When Supply Chains Become Autonomous

by Carol Long (Harvard University), David Simchi-Levi (MIT), Andre P. Calmon (Georgia Tech), Flavio P. Calmon (Harvard University)

Summary:

A testbed built around one of management education's most enduring simulations, the MIT Beer Distribution Game, has shown that the latest generation of generative AI models can now autonomously manage supply chains. Systems using advanced reasoning models like GPT-5 and Llama 4 adapted to changing conditions, minimized costs, and overcame the bullwhip effect. But managers should be aware that success depends on model selection, guardrails, curated data sharing, and prompt design. Such autonomous AI agents will allow human managers to focus on higher-value functions.

Less than a year ago, it seemed like that day when generative AI would bring about a new era of supply chain autonomy—one where AI could adeptly make all the inventory and logistics decisions—was still far off. But to the astonishment of many experts, including us, that day has arrived—at least in the lab.

In this article, we will share our findings on the capabilities of current generative AI models to manage supply chains autonomously and provide some high-level advice on how to build such a system.

Automated vs. Autonomous Systems

For a decade, supply chain leaders have raced to automate processes by deploying robots, building digital twins, and designing optimized data-driven, inventory-management policies. This wave of automation has enabled faster operations, reduced errors, and led to supply chains that operate according to carefully designed sets of rules. Yet automation has a ceiling. Humans still write the rules, coordinate across functions, and make management decisions. Automated supply chains adapt by applying the given rules but cannot learn, reason, or manage the fundamental tradeoffs that define supply chain operations. In contrast, supply-chain-management systems powered by generative AI could have the capability to operate autonomously.

Using a simulation model that we built in our lab, we tested whether newly released gen AI reasoning models could manage supply chains autonomously, coordinating demand forecasting, inventory planning, and replenishment decisions across multiple functions with minimal human oversight. The results were striking. A system comprised of multiple agents—each powered by the same state-of-the art gen AI model like either GPT-5 or Llama 4—that shared information turned in a stellar performance. Such systems even outperformed more than 100 undergraduate students at Georgia Tech's Scheller College of Business, cutting total supply chain costs, which consist of backorder penalties for unfulfilled demand and holding costs for excessive inventory, by as much as 67% compared to the performance of the students. By comparison, older-generation large language models, including those many firms still use, often fail catastrophically in our simulations, generating supply chain costs up to five times higher than human teams achieve.

We also found that such a system can learn and adapt as conditions change with minimal human intervention: It can learn from their environment, anticipate bottlenecks, and adjust strategies in real time. This is the first evidence that gen AI can handle the cross-functional complexity that human supply-chain managers navigate daily.

Our Simulation Model

Most companies don't train their own AI models. They use frontier models like GPT-5, Claude, and Llama 4 off the shelf, accessed through standard interfaces with minimal customization. The question, therefore, that we explored wasn't how to build better models. It was how to effectively deploy models that already exist.

Our research asked the following fundamental questions: When gen AI models are used as is, with natural language prompts and no model modification, can autonomous agents effectively manage complex supply chain operations? And what strategies must supply chain managers master to orchestrate these off-the-shelf models successfully?

We built the first autonomous supply chain testbed around one of management education's most enduring simulations: the MIT Beer Distribution Game. For nearly 70 years, this deceptively simple exercise has humbled MBA students and seasoned executives alike. Developed in the 1950s by Jay Forrester to explain puzzling production swings at General Electric, the game captures the essential dynamics of any supply chain: information delays, coordination failures, and the human tendency to overreact under uncertainty.

The game works as follows. Four players—retailer, wholesaler, distributor, and factory—form a serial supply chain. Each week, every player makes one decision: how much to order from their upstream partner. The goal is straightforward: meet customer demand at the lowest total cost, balancing inventory expenses against costly backorders. The structure makes this deceptively difficult. Players operate in silos and cannot communicate. Only the retailer sees actual end-customer demand. Built-in shipping and ordering delays magnify uncertainty. When humans play, the result is almost always the same: chaos.

A minor spike in demand cascades into disaster. The retailer, seeing a short-term uptick, orders slightly more as a buffer. The wholesaler interprets this larger order as a lasting surge and orders even more in turn. The distributor and factory amplify the signal further. This chain reaction, known as the bullwhip effect, transforms small fluctuations into massive swings in inventory and cost. Delayed shipments eventually flood the system, leaving everyone drowning in excess stock. This cycle is nearly impossible to escape, even when you know it's coming.

In our testbed, four autonomous AI agents, each powered by one type of large language model such as GPT-5, operated the same supply chain under identical constraints. They faced the same information silos, delays, and pressure to avoid stockouts. Like humans, they had to anticipate demand, manage inventory, and coordinate implicitly across the chain. Unlike humans, they can be systematically orchestrated, with their decision-making guided by policies, data sharing, and carefully designed prompts.

We ran hundreds of simulations, testing different models and inference-time methods—techniques that optimize how models are used rather than how they are trained. These methods included designing better prompts, controlling what data to share, and implementing guardrails to limit the range of acceptable actions. We benchmarked AI performance against that of humans: We used data from 12 Georgia Tech cohorts with more than 100 undergraduate students in total who played the Beer Game over the past two years, all operating under the same system conditions as the gen AI testbed. In our best-performing setup—using Llama 4 Maverick 17B with optimized prompts, data-sharing rules, and guardrails—the AI agents reduced costs by as much as 67% relative to the student teams.

How Different Models Performed

Our experiments revealed a sharp divide in the capability of current gen AI models. The release of GPT-5-class models in the summer of 2025 marked the widespread availability of a new generation of reasoning models that fundamentally differ from their non-reasoning predecessors. Earlier non-reasoning models solved problems by matching patterns from their training data; they could predict how to respond to queries linearly but were limited in their capability for making decisions with clear, structured logic. The new generation of reasoning models break down complex problems into manageable steps, solving them through explicit logical reasoning; they are guided by plan-execute-reflect loops, where reasoning continuously updates the plan as the model works toward a solution, enabling truly adaptive decision-making.

Reasoning models outperformed non-reasoning models by wide margins in our testbed. In addition, providing each model with the right information—tailored to supplement its capabilities—and policies that constrain the range of acceptable decisions improved performance across the board.

Important Factors Affecting Performance

Four factors determine whether autonomous gen AI agents succeed or fail in supply chains.

1. A capable, reliable model. Model selection matters most—no amount of orchestration can fix a model that cannot understand the task or follow instructions. An agent's core reasoning capability directly drives supply chain costs and stability. Less-capable models amplify system noise (i.e., misleading signals about real demand) into costly bullwhip effects while more capable models can dampen it.

To test reliability of LLM models, we conducted many identical runs of the simulation for each model. In our decentralized setup—one in which no information is shared across the gen AI agents—we found many popular models like Llama 3.3 70B and GPT-4o mini are highly inefficient; they produce pronounced bullwhip effects, and their costs are an order of magnitude higher than those of human teams. All models showed instability in terms of their performance across identical runs (i.e., their outputs were unpredictable, inconsistent, and often degraded in quality or reliability over time), with total costs varying from 13% to 46% of the mean. Llama 4 Maverick 17B exhibited the greatest variability.

Worse, some models simply failed to follow instructions, causing systemic breakdowns. In our trials, models like Microsoft's Phi-4 and DeepSeek-R1-0528 failed to generate their decision in the required format in over 25% of cases.

However, the latest generation of models with advanced reasoning capabilities produced a clear leap in performance. For example, upgrading agents from GPT-4o mini to GPT-5 mini cut total supply-chain costs by 70%. Similarly, the newer and smaller Llama 4 Maverick 17B model dramatically outperformed its much larger predecessor, Llama 3.3 70B, reducing costs by 82%.

The superior performance of advanced reasoning models can be attributed to the policy they adopt to make decisions. A striking observation was that the newer reasoning models frequently applied the classic order-up-to policy—raising inventory positions to a target level—whereas older reasoning models often failed to articulate a coherent rationale for their decisions.

2. Guardrails to limit costly errors. Policies that constrain a gen AI agent's range of possible actions can materially improve both efficiency and reliability. For instance, a policy might cap order quantities or prevent new orders once inventory exceeds a set threshold.

In our experiment, a simple budget constraint proved highly effective. Each gen AI agent was given a fixed budget; orders could not exceed the available funds. In the real world, this hard guardrail works by preventing human purchasing agents from making panic buys. When an agent faces a stockout and attempts to place a massive order, the budget acts as a brake, forcing a more measured response and curtailing the amplification of misleading demand signals up the chain that can lead to bullwhip effects.

The results were dramatic: Total costs dropped 25% for GPT-5 mini, 39% for GPT-4o mini, and 41% for Llama 4 Maverick 17B. For capable models like Llama 4 Maverick 17B whose performance without the guardrail had been unstable, variation in performance across runs fell from 46% to 37%.

3. Curated data shared through a central orchestrator. LLMs don't reason like humans. The data that helps your team can distract an AI agent, leading to worse decisions and higher costs. So, be selective and test what data you share with an AI agent. For more capable gen AI models, less is often more.

To test how information sharing affects agent performance, we introduced a central "orchestrator," an agent with full visibility across the supply chain, responsible for sharing specific, curated data with the agents playing the game. We tested two information-sharing strategies, where the orchestrator shares information but makes no decision, and found that more data is not always better.

- **Scenario 1: Share real-time customer demand.** When the orchestrator shared only the current week's end-customer demand, performance improved across the board. Total costs fell by approximately 18% for GPT-5 mini, 25% for Llama 4 Maverick 17B, and 38% for GPT-4o mini.
- **Scenario 2: Share demand history and analysis.** When we also provided a five-week demand history and a volatility analysis, the results were mixed. This richer data significantly helped less-capable models (costs for GPT-4o mini fell by 69%). But for more capable models, the extra information was a distraction, and they performed worse than when they only received information on real-time demand.

Notably, other data points that typically help humans—such as inventory position or pipeline inventory—offered little benefit and often made the bullwhip effect worse.

4. Fine-tune performance with better prompts. Prompt design can significantly improve the performance of less-capable models, but it may offer limited benefit for more-capable models. For more-capable models, robust guardrails and curated data matter more.

Because LLMs are probabilistic, the way you frame a task matters. Reframing the objective (i.e., the instruction given to LLM) from the general goal of "minimize total costs" to the more specific "minimize the weighted average of backlog and holding costs" produced large gains for less-capable models, cutting costs by 44% for GPT-4o mini and 33% for GPT-4.1 mini. For more capable models, the effect was negligible.

A New Paradigm for Supply Chain Management

Our autonomous supply chain testbed demonstrates that gen AI agents, when provided with tailored information and policies that constraint the range of actions and orchestrate the flow of information among them, are capable of managing multi-function supply chain systems. It means that the capabilities of gen AI models have progressed to the point that autonomous systems—those that learn, adapt, and coordinate across functions in real time—are achievable and can replace both human-managed systems and automated systems that follow human-designed rules.

Critically, this approach has minimal development costs. Unlike traditional AI implementations that required expensive model retraining and specialized data science teams, properly configured gen AI agents can deliver substantial value straight out of the box. Moreover, barriers to adoption have all but disappeared. With the release of OpenAI’s AgentKit in October, even non-technical teams can design and deploy autonomous agents without writing a single line of code. World-class supply chain management is becoming a plug-and-play capability, accessible to any business that understands how to guide the gen AI agents with the right data and policies.
The implications extend beyond cost reduction. When autonomous agents handle operational coordination, human managers can redirect their expertise toward strategic challenges: network redesign, supplier relationships, cross-functional integration across the supply chain, finance, and marketing and sales. Thus, the role of supply chain leadership shifts from operator to orchestrator—from designing rigid rules to guiding intelligent agents.

The testbed also reveals a broader opportunity. Because gen AI agents can execute supply chain simulations in minutes rather than weeks, organizations can now rapidly test policies, benchmark strategies, and identify optimal approaches at unprecedented speed. This transforms supply chain strategy from experience-based intuition to data-driven experimentation.

This technological breakthrough arrives amid unprecedented volatility—from black-swan events to geopolitical shocks and fragile global networks that traditional forecasting models can’t handle. In this environment, gen AI’s ability to reason, simulate, and adapt dynamically makes it not just a technological advantage but a strategic imperative.

How to Embark on the Journey

Executives should take three steps to start experimenting with these new systems:

First, audit your AI infrastructure. Identify which models currently power your supply chain systems. Many firms still rely on older, non-reasoning gen AI models that our research shows will fail at autonomous coordination of information across functions. They will have to upgrade to reasoning models.

Second, start with constrained pilots. Deploy autonomous agents in bounded environments with clear guardrails. Test budget constraints, experiment with information sharing, and measure performance against human benchmarks. The methodology we used in our simulations llprovides a template for this experimentation. If your company operates a digital twin to simulate decisions in your supply chain, foow our methodology and embed gen AI agents in the digital twin. Doing so lets you quickly test, learn, and pinpoint how and what delivers real impact for your business.

Third, build orchestration capabilities. The autonomous supply chain requires a new skillset: curating data flows between agents, designing policies that prevent systemic failures, and crafting prompts that align agent behavior with business objectives. These capabilities will differentiate leaders from followers.

. . .

Our experiments suggest that the age of autonomous supply chains is at hand. Success will require more than deploying powerful models. It will demand a new form of leadership that orchestrates intelligence rather than executes tasks, one that designs systems for learning rather than compliance.

Appendix: Exceeding Human Performance

Can autonomous GenAI agents match human performance? We benchmarked the performance of gen AI agents against 100+ human participants who played the Beer Game over the past two years, under the same system conditions as the gen AI testbed. Figure 1 summarizes the effects of the four factors that determine the success or failure of the autonomous supply chains.

The top panel shows that non-reasoning models (GPT-4o mini) performed poorly out of the box, generating costs 2x higher than that of human teams. However, when curated information is shared via a central orchestrator, the model surpassed human performance by 33%.

The bottom panel demonstrates that reasoning models (GPT-5 mini and Llama 4 Maverick 17B) already outperformed human teams in baseline tests. When enhanced with inference-time techniques—such as information sharing, policy constraints (budget guardrails), and refined prompting—these agents delivered 50% to 67% reductions in total supply chain costs compared to human teams operating under identical conditions.

These results indicate that gen AI agents powered by state-of-the-art reasoning models can manage supply chains with proficiency that exceeds human teams, at least within the testbed.

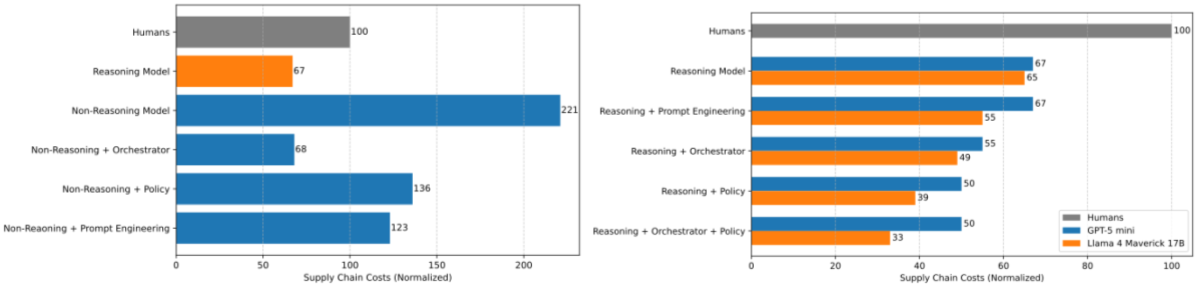

Fig 1. *Performance gains of gen AI agents via model selection and inference-time techniques.* Non-reasoning models (left, GPT-4o mini) required policy constraints, orchestration, and prompt engineering to close the performance gap with humans. In contrast, reasoning models (right, GPT-5 mini and Llama 4 Maverick 17B) started above human-level performance, and, when optimized with the same techniques, achieved up to a 67% reduction in cost relative to human teams.