26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Session M02 – Invited speaker

A foundational framework for generative simulation models: pathway to generative digital twins for supply chain

S Chotaliya[1], J Fowler[1], G Pedrielli[1], D Bayba[2], P Saini[2] and K Yu[1]

[1]Arizona State University, Tempe, AZ, USA; [2]Intel Corporation, Chandler, AZ, USA
Email: john.fowler@asu.edu

Abstract: Designing high-fidelity simulation models from natural language descriptions remains a significant challenge, particularly in complex domains like supply chains. Pretrained large language models (LLMs), when used without domain-specific adaptation, often fail to translate unstructured inputs into executable model representations. In this work, we present a fine-tuned LLM-based pipeline specifically tailored for simulation model generation. We construct a high-quality dataset using a novel data generation process that captures diverse supply chain scenarios. The fine-tuned LLM first converts human-like natural language scenario descriptions into structured representations, then translates these into executable code for a modular Python-based simulation engine built to support a wide range of supply chain configurations. Quantitative and qualitative evaluations show that the fine-tuned model consistently generates high-fidelity discrete event simulation models, significantly outperforming pretrained LLMs in terms of structural accuracy, simulation behavior, and its ability to robustly extract relevant information from linguistically variable natural language descriptions. Moreover, the approach also holds potential applicability to other simulation paradigms like System Dynamics and Agent-Based Simulation.

Keywords: *Discrete event simulation, supply chain, digital twin, generative AI, large language models*

1. INTRODUCTION

Discrete-event simulation (DES) is widely used in supply chain management for analyzing policies, testing resilience, and optimizing resource use (Terzi et al. 2004). Despite its utility, constructing a DES model is labor-intensive, requiring both domain and coding expertise. As discussed in (Chen et al. 2021), Generative AI techniques like Large Language Model (LLM) offer new opportunities to automate this process through natural language understanding and code generation. Yet, LLMs struggle with structured domain logic and deterministic output, limiting their current use in simulation (Jackson et al. 2024). This work introduces an end-to-end framework that transforms natural language scenario descriptions into executable DES models, lowering the barrier to adoption and accelerating the creation of scalable digital twins.

2. METHODS

To address the challenges of capturing complex supply chain dynamics, we developed a structured methodology that integrates representation learning, simulation modeling, and tailored fine-tuning of large language models.

2.1. Architecture Overview

The architecture follows a pipeline design where scenario inputs are systematically transformed into structured representations and simulation-ready models.

- Structured Representation Generation – Scenario descriptions are converted into structured representations that encode raw materials, suppliers, procurement schemes, assemblers, and customer demand.
- Simulation Code Generation – Representations are fed into a modular simulation engine that generates SimPy-based DES models. The engine performs feasibility checks to ensure logical validity (e.g., inventory threshold < maximum inventory, raw material–supplier linkage).

Since all simulation paradigms are built on fundamental building blocks (e.g., events in DES, stocks and flows in System Dynamics, agents and rules in Agent-Based Simulation), the proposed approach has the potential to generalize beyond DES by adapting its structured representation to paradigm-specific primitives.

© MSSANZ and The Authors, CC-BY

2.2. Data Generation

A schema capturing supply chain logic was defined, including inventory dynamics, procurement strategies, and customer demand. Randomized, constrained generation produced valid representations. A subset was manually annotated with natural language and then expanded via LLM-based augmentation to create a large dataset of input–output pairs.

2.3. Model Selection and Fine-Tuning

Several language models, such as Phi-2 (Abdin et al. 2024), Mistral (Albert et al. 2023), GPT-4 (Achiam et al. 2023) and GPT-4.1 mini (OpenAI 2025) were assessed for context length, fine-tuning availability, and cost. GPT-4.1 mini was selected for its 128K token window and API fine-tuning support. It was fine-tuned on 900 training and 100 validation samples for two epochs, imparting domain-specific knowledge.

3. RESULTS

The architecture was evaluated through both quantitative and qualitative experiments.

3.1. Quantitative Analysis

In quantitative tests, two synthetic supply chain scenarios of differing complexity were simulated using the proposed framework and compared against validated benchmark models implemented in SimPy and Arena. Results showed that the fine-tuned GPT-4.1 mini achieved close alignment with benchmark outputs across critical performance measures, including raw material inventories, finished goods levels, and daily cash balances. In contrast, GPT-4, despite its strong general reasoning ability, required multiple rounds of corrective prompting and still diverged from the benchmark in several dimensions. Confidence interval analysis demonstrated that the fine-tuned model achieved stable and repeatable results across replications, highlighting the importance of domain-specific adaptation.

3.2. Qualitative Analysis

Qualitative analysis further underscored the benefits of fine-tuning. Code generated by GPT-4.1 mini exhibited modular and object-oriented design, making it more extensible to larger supply chain configurations, while GPT-4 outputs were often incomplete or entangled with spurious dependencies. The fine-tuned model also successfully extracted implicit details from scenario descriptions and integrated them into coherent code, whereas the baseline GPT-4 frequently omitted important constraints or produced logically inconsistent simulations. The combination of improved fidelity, modularity, scalability, and robustness makes the fine-tuned GPT-4.1 mini a strong candidate for practical deployment in supply chain simulation environments.

4. CONCLUSION

This work demonstrates that generative AI can substantially reduce the barriers associated with simulation modeling, enabling rapid development of DES models directly from natural language. The contributions include a domain-specific data generation pipeline, a modular simulation engine, and a fine-tuned LLM that together form a scalable architecture for supply chain modeling. Future efforts will expand the dataset to include more varied supply chain structures, extend the modeling capabilities as well as generalizability to other simulation paradigms, and explore privacy-preserving fine-tuning approaches for enterprise contexts. Ultimately, this work highlights the transformative potential of combining simulation with generative AI, creating a pathway toward intelligent, automated, and scalable digital twin systems for modern supply chains.

REFERENCES

Abdin M, Aneja J, Behl H, Bubeck S, Eldan R, Gunasekar S, et al. (2024) “Phi-4 technical report.” *arXiv preprint* arXiv:2412.08905.

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. (2023) “GPT-4 technical report.” *arXiv preprint* arXiv:2303.08774.

Albert QJ, Sablayrolles A, Mensch A, Bamford C and Chaplot DS (2023) “Mistral 7B.” *arXiv preprint*.

Chen M, Tworek J, Jun H, Yuan Q, de Oliveira Pinto HP, Kaplan J, et al. (2021) “Evaluating large language models trained on code.” *arXiv preprint* arXiv:2107.03374.

OpenAI (2025) April. “Introducing GPT-4.1 in the API.” Accessed: 2025-04-18.

Terzi S and Cavalieri S (2004) “Simulation in the supply chain context: a survey”. *Computers in Industry* 53(1):3–16.

© MSSANZ and The Authors, CC-BY