Magazine Article / Supply Chain Management

How Generative AI Improves Supply Chain Management

It can cut decision-making time from days to minutes and dramatically improve results. *by Ishai Menache, Jeevan Pathuri, David Simchi-Levi, and Tom Linton*

From the Magazine (January–February 2025) / Reprint R2501F

Alana Paterson

Companies face a variety of complex challenges in designing and optimizing their supply chains. Increasing their resilience, reducing costs, and improving the quality of their planning are just a few of them. Over the past few decades, advances in information technologies have allowed firms to move from decision-making on the basis of intuition and experience to more automated and data-driven methods.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

As a result, businesses have seen efficiency gains, substantial cost reductions, and improved customer service.

Unfortunately, business planners and executives still need to spend considerable time and effort to understand the recommendations coming out of their systems, analyze various scenarios, and conduct what-if analyses. Updating the supply-chain-management tools' mathematical models to reflect changes in the business environment is time-consuming as well. To address these issues, planners and executives have had to pull in data science teams or the technology providers to explain results or make changes in the system.

Recent advances in large language models (LLMs), a type of generative AI, are now making it possible to perform these activities without such support, reducing the time to make decisions from days and weeks to minutes and hours and dramatically increasing planners' and executives' productivity and impact. In this article we'll explore how LLMs can be used to generate insights from data that will give executives a better understanding of the state of their supply chains, answer what-if questions, and update supply-chain-management tools in order to take into account the current business environment. We'll also highlight the challenges that companies must overcome to adopt LLMs and the opportunities for expanding their applications in the future.

The experiences we will share are drawn mostly from Microsoft's employment of an LLM-based system to manage the supply of servers and other hardware to more than 300 data centers around the world in support of its cloud services. Microsoft piloted its LLM-based system from March 2023 to October 2023 before fully deploying it in November 2023. Since then, the system has had a notable impact on efficiency and productivity, measured by incident response time and speed in making

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

decisions, and those gains are expected to increase as the system is refined even more over time. The capabilities that we discuss are not dependent on the use of a Microsoft product, though; a wide variety of the high-quality LLMs available today could be used to achieve them.

Now let's explore the benefits that LLMs can provide.

[1]

Data Discovery and Insights

Consider a classic supply chain with a certain number of suppliers of raw materials, factories for producing products, and retailers that sell those products. With an LLM a planner can ask in plain language questions such as "How much raw material of type T does supplier S have today?" or "What is the cheapest option for shipping items from factory F to retailer R?" The LLM can translate those questions into data science queries that in turn are fed into the company's data repository (for example, a SQL database) and then provide an answer in a complete sentence. Significantly from a privacy perspective, the LLM can be utilized as a cloud service, which means that propriety data does not need to be transferred to a third-party LLM.

Beyond serving as a tool to understand the current state of a company's supply chain, the LLM can be used to explain decisions made by the supply chain system and provide additional insights, such as information about trends. For example, a planner could ask questions about recent trends—"Which factory was the most productive last week?" or "What was the number or percentage of instances last month when the total shipping cost exceeded $50,000?" In what follows, we provide concrete examples from early adopters of LLMs for data discovery and insights.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

Tracking shifting demand. Cloud computing is a multibillion-dollar business that requires providers such as Amazon, Microsoft, Google, and others to make large investments in building data centers, equipping them with hardware, and operating them so that their capacity is readily available. They must constantly satisfy the growing demand for those services while minimizing hardware and operational costs. To that end, cloud service providers periodically make hardware-deployment decisions that take into account many cost considerations, such as shipping and the depreciation of hardware, and operational considerations, such as hardware compatibility, inventory, and personnel available to carry out server deployments.

At Microsoft the demand for servers comes from internal business groups that own different cloud offerings (for example, Azure Storage, Azure Virtual Machines, and Microsoft 365). A demand is specified via a request, which includes the type and number of servers required, the region where the servers should be docked, and the ideal dock date. Using those requests as inputs, the supply chain team regularly generates a single demand plan. Microsoft's engineers periodically run a computerized optimization tool to produce a fulfillment plan that assigns the actual hardware from supply warehouses and specifies when they will be shipped to the data centers. Microsoft planners oversee the fulfillment plan. Their tasks include confirming that the fulfillment plan meets the businesses' needs and ensuring that the servers have been deployed according to the fulfillment plan. The deployed servers are typically utilized by the business group for multiple years until they are decommissioned.

The LLM can be used to explain decisions made by the supply chain system and provide additional insights, such as information about trends.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

Planners also monitor changes in the demand plan, called the demand drift, on a monthly basis to ensure that the revised plan fulfills all customer requirements and falls within budget guidelines. The task of evaluating the demand drift is traditionally done by the planners, who often involve data scientists and engineers from different business units in the process. Once the changes are understood, the planners prepare an executive summary to explain the changes for each region.

LLM-based technology now does all this. It automatically generates an email report that details who made each change and the reason for doing so. It also points out potential errors that planners can review. For example, suppose that demand (the total number of servers) in the new plan is lower than in the old. The email can point to the exact reason the demand decreased—for example, the introduction of a new, more-efficient generation of hardware that allows fewer servers to be used. This LLM tool allows planners to complete the demand-drift analysis on their own in minutes; previously, it would take them about a week.

Enforcing contracts. In the automotive industry, original-equipment manufacturers (OEMs) such as Ford, Toyota, and General Motors have thousands of suppliers and multiple contracts with each one. These contracts specify the details of the price paid by the OEM, quality requirements, lead times, and the resiliency measures suppliers must take to ensure supply. After feeding the LLM data from thousands of contracts, one OEM was able to identify price reductions it was entitled to for surpassing a certain volume threshold. Its procurement team had overlooked that opportunity because of the complexities and the number of the contracts. The result was millions of dollars in procurement savings.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

[2]

Answers to What-If Questions

An LLM allows planners to ask detailed questions. Here are a few examples: "What would be the additional transportation cost if overall product demand increased by 15%?" "What would be the additional procurement cost if retailer R uses products only from factory F?" "Can we fulfill all demand if we shut down factory F?" "How much would the total cost of producing product P be reduced if the cost of type M raw material were $1 less per unit?"

Here's how an LLM can answer questions like these accurately and efficiently. Many optimization tasks are written in the form of mathematical programs, which take into account the structure of the supply chain and all the business requirements and generate effective supply chain recommendations. An LLM doesn't replace the mathematical model; rather, it complements it. Specifically, it translates a human query into a mathematical code that is a small change to the original mathematical model used to produce the plan. For example, mandating that a retailer use products from a particular factory can be done by adding a mathematical requirement, or "constraint," that prohibits other factories from sending products to that retailer. This small change in the mathematical model is then fed to the supply chain tool to produce a modified plan, which is used only for comparison with the existing one. As before, the output of the new mathematical model is then passed through the LLM to produce the answer in human language. (To learn about this approach of using an LLM to obtain current information about the supply chain and to ask what-if questions, you can find Microsoft's open-source code and benchmarks at github.com/microsoft/OptiGuide.)

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

Consider how planners in Microsoft's cloud service operation use this capability to create a fulfillment plan that assigns servers to be shipped from warehouses to data centers. For each request, the main decisions consist of (1) the server type and the warehouse that will be used to fulfill the demand, (2) the shipping date, and (3) the docking target of the servers (the specific data center and the specific location within it). The goal is to minimize the total cost of the multiple components, such as the shipping costs and the estimated opportunity cost of delaying server deployments beyond the ideal date.

Alana Paterson photographed offshore tankers and container ships for a story about the effects of ocean noise on sea life.

When planners receive the output of the optimization tool, they can confirm that it meets business needs and ensure that it is executed accordingly. However, the underlying optimization issues are so complex that it can be difficult, if not impossible, to immediately understand the reasoning behind each decision. Consequently, planners often reach out to the engineers and data scientists who

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

developed the optimization tool to obtain additional insights. The planners and the engineers often need multiple rounds of interactions in order to fully explore an issue or a what-if scenario, which might result in a delay of days. Now the LLM-based system allows planners to obtain in a few minutes answers to questions such as "What is the percentage increase in cost if we fulfill a specific order by the requested date versus another date?" and "What would be the cost increase if we deactivated a certain warehouse for one week?"

[3]

Interactive Planning

Planners can use LLM technology to update the mathematical models of a supply chain's structure and the business requirements to reflect the current business environment. Further, an LLM can update planners on a change in business conditions.

Let's say planners have received real-time information that a specific manufacturing facility is down for seven days owing to a winter storm. To update the sales and operation plan (S&OP) to account for the disruption without the assistance of an LLM, the planners would have to engage the IT and data science teams to make the necessary plan adjustments—a process that might be time-consuming. With the aid of an LLM, however, planners can directly ask the system to generate a new plan that avoids using the disrupted facility. If the new plan will not be able to satisfy all forecasted demand, the LLM-assisted planning tool will not only generate an updated S&OP and the corresponding costs (for example, procurement and transportation costs) but also identify the demand that cannot be supplied and the impact on profitability.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

In the next few years LLM-based technology will support end-to-end decision-making scenarios.

The need to change the supply plan may also be driven by LLM-based technology. For example, after analyzing shipment data from a specific supplier, it may generate an alarm that the lead time from the supplier has increased significantly over the past few months. Further, the LLM-based technology will predict the likely timing of the next shipment and will communicate that to the planner. Recognizing that the increase in delivery lead time will adversely affect the service level in a specific region unless corrective actions are taken, the planner may ask the LLM-based system to rerun the planning tool with the new information and generate a new plan. That plan, which the LLM conveys to planners in plain human language, may call for expedited shipments from the supplier or the transfer of inventory from a warehouse the company has in a different region to the affected region.

Using LLMs in the ways discussed here is still relatively new. We envision that in the next few years LLM-based technology will support end-to-end decision-making scenarios. For example, users may be able to describe in plain language the decision-making problem that they wish to solve. It could be a specific production problem (given a complex network of manufacturing facilities, where and when to produce a certain product) or an inventory-allocation problem (given limited inventory at a warehouse, how should it be allocated to various stores to meet demand most efficiently). The technology today is capable of generating such (mathematical) models and a recommendation, but validating that the model correctly represents the business environment is currently still a challenge.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

[4]

Overcoming Barriers

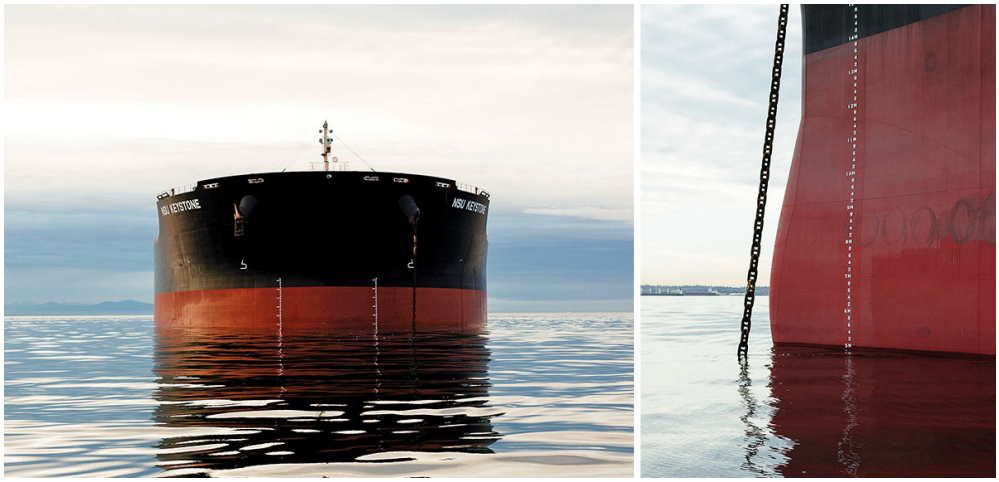
As companies begin to adopt LLMs in supply chain management, they will need to overcome a variety of obstacles to deploying them effectively.

Adoption and training. Using an LLM to optimize supply chains requires very precise language. For example, if a user asks, "Can we utilize factory F better?" the word "better" can be interpreted in multiple ways: lower costs, higher throughput, a more balanced throughput over time, and so on. Each interpretation leads to different decisions. Therefore, it is crucial that the people using the system are given training. Planners may need to be trained to ask more-precise questions, and managers and executives may require schooling in the capabilities and limitations of LLM-based technology.

For those reasons, Microsoft is gradually deploying this new technology, and the tool described earlier for answering what-if questions supports only a set of common questions. The company will monitor user interactions, accuracy, and fallback mechanisms and then expand coverage over time. Planners have received training about the technology and have been given the set of questions that the tool currently supports.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.



Alana Paterson

Dell Technologies also recognizes the importance of upskilling its workforce to use LLMs in its supply chain. The company's early experience with using AI through a partnership with a supply-chain-application provider has created both a hunger for what is possible and an urgency to prepare people to manage AI. "Training our human workforce to ask the right questions of generative AI is proving to be more of a challenge than the technology itself," Sasha Pailet Koff, a senior vice president, told one of us (Tom). She added: "Developing leaders who can manage generative AI versus a human workforce is critical. Only then can companies choose the right application workflows best suited for AI decision-making."

Verification. LLM technology may occasionally produce a wrong output. Thus, a general challenge is to keep the technology "inside the rails" —namely, identify mistakes and recover from them. Companies are currently dealing with this challenge by providing the LLMs with rich domain-specific examples that increase the accuracy of their outputs and adding mechanisms for identifying ahead of time queries that the technology does not support. For instance, if a nonsupported question is asked, the LLM-based system provides a default answer such as "Unfortunately, I cannot help with this question; here is the list of

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

supported questions." Naturally, the difficulty of verifying accuracy will increase with the complexity of the output. For example, suppose we let the LLM generate an entire mathematical program to create an optimized fulfillment plan from scratch. How would the system verify that it is correct? And how can we ensure that the program will produce an optimal plan in a reasonable amount of time? These are still open questions that require further research.

New workforce roles. As LLM technology leads to a high degree of automation, the role of executives and planners will change. Instead of engaging in a time-consuming decision-making process that is prone to human errors, planners will be able to apply LLM technology to generate more insights into, and to explain the recommendations of, their supply-chain-planning technology. This will lead to a higher level of user trust and significant adoption of the tools' recommendations. Similarly, in procurement, employees will need to spend much less time generating new contracts; LLMs will be able to design contracts for a specific product category and provide information about the past performance of various suppliers to help executives choose the appropriate one.

Put another way, a workforce that uses LLM-based tools will be able to shift its focus from day-to-day repeated tasks to value-added activities, such as thinking strategically about various supply chain activities or collaborating internally across function areas and externally with suppliers and customers. For example, demand planners can collaborate with trade planners (who are responsible for marketing, pricing, and discounting) to understand the impact of trade on the demand forecast. In our experience, that sort of collaboration does not currently exist in most organizations. The challenge, of course, is to ensure that leadership breaks down the walls that exist

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

between functional areas and modifies business processes to enable collaboration.

...

Despite these challenges, we are confident that LLM-based technology will transform supply chain management in the near future—enhancing its efficiency, resiliency, productivity, and accuracy. It will complement today's supply chain technologies, allowing planners to interact directly with their supply chain tools without the need for data scientists or engineers. Firms will be able to automate a significant number of supply chain processes and even create new ones, such as by integrating the trade and forecasting processes. Indeed, that integration would result in a closed-loop supply-chain-management system, where the trade, supply chain, and finance functions collaborate to develop a supply plan that adheres to all business and financial objectives and requirements. Within a few years LLM-based technology could truly revolutionize supply chain management.

A version of this article appeared in the January–February 2025 issue of Harvard Business Review.

IM

Ishai Menache is the partner research manager of the machine learning and optimization group at Microsoft Research.

JP

Jeevan Pathuri is a general manager and the partner director of software engineering in Microsoft's cloud supply chain group.

David Simchi-Levi is the William Barton Rogers Professor at the Massachusetts Institute of Technology, the head of the MIT Data Science Lab, and an Accenture Luminary.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.

Tom Linton is a senior adviser to McKinsey & Company and was previously the chief procurement and supply chain officer at Flex.

Copyright © 2025 Harvard Business School Publishing. All rights reserved.

This document is authorized for use only by PONTICIFIA UNIVERSIDAD CATOLICA DEL PERU Juan Rojas (jsrojas@pucp.edu.pe). Copying or posting is an infringement of copyright. Please contact customerservice@harvardbusiness.org or 800-988-0886 for additional copies.