute for DCT, but as a causal explanation logic that is explicitly aligned with the sensing, seizing, and transforming capabilities articulated in DCT. While DCT explains why organisations adapt through these capabilities, the CMO framework clarifies how such capabilities are enacted within specific supply chain settings.

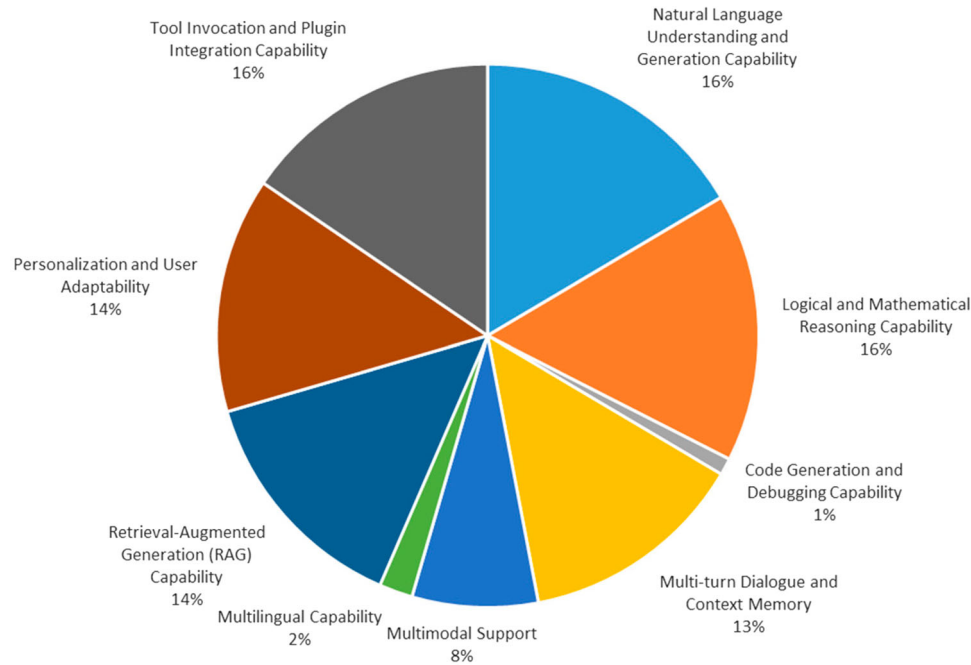


Figure 5. Distribution of LLMs competency in the literature on the application of LLMs in SCM.

Table 3. Descriptive characteristics of reviewed studies.

Dimension	Category	No. of studies	Key observations
Methodological approach	Analytical modelling	28	Primarily develop forecasting, inventory, routing, and scheduling models enhanced by LLMs or Transformer-based techniques.
	Simulation / computational modelling	14	Rely on simulation, agent-based modelling, and digital twins to analyse dynamic supply chain behaviours.
	Empirical study	9	Employ firm data, operational records, or textual data to validate LLM-enabled SCM applications.
	Conceptual / framework-based	6	Propose LLM-enabled SCM frameworks or agentic systems with limited empirical validation.
	Survey-based / mixed methods	4	Only a small number of studies examine adoption, perceptions, or organisational impacts using surveys or mixed methods.
Primary journals	<i>International Journal of Production Research</i>	8	Articles in this journal mainly focus on analytical and optimisation-oriented SCM problems supported by modelling and simulation.
	<i>IEEE Access</i>	6	Publications emphasise algorithm design, system implementation, and performance evaluation in applied SCM contexts.
	<i>Sustainability</i>	2	Apply LLMs to address sustainability-oriented SCM challenges such as green sourcing and low-carbon operations.
	<i>Expert Systems with Applications</i>	2	Focus on intelligent decision-support systems that enhance reasoning and decision quality in SCM.
	<i>Resources, Conservation and Recycling</i>	2	Apply LLM-related techniques to circular supply chain and resource-efficiency problems.
	<i>Annals of Operations Research</i>	2	Emphasise rigorous analytical modelling and optimisation to examine SCM coordination and decision mechanisms.
	Other journals (39 journals)	39	The remaining studies are widely dispersed across journals, reflecting the interdisciplinary and emerging nature of the field.
Geographical focus	China	27	A large share of studies originates from China, reflecting strong AI research activity and manufacturing-intensive SCM contexts.
	Europe	18	European studies frequently address logistics, sustainability, and coordination problems using LLM-enabled approaches.
	North America	12	Research from North America often focuses on data-driven SCM, retail applications, and knowledge-intensive decision support.
	Other / multi-country	4	Contributions from other regions remain limited within the reviewed corpus.

Specifically, the context (C) component corresponds to sensing, as it captures how supply chain conditions, problem framings, and research motivations shape the identification and interpretation of disruptions, uncertainties, and emerging signals. The mechanism (M) component aligns with seizing, as it explains how

LLM-enabled interventions, such as reasoning, retrieval-augmented generation, conversational interaction, and tool use, are mobilised to structure decision alternatives and support timely action under uncertainty. The outcome (O) component reflects transforming, as it captures how repeated and coordinated use of LLMs leads to

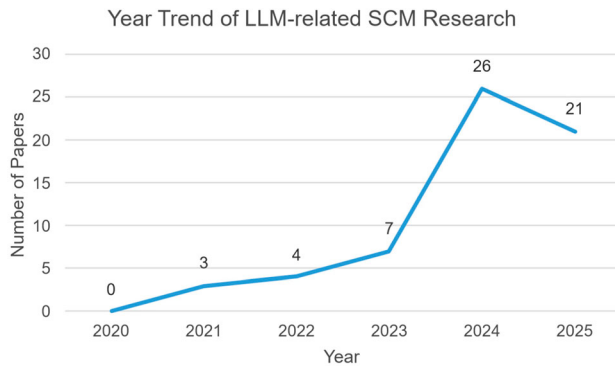


Figure 6. Year trend of LLM-related SCM research.

changes in routines, coordination practices, and process configurations over time.

Through this one-to-one alignment, the CMO framework provides the causal structure through which dynamic capabilities are operationalised in practice. Rather than attributing adaptation to a single mechanism, the framework shows how sensing, seizing, and transforming unfold as an integrated process shaped by contextual conditions, enacted through LLM-enabled mechanisms, and realised in observable outcomes.

3.4.2. Cluster mapping and thematic extraction

To explore the motivations, mechanisms, and outcomes associated with the adoption of LLMs in SCM, this study combines bibliometric text analysis with the CMO framework. Using VOSviewer, a keyword co-occurrence analysis was conducted on the selected corpus of 36 articles. Figure 6 presents the cluster diagram of selected articles showcasing important clusters. Smaller clusters around them show key linkages associated with (b) Supply Chain Management and (c) Large Language Model.

As shown in Figure 6(a), the keyword co-occurrence map reveals three distinct clusters, each representing a thematic focus aligned with specific SCOR process stages. The red cluster centres around terms such as supply chain management, artificial intelligence, and generative AI, highlighting how LLMs are being examined as overarching coordination tools, particularly in the Make stage. This reflects a strong emphasis on mechanisms such as AI-driven control and decision-making that are triggered within complex manufacturing contexts. The green cluster includes keywords like forecasting, transformer modelling, and decision support, indicating that LLMs are increasingly applied in Plan and Deliver stages for predictive modelling, risk sensing, and simulation. The blue cluster, anchored by terms like natural language processing, information extraction, and large language models, points to applications in semantic automation tasks such as supplier matching and returns processing, aligned with

Table 4. Research themes for LLMs in SCM process.

Theme	SCOR process	LLM contribution
Forecasting and simulation	Plan, Deliver	Predictive modelling, risk sensing, scenario generation
Semantic automation	Source, Return	Information extraction, supplier matching, after-sales processing
Cognitive coordination	Make	Scheduling, traceability, adaptive process control

the Source and Return stages. The directional links and density of connections between clusters also suggest a significant degree of overlap and thematic interaction, with concepts like decision-making appearing as central mechanisms across multiple contexts (Figure 7).

Based on the above analysis, the authors identified different potential themes through text mining and preliminary review of the retrieved articles. After several brainstorming sessions among the authors, the key thematic focuses were finally determined. These are: (1) Forecasting and Simulation, (2) Semantic Automation, and (3) Cognitive Coordination, which shows in Table 4.

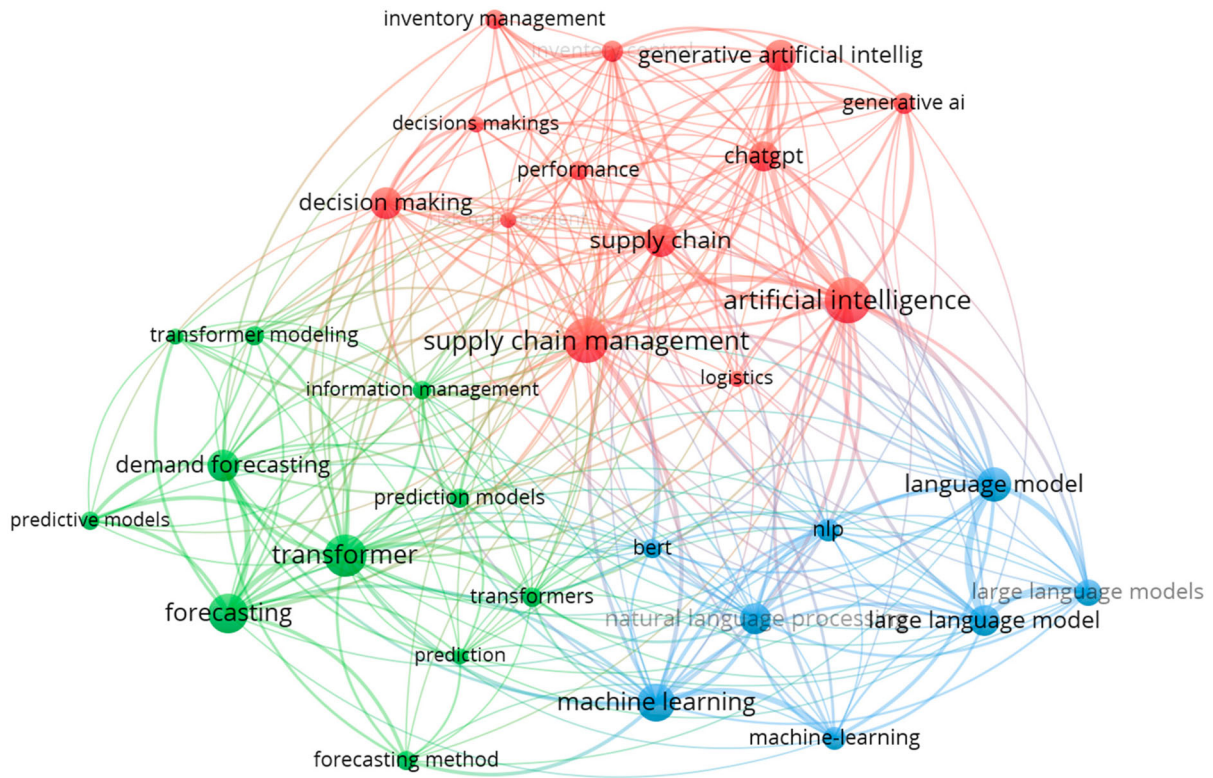
The first cluster is centred on forecasting, predictive models, and transformer-based learning, indicating that much research has been focused on the use of LLMs in the Plan and Deliver stages. The second cluster highlights the role of natural language processing and information extraction, pointing to applications in the Source and Return stages. The third cluster encompasses decision-making, scheduling, and AI-driven control, suggesting relevance to manufacturing coordination in the Make stage.

3.4.3. Theme 1: forecasting and simulation in plan and deliver

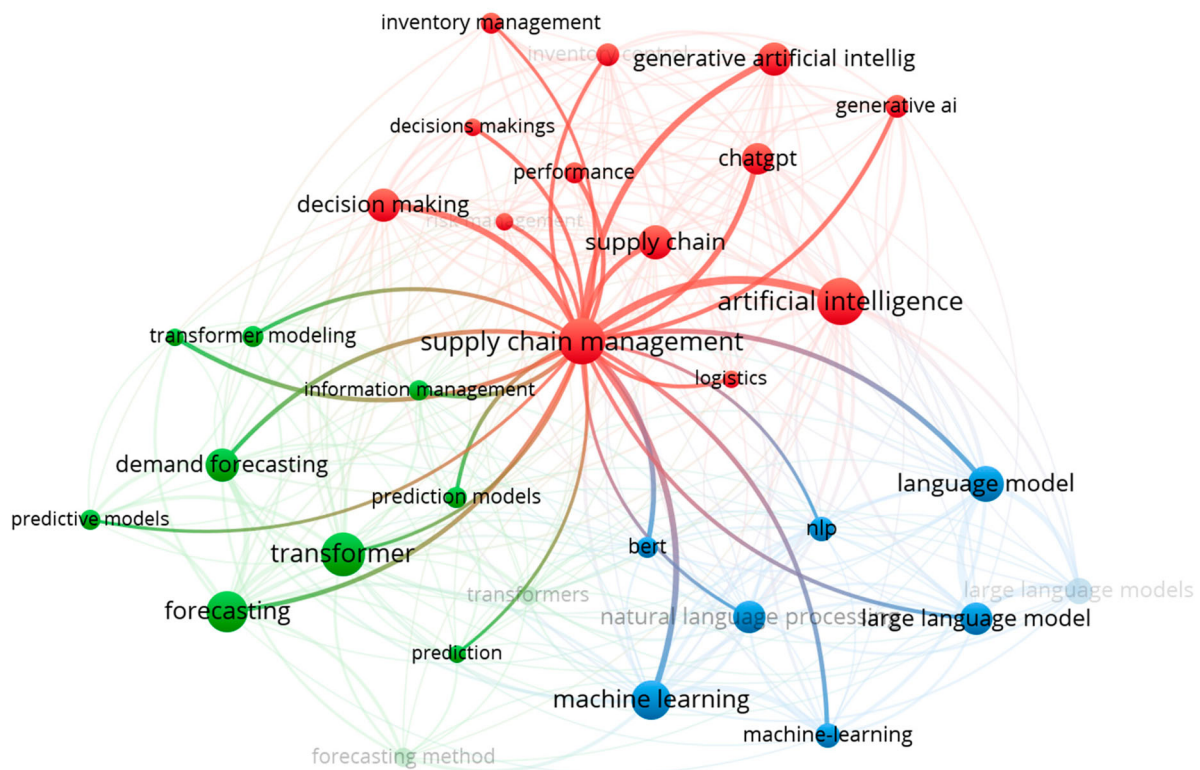
This theme reflects the widespread use of LLMs in enhancing planning capabilities and delivery coordination. It encompasses demand prediction, risk classification, and real-time scenario modelling. These functions align with the Plan and Deliver processes in the SCOR model.

3.4.3.1. Context: motivation for forecast enhancement.

In contemporary supply chains, forecasting inaccuracy and limited responsiveness remain key challenges. From a theoretical perspective, traditional forecasting in SCM often relies on static, rule-based models that lack flexibility when handling uncertainty or unstructured information (Cauli et al. 2024; Jackson, Jesus Saenz, and Ivanov 2024a; Long et al. 2025). This limits their value in dynamic planning environments. Practically, firms struggle with inaccurate demand forecasts, disconnected signals, and delays in scenario planning, especially in volatile markets (AlMahri, Xu, and Brintrup 2025; Ni

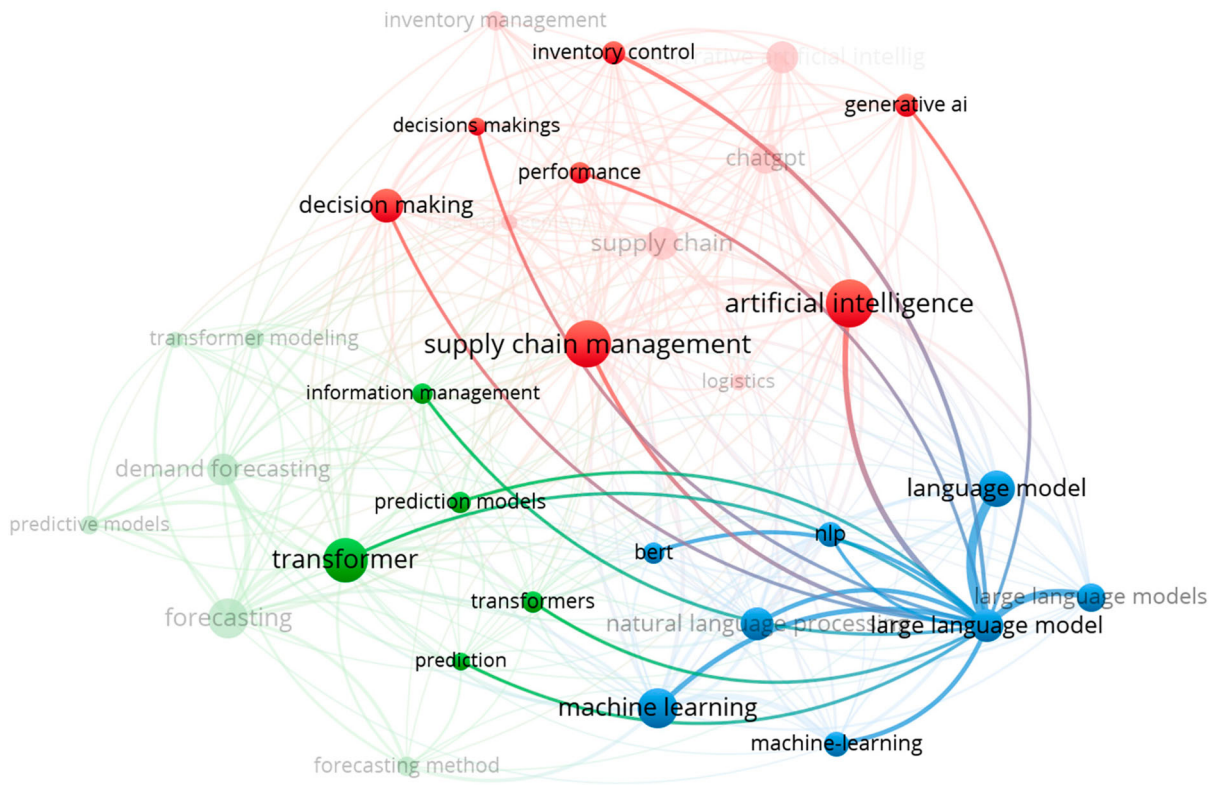


(a) Keyword Co-occurrence Network of LLM Research in SCM



(b) SCM-Centred Keyword Linkage Network

Figure 7. Cluster diagram of selected articles. (a) Keyword Co-occurrence Network of LLM Research in SCM. (b) SCM-Centred Keyword Linkage Network. (c) LLM-Centred Keyword Linkage Network.



(c) LLM-Centred Keyword Linkage Network

Figure 7. Continued.

et al. 2024). LLMs offer practical advantages by integrating diverse data sources and enabling contextual reasoning across planning and delivery functions. From an innovation perspective, recent research highlights the use of LLMs to support strategic simulation and predictive modelling that goes beyond statistical extrapolation (Hu et al. 2025). These models enable proactive planning and dynamic adjustment, reflecting a shift from reactive forecasts to intelligent foresight systems that continuously adapt to evolving supply chain conditions.

3.4.3.2. Mechanism: predictive and semantic techniques.

In planning and delivery contexts, LLMs function as analytical enablers that allow organisations to recognise emerging signals from complex data environments and translate them into timely planning and logistics decisions. In the planning phase, LLMs enhance a firm's sensing capability by analysing vast amounts of unstructured data, from legal texts to social media sentiment, to detect emerging market trends and risk signals (Fan, Wu, and Yang 2025; Wang et al. 2025a). These models seize strategic opportunities by enabling scenario-based simulations and supporting decisions under uncertainty, often through integration with agent-based AI or game-theoretic models (Mishra and Senapati 2025; Xin, Zhu,

and Peng 2024). Additionally, LLMs contribute to transforming static planning processes into adaptive, semantically intelligent systems that offer proactive risk mapping and pricing strategies at national and global levels (Breitung and Müller 2025; Li, Ko, and Ameri 2025). In the delivery process, LLMs are used to seize logistical decisions through multimodal routing optimisation that balances criteria such as time, cost, and emissions (Hu et al. 2025; Yazdinejad et al. 2023). They support sensing of real-time disruptions by parsing delivery notes, customer feedback, and geospatial data from IoT and remote sensing systems (Bao et al. 2024; Xie et al. 2025). These mechanisms help transform logistics from rigid to agile systems, where routing and scheduling can be dynamically adjusted to handle last-mile variability or crisis events.

3.4.3.3. Outcome: improved planning and delivery performance.

LLMs significantly improve forecasting accuracy by integrating structured and unstructured data for demand, pricing, and risk prediction (Caetano, Oliveira, and Ramos 2025; Zheng and Brintrup 2025). This leads to more responsive inventory strategies and better resource allocation, especially in volatile contexts such as disaster logistics and energy scheduling (Mishra and

Senapati 2025). At the strategic level, LLMs enhance cross-departmental coordination and risk detection by enabling shared semantic understanding across business functions (Wang et al. 2025a; Xin, Zhu, and Peng 2024). This supports the formation of semantically intelligent, forward-looking planning systems that are capable of navigating complexity with agility.

LLMs enhance logistical intelligence by improving route planning and enabling real-time monitoring of assets through visual and contextual data fusion (Bao et al. 2024; Xie et al. 2025). This enables more accurate and adaptive delivery schedules, particularly within last-mile logistics (Qu et al. 2024). By aligning logistical operations with real-time insights, LLMs facilitate a shift from point-based optimisation to system-wide logistics orchestration, contributing to both efficiency gains and carbon footprint reduction (Hu et al. 2025; Yazdinejad et al. 2023).

While the reviewed studies highlight the strengths of LLMs in enabling more adaptive and context-aware planning, several limitations remain insufficiently addressed. A notable concern is the frequent reliance on idealised or simulated datasets, which may not capture the complexity of real-time, multi-source supply chain signals. Additionally, many forecasting models fail to quantify uncertainty or provide confidence intervals, limiting their operational reliability in volatile environments. Furthermore, comparative studies with traditional statistical or hybrid AI approaches are rare, leaving open questions regarding the marginal value added by LLMs. These observations suggest that future research should move beyond proof-of-concept implementations and focus on robust, comparative, and context-sensitive evaluation frameworks.

3.4.4. Theme 2: semantic automation in source and return

This theme highlights the growing role of LLMs in automating language-intensive tasks in supplier management and after-sales processes. The focus is on semantic processing, compliance assessment, and intelligent service workflows. The associated SCOR stages are Source and Return.

3.4.4.1. Context: demand for data interpretation and responsiveness. In sourcing, organisations face challenges in identifying suitable suppliers, managing multilingual contracts, and ensuring regulatory compliance (AlMahri, Xu, and Brintrup 2025). Similarly, return processes often suffer from fragmented customer feedback and inconsistent service logic. Theoretically, sourcing and after-sales processes remain underexplored in terms of their semantic and interpretive complexity (Sliž 2024).

Most models overlook the cognitive effort required for supplier evaluation or customer feedback analysis. In practice, challenges arise from multilingual documents, unstructured contracts, and inconsistent return data (Aylak 2025), leading to inefficiencies in supplier selection and service handling. LLMs address these problems by enabling automated text understanding, information extraction, and context-aware decision support. From an innovation perspective, LLMs have begun transforming these functions by integrating with intelligent agents and adaptive rule systems (Li and Starly 2024). This enables personalised procurement recommendations and service workflows, allowing firms to move from transactional automation to cognitive and context-sensitive operations.

3.4.4.2. Mechanism: automated semantic matching and service adaptation. Within sourcing and return processes, LLMs enable organisations to move beyond information extraction toward actionable interpretation, allowing supplier intelligence and service insights to be absorbed into adaptive procurement and after-sales routines. In sourcing, LLMs demonstrate powerful seizing capabilities by extracting actionable insights from supplier profiles, environmental audits, and multilingual compliance reports using semantic evaluation techniques (Li, Ko, and Ameri 2025; Li and Starly 2024). This allows organisations to seize supplier intelligence faster and more transparently. By recommending low-carbon sourcing decisions based on historical data and green metrics (Aylak 2025), LLMs actively promote sustainable procurement. They also help transform traditional sourcing models by integrating real-time data fusion and value-based supplier evaluation into dynamic decision loops. Within the Return phase, LLMs contribute to sensing customer dissatisfaction or defects by interpreting natural language feedback and emotional tone, enhancing root-cause analysis (Sliž 2024). These insights are used to seize optimisation opportunities in warranty management and reverse logistics workflows. Through chatbots and automated service agents, LLMs also transform after-sales operations from reactive to proactive systems, where standardised, AI-driven workflows support consistent, efficient, and customer-aligned service delivery (Gezdur and Bhattacharjya 2025).

3.4.4.3. Outcome: enhanced sourcing and service capabilities. The outcomes of LLM adoption in these areas include greater sourcing transparency, faster supplier identification, and improved alignment with environmental goals. In sourcing, LLMs strengthen supplier discovery and evaluation through multilingual information

extraction and semantic matching, increasing both transparency and responsiveness (Aylak 2025). These capabilities are particularly impactful in expanding procurement access for SMEs and in embedding sustainability considerations into sourcing decisions (Li, Ko, and Ameri 2025; Li and Starly 2024). Firms are increasingly able to balance cost-effectiveness with environmental impact, driving a shift toward fair, intelligent, and low-carbon procurement ecosystems. LLMs are redefining the return process through the automation of service workflows, including complaint handling, warranty claims, and failure diagnostics (Gezdur and Bhattacharjya 2025). These improvements result in faster response times, lower service costs, and greater customer satisfaction (Raman et al. 2024; Sliz 2024). Through advanced natural language understanding, LLMs support a shift from reactive problem-solving to proactive service management, improving the resilience and consistency of reverse logistics systems.

Despite growing attention to the automation of sourcing and returns via LLMs, current studies tend to emphasise technical feasibility over practical viability. Most works operate under the assumption of high-quality, multilingual textual data but can rarely hold in fragmented or low-resource supply chain contexts. Moreover, semantic automation systems often lack explainability, raising concerns about trust and compliance in regulatory-sensitive industries. This highlights an urgent need for research that explores boundary conditions, interpretability, and the integration of LLMs with existing procurement and service infrastructures.

3.4.5. Theme 3: cognitive coordination in make

This theme addresses how LLMs are applied to support coordination, real-time monitoring, and rule standardisation in manufacturing environments. The applications are most prominent in the Make process of the SCOR model.

3.4.5.1. Context: need for adaptive scheduling and process control. Manufacturing processes often struggle with integration across departments, inflexible scheduling routines, and delayed response to unexpected events. From a theoretical viewpoint, manufacturing coordination has traditionally been addressed through deterministic scheduling and control frameworks (Oh 2024; Xie et al. 2025). These approaches often ignore adaptive learning and semantic reasoning. In practice, operational disruptions, last-minute changes, and fragmented information create persistent coordination problems (Hu et al. 2025). LLMs contribute by supporting context-driven scheduling, intelligent documentation, and real-time response generation. From an innovation angle,

LLMs are increasingly used in cognitive manufacturing environments, where they interact with sensors, visual data, and machine logs to guide process adjustments (Cauli et al. 2024; Jannelli et al. 2025; Knott, Perez-Cruz, and Defraeye 2023). This reflects a shift from rule-based execution to adaptive and autonomous coordination, aligned with dynamic capability principles such as sensing and reconfiguring in real time.

3.4.5.2. Mechanism: intelligent scheduling and real-time feedback processing.

In manufacturing environments, LLMs support coordination by continuously interpreting operational signals and embedding this understanding into adaptive scheduling and process reconfiguration. LLMs enhance sensing in the Make phase by enabling real-time monitoring of shop floor conditions, detecting defects through integration with Vision Transformers, and interpreting unstructured machine-generated text or image data (Cauli et al. 2024; Knott, Perez-Cruz, and Defraeye 2023). Through reinforcement learning and language-based planning, LLMs seize control of intelligent scheduling and environment-aware process regulation (Hu et al. 2025; Leite et al. 2025). These capabilities facilitate transformation by enabling systems that learn from anomalies, adapt workflows on the fly, and continuously optimise operational rules, marking a shift toward cognitive manufacturing.

3.4.5.3. Outcome: agile, transparent, and intelligent operations.

The implementation of LLMs in manufacturing results in improved traceability, higher quality assurance, and greater process transparency. LLMs are driving a transformation toward cognitive and autonomous manufacturing systems. Their impact is evident in areas such as real-time scheduling, quality assurance, and green production. When integrated with visual models, LLMs support automated component recognition (Cauli et al. 2024), defect detection (Hu et al. 2025; Ni et al. 2024), and traceability, enabling higher process transparency and compliance assurance (Oh 2024; Xie et al. 2025). At the operational level, LLMs reduce manual effort and increase system responsiveness, especially in environments characterised by fluctuating demand (Liu et al. 2023; Löwhagen, Schwendener, and Netland 2025).

Although cognitive coordination via LLMs is emerging as a transformative approach in manufacturing, several methodological and empirical gaps persist. Existing studies often focus on narrow use cases, such as documentation or defect detection, without exploring full-system integration or human-machine collaboration in real-time shop floors. Moreover, the adaptability of LLMs in highly time-sensitive or safety-critical scenarios remains largely untested. Another limitation is the

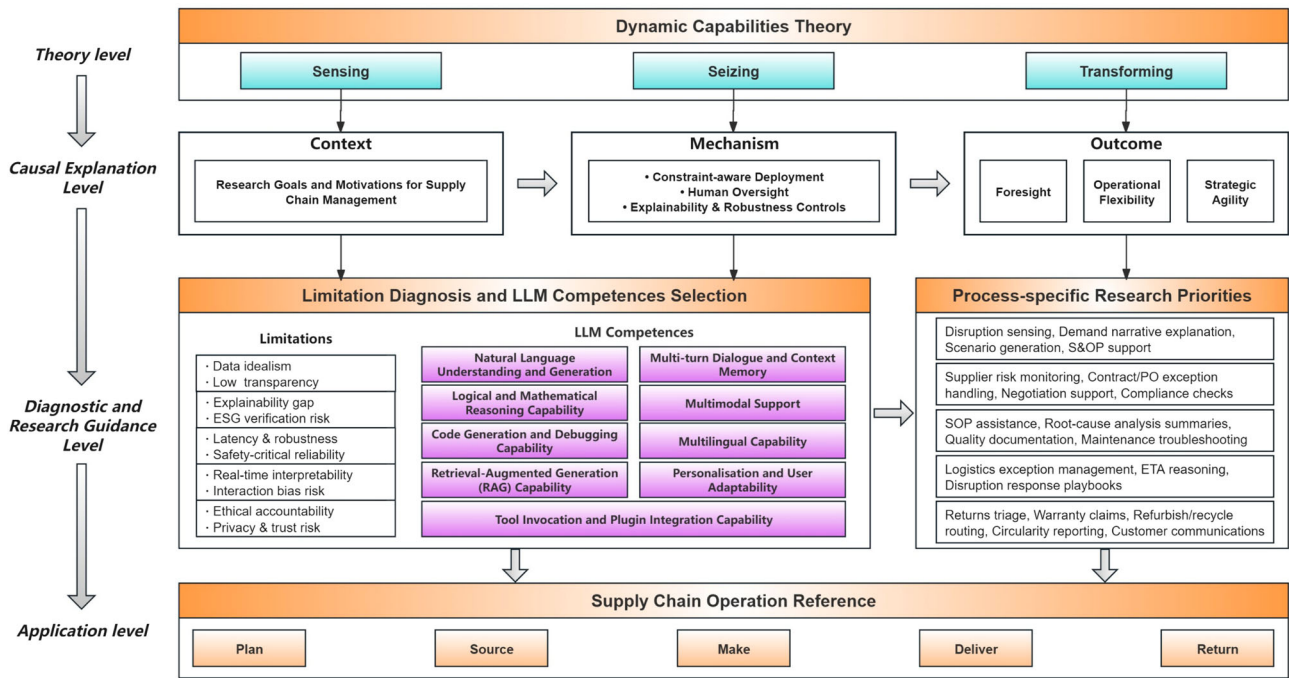


Figure 8. The proposed LLM-SCM framework.

lack of standard benchmarks for assessing scheduling intelligence, making reproducibility and comparison difficult. Future investigations should prioritise longitudinal studies in industrial environments, incorporate diverse sensor data, and address the challenges of latency, robustness, and system interoperability.

3.5. Conceptual framework

Table 5 presents a summary of the thematic analysis. Based on thematic analysis, this study proposes an integrated framework that explains how LLMs can enable dynamic capabilities and reveal critical limitations across supply chain processes to bridge the theoretical and operational gaps identified in current literature. As illustrated in Figure 8, the framework adopts a four-level architecture that links theory, causal explanation, diagnostic analysis, and operational application in a structured manner. The four levels comprise a theory level, a causal explanation level based on the CMO framework, a diagnostic and research guidance level, and an application level aligned with the SCOR supply chain processes.

The theory level is grounded in DCT and focuses on three core organisational capabilities: Sensing, Seizing, and Transforming. Within the context of LLM-enabled supply chain management, sensing provides guidance on how LLMs can support the identification and interpretation of disruptions, weak signals, and emerging patterns in supply and demand. Seizing relates to how insights generated by LLMs can be translated into timely and

informed decisions under operational constraints. Transforming captures the longer-term role of LLMs in reshaping routines, coordination mechanisms, and decision processes across supply chain activities.

These dynamic capabilities inform the causal explanation level, which is structured using the CMO logic. At this level, sensing primarily guides the identification and framing of research contexts (C), including research motivations and problem settings in supply chain management. Seizing informs the analysis of mechanisms (M), explaining how LLMs operate in practice and how their reasoning, retrieval, and interaction capabilities are enacted within supply chain decision processes. Transforming is reflected in the interpretation of outcomes (O), capturing how repeated and coordinated use of LLMs can lead to changes in performance, adaptability, and organisational practices.

Building on this causal explanation, the diagnostic and research guidance level combines insights from the context and mechanism components of the CMO framework to support the identification of key limitations in current LLM-enabled supply chain applications. Across SCOR, data idealism is most severe in Plan, Source, and Return due to fragmented and uncertain data; explainability is most severe in Plan, Make, and Deliver where decisions are high-stakes and must be defensible; and integration/governance limitations are most acute in Deliver and Enable where real-time execution and accountability are essential. At this level, diagnosed limitations are used to guide the selection of appropriate LLM competences

Table 5. CMO-based literature summary.

SCOR	Theory-driven	Context (Motivations) Practice-driven	Innovation-driven	Mechanism (Application)	Outcome (Impact)	Reference
Plan	Structure dependence	Forecast inaccuracy	Semantic-driven planning chain	Unstructured data-driven prediction	Enhanced prediction accuracy	Fan, Wu, and Yang (2025); Wang et al. (2025a); Mishra and Senapati (2025); Xin, Zhu, and Peng (2024); Breitung and Müller (2025); Li, Ko, and Ameri (2025)
				Multimodal fusion forecasting Strategy simulation & resource planning Risk classification & identification Crisis-responsive decision support Intelligent supplier matching	Optimised resource allocation Improved risk identification Strengthened system collaboration Multi-dimensional risk perception Improved supplier identification	
Source	Semantic absence	Supplier identification delays	Green intelligent sourcing	Green sourcing optimisation Compliance & risk detection Cybersecurity & data tracking Intelligent scheduling	Enhanced green procurement capacity Strengthened network & data security	
Make	Integration deficiency	Rigid scheduling	Smart manufacturing coordination	Process & environmental optimisation Product traceability & identification	Improved scheduling efficiency	Cauli et al. (2024); Knott, Perez-Cruz, and Defraeye (2023); Hu et al. (2025); Leite et al. (2025)
				Route planning & delivery optimisation	Enhanced traceability & security Process automation & intelligence Data fusion & quality control Adaptive rules & standardised workflows Improved logistics & inventory efficiency	
Deliver	Coordination blind spot	Complex delivery	Integrated delivery system	Inventory synergies & causal analysis	Enhanced service quality & satisfaction Reduced cost & inventory risk Optimised resource scheduling Optimised customer experience Enhanced process transparency	
Return	Framework disconnection	Lagging risk control	Cognitive support system	After-sales process optimisation Customer interaction & rule adaptation	Strengthened learning & talent development	Sliz (2024)

and to indicate feasible research priorities under different SCOR process conditions. These priorities represent the outcome (O) of the diagnostic process, translating analytical insights into actionable research guidance.

Finally, the application level situates the diagnostic and research guidance within the SCOR processes of Plan, Source, Make, Deliver, and Return. This level illustrates how limitation-aware LLM competences and research priorities can be applied across different supply chain processes, thereby linking theoretical reasoning and causal analysis to process-specific application contexts.

4. Application pathway and development trend of LLM in SCM

This section builds upon the proposed LLM-SCM framework to further explore the application pathways of LLMs in SCM, considering the three core strategies of Industry 5.0, – Human-Centric, Resilience, and Sustainability (TWI 2022). A corresponding pathway framework outlining this application roadmap is developed. In addition, based on thematic analysis and the constructed pathway framework, the study identifies key development trends for SCM empowered by LLMs.

4.1. LLM-SCM application pathway framework

As shown in the Figure 9, the future application pathway of LLMs in SCM can be systematically structured through a CMO framework. This structure offers a clear logic for aligning LLM deployment with supply chain transformation goals.

4.1.1. Step 1 – identify the problem and motivation (context)

Organisations must first clarify the types of challenges they aim to address. These include theory-driven limitations such as semantic gaps and coordination issues, practice-driven challenges like supplier identification delays and risk control lag, and innovation-driven pursuits such as cognitive support systems or semantic-driven planning chains. These motivations provide the grounding context for LLM adoption.

4.1.2. Step 2 – enable dynamic capabilities through LLMs (mechanism)

In response to the evolving demands of modern supply chains, marked by disruption, technological shifts, and global competition, organisations require adaptive and learning-oriented capabilities. Here, the Dynamic Capabilities Theory offers a foundation, emphasising the importance of sensing, seizing, and transforming. LLMs,

with their competencies in natural language understanding, multimodal reasoning, RAG, and multilingual interaction, can be integrated into SCOR processes to activate these dynamic capabilities. Specifically, LLMs are proving transformative in Planning, where they support scenario simulation and adaptive reconfiguration, helping firms build resilience in the face of uncertainty (Fan, Wu, and Yang 2025; Wang et al. 2025a). In Sourcing, models enable automated contract drafting, supplier evaluation, and even assist in multilingual negotiations, while also promoting human involvement through collaborative AI interfaces (Aylak 2025; Li and Starly 2024). For the Make process, LLMs, when integrated with IoT or digital twins, help manufacturers reconfigure workflows and respond dynamically, which benefits both continuity and sustainability (Bao et al. 2024; Xie et al. 2025). In Delivery, they offer tools for handling exceptions, refining communication with customers, and developing flexible fulfilment options, all of which feed into greater responsiveness (Bao et al. 2024; Xie et al. 2025). In the Return phase, LLMs assist with both customer-facing interaction and reverse logistics planning, helping organisations reduce waste and improve circularity (Sliz 2024). This synergy allows firms to sense weak signals across data sources, seize new opportunities through automation and augmentation, and transform decision-making practices at scale.

4.1.3. Step 3 – achieve future-oriented SCM outcomes (outcome)

Guided by the core values of Industry 5.0, including human-centric, resilience, and sustainability, organisations can steer LLM integration toward strategic outcomes. These include insight-driven planning, ethical sourcing, collaborative smart making, experience-enhanced delivery, and transparent return services. By embedding LLMs into the operational core, firms move beyond efficiency gains to build supply chains that are adaptive, responsible, and future-ready.

This three-step logic offers researchers and practitioners a coherent pathway to plan, prioritise, and implement LLM-driven transformation in supply chain management, supporting both immediate digital enhancement and long-term strategic alignment.

4.1.4. Case example

To better demonstrate the practical viability of this three-step pathway, we refer to a recent exploratory study (Jackson, Jesus Saenz, and Ivanov 2024b) that applied LLMs to automate simulation modelling in warehouse logistics. This case offers a representative prototype that aligns well with our LLM-SCM application pathway framework.

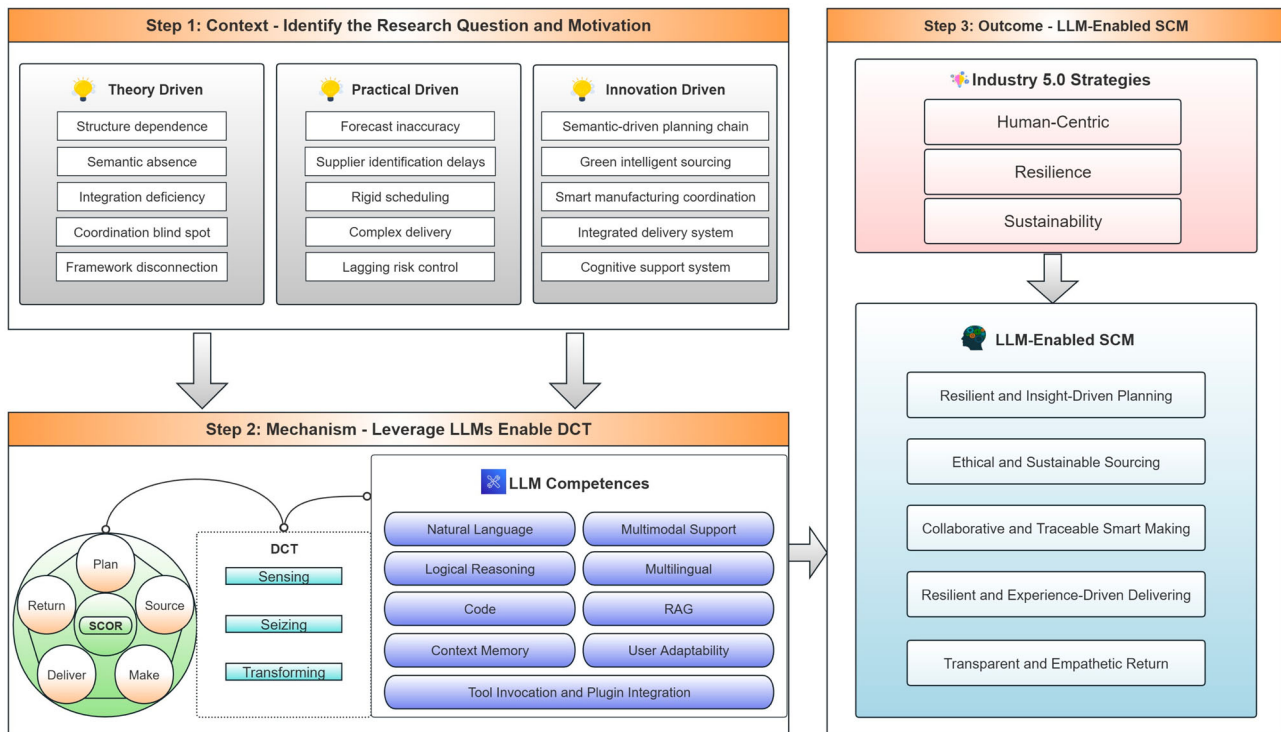


Figure 9. The proposed LLM-SCM application pathway framework.

In Step 1, the core challenge addressed in this study lies in the high technical barrier of building warehouse simulation models, a task that traditionally requires expert knowledge and manual configuration. This constraint limits the ability of organisations to rapidly experiment with alternative planning, sourcing, delivery configurations, or return flows in response to demand volatility, supply disruptions, or operational redesign. From a DCT perspective, this challenge reflects a limitation in organisational sensing and seizing capabilities, as decision-makers struggle to translate emerging operational signals into timely analytical insights.

In Step 2, the framework developed in the study operationalises the three core dynamic capabilities as follows:

- **Sensing:** The LLM ingests unstructured natural-language descriptions of logistics systems, such as storage layouts, inventory policies, picking rules, or queuing mechanisms, and interprets them using a GPT-based architecture. This enables the system to extract contextual knowledge directly from managerial or operational narratives, lowering the cognitive and technical barriers to system modelling. In SCM terms, this capability supports planning and sourcing decisions by rapidly capturing changes in demand patterns, supplier flows, or resource constraints.
- **Seizing:** Based on this interpreted input, the LLM generates executable Python code that instantiates a

discrete-event simulation model of the logistics system. The model is executed dynamically, producing performance metrics such as throughput, utilisation, lead times, and congestion. This capability enables rapid evaluation of alternative Plan–Make–Deliver configurations, allowing decision-makers to assess trade-offs and respond quickly to operational opportunities or disruptions without relying on lengthy model-development cycles.

- **Transforming:** In the transforming stage, human experts evaluate simulation outputs and iteratively refine the natural-language inputs to improve model fidelity or explore new scenarios. This human–AI feedback loop gradually reshapes organisational routines for analysis and decision support. Over time, logistics simulation evolves from an expert-centric activity into a language-driven, collaborative capability, enabling continuous reconfiguration of warehousing, fulfilment, and return processes. This represents a structural transformation in how analytical capabilities are embedded within supply chain operations.

In Step 3, the application yields outcomes that align closely with Industry 5.0 objectives. First, it empowers non-technical users to participate directly in modelling and decision-making processes (human-centric); Second, it enables rapid reconfiguration of storage layouts, picking strategies and return flows to respond to demand

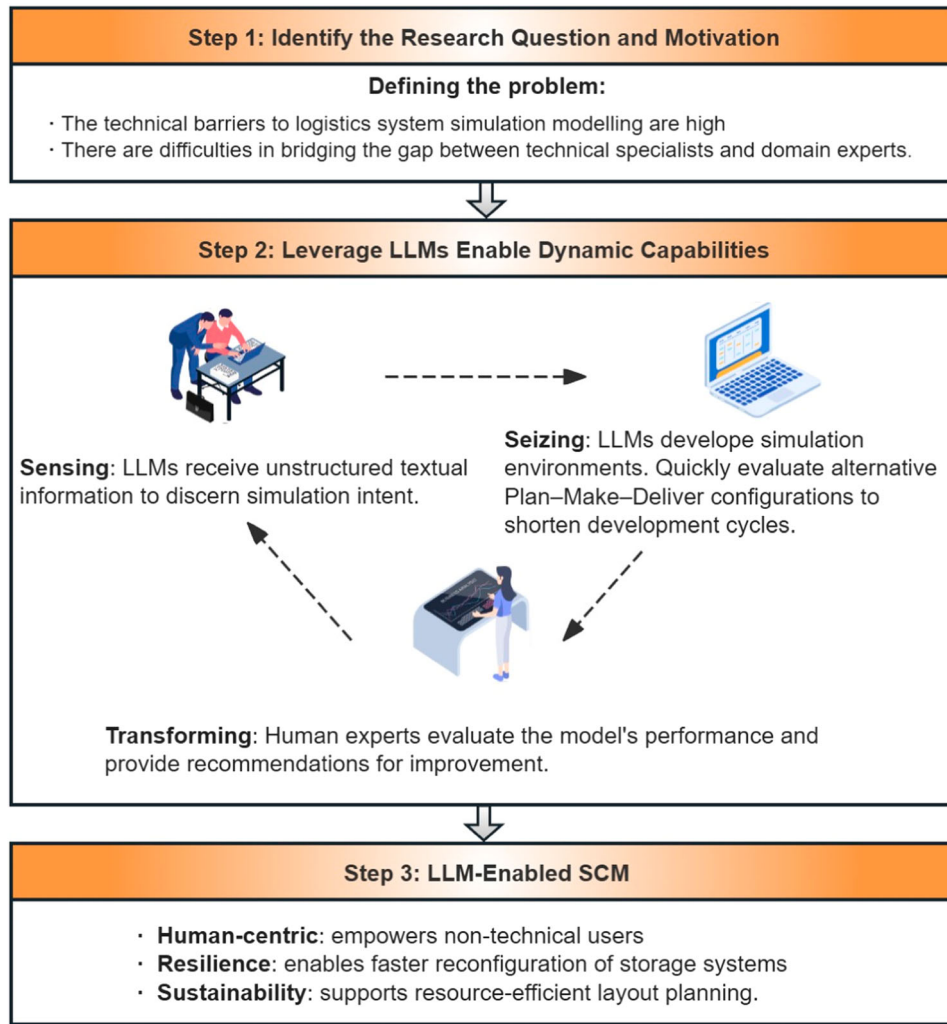


Figure 10. Practical case workflow under the LLM-SCM application pathway framework.

volatility and disruptions (resilience); Third, it facilitates resource-efficient layout planning, such as optimising space utilisation, reducing handling effort and evaluating circular logistics scenarios in return operations (sustainability). The case thus exemplifies how the proposed implementation logic can transition from concept to operational insight, underscoring the transformative potential of LLMs in SCM.

A visual summary of the system workflow is provided in Figure 10 to support comprehension.

4.2. Development trend

4.2.1. Resilient and insight-driven planning

Future research focus in Plan could be how to combine LLM reasoning with probabilistic uncertainty and forecasting models without misleading confidence and what explanation formats actually improve S&OP alignment (counterfactuals, causal graphs, provenance-based narratives). Enterprises increasingly face challenges such

as high demand uncertainty, delayed responsiveness to external shocks, and fragmented coordination among cross-functional teams (Fan, Wu, and Yang 2025; Wang et al. 2025a). These conditions hinder supply chains' ability to anticipate disruptions, distribute resources efficiently, and align operational strategies with dynamic market demands. LLMs offer new capabilities to support semantic sensing and predictive simulation, particularly in integrating unstructured data into forecasting workflows (Fan, Wu, and Yang 2025). Furthermore, the transforming capability of LLMs lies in their potential to continuously learn from diverse planning scenarios and adjust prediction logic based on stakeholder feedback. However, current systems remain limited by challenges in forecast explainability, semantic interoperability, and deployment robustness, particularly under high-stakes, time-sensitive planning (Caetano, Oliveira, and Ramos 2025; Hamidi, Hosseini, and Hosseini 2025).

When effectively embedded LLMs into SCM planning, operationally, they improve demand forecasting

precision, reduce planning latency, and enhance cross-functional collaboration (Fosso Wamba et al. 2024; Simchi-Levi, Mellou, and Pathuri 2025). Strategically, they increase organisational foresight and scenario readiness, allowing firms to anticipate structural risks and model alternative futures (Xie et al. 2025). These contributions support the resilience pillar of Industry 5.0 by enabling early signal detection, collaborative planning, and dynamic response modelling across internal and external supply chain actors.

However, ethical considerations remain. Emotion-driven or sentiment-based forecasting systems may suffer from hallucinated correlations or overfitting (Zouaghi et al. 2025), leading to mismatches between perceived and actual demand. Hence, future research should pursue transparent and auditable forecasting pipelines, embedding multi-stakeholder governance, fairness constraints, and domain-specific knowledge bases.

4.2.2. *Ethical and sustainable sourcing*

How to design traceable supplier intelligence pipelines (sources, timestamps, confidence, contradiction handling) and how to evaluate and mitigate bias in LLM-based supplier assessments and contract recommendations could be a future research focus. Organisations continue to face persistent obstacles in the interpretation of complex contracts, traceability of upstream suppliers, and the ethical use of Environmental, Social, and Governance (ESG) data (Aylak 2025; Li and Starly 2024). Challenges also arise in managing sensitive information, verifying supplier claims, and mitigating greenwashing risks (Liu and Chou 2025; Wang et al. 2025a). These contextual complexities limit the ability of firms to sense risks, seize ethical opportunities, and transform procurement systems in line with sustainable development goals.

LLMs offer novel mechanisms to address sourcing challenges by enabling semantic reasoning over multilingual procurement data, supporting contract clause interpretation, and facilitating green supplier discovery. Furthermore, LLMs seize sourcing opportunities by aligning procurement decision-making with sustainability targets. LLMs also contribute to transforming procurement workflows by generating auto-complete or summarised procurement reports, detecting clause-level anomalies, or recommending edits to reinforce alignment with ethical policies.

Despite these capabilities, key barriers remain. These include limited auditable explainability in sensitive decision scenarios, absence of standardised semantic corpora for sustainability verification, and high risks of ESG misuse or manipulation in LLM output, especially when firms repurpose AI-generated recommendations to support pre-existing greenwashing strategies (TWI 2022).

When responsibly deployed, LLMs can reduce manual workload in contract review, improve data interpretability, and accelerate supplier screening under green criteria. On the other hand, deploying LLMs in SCM supports the development of transparent, accountable, and resilient procurement systems aligned with Industry 5.0's ethics and sustainability pillars.

However, the increased reliance on LLMs in sourcing decisions also raises ethical concerns, such as the risk of greenwashing (TWI 2022), where suppliers may manipulate ESG narratives. To mitigate such risks, future research should explore the construction of domain-specific green corpora and audit-ready LLM pipelines to ensure verifiable sourcing transparency. Future research should prioritise the creation of domain-specific green corpora, auditable LLM training pipelines, and compliance-aware language models capable of verifying and justifying sustainability claims. In parallel, the development of standardised interfaces to integrate LLMs into Supplier Relationship Management (SRM) systems and ethical compliance dashboards is critical to ensuring operational scalability.

4.2.3. *Collaborative and traceable smart making*

Future research could focus on how can LLMs safely interface with operational technology/industrial IoT and structured control rules without unsafe recommendations or what governance patterns best prevent leakage of proprietary manufacturing knowledge. The Make process in supply chains has become increasingly complex, requiring high levels of real-time responsiveness, traceability, and human–AI collaboration. Smart factories often struggle with production scheduling volatility, quality inspection delays, and loss of process knowledge due to fragmented systems and siloed workflows (Cauli et al. 2024; Hu et al. 2025; Xie et al. 2025). Moreover, small-batch customisation, sustainability goals, and digital transformation initiatives have intensified the need for intelligent, flexible manufacturing environments that can adapt dynamically to disruptions and variance (Cauli et al. 2024; Fan and Chiu 2024). These challenges create an urgent need for systems that can sense operational disruptions, seize optimisation opportunities, and continuously transform shop-floor logic in real time.

LLMs exhibit strong sensing capabilities through integration with visual transformers and machine learning models (Knott, Perez-Cruz, and Defraeye 2023; Yazdinejad et al. 2023). LLMs also support seizing actions by generating adaptive responses. Their semantic reasoning capabilities enhance machine–human interaction, acting as AI co-pilots that mediate between system logic and operator queries, improving production transparency and traceability. In terms of transforming, LLMs

enable knowledge transfer and dynamic workflow generation through integration with manufacturing knowledge graphs, digital twins, and MES (Manufacturing Execution Systems).

Despite these promises, practical challenges persist, including weak compatibility with manufacturing logic, inadequate handling of proprietary operational data, and instability in output under time-critical constraints. LLMs also lack standardised performance metrics to evaluate their suitability in high-risk manufacturing environments.

Deployed appropriately, LLMs can reduce quality inspection time, enhance traceability, and streamline anomaly detection in complex production environments. Strategically, they promote a shift toward adaptive, intelligent manufacturing systems that support collaborative decision-making and self-learning workflows, reinforcing the human-centric and resilience objectives of Industry 5.0.

Future research should focus on building secure, explainable LLM agents tailored for production environments, including error annotation tools, real-time audit trails, and role-based human-AI collaboration interfaces. Additionally, a pressing need exists to develop evaluation frameworks that measure LLM performance in terms of scheduling accuracy, explainability, and knowledge generalisation within dynamic production contexts.

4.2.4. *Resilient and experience-driven delivering*

Future research focus could be how to make LLM decision support robust to rapidly changing conditions (real-time retrieval, continuous validation) or how to evaluate LLMs in live operations without unacceptable service risk. In the Deliver process in supply chain management, Challenges persist in route optimisation, inventory reallocation, and context-aware service adaptation, particularly in response to rapid demand fluctuations, urban congestion, or delivery exceptions (Aylak 2025; Hu et al. 2025). Traditional logistics systems, heavily reliant on static optimisation models, often fail to integrate dynamic customer feedback, real-time sensor data, and exception-driven reasoning, hindering their ability to sense disruptions early and respond with agility.

LLMs enhance sensing capabilities by interpreting multi-modal, unstructured signals to identify delivery risks, rerouting needs, or customer dissatisfaction (Aljaž 2024; Bao et al. 2024). LLMs also support seizing by enabling semantic route optimisation and proactive customer engagement. LLMs play a transforming role in reshaping logistics architectures from reactive delivery to anticipatory service ecosystems. Through integration with transportation management systems (TMS) and warehouse management systems (WMS), LLMs facilitate

continuous learning from historical disruptions, co-develop routing heuristics with human planners, and automate exception-handling protocols with embedded ethics and emotional reasoning.

However, significant deployment challenges remain, including limited integration with real-time structured inventory systems, insufficient interpretability of LLM outputs in high-pressure delivery windows, and ethical concerns related to bias in customer interaction or cross-cultural misinterpretation (TWI 2022).

Properly implementing LLMs can improve delivery accuracy, reduce customer service response times, and enable adaptive routing in volatile environments. Additionally, they support the transition to resilient, human-centric logistics networks that are both experience-driven and systemically robust, contributing to Industry 5.0's values of customer empowerment, transparency, and sustainable responsiveness. Future research should investigate LLM-aided semantic reasoning chains across logistics flows, combining structured and unstructured data streams. Moreover, developing auditing and explainability mechanisms for LLMs in customer-facing applications is critical to prevent service bias, manage tone, and build trust in emotionally charged or multicultural service environments.

4.2.5. *Transparent and empathetic return*

How can LLMs extract reliable structure from messy return narratives while quantifying uncertainty and how to prevent “plausible but wrong” sustainability narratives in ESG/circularity reporting could be future research focus in Return process. In this process, enterprises often face persistent issues such as unpredictable return volumes, low traceability of returned items, inefficient service processing, and weak customer interaction mechanisms (Sliz 2024). These limitations hinder the ability to effectively process feedback, manage warranties, and extract operational insights from return behaviours, raising both operational and ethical concerns.

LLMs can parse customer feedback across channels to identify patterns in return reasons, dissatisfaction signals, or service quality gaps. Additionally, LLMs automate common service responses and provide dynamic support in multilingual environments, generating empathetic and situation-aware replies that enhance customer satisfaction. By continuously learning from return interactions, LLMs can help reshape reverse logistics workflows. When integrated with rule-based engines, decision trees, or dynamic knowledge graphs, they support adaptive service reconfiguration.

However, critical challenges persist. LLM-generated recommendations risk emotional misjudgement, fraudulent misclassification, or privacy violations, especially

when handling sensitive personal data. There is also a lack of explainable reasoning for return denials, inadequate trust safeguards, and legal ambiguity regarding accountability in AI-led service decisions (TWI 2022).

When responsibly applied, LLMs can reduce return processing time, improve service standardisation, and enhance issue resolution accuracy. Moreover, LLMs enable firms to develop empathetic, transparent, and ethically aligned after-sales systems, advancing Industry 5.0's focus on human dignity, trust, and inclusiveness in digitally mediated services.

The application of LLMs in return classification and dispute handling also introduces potential ethical pitfalls, including fraudulent misclassification and privacy breaches (TWI 2022). As a safeguard, there is an urgent need to develop trust frameworks and domain-tuned safeguards to ensure that automated return systems remain fair, secure, and accountable. Future research should focus on embedding explainable AI and bias detection tools into LLM service pipelines, while defining the legal and ethical boundaries of AI-enabled customer interaction. Creating standard return corpora, multilingual service tone monitors, and transparency dashboards will further reinforce trustworthiness in reverse service automation.

4.2.6. Explainability and transparency

While LLMs exhibit strong transformative potential across all SCOR processes, from planning and sourcing to delivery and return, a persistent and cross-cutting challenge lies in their explainability and transparency (Simchi-Levi, Mellou, and Pathuri 2025). As discussed in previous subsections, each domain encounters specific concerns, such as unverifiable supplier claims, opaque demand forecasts, and ambiguous justifications in automated customer interactions. These issues are not isolated; rather, they point to a systemic barrier in the trustworthy deployment of LLMs in SCM.

The black-box nature of many foundation models limits their interpretability, especially when operational decisions have ethical, legal, or strategic implications. This can erode user trust, complicate accountability, and hinder regulatory compliance. Moreover, without transparent reasoning mechanisms, it becomes difficult for supply chain professionals to audit, validate, or refine model outputs in real time.

Recent research offers several promising directions to address these challenges. RAG and agent-based systems can ground LLM outputs in verifiable sources (Li, Ko, and Ameri 2025), enabling more transparent and traceable reasoning. These techniques are particularly beneficial for applications like sourcing documentation or regulatory reporting, where explainability is

critical. In addition, well-structured prompt engineering strategies, such as chain-of-thought (Zhang et al. 2025), can guide the model to articulate intermediate reasoning steps, enhancing human interpretability and auditability.

5. Conclusions

This review systematically examined how LLMs are being applied in SCM and, on this basis, proposed an integrated LLM-SCM application pathway framework. The framework maps the technical capabilities of LLMs to the five SCOR processes, Plan, Source, Make, Deliver, and Return, and organises their application pathways through the CMO lens. By linking these pathways with the dynamic capability perspective, the framework explains how LLMs can enhance sensing, seizing, and transforming routines across supply chains. The analysis further connects these outcomes with the priorities of Industry 5.0: human-centricity, resilience, and sustainability.

The review identifies three major thematic clusters in current research,— forecasting and simulation, semantic automation, and cognitive coordination,— which correspond to different stages of the supply chain. From these clusters, several development trends were derived: resilient and insight-driven planning, ethical and sustainable sourcing, collaborative and traceable smart making, resilient and experience-driven delivering, and transparent and empathetic return. Each trend highlights both opportunities and persistent barriers, such as the need for explainable forecasting models, domain-specific corpora for sustainable sourcing, evaluation metrics for manufacturing agents, integration with real-time logistics systems, and trustworthy mechanisms in reverse supply chain services.

The proposed framework offers both theoretical and practical implications. Theoretically, it provides a structured lens for aligning LLM competencies with established SCM process logic, thereby offering a foundation for future conceptual development. Practically, it supports supply chain practitioners, digital transformation managers, and policymakers in identifying where LLMs can address pressing challenges such as demand volatility, supplier transparency, production adaptability, logistics disruptions, and return complexity.

Looking forward, future research may further validate and extend this framework by conducting empirical studies on LLM deployment within specific industry sectors, integrating LLMs with domain-specific knowledge graphs and simulation systems, and developing evaluation benchmarks for ethical risk, interpretability, and system resilience. The advancement of such work will be essential to fully harness the transformative potential of

LLMs in building smarter, more responsive, and socially responsible supply chains.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was undertaken with funding from the ESRC through the 'UKRI Ideas to Address COVID-19' call (Grant reference: ES/W001195/1).

Notes on contributors

Song Zhe was born in Hebei, China. He is currently a Ph.D. candidate at the Faculty of Engineering and Applied Sciences, Cranfield University, Cranfield, UK. His research focuses on the application of Artificial Intelligence (AI) and Large Language Models (LLMs) in construction supply chain and warehouse management. His work aims to develop data-driven and risk-aware warehouse management systems, integrating inventory optimisation, supply chain risk management, and real-time analytics to enhance operational efficiency and decision-making under uncertainty.

Lichao Yang was born in Shanxi, China. He received M.Sc. and Ph.D. degrees in automotive mechatronics and manufacturing from Cranfield University, Cranfield, U.K., in 2018 and 2021. He is currently a research fellow in computer vision and artificial intelligence at the Centre for Life-cycle Engineering and Management, Cranfield University, Cranfield, U.K. His research interests include computer vision, image processing, deep neural networks, machine learning, and human behaviour analysis.

Ying Xie was born in Hubei, China. She received her PhD in Management Science from Coventry University, UK. Ying currently serves as a Professor in Supply Chain Analytics at Cranfield University, UK. Her research skills cover machine learning, big data analytics, natural language processing (NLP), operational research, supply chain modelling and simulation. She specialises in maritime transport, clean energy supply chain, sustainable aviation supply chain and healthcare supply chain. Her work supports businesses in enhancing data driven decision-making and improving operational efficiency.

Yifan Zhao was born in Zhejiang, China. He received a Ph.D. degree in Automatic Control and System Engineering from the University of Sheffield, Sheffield, UK, in 2007. He is currently the Professor of Data Science at Cranfield University, Cranfield, UK. His research focuses on solving Inverse Problems based on data science including Artificial Intelligence (AI), computer vision, pattern recognition, signal processing, and nonlinear system identification.

References

- Ahmed, I., M. Alkahtani, Q. S. Khalid, and F. M. Alqahtani. 2024. "Improved Commodity Supply Chain Performance through AI and Computer Vision Techniques." *IEEE Access* 12:24116–24132. <https://doi.org/10.1109/ACCESS.2024.3361756>.
- Aljaż, T. 2024. Enhancing Retail Operations: Integrating Artificial Intelligence into the Theory of Constraints Thinking Process to Solve Shelf Issue.
- AlMahri, S., L. Xu, and A. Brintrup. 2025. "Enhancing Supply Chain Visibility with Knowledge Graphs and Large Language Models." *International Journal of Production Research*: 1–32. <https://doi.org/10.1080/00207543.2025.2575841>.
- Amellal, I., A. Amellal, H. Seghioeur, and M. R. Ech-Charrat. 2024. "An Integrated Approach for Modern Supply Chain Management: Utilizing Advanced Machine Learning Models for Sentiment Analysis, Demand Forecasting, and Probabilistic Price Prediction." *Decision Science Letters* 13 (1): 237–248. <https://doi.org/10.5267/j.dsl.2023.9.003>.
- Anthropic. 2024. *The Claude 3 Model Family: Opus, Sonnet, Haiku*. San Francisco, CA: Anthropic. <https://www.anthropic.com/news/claude-3-family>
- Automating Supply Chain Disruption Monitoring via an Agentic AI Approach. n.d.
- Aylak, B. L. 2025. "SustAI-SCM: Intelligent Supply Chain Process Automation with Agentic AI for Sustainability and Cost Efficiency." *Sustainability* 17 (6): 2453. <https://doi.org/10.3390/su17062453>.
- Bai, W., J. Zhu, J. Zhao, W. Cai, and K. Li. 2022. "An Unsupervised Multi-dimensional Representation Learning Model for Short-Term Electrical Load Forecasting." *Symmetry* 14 (10): 1999. <https://doi.org/10.3390/sym14101999>.
- Bao, S., M. Zhang, R. Xie, D. Huang, and J. Kong. 2024. "Enhancing a You Only Look Once-Plated Detector via Auxiliary Textual Coding for Multi-scale Rotating Remote Sensing Objects in Transportation Monitoring Applications." *Applied Sciences* 14 (19): 9074. <https://doi.org/10.3390/app14199074>.
- Beltagy, I., M. E. Peters, and A. Cohan. 2020. Longformer: The Long-Document Transformer (arXiv:2004.05150). arXiv. <https://doi.org/10.48550/arXiv.2004.05150>.
- Beyer Díaz, S., K. Coussement, A. De Caigny, L. F. Pérez, and S. Creemers. 2024. "Do the US President's Tweets Better Predict oil Prices? An Empirical Examination Using Long Short-Term Memory Networks." *International Journal of Production Research* 62 (6): 2158–2175. <https://doi.org/10.1080/00207543.2023.2217286>.
- Blome, C., T. Schoenherr, and D. Rexhausen. 2013. "Antecedents and Enablers of Supply Chain Agility and Its Effect on Performance: A Dynamic Capabilities Perspective." *International Journal of Production Research* 51 (4): 1295–1318. <https://doi.org/10.1080/00207543.2012.728011>.
- Bo, L., and J. Xu. 2024. "Enhancing Supply Chain Efficiency Resilience Using Predictive Analytics and Computational Intelligence Techniques." *IEEE Access* 12: 183451–183465. <https://doi.org/10.1109/ACCESS.2024.3507161>.
- Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, et al. 2022. On the Opportunities and Risks of Foundation Models (arXiv:2108.07258). arXiv. <https://doi.org/10.48550/arXiv.2108.07258>.
- Breitung, C., and S. Müller. 2025. "Global Business Networks." *Journal of Financial Economics* 166:104007. <https://doi.org/10.1016/j.jfineco.2025.104007>.
- Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. Language Models Are Few-Shot Learners.

- Caetano, R., J. M. Oliveira, and P. Ramos. 2025. "Transformer-Based Models for Probabilistic Time Series Forecasting with Explanatory Variables." *Mathematics* 13 (5): 814. <https://doi.org/10.3390/math13050814>.
- Cano-Marín, E., D. Ribeiro-Soriano, A. Mardani, and C. Blanco Gonzalez-Tejero. 2023. "Exploring the Challenges of the COVID-19 Vaccine Supply Chain Using Social Media Analytics: A Global Perspective." *Sustainable Technology and Entrepreneurship* 2 (3): 100047. <https://doi.org/10.1016/j.stae.2023.100047>.
- Cauli, N., M. Murgia, D. Reforgiato Recupero, and G. Scarpi. 2024. "Leveraging Transformers Architectures and Augmentation for Efficient Classification of Fasteners and Natural Language Searches." *Discover Computing* 27:1. <https://doi.org/10.1007/s10791-024-09443-8>.
- Chiu, M.-C., P.-Y. Tai, and C.-Y. Chu. 2024. "Developing a Smart Green Supplier Risk Assessment System Integrating Natural Language Processing and Life Cycle Assessment Based on AHP Framework: An Empirical Study." *Resources, Conservation and Recycling* 207:107671. <https://doi.org/10.1016/j.resconrec.2024.107671>.
- Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, et al. 2022. Scaling Instruction-Finetuned Language Models (arXiv:2210.11416). arXiv. <https://doi.org/10.48550/arXiv.2210.11416>.
- Coltey, E., D. Alonso, S. Vassigh, and S.-C. Chen. 2022. "Towards an AI-Driven Marketplace for Small Businesses during COVID-19." *SN Computer Science* 3 (6): 441. <https://doi.org/10.1007/s42979-022-01349-w>.
- DALL-E 2. 2022. <https://openai.com/index/dall-e-2/>.
- DeepSeek-AI. 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in Large Language Models via Reinforcement Learning*. arXiv preprint arXiv:2501.12948. <https://doi.org/10.48550/arXiv.2501.12948>.
- Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, et al. 2021. An Image Is Worth 16(16 Words): Transformers for Image Recognition at Scale (arXiv:2010.11929). arXiv. <https://doi.org/10.48550/arXiv.2010.11929>.
- European Commission. 2021. *Industry 5.0: Towards a Sustainable, Human Centric and Resilient European Industry*. Luxembourg: Publications Office of the European Union. <https://doi.org/10.2777/308407>.
- Fan, S., Y. Wu, and R. Yang. 2025. "Measuring Firm-Level Supply Chain Risk Using a Generative Large Language Model." *Finance Research Letters* 77:107111. <https://doi.org/10.1016/j.frl.2025.107111>.
- Fan, S.-K. S., and S.-H. Chiu. 2024. "A new ViT-Based Augmentation Framework for Wafer map Defect Classification to Enhance the Resilience of Semiconductor Supply Chains." *International Journal of Production Economics* 273:109275. <https://doi.org/10.1016/j.ijpe.2024.109275>.
- Fedus, W., B. Zoph, and N. Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity.
- Fontoura, L., Luiz De Mattos Nascimento, D. Neto, J. V. and R. G. Gusmão Caiado. 2025. "Energy Gen-AI Technology Framework: A Perspective of Energy Efficiency and Business Ethics in Operation Management." *Technology in Society* 81:102847. <https://doi.org/10.1016/j.techsoc.2025.102847>.
- Fosso Wamba, S., C. Guthrie, M. M. Queiroz, and S. Minner. 2024. "ChatGPT and Generative Artificial Intelligence: An Exploratory Study of key Benefits and Challenges in Operations and Supply Chain Management." *International Journal of Production Research* 62 (16): 5676–5696. <https://doi.org/10.1080/00207543.2023.2294116>.
- Garg, R., A. W. Kiwelekar, and L. D. Netak. 2021. "Logistics and Freight Transportation Management: An NLP Based Approach for Shipment Tracking." *Pertanika Journal of Science and Technology* 29 (4). <https://doi.org/10.47836/pjst.29.4.28>.
- Gezdur, A., and J. Bhattacharjya. 2025. "Innovators and Transformers: Enhancing Supply Chain Employee Training with an Innovative Application of a Large Language Model." *International Journal of Physical Distribution & Logistics Management*. <https://doi.org/10.1108/IJPDLM-12-2023-0492>.
- Ghimire, S., T. Nguyen-Huy, M. S. AL-Musaylh, R. C. Deo, D. Casillas-Pérez, and S. Salcedo-Sanz. 2023. "Integrated Multi-head Self-attention Transformer Model for Electricity Demand Prediction Incorporating Local Climate Variables." *Energy and AI* 14:100302. <https://doi.org/10.1016/j.egyai.2023.100302>.
- GPT-4 V(ision) system card. 2024. <https://openai.com/index/gpt-4v-system-card/>.
- Guo, W., Z. Ren, W. Du, and T. Weng. 2025. "TFT-MPIR: An end-to-end Multi-period Inventory Replenishment Strategy Based on Temporal Fusion Transformer." *Expert Systems with Applications* 261:125464. <https://doi.org/10.1016/j.eswa.2024.125464>.
- Hamidi, H., M. Hosseini, and S. S. Hosseini. 2025. "Design of a Framework Using Bidirectional Encoder Representations from Transformers to Understanding Panic Buying Behavior during the COVID-19 Pandemic." *International Journal of Engineering* 38 (1): 247–261. <https://doi.org/10.5829/ije.2025.38.01a.22>.
- Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models (arXiv:2106.09685). arXiv. <https://doi.org/10.48550/arXiv.2106.09685>.
- Hu, X., J. An, X. Gui, L. Chen, S. Tang, and X. He. 2025. "TDCR: Transformer Based Decision Conflict Resolution Model for Collaborative Scheduling." *Neurocomputing* 632:129760. <https://doi.org/10.1016/j.neucom.2025.129760>.
- Huang, C., G. Nan, H. Qiu, M.-L. Tseng, and K.-J. Wu. 2025. "Launching Smart Circular Supply Chain Practices toward Sociotechnological Synergy: An Integrative Influential Fast Fashion Model." *Resources, Conservation and Recycling* 212:107935. <https://doi.org/10.1016/j.resconrec.2024.107935>.
- Ivanov, D. 2022. "Viable Supply Chain Model: Integrating Agility, Resilience and Sustainability Perspectives—Lessons from and Thinking beyond the COVID-19 Pandemic." *Annals of Operations Research* 319 (1): 1411–1431. <https://doi.org/10.1007/s10479-020-03640-6>.
- Ivanov, D., A. Dolgui, J. V. Blackhurst, and T.-M. Choi. 2023. "Toward Supply Chain Viability Theory: From

- Lessons Learned through COVID-19 Pandemic to Viable Ecosystems.” *International Journal of Production Research* 61 (8): 2402–2415. <https://doi.org/10.1080/00207543.2023.2177049>.
- Jackson, I., M. Jesus Saenz, and D. Ivanov. 2024a. “From Natural Language to Simulations: Applying AI to Automate Simulation Modelling of Logistics Systems.” *International Journal of Production Research* 62 (4): 1434–1457. <https://doi.org/10.1080/00207543.2023.2276811>.
- Jackson, I., M. Jesus Saenz, and D. Ivanov. 2024b. “From Natural Language to Simulations: Applying AI to Automate Simulation Modelling of Logistics Systems.” *International Journal of Production Research* 62 (4): 1434–1457. <https://doi.org/10.1080/00207543.2023.2276811>.
- Jaegle, A., F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. 2021. Perceiver: General Perception with Iterative Attention (arXiv:2103.03206). arXiv. <https://doi.org/10.48550/arXiv.2103.03206>.
- Jahangir, M. S., and J. Quilty. 2024. Sequence-to-Sequence Deep Learning for Urban Water Demand Forecasting. *The 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024)*, 41. <https://doi.org/10.3390/engproc2024069041>.
- Jannelli, V., S. Schöpf, M. Bickel, T. Netland, and A. Brintrup. 2025. “Agentic LLMs in the Supply Chain: Towards Autonomous Multi-agent Consensus-Seeking.” *International Journal of Production Research*: 1–31. <https://doi.org/10.1080/00207543.2025.2604311>.
- Jiang, X., Y. Gou, M. Jiang, L. Luo, and Q. Zhou. 2025. “Photovoltaic Power Forecasting with Weather Conditioned Attention Mechanism.” *Big Data Mining and Analytics* 8 (2): 326–345. <https://doi.org/10.26599/BDMA.2024.9020066>.
- Jing, D., Y. Yang, Z. Gu, R. Feng, Y. Li, and H. Jiang. 2024. “Multi-feature Fusion in Graph Convolutional Networks for Data Network Propagation Path Tracing.” *Electronics* 13 (17): 3412. <https://doi.org/10.3390/electronics13173412>.
- Keskar, N. S., B. McCann, L. R. Varshney, C. Xiong, and R. Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation (arXiv:1909.05858). arXiv. <https://doi.org/10.48550/arXiv.1909.05858>.
- Ketchen, D. J., and L. C. Giunipero. 2004. “The Intersection of Strategic Management and Supply Chain Management.” *Industrial Marketing Management* 33 (1): 51–56. <https://doi.org/10.1016/j.indmarman.2003.08.010>.
- Kitae, N., I. Kaiser, and A. Levskaya. 2020. Reformer: The Efficient Transformer (arXiv:2001.04451). arXiv. <https://doi.org/10.48550/arXiv.2001.04451>.
- Knott, M., F. Perez-Cruz, and T. Defraeye. 2023. “Facilitated Machine Learning for Image-Based Fruit Quality Assessment.” *Journal of Food Engineering* 345:111401. <https://doi.org/10.1016/j.jfoodeng.2022.111401>.
- Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (arXiv:1909.11942). arXiv. <https://doi.org/10.48550/arXiv.1909.11942>.
- Leite, M. V., J. M. Abe, M. L. H. Souza, and I. De Alencar Nääs. 2025. “Enhancing Environmental Control in Broiler Production: Retrieval-Augmented Generation for Improved Decision-Making with Large Language Models.” *AgriEngineering* 7 (1): 12. <https://doi.org/10.3390/agriengineering7010012>.
- Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (arXiv:1910.13461). arXiv. <https://doi.org/10.48550/arXiv.1910.13461>.
- Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
- Li, D., and Q. Fu. 2024. “Deep Learning Model-Based Demand Forecasting for Secondary Water Supply in Residential Communities: A Case Study of Shanghai City, China.” *IEEE Access* 12:38745–38757. <https://doi.org/10.1109/ACCESS.2023.3288817>.
- Li, X., J. Shen, D. Wang, W. Lu, and Y. Chen. 2024. “Multi-modal Transform-Based Fusion Model for new Product Sales Forecasting.” *Engineering Applications of Artificial Intelligence* 133:108606. <https://doi.org/10.1016/j.engappai.2024.108606>.
- Li, Y., H. Ko, and F. Ameri. 2025. “Integrating Graph Retrieval-Augmented Generation with Large Language Models for Supplier Discovery.” *Journal of Computing and Information Science in Engineering* 25 (2), <https://doi.org/10.1115/1.4067389>.
- Li, Y., and B. Starly. 2024. “Building a Knowledge Graph to Enrich ChatGPT Responses in Manufacturing Service Discovery.” *Journal of Industrial Information Integration* 40:100612. <https://doi.org/10.1016/j.jii.2024.100612>.
- Lin, Y., W. Yan, Y. Zhang, Hua Zhang, Haiyang Zhang, and D. Wang. 2025. “Emergency Drug Demand Forecasting in Earthquakes with XGBoost and AFT-LSTM.” *Sustainability* 17 (5): 1910. <https://doi.org/10.3390/su17051910>.
- Liu, C., W. Lin, Y. Feng, Z. Guo, and Z. Xie. 2023. “ATC-YOLOv5: Fruit Appearance Quality Classification Algorithm Based on the Improved YOLOv5 Model for Passion Fruits.” *Mathematics* 11 (16): 3615. <https://doi.org/10.3390/math11163615>.
- Liu, C.-Y., and J.-S. Chou. 2025. “Automated Legal Consulting in Construction Procurement Using Metaheuristically Optimized Large Language Models.” *Automation in Construction* 170:105891. <https://doi.org/10.1016/j.autcon.2024.105891>.
- Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach (arXiv:1907.11692). arXiv. <https://doi.org/10.48550/arXiv.1907.11692>.
- Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (arXiv:2103.14030). arXiv. <https://doi.org/10.48550/arXiv.2103.14030>.
- Long, Y., S. Kroeger, M. F. Zaeh, and A. Brintrup. 2025. “Leveraging Synthetic Data to Tackle Machine Learning Challenges in Supply Chains: Challenges, Methods, Applications, and Research Opportunities.” *International Journal of Production Research*: 1–22. <https://doi.org/10.1080/00207543.2024.2447927>.
- Löwhagen, N., P. Schwendener, and T. Netland. 2025. “Can a Troubleshooting AI Assistant Improve Task Performance in Industrial Contexts?” *International Journal of Production Research*: 1–22. <https://doi.org/10.1080/00207543.2025.2527368>.

- Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space (arXiv:1301.3781). arXiv. <https://doi.org/10.48550/arXiv.1301.3781>.
- Mishra, L. N., and B. Senapati. 2025. "Retail Resilience Engine: An Agentic AI Framework for Building Reliable Retail Systems with Test-Driven Development Approach." *IEEE Access* 13:50226–50243. <https://doi.org/10.1109/access.2025.3552592>.
- Nahavandi, S. 2019. "Industry 5.0—A Human-Centric Solution." *Sustainability* 11 (16): 4371. <https://doi.org/10.3390/su11164371>.
- Ni, Z., X. Wang, Z. Zhang, and L. Wang. 2024. "Development of Major Process Accident Indicators Based on Industrial Internet." *Journal of Loss Prevention in the Process Industries* 92:105418. <https://doi.org/10.1016/j.jlp.2024.105418>.
- Oh, J. 2024. "Developing a Model for Extracting Actual Product Names from Order Item Descriptions Using Generative Language Models." *IEEE Access* 12:122695–122701. <https://doi.org/10.1109/access.2024.3437197>.
- Onoufriou, G., M. Hanheide, and G. Leontidis. 2023. "Premotion Net, a Multi-timeline Transformer Network Architecture towards Strawberry Tabletop Yield Forecasting." *Computers and Electronics in Agriculture* 208:107784. <https://doi.org/10.1016/j.compag.2023.107784>.
- OpenAI. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. <https://doi.org/10.48550/arXiv.2303.08774>.
- Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, et al. 2022. Training language models to follow instructions with human feedback.
- Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, ... D. Moher. 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *BMJ* 372, <https://doi.org/10.1136/bmj.n71>.
- Painuly, S., and S. Sharma. 2024. Natural Language Processing Techniques for e-Healthcare Supply Chain Management System. 2024 *International Conference on Emerging Systems and Intelligent Computing (ESIC)*, 11–16. <https://doi.org/10.1109/ESIC60604.2024.10481580>.
- Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: SAGE Publications. <https://uk.sagepub.com/en-gb/eur/realistic-evaluation/book205276>.
- Peng, B., E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, et al. 2023. RWKV: Reinventing RNNs for the Transformer Era (arXiv:2305.13048). arXiv. <https://doi.org/10.48550/arXiv.2305.13048>.
- Phuengrod, S., P. Wannapiroon, and P. Nilsook. n.d. The Intelligent Collaborative Supply Chain Management with Large Language Models ...
- Pichai, S., and D. Hassabis. 2024. Our Next-generation Model: Gemini 1.5. Google. <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/>.
- Prakash, S., and Sandeep. 2013. Supply Chain Operations Reference (Scor) Model: An Overview and a Structured Literature Review of its Application. <https://doi.org/10.13140/2.1.1855.3122>.
- Press, O., N. A. Smith, and M. Lewis. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation (arXiv:2108.12409). arXiv. <https://doi.org/10.48550/arXiv.2108.12409>.
- Qu, H., H. Li, L. You, R. Zhu, J. Yan, P. Santi, C. Ratti, and C. Yuen. 2024. "ChatEV: Predicting Electric Vehicle Charging Demand as Natural Language Processing." *Transportation Research Part D: Transport and Environment* 136:104470. <https://doi.org/10.1016/j.trd.2024.104470>.
- Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision (arXiv:2103.00020). arXiv. <https://doi.org/10.48550/arXiv.2103.00020>.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language Models are Unsupervised Multi-task Learners.
- Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.
- Raman, R., A. Sreenivasan, M. Suresh, A. Gunasekaran, and P. Nedungadi. 2024. "AI-driven Education: A Comparative Study on ChatGPT and Bard in Supply Chain Management Contexts." *Cogent Business & Management* 11 (1): 2412742. <https://doi.org/10.1080/23311975.2024.2412742>.
- Ramesh, A., M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. 2021. Zero-Shot Text-to-Image Generation (arXiv:2102.12092). arXiv. <https://doi.org/10.48550/arXiv.2102.12092>.
- Sangher, K. S., A. Singh, and H. M. Pandey. 2024. "LSTM and BERT Based Transformers Models for Cyber Threat Intelligence for Intent Identification of Social Media Platforms Exploitation from Darknet Forums." *International Journal of Information Technology* 16 (8): 5277–5292. <https://doi.org/10.1007/s41870-024-02077-5>.
- Satirapiwong, K., and T. Siriborvornratanakul. 2021. "Information Extraction for Different Layouts of Invoice Images." *The Imaging Science Journal* 69 (5–8): 417–429. <https://doi.org/10.1080/13682199.2022.2157367>.
- Shazeer, N., A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer (arXiv:1701.06538). arXiv. <https://doi.org/10.48550/arXiv.1701.06538>.
- Shishehgharkhaneh, M. B., R. C. Moehler, Y. Fang, A. A. Hijazi, and H. Aboutorab. 2024. "Transformer-Based Named Entity Recognition in Construction Supply Chain Risk Management in Australia." *IEEE Access* 12:41829–41851. <https://doi.org/10.1109/ACCESS.2024.3377232>.
- Simchi-Levi, D., K. Mellou, and J. Pathuri. 2025. Large Language Models for Supply Chain Decisions.
- Singh, S. K., and A.-N. El-Kassar. 2019. "Role of big Data Analytics in Developing Sustainable Capabilities." *Journal of Cleaner Production* 213:1264–1273. <https://doi.org/10.1016/j.jclepro.2018.12.199>.
- Skórňóg, D., and M. Kmiecik. 2023. "Supporting the Inventory Management in the Manufacturing Company by ChatGPT." *Logforum* 19 (4): 535–554. <https://doi.org/10.17270/J.LOG.2023.917>.
- Sliz, P. 2024. "The Role of ChatGPT in Elevating Customer Experience and Efficiency in Automotive After-Sales Business Processes." *Applied System Innovation* 7 (2): 29. <https://doi.org/10.3390/asi7020029>.
- Sora. n.d. Accessed 25 June 2025. <https://openai.com/sora/>.
- Stewart, G. 1997. "Supply-Chain Operations Reference Model (SCOR): The First Cross-Industry Framework for Integrated Supply-Chain Management." *Logistics Information Management* 10 (2): 62–67. <https://doi.org/10.1108/09576059710815716>.

- Sun, Y., L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei. 2023. Retentive Network: A Successor to Transformer for Large Language Models (arXiv:2307.08621). arXiv. <https://doi.org/10.48550/arXiv.2307.08621>.
- Taiwo, R., A.-M. Yussif, and T. Zayed. 2025. "Making Waves: Generative Artificial Intelligence in Water Distribution Networks: Opportunities and Challenges." *Water Research X* 28:100316. <https://doi.org/10.1016/j.wroa.2025.100316>.
- Tay, Y., M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, et al. 2023. UL2: Unifying Language Learning Paradigms (arXiv:2205.05131). arXiv. <https://doi.org/10.48550/arXiv.2205.05131>.
- Teece, D. J., G. Pisano, and A. Shuen. 1997. "Dynamic Capabilities and Strategic Management." *Strategic Management Journal* 18 (7): 509–533. [https://doi.org/10.1002/\(sici\)1097-0266\(199708\)18:7<509::aid-smj882>3.0.co;2-z](https://doi.org/10.1002/(sici)1097-0266(199708)18:7<509::aid-smj882>3.0.co;2-z).
- Thilakarathna, R. H., M. N. Dharmawardana, and T. Rupasinghe. 2015. "The Supply Chain Operations Reference (SCOR) Model: A Systematic Review of Literature from the Apparel Industry." *SSRN Electronic Journal*, <https://doi.org/10.2139/ssrn.2699886>.
- Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. 2021. Training data-efficient image transformers & distillation through attention (arXiv:2012.12877). arXiv. <https://doi.org/10.48550/arXiv.2012.12877>.
- Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, et al. 2023. LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971). arXiv. <https://doi.org/10.48550/arXiv.2302.13971>.
- TWI. 2022. What Is Industry 5.0? (Top 5 Things You Need To Know). <https://www.twi-global.com/technical-knowledge/faqs/industry-5-0.aspx>.
- Vallés-Pérez, I., E. Soria-Olivas, M. Martínez-Sober, A. J. Serrano-López, J. Gómez-Sanchís, and F. Mateo. 2022. "Approaching Sales Forecasting Using Recurrent Neural Networks and Transformers." *Expert Systems with Applications* 201:116993. <https://doi.org/10.1016/j.eswa.2022.116993>.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need.
- Wan, W., Y. Liu, X. Han, and H. Wang. 2021. "Evaluation Model of Power Operation and Maintenance Based on Text Emotion Analysis." *Mathematical Problems in Engineering* 2021:1–8. <https://doi.org/10.1155/2021/2824689>.
- Wang, S., B. Z. Li, M. Khabsa, H. Fang, and H. Ma. 2020. Linformer: Self-attention with Linear Complexity (arXiv:2006.04768). arXiv. <https://doi.org/10.48550/arXiv.2006.04768>.
- Wang, Weiqing, Y. Chen, L. Wang, and Y. Xiong. 2025a. "Developing the Value of Legal Judgments of Supply Chain Finance for Credit Risk Prediction through Novel ACWGAN-GPSA Approach." *Transportation Research Part E: Logistics and Transportation Review* 196:104020. <https://doi.org/10.1016/j.tre.2025.104020>.
- Wang, Wenyang, Y. Luo, M. Ma, J. Wang, and C. Sui. 2025b. "A Novel Forecasting Framework Leveraging Large Language Model and Machine Learning for Methanol Price." *Energy* 320:135123. <https://doi.org/10.1016/j.energy.2025.135123>.
- Wieland, A. 2021. "Dancing the Supply Chain: Toward Transformative Supply Chain Management." *Journal of Supply Chain Management* 57 (1): 58–73. <https://doi.org/10.1111/jscm.12248>.
- Wu, Q., G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, et al. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-agent Conversations.
- Xie, Y., J. Zheng, A. Gou, F. Sattar, and L. Liao. 2025. "Log End Face Feature Extraction and Matching Method Based on Swin Transformer V2." *Forests* 16 (1): 124. <https://doi.org/10.3390/f16010124>.
- Xin, B., L. Zhu, and W. Peng. 2024. "Competitive Advantage in Online Grocery Supply Chains: Leveraging GPT Technology." *Management Decision*, <https://doi.org/10.1108/md-03-2024-0581>.
- Yang, H., M. Zhao, L. Yuan, Y. Yu, Z. Li, and M. Gu. 2023. "Memory-efficient Transformer-Based Network Model for Traveling Salesman Problem." *Neural Networks* 161: 589–597. <https://doi.org/10.1016/j.neunet.2023.02.014>.
- Yazdinejad, A., E. Rabieinejad, T. Hasani, and G. Srivastava. 2023. "A BERT-Based Recommender System for Secure Blockchain-Based Cyber Physical Drug Supply Chain Management." *Cluster Computing* 26 (6): 3389–3403. <https://doi.org/10.1007/s10586-023-04088-6>.
- Zaheer, M., G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, et al. 2020. Big Bird: Transformers for Longer Sequences.
- Zhang, C., Q. Xu, Y. Yu, G. Zhou, K. Zeng, F. Chang, and K. Ding. 2025. "A Survey on Potentials, Pathways and Challenges of Large Language Models in new-Generation Intelligent Manufacturing." *Robotics and Computer-Integrated Manufacturing* 92:102883. <https://doi.org/10.1016/j.rcim.2024.102883>.
- Zhang, G. P., Y. Xia, and M. Xie. 2024. "Intermittent Demand Forecasting with Transformer Neural Networks." *Annals of Operations Research* 339 (1–2): 1051–1072. <https://doi.org/10.1007/s10479-023-05447-7>.
- Zheng, G., and A. Brintrup. 2025. "Enhancing Supply Chain Visibility with Generative AI: An Exploratory Case Study on Relationship Prediction in Knowledge Graphs." *International Journal of Production Research*: 1–23. <https://doi.org/10.1080/00207543.2025.2543964>.
- Zhou, H., L. Shao, and H. Zhang. 2024b. "SRRNet: A Transformer Structure with Adaptive 2-D Spatial Attention Mechanism for Cell Phone-Captured Shopping Receipt Recognition." *IEEE Transactions on Consumer Electronics* 70 (1): 3289–3298. <https://doi.org/10.1109/TCE.2022.3229438>.
- Zhou, Q., J. Cui, Lan Wang, S. Sun, H. Jiang, J. Li, A. J. Y. Chen, et al. 2024a. "A Waste Extended Input-Output-Based Transformer-LSTM Method for Analyzing Hazardous Waste Reduction Patterns: A Case Study of Shanghai." *Journal of Cleaner Production* 458:142435. <https://doi.org/10.1016/j.jclepro.2024.142435>.
- Zou, Y., H. Wu, Y. Yin, L. Dhamotharan, D. Chen, and A. K. Tiwari. 2024. "An Improved Transformer Model with Multi-head Attention and Attention to Attention for low-Carbon Multi-depot Vehicle Routing Problem." *Annals of Operations Research* 339 (1–2): 517–536. <https://doi.org/10.1007/s10479-022-04788-z>.
- Zouaghi, I., S. Beldjoudi, A. Gunasekaran, S. Fosso Wamba, and S. Sahrane. 2025. "Balancing Performance and Innovation in AI-Driven Supply Chains through Temperature-Scaled Hallucination Control." *International Journal of Production Research*: 1–38. <https://doi.org/10.1080/00207543.2025.2601264>.

Appendices

Appendix A. Transformer architecture

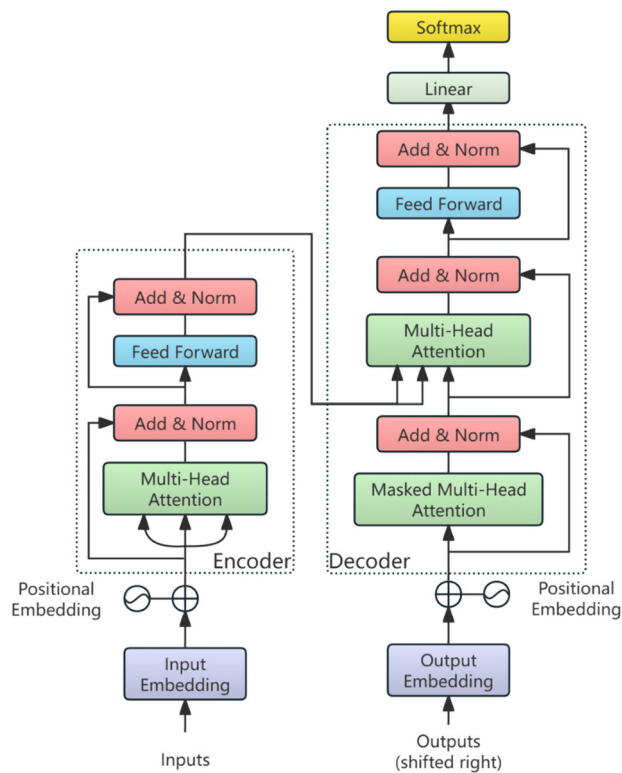


Figure A1. Transformer architecture illustrating the encoder-decoder structure, with multi-head attention, feed-forward layers, and positional encoding.

Appendix B. Transformer variants

Table A1. Five popular Transformer variants.

	Group	Architecture name	Year	Key optimisation	Representative model
A	Text-focused Transformers	BERT (Bidirectional Encoder Representations from Transformers)	2018	Bidirectional with Masked Language Model (MLM) pre-training	BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), ALBERT (Lan et al. 2020)
		GPT (Generative Pretrained Transformer)	2018	One-way autoregressive (Auto-regressive) for text generation	GPT-2 (Radford et al. 2019), GPT-3 (Brown et al. 2020), GPT-4 (OpenAI Technical Report Chatgpt-4 2023)
		T5 (Text-To-Text Transfer Transformer)	2019	Unified task paradigm: text-to-text (Seq2Seq)	T5 (Raffel et al. 2020), FLAN-T5 (Chung et al. 2022)
		UL2 (Unifying Language Learning)	2022	Incorporate BERT, GPT, and T5 training methods to improve adaptability	UL2 (Tay et al. 2023)
		CTRL (Conditional Transformer Language)	2019	Controlled generation, user input to control text style	CTRL (Keskar et al. 2019)
B	Transformers for Reasoning and Optimisation	MoE (Mixture of Experts Transformer)	2021	Activates only some of the parameters to improve computational efficiency.	Switch Transformer (Fedus, Zoph, and Shazeer 2022), DeepSeek (DeepSeek-AI 2025)
		LoRA (Low-Rank Adaptation)	2021	Parameter efficient fine-tuning (PEFT) to reduce computational costs	LoRA (Hu et al. 2021), QLoRA (Dettmers et al. 2023)
		Reformer (Efficient Transformer)	2020	Reduce storage requirements and use LSH attention	Reformer (Kitaev, Kaiser, and Levskaya 2020)
		Linformer (Linear Transformer)	2020	Linear attention computation to reduce complexity	Linformer (Wang et al. 2020)
		BigBird (Sparse Attention Transformer)	2020	Sparse Attention, Long Text Support	BigBird (Zaheer et al. 2020)
C	Multimodal Transformer	Perceiver (General Perception Model)	2021	Universal input (text/image/video) to extend Transformer applicability	Perceiver (Jaegle et al. 2021)
		ViT (Vision Transformer)	2021	Suitable for image classification, direct processing of pixel blocks	ViT (Dosovitskiy et al. 2021), DeiT (Touvron et al. 2021)
		Swin Transformer (Shifted Window Transformer)	2021	Hierarchical window attention for computational efficiency	Swin Transformer (Liu et al. 2021)
		CLIP (Contrastive Language-Image Pretraining)	2021	Joint text-image training for cross-modal retrieval	CLIP (Radford et al. 2021)
		DALL-E (Text-to-Image Transformer)	2021	Generate images, text-based descriptions	DALL-E (Ramesh et al. 2021), DALL-E 2 (2022)
D	Long Context Transformer	GPT-4 V (Vision-capable Transformer)	2023	Text + image input for enhanced visual understanding	GPT-4 V (GPT-4 V (Vision) System Card 2024)
		Sora (Text-to-Video Transformer)	2024	Text to Video Generation	Sora (n.d.)
		ALiBi (Attention Linear Biases)	2021	Enhanced long text handling to avoid fixed position encoding	ALiBi (Press, Smith, and Lewis 2022)
		RWKV (RNN + Transformer Hybrid)	2023	Combining RNN and Transformer features to support long sequences	RWKV (Peng et al. 2023)
		Ring Attention	2024	Supports long text with millions of tokens	Claude 3 (Anthropic 2024), Gemini 1.5 (Pichai and Hassabis 2024)
E	Task Adaptation Transformer	RetNet (Retention Transformer)	2023	Memory enhancement to improve long-term interaction	RetNet (Sun et al. 2023)
		Instruction-Tuned Transformer	2022	Instruction Tuning for dialogue	InstructGPT (Ouyang et al. 2022), FLAN-T5 (Chung et al. 2022)
		RLHF (Reinforcement Learning from Human Feedback)	2022	Human feedback reinforces learning to optimise AI alignment	ChatGPT (Ouyang et al. 2022)
		RAG (Retrieval-Augmented Generation)	2020	Incorporate knowledge retrieval to improve the quality of responses	RAG (Lewis et al. 2020)
		Self-Refinement Transformer	2023	Model self-optimisation to reduce hallucinations	Claude 3 (Anthropic 2024)

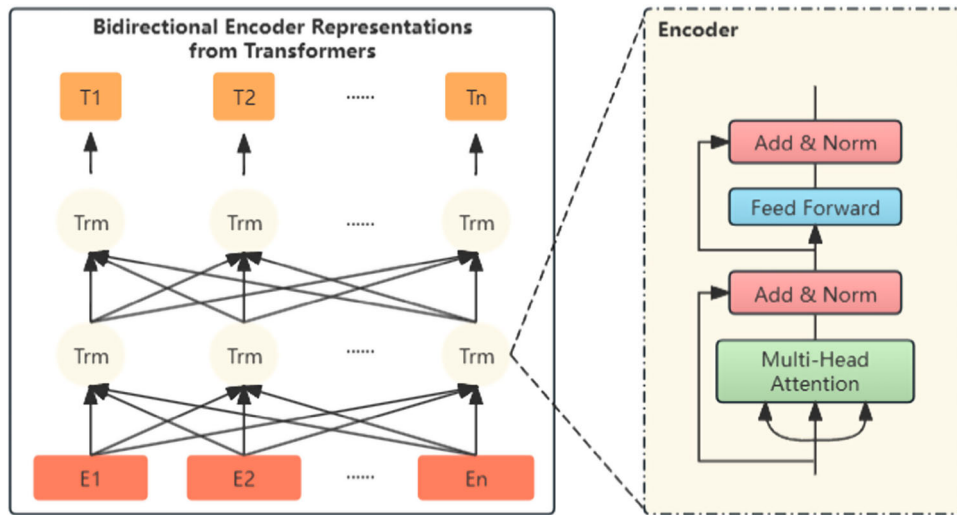


Figure A2. BERT adopts only the encoder component of the original Transformer architecture and emphasises the use of a bidirectional attention mechanism. By incorporating two pretraining tasks – Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) – BERT.

Appendix C. Representative model

Classic NLP transformer

Transformer models for natural language processing (NLP) are commonly structured into three types: encoder-only, decoder-only, and encoder-decoder. The encoder applies self-attention and feedforward layers to extract deep feature representations from the input. The decoder uses unidirectional self-attention and cross-attention to integrate the encoder's output and produce the final output. The encoder-decoder architecture captures context by using bidirectional self-attention in the encoder and cross-attention in the decoder to generate high-quality text (Zhang et al. 2025). Encoder-only models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) are bidirectional and optimised for language understanding via pretraining tasks such as masked language modelling (MLM) and next sentence prediction (NSP). Decoder-only models, such as GPT (Brown et al. 2020) and LLaMA (Touvron et al. 2023), follow an autoregressive pattern and excel in text generation. Encoder-decoder architectures, like T5 (Raffel et al. 2020) and BART (Bidirectional and Auto-Regressive Transformers) (Lewis et al. 2019), are well-suited for sequence-to-sequence tasks including translation and summarisation. These structures serve as foundational designs for numerous downstream models and adaptations. In supply chain operations, these models are used for information extraction from contracts, automatic document summarisation, and generation of communication templates for suppliers or customers.

Calculated optimisation transformer

To address the limitations of the standard Transformer in terms of computational efficiency, memory usage, and model size, researchers have introduced various structural optimisations. Parameter activation methods like Mixture of Experts (MoE) (Shazeer et al. 2017) and Switch Transformer (Fedus, Zoph, and Shazeer 2022) activate only subsets of the network, enabling trillion-parameter scaling (Shazeer et al. 2017). Efficient tuning approaches such as LoRA (Hu et al. 2021) reduce

training costs by updating limited parameters. Sparse attention models, such as Longformer (Beltagy, Peters, and Cohan 2020) and BigBird (Zaheer et al. 2020), enable longer input handling with reduced computation. Finally, architectural replacements like Reformer (Kitaev, Kaiser, and Levskaya 2020) and Linformer (Wang et al. 2020) reduce attention complexity. Their logical reasoning capabilities are especially relevant for demand forecasting, cost optimisation, and multilingual logistics scenarios, enabling support across complex international procurement environments.

Multimodal transformer

To support multimodal tasks such as image-text matching, text-to-image generation, and visual question answering, researchers have introduced a range of structural and processing enhancements to the original Transformer architecture. These core optimisations include extending input representations, designing modality alignment mechanisms, adjusting fusion strategies, and developing unified modelling paradigms. Multimodal Transformers extend the application of Transformers from text to other modalities such as images, video, and audio by modifying input formats and incorporating multimodal encoder fusion mechanisms. Models like Vision Transformer (ViT) (Dosovitskiy et al. 2021) process image patches as tokens using self-attention. CLIP (Radford et al. 2021) aligns image-text pairs via dual encoders for zero-shot classification. DALL-E (Ramesh et al. 2021) and GPT-4 V (GPT-4 V (Ision) System Card 2024) generate images from text, while Sora (n.d.) enables text-to-video generation. These models employ cross-modal attention, shared embeddings, or unified tokenisation to support tasks like visual question answering, captioning, and multimodal generation. Architectures like CLIP and GPT-4 V fuse image-text representations to support tasks such as OCR, visual inspection, and anomaly detection in packaging or goods handling. This is valuable in warehouse operations and quality assurance.

Long context & memory enhancement transformer

To address the computational bottlenecks and limited memory capacity of standard Transformers in handling long-context

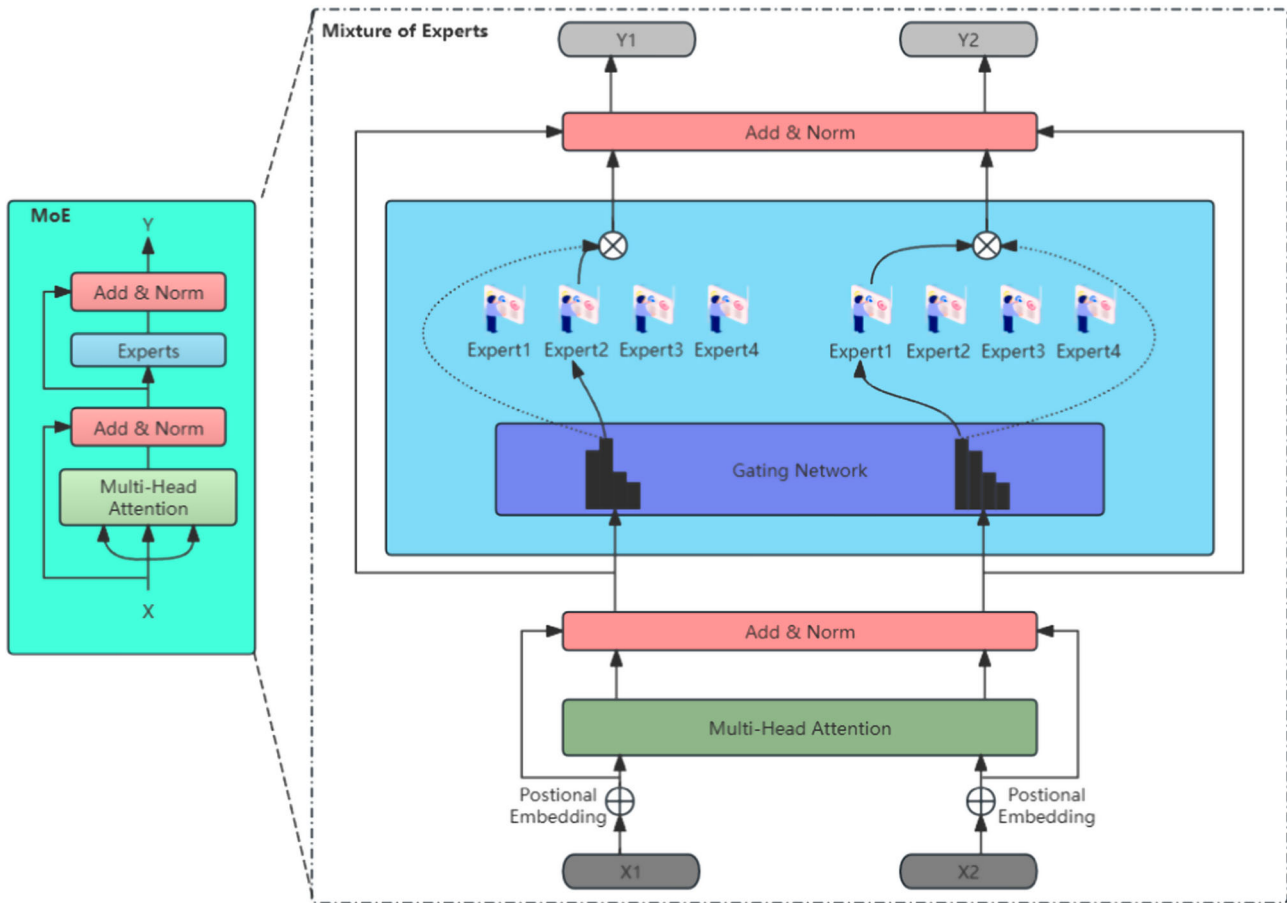


Figure A3. Mixture of Experts (MoE) architecture, which replaces the standard feed-forward layer with multiple expert subnetworks controlled by a learned gating mechanism. During inference, only a small subset of experts is activated, allowing the model to scale to trillions of parameters while maintaining computational tractability.

information, researchers have introduced targeted structural optimisations. These enhancements focus on redesigning the attention mechanism, replacing traditional positional encoding, improving state storage, and incorporating recurrent structures. Sparse attention architectures like Longformer (Beltagy, Peters, and Cohan 2020) and BigBird (Zaheer et al. 2020) reduce overhead via local/global windows. Recurrent-state models such as RWKV (Peng et al. 2023) combine RNN-like memory with Transformer parallelism, supporting streaming inference. Transformer-XL introduces segment-level recurrence and relative positional encodings, while RetNet (Sun et al. 2023) uses recurrence-based propagation to enable efficient long-form processing. These models extend Transformer usability to legal texts, code, and multi-document QA. This enables long-range supplier evaluation, production tracking, and scenario-based planning in multi-echelon supply chains.

Task adaptation and self-supervised learning transformer

Transformers for task adaptation and self-supervised learning are designed to improve generalisation, unsupervised

pretraining, and alignment with human intent. Although these methods are not architectural variants themselves, they reflect key strategies for task adaptation and alignment in the evolution of Transformer-based models. Instruction tuning enables models like FLAN-T5 (Chung et al. 2022) to follow natural-language prompts across diverse tasks. Reinforcement Learning from Human Feedback (RLHF), as used in InstructGPT and ChatGPT (Ouyang et al. 2022), aligns output behaviour with human preferences through reward modelling and Proximal Policy Optimisation. Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) integrates external knowledge during decoding, improving factual accuracy and transparency. These strategies enhance model reliability without modifying the core Transformer architecture. Their flexibility makes them ideal for decision support systems that personalise procurement recommendations or automate planning workflows. They are also applied in dynamic report generation and interaction with ERP or MRP systems.

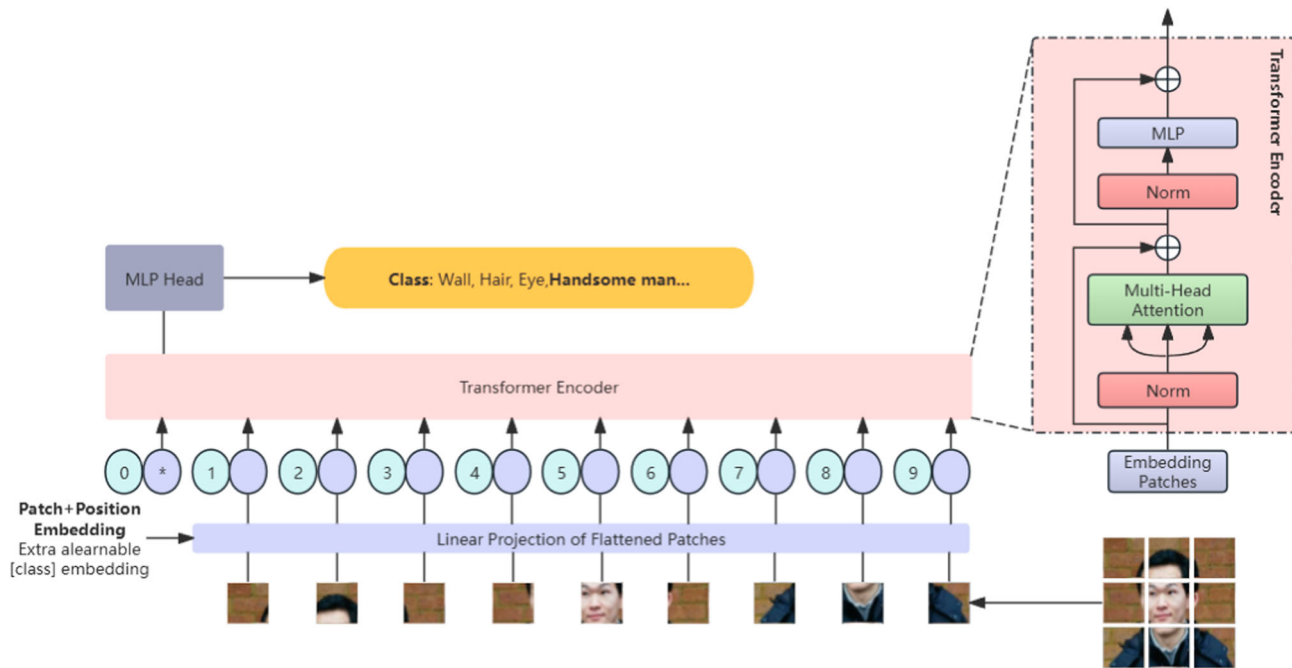


Figure A4. Vision Transformer (ViT), which treats images as sequences by partitioning them into fixed-size patches. Each patch is linearly projected into an embedding and combined with positional encodings before being processed by self-attention layers, enabling pure Transformer-based vision processing without Convolutional Neural Networks (CNNs). An MLP head performs final classification, while internal MLP blocks enhance nonlinear feature representation.

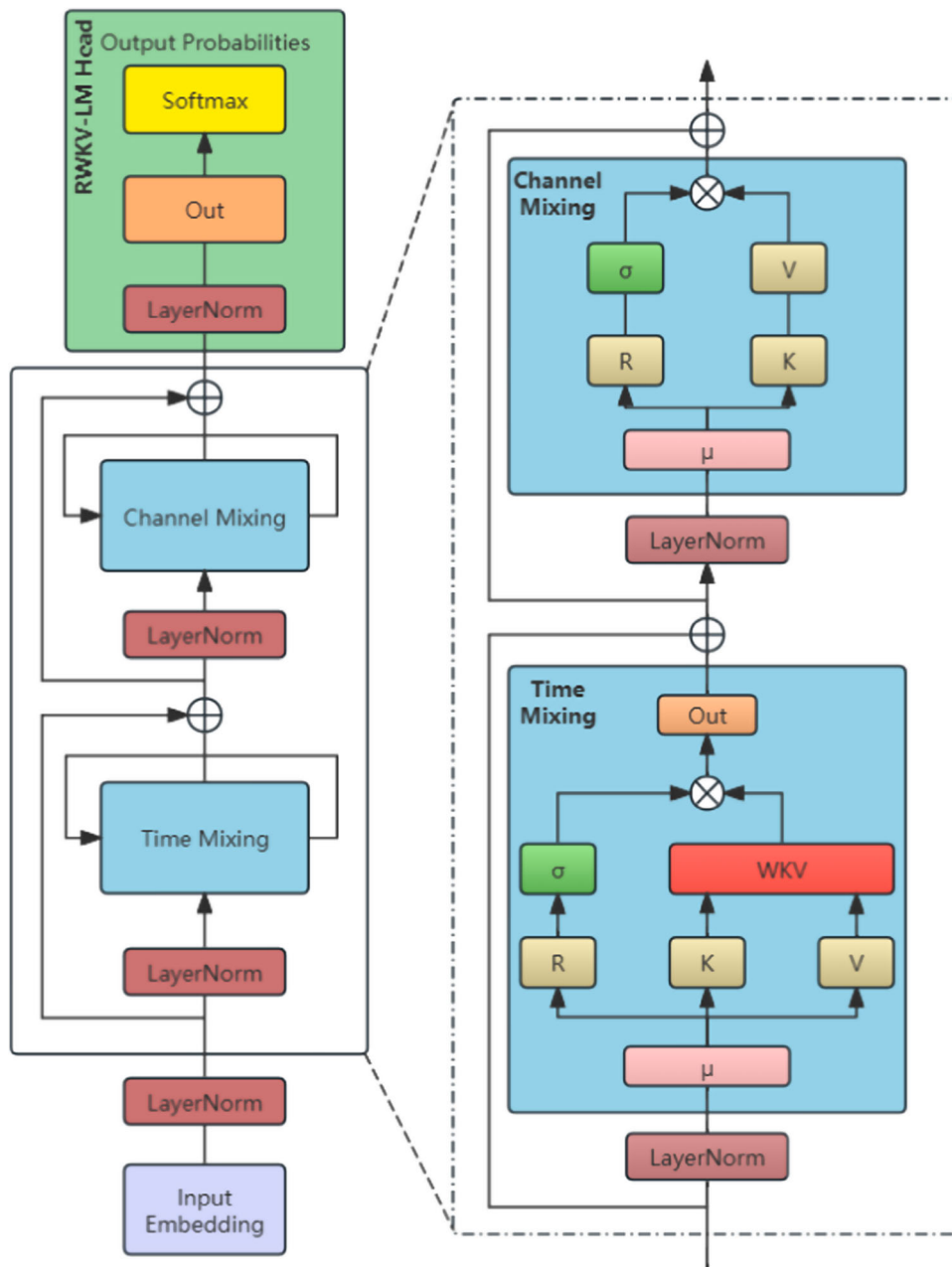


Figure A5. The Receptance Weighted Key Value (RWKV) model combines Transformer-style parallel training with an RNN-inspired recurrent state-passing mechanism. Through gating functions and decay-based memory, RWKV supports streaming input and captures long-range dependencies with reduced inference complexity, enabling efficient processing of contexts spanning millions of tokens without explicit self-attention.

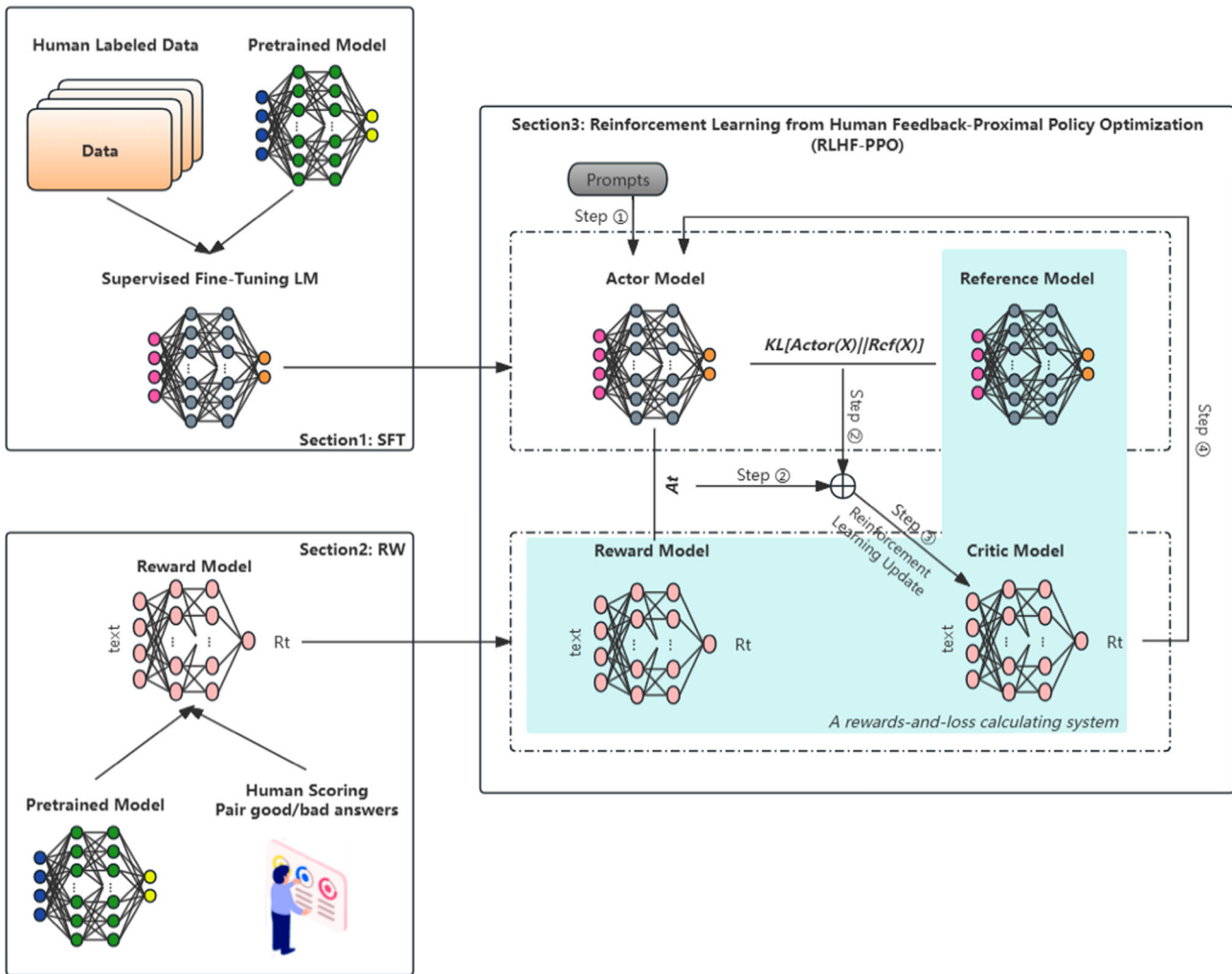


Figure A6. Reinforcement Learning from Human Feedback (RLHF), which enhances alignment by combining Supervised Fine-Tuning (SFT) with Proximal Policy Optimisation (PPO). Human-labelled preferences guide reward modelling, enabling the Transformer to generate output that better reflect human expectations, reduce hallucinations, and adhere to ethical standards.

Appendix D. LLMs competency matrix

	Model architecture and technical features							Training and data characteristics				
	Model architecture type*	Parameter scale	Context window size	Activated parameters (for MoE Models)	Tokenizer type	Model compression and distillation	Inference and training efficiency	Training data scale	Training data sources	Data diversity and quality	Training optimisation techniques	Model fine-tuning and instruction tuning capability
PaLM 1	E	540B	512	No	SentencePiece	None	Relatively High	Relatively large	Internet + Open Data	Medium	Optimized SGD	Limited
PaLM 2	C,E	Undisclosed	4K	No	SentencePiece	None	Efficiently	Relatively large	Internet + Open Data	Relatively High	Optimized SGD	Limited
Gemini 1	C,E	Undisclosed	32K	Partial activation	SentencePiece	None	Relatively High	Large scale	Internet + multimodal data	High	Optimized SGD + RLHF	Supported
Gemini 1.5	C,E	Undisclosed	1M	Partial activation	SentencePiece	None	Extremely high	Large scale	Multimodal data + Internet data	Extremely high	Optimized SGD + RLHF	Supported
Gemini 2.0	C,E	Undisclosed	2M	Partial activation	SentencePiece	None	Extremely high	Large scale	Multimodal data + Internet data	Extremely high	Optimized SGD + RLHF	Supported
Gemma2	C	Undisclosed	128K	No	Improved BPE	None	Efficiently	Relatively large	Open Source data + professional corpus	High	Optimized SGD + RLHF	Supported
ChatGPT-o1	E	Undisclosed	32K	No	Improved BPE	None	Efficiently	Large scale	Internet + Open Data	High	RLHF	Supported
ChatGPT-o1-mini	E	Undisclosed	32K	No	Improved BPE	None	Extremely high	Medium scale	Internet + Open Data	High	RLHF	Supported
GPT-3.5	E	175B	4 K/16K	No	BPE	None	Medium	300B+	Internet + Open Data	Relatively High	RLHF	Supported
GPT-3.5 Turbo	E	175B	16K	No	BPE	None	Efficiently	300B+	Internet + Open Data	High	RLHF	Supported
GPT-4	C,D	Undisclosed	8 K/32K	No	BPE	None	Relatively High	Hundreds of billions	Internet + Open Data + Professional corpus	Very high	RLHF	Supported
GPT-4o	C,D,E	Undisclosed	128K	No	Improved BPE	None	Extremely high	Hundreds of billions	Multi-source data + multimodal data	Extremely high	RLHF + multimodal tuning	Supported
LLaMA 3 8B	E	8B	128K	No	SentencePiece	None	Efficiently	15.6 T tokens	Internet + multimodal data	High	Optimized SGD + RLHF	Supported
LLaMA 3 70B	E	70B	128K	No	SentencePiece	None	Efficiently	15.6 T tokens	Internet + multimodal data	High	Optimized SGD + RLHF	Supported

(continued).

Table A2. Continued.

	Model architecture and technical features							Training and data characteristics				
	Model architecture type*	Parameter scale	Context window size	Activated parameters (for MoE Models)	Tokenizer type	Model compression and distillation	Inference and training efficiency	Training data scale	Training data sources	Data diversity and quality	Training optimisation techniques	Model fine-tuning and instruction tuning capability
LLaMA 3 405B	E	405B	128K	No	SentencePiece	None	Extremely high	15.6 T tokens	Internet + multimodal data	Extremely high	Optimized SGD + RLHF	Supported
Claude 3 Opus	C,D,E	Undisclosed	1M	No	Improved BPE	None	Extremely high	Large scale	Multimodal data + Internet	Extremely high	RLHF + Optimized SGD	Supported
Claude 3 Sonnet	C,D,E	Undisclosed	200K	No	BPE	None	Efficiently	Large scale	Multimodal data + Internet	High	RLHF	Supported
Claude 3.5 Sonnet	C,D,E	Undisclosed	200K	No	BPE	None	Extremely high	Large scale	Multimodal data + Internet	Extremely high	RLHF	Supported
Mistral 7B	B	7.3B	8K	No	BPE	None	Efficiently	1.4 T tokens	Internet + Open Data	High	Optimized SGD	Supported
Mixtral 8 × 22B	B	141B	64K	39B	BPE	None	Extremely high	Multi-T level	Multimodal data + Internet	Extremely high	Optimized SGD + RLHF	Supported
Mistral Small 3	B	24B	32K	No	BPE	None	Extremely high	1 T tokens	Internet + Open Data	High	Optimized SGD	Supported
Mixtral (MoE)	B,D	141B	64k	39B	BPE	None	Extremely high	Multi-T level	Multimodal data + Internet	Extremely high	Optimized SGD + RLHF	Supported
BERT	A	110M	512	No	Wordpiece	Supported	Medium	3.3B tokens	Wikipedia + BooksCorpus	Medium	Adam	Supported
RoBERTa	A	355M	512	No	BPE	Supported	Efficiently	160GB text	Common Crawl	High	Adam	Supported
DistilBERT	A	66M	512	No	BPE	Supported	Efficiently	82GB text	Wikipedia + BooksCorpus	Medium	Adam	Supported
BLOOM	E	176B	2K	No	BPE	None	Medium	1.6 T tokens	Multilingual data	High	Optimized SGD	Supported
BLOOMZ	E	176B	2K	No	BPE	None	Medium	1.6 T tokens	Multilingual data	High	Optimized SGD	Supported
ChatGLM	E	6.2B	2K	No	SentencePiece	None	Efficiently	1 T tokens	Chinese + English data	High	Optimized SGD	Supported
ChatGLM2	E	6B	32K	No	SentencePiece	None	Efficiently	1.4 T tokens	Chinese English bilingual data	High	Optimized SGD	Supported
ChatGLM3	D	6B	128K	No	SentencePiece	None	Extremely high	2 T tokens	Multilingual data	Extremely high	RLHF + multi-round optimisation	Supported

(continued).

Table A2. Continued.

	Model architecture and technical features							Training and data characteristics				
	Model architecture type*	Parameter scale	Context window size	Activated parameters (for MoE Models)	Tokenizer type	Model compression and distillation	Inference and training efficiency	Training data scale	Training data sources	Data diversity and quality	Training optimisation techniques	Model fine-tuning and instruction tuning capability
ChatGLM4	D,E	Undisclosed	128K	No	SentencePiece	None	Extremely high	10 T tokens	Multimodal data	Extremely high	RLHF	Supported
Falcon3 1B	E	1B	8K	No	BPE	None	Efficiently	< 100B tokens	High-quality text	Medium	Optimized SGD	Supported
Falcon3 3B	E	3B	32K	No	BPE	None	Efficiently	100B tokens	High quality data + code	High	Optimized SGD	Supported
Falcon3 10B	E	10B	32K	No	BPE	None	Extremely high	14 T tokens	Multilingual data	High	Optimized SGD	Supported
DeepSeek-v3	B,D	671B	100K+	37B	SentencePiece	None	Extremely high	20 T tokens	Internet + multimodal data	Extremely high	RLHF + self-optimisation	Supported
DeepSeek-R1	B,D	671B	100K+	37B	SentencePiece	None	Extremely high	20 T tokens	Internet + multimodal data	Extremely high	RLHF + self-optimisation	Supported
Qwen1.5-32B	E	32B	32K	No	BPE	None	Efficiently	15 T tokens	Chinese corpus + Internet	High	Optimized SGD + RLHF	Supported
Qwen2-7B	C,E	7B	128K	No	BPE	None	Efficiently	12 T tokens	Chinese + English data	High	Optimized SGD + RLHF	Supported
Qwen2.5-32B	C,E	32.5B	128K	No	BPE	None	Efficiently	18 T tokens	Internet + Chinese Corpus	High	Optimized SGD + RLHF	Supported
Grok-1	E	314B	8K	7.85B	Unknown	None	Relatively High	Large scale	Internet data	High	Unknown	Supported
Grok-2	C,E	Unknown	Unknown	Unknown	Unknown	Unknown	High	Unknown	Unknown	Unknown	Unknown	Unknown
Grok-3	C,E	Unknown	Unknown	Unknown	Unknown	Unknown	Extremely high	Large scale	Extended data sets	Unknown	Unknown	Unknown

Natural language understanding and generation capability	Logical and mathematical reasoning capability	Code generation and debugging capability	Competency dimensions						Extended	Safety and ethical dimensions						API availability	Hardware compatibility	Mobile and edge device support	Model openness and open-source status	Enterprise support and service level agreement (SLA)
			Core																	
			Multi-turn dialogue and context memory	Multilingual capability	Retrieval-augmented generation (RAG) capability	Personalisation and user adaptability	Tool invocation and plugin integration capability	Bias and discrimination control	Harmful content filtering	Fact consistency and authenticity	Data privacy and security	Model alignment	Local deployment capability							
65	55	45	512Kwindow	60	Not Supported	Limited	Not Supported	Fair	Relatively Strict	Medium	Standards	Fair	Not Supported	Available	TPU	Not Supported	Closed Source	Supported		
78	70	60	4Kwindow	75	Supported	Limited	Supported	Relatively Strong	Strict	Relatively High	Standards	Relatively Strong	Not Supported	Available	TPU	Not Supported	Closed Source	Supported		
85	80	75	32Kwindow	83	Supported	Relatively Strong	Supported	Relatively Strong	Strict	High	Standards	Strong	Not Supported	Available	TPU	Not Supported	Closed Source	Supported		
88	85	80	1Mwindow	86	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Not Supported	Available	TPU	Not Supported	Closed Source	Supported		
90	88	82	2Mwindow	89	Supported	Extremely Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Not Supported	Available	TPU	Not Supported	Closed Source	Supported		
85	82	78	128Mwindow	84	Supported	Relatively Strong	Supported	Relatively Strong	Strict	High	Standards	Relatively Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported		
88	85	80	32Kwindow	87	Supported	Limited	Not Supported	Relatively Strong	Strict	High	SOC 2 compliance	com- Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
80	78	75	32Kwindow	80	Supported	Limited	Not Supported	Relatively Strong	Strict	High	SOC 2 compliance	com- Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
70	65	48	16Kwindow	75	Supported	Limited	Supported	Fair	Relatively Strict	Medium	SOC 2 compliance	com- Medium	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
72	68	50	16Kwindow	78	Supported	Limited	Supported	Relatively Strong	Strict	Relatively High	SOC 2 compliance	com- Relatively Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
86	85	67	32Kwindow	85	Supported	Relatively Strong	Supported	Strong	Strict	High	SOC 2 compliance	com- Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
90	88	70	128Kwindow	89	Supported	Strong	Supported	Extremely Strong	Strict	Very high	SOC 2 compliance	com- Extremely Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
66.7	57.2	72.6	128Kwindow	68.9	Supported	Limited	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported		
79.3	83.7	80.5	128Kwindow	81.2	Supported	Limited	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported		
85.2	89	89	128Kwindow	91.6	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported		
90	87	85	1Mwindow	88	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
81.5	78	75	200Kwindow	82	Supported	Strong	Supported	Extremely Strong	Strict	High	Standards	Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
90.4	85	83	200Kwindow	86	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Not Supported	Available	GPU/TPU	Not Supported	Closed Source	Supported		
70	68	65	8Kwindow	72	Supported	Limited	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported		
85	90.8	80	64Kwindow	82	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported		
81	79.1	76	32Kwindow	80	Supported	Strong	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported		

(continued).

Competency dimensions																		
Competency dimensions									Safety and ethical dimensions					Application and deployment dimensions				
Natural language understanding and generation capability	Logical and mathematical reasoning capability	Code generation and debugging capability	Core		Extended				Harmful content filtering	Fact consistency and authenticity	Data privacy and security	Model alignment	Local deployment capability	API availability	Hardware compatibility	Mobile and edge device support	Model openness and open-source status	Enterprise support and service level agreement (SLA)
			Multi-turn dialogue and context memory	Multilingual capability	Retrieval-augmented generation (RAG) capability	Personalisation and user adaptability	Tool invocation and plugin integration capability	Bias and discrimination control										
85	90.8	80	64window	82	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
None	None	None	512window	None	Not Supported	Limited	Not Supported	Medium	Relatively Strict	Medium	Standards	Strong	Supported	Available	CPU/GPU	Supported	Open Source	Not Supported
88.5	None	None	512window	None	Not Supported	Limited	Not Supported	Medium	Relatively Strict	High	Standards	Strong	Supported	Available	CPU/GPU	Supported	Open Source	Not Supported
79	None	None	512window	None	Not Supported	Limited	Not Supported	Medium	Relatively Strict	High	Standards	Medium	Supported	Available	CPU/GPU	Supported	Open Source	Not Supported
39	None	15	2Kwindow	None	Supported	Limited	Supported	Relatively Strong	Strict	Medium	Standards	Medium	Supported	Available	GPU/TPU	Supported	Open Source	Supported
40	None	16	2Kwindow	None	Supported	Limited	Supported	Relatively Strong	Strict	Medium	Standards	Medium	Supported	Available	GPU/TPU	Supported	Open Source	Supported
54	None	None	2Kwindow	50	Supported	Limited	Supported	Relatively Strong	Strict	Medium	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
60	None	None	32Kwindow	55	Supported	Limited	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TUP	Supported	Open Source	Supported
70	None	None	128Kwindow	75	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
85	92	84	128Kwindow	87	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
54.4	None	None	8Kwindow	60	Supported	Limited	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
65	72	69	32Kwindow	70	Supported	Limited	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
73.1	83	74	32Kwindow	78	Supported	Strong	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
87.1	90.2	82.6	100K window	+ 89	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
90.2	92	85	100K window	+ 91	Supported	Strong	Supported	Extremely Strong	Strict	Very high	Standards	Extremely Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
73.4	77.4	37	32Kwindow	80	Supported	Strong	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
65	72	70	128Kwindow	68	Supported	Limited	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
78	85	78	128Kwindow	83	Supported	Strong	Supported	Relatively Strong	Strict	High	Standards	Strong	Supported	Available	GPU/TPU	Supported	Open Source	Supported
73	50	63.2	8Kwindow	Unknown	Supported	Limited	Not Supported	Unknown	Unknown	Medium	Open Source	Unknown	Supported	Unknown	GPT/TPU	Not Supported	Open Source	Unknown
Estimated 80	Estimated 60	Estimated 70	Unknown	Unknown	Supported	Limited	Unknown	Unknown	Unknown	Medium	Standards	Unknown	Not Supported	Available	GPU	Not Supported	Closed Source	Unknown
Unknown	Unknown	Estimated 85	Unknown	Unknown	Supported	Strong	Unknown	Unknown	Unknown	High	Standards	Unknown	Not Supported	Available	GPU	Not Supported	Closed Source	Unknown

Appendix E. Reviewed studies summary table

Author	Year	LLM type	SCM process	Key findings
W. Wang et al.	2025	BERT	Plan	Used BERT to extract legal-text insights to enhance credit risk prediction in supply chain finance.
Mishra & Senapati	2025	LLM-based Agentic AI	Plan→ Deliver	LLM-driven TDD automated retail decision-making and demand forecasting with expert-level accuracy.
Y. Li et al.	2025	RAG-LLM, KGs	Source	Combined LLMs with knowledge graphs to enhance supplier discovery and Q&A interpretability.
Huang et al.	2025	ChatGPT 4.0	Plan→ Return	Used ChatGPT to extract expert insights for stakeholder-driven sustainable supply chain strategies.
X. Hu et al.	2025	Transformer-based model	Make	Transformer-based model resolved decision conflicts in manufacturing scheduling for improved coordination.
S. Fan et al.	2025	GPT	Plan/Deliver	Used GPT to extract firm-level supply chain risk from site visit transcripts for financial forecasting.
W. Wang, et al.	2025	LLaMA3	Source	LLM-based hybrid model improved methanol price forecasting for supply chain risk management.
Breitung & Müller	2025	GPT-3	Plan/Source	Used GPT-3 to map global firm networks and identify supplier–customer–competitor relationships.
Phuengrod et al.	2025	LLM based	Plan	LLM-driven analytics supported research–industry alignment and project matching in innovation supply chains.
Lin et al.	2025	LSTM-based	Plan	LSTM-based model forecasted emergency drug demand for disaster-responsive pharmaceutical supply chains.
Aylak	2025	Transformer-based	Make/Deliver/Return	Transformer-based SCM system optimised logistics and reduced emissions through intelligent automation.
Gezdur & Bhattacharjya	2025	LLM based	Plan→ Return	LLMs enhanced supply chain agility, decision support, and knowledge management via unstructured data insights.
Xie et al.	2025	Swin Transformer V2	Make/Deliver	Swin Transformer improved log identification accuracy for traceability in timber supply chains.
Hamidi et al.	2025	NLP archiecture	Plan	Topic modelling of social media revealed demand shifts and panic buying patterns for responsive SCM.
C.-Y. Liu & Chou	2025	Llama3 fine-tuned	Plan/Source	LLM fine-tuned for legal text analysis improved procurement compliance in construction SCM.
Leite et al.	2025	RAG-LLM	Plan/Make	RAG-enhanced LLM improved decision-making and environmental control in livestock supply chains.
Taiwo et al.	2025	RAG-LLM	Plan/Make/Deliver	Used GenAI and RAG to optimise water network planning, quality monitoring, and maintenance.
Fontoura et al.	2025	GLM	Plan	GLMs supported ethical, energy-efficient planning in supply systems under BE-CSR principles.
Guo et al.	2025	LLM based	Make/Deliver	Used an LLM-aided E2E model to optimise multi-period inventory replenishment and cost control.
Jiang et al.	2025	Attention-based LLM	Plan	Attention-based LLM framework improved day-ahead PV forecasting accuracy under varying conditions.
Jannelli et al.	2025	LLM-based Agentic AI	Plan → Deliver	Automated consensus-seeking via LLM agents reduces the bullwhip effect and outperforms traditional inventory coordination approaches.
Fosso Wamba et al.	2024	ChatGPT / Generative AI	Plan → Make → Deliver → Return	Empirical survey evidence shows GenAI adoption improves cost-efficiency and customer experience but faces data and organisational challenges.
Sliz	2024	ChatGPT	Return	Applied ChatGPT to streamline after-sales processes, reducing lead times and improving service quality.
Shishehgharkhaneh et al.	2024	BERT, RoBERTa	Plan	Transformer-based NER enhanced supply chain risk detection from news data in construction sector.
Xin et al.	2024	GPT	Plan	GPT enhanced demand forecasting, pricing, and collaboration in online grocery supply chains.
Y. Li & Starly	2024	ChatGPT, KGs	Source	Integrated LLMs with MSKG to improve manufacturer recommendation and service identification.
S.-K. S. Fan & Chiu	2024	Vision Transformer	Make	Transformer-based vision model improved defect detection and sustainability in semiconductor manufacturing.
Bao et al.	2024	Swin Transformer	Deliver	Swin Transformer improved multi-scale logistics detection in agricultural supply via remote sensing.
D. Li & Fu	2024	LSTM-Transformer	Plan	Hybrid LSTM-Transformer model improved long-term sequence prediction for SCM-relevant forecasting.
Beyer Díaz et al.	2024	LSTM-BERT	Plan	Used BERT-enhanced LSTM to improve oil price prediction for supply planning insights.
Ni et al.	2024	LLM based	Make	LLM-supported metrics enabled smarter safety risk monitoring in manufacturing supply chains.
Cauli et al.	2024	Vision Transformer	Make/Source	Vision Transformer and QA system enabled automated part recognition and selection in industrial SCM.
Sangher et al.	2024	BERT	Plan	LLM-based models detected cybercrime patterns from dark web to secure digital supply chains.
Qu et al.	2024	T5	Plan	LLM-based framework predicted EV charging demand by fusing spatiotemporal and environmental data.

Jing et al.	2024	BERT	Plan	BERT-GCN fusion improved tamper detection in electricity supply chain data propagation.
X. Li et al.	2024	Multimodal LLM	Plan	Used LLM-enhanced multi-modal modelling to improve new product sales forecasting accuracy.
Raman et al.	2024	ChatGPT	Plan→Return	ChatGPT outperformed Bard in SCM Q&A tasks, improving answer accuracy via iterative queries.
Aljaž	2024	ChatGPT	Make/Deliver	Used ChatGPT within TOC TP to diagnose root causes of shelf stockouts and support retail decision-making.
Jackson et al.	2024	GPT-3 Codex	Plan	Used GPT-3 Codex to auto-generate logistics simulation models from natural language inputs.
Oh	2024	GPT-4	Source	Used GPT-4 prompts to extract accurate product names for improved sourcing data quality.
Chiu et al.	2024	LLM based	Source	Used NLP and LCA to assess green supplier risks and support sustainable sourcing decisions.
Q. Zhou et al.	2024	Transformer-LSTM	Plan	Used LLM-based forecasting to model hazardous waste trends and guide industrial SCM policy.
H. Zhou et al.	2024	Transformer based	Source	Used transformer-based OCR to extract structured data from shopping receipts for offline demand analysis.
Bo & Xu	2024	Transformer-PSO	Plan	Combined Transformer and PSO to enhance demand forecasting and reduce SCM costs.
Ahmed et al.	2024	Multimodal LLM	Source	Used DETR to detect crop diseases, improving upstream quality control in agri-supply chains.
G. P. Zhang et al.	2024	Transformer based	Plan	Transformer outperformed classical and neural models in forecasting sparse, intermittent demand.
Jahangir & Quilty	2024	Transformer based	Plan	Transformer outperformed LSTM in 24-hour water demand forecasting across urban districts.
Knott et al.	2023	Vision Transformer	Make	Used lightweight ViT to assess fruit quality in low-resource, decentralised food supply chains.
Cano-Marin et al.	2023	BERTopic	Deliver	Used NLP on social media to reveal global risks in COVID-19 vaccine supply chains.
Onoufriou et al.	2023	Multi-timeline Transformer	Plan	Multi-timeline Transformer improved yield forecasting under data scarcity in agri-supply chains.
Yazdinejad et al.	2023	BERT	Source	Used BERT to recommend alternative drugs in blockchain-based pharma supply chains.
Skórnióg & Kmicik	2023	ChatGPT	Plan	ChatGPT outperformed auto.arima in production forecasting despite workflow limitations.
Yang et al.	2023	Tspformer (Transformer based)	Deliver	Proposed Tspformer to optimise delivery routing via efficient transformer-based combinatorial search.
Ghimire et al.	2023	multi-head attention Transformer	Plan	Used multi-head attention Transformer (TNET) to improve power demand forecasting accuracy.
Vallés-Pérez et al.	2022	Seq2Seq-Transformer	Plan	Transformer-based model enabled efficient multi-store sales forecasting without retraining.
Bai et al.	2022	Transformer based	Plan	Used unsupervised MultiDBN-T to improve multivariate load forecasting accuracy by over 40%.
Zou et al.	2022	Transformer-AOA	Deliver	Used TAOA (Transformer + AOA + RL) to optimise low-carbon MDVRP with best delivery cost-efficiency.
Coltey et al.	2022	LLM-driven NLP	Plan/Source	Used LLM-driven NLP to support sourcing, regulation guidance, and crisis-aware supply decisions for SMEs.
Garg et al.	2021	LLM based	Deliver	Proposed LLM-enabled semantic queries to improve logistics tracking and transport decision-making.
Satirapiwong & Siriborvornratanakul	2021	LLM-driven NLP	Source	Automated invoice information extraction using NLP to support SCM document processing.
Wan et al.	2021	LLM-driven NLP	Source	Used text mining to evaluate suppliers from dialogue data, improving sourcing decisions.



DSPACE

<https://dspace.org/>

Large language models in supply chain management: a systematic literature review and application framework

Song, Zhe; Xie, Ying; Yang, Lichao; Zhao, Yifan

2026-12-31

Taylor & Francis

<https://doi.org/10.1080/00207543.2026.2641103>

Song Z, Xie Y, Yang L, Zhao Y. (2026) Large language models in supply chain management: a systematic literature review and application framework. International Journal of Production Research, Available online 13 March 2026