Human-AI Complementarity in Inventory Control with LLMs
[Tianyi: AI Agents for Supply Chain: Human-AI Complementarity]

Abstract

Operations research (OR) algorithms for inventory and supply chain decisions have been widely deployed and have had substantial practical impact. However, these algorithms are often rigid, relying on historical data and fixed modeling assumptions that can break down under distribution shifts or when key contextual knowledge is missing. As a result, they are frequently used as decision-support tools, with human experts making the final call and applying contextual judgment to accept or override algorithmic recommendations. Recent advances in large language models (LLMs) offer a less rigid form of reasoning that may be leveraged to improve existing decision-making frameworks.

We study how LLMs can improve inventory decisions, and how they interact with both traditional OR algorithms and human decision makers. We first construct a set of benchmark instances in a multi-period inventory setting, composing of both synthetic instances and ones grounded in real demand data. The instances are designed to stress rigid methods via ad hoc demand change points, seasonal demand shifts, and unpredictable lead times. On these instances, LLM-based approaches improve performance relative to an OR algorithm. [Will: TODO: be more specific in last sentence]

We then run a classroom experiment that embeds LLMs into a human-in-the-loop decision pipeline, focusing on the real-data benchmark instances. We find that humans achieve higher profits when equipped with natural-language LLM recommendations than when equipped with a formulaic OR algorithm recommendation, indicating that LLMs add value even when a human makes the final decision. At the same time, humans outperform fully automated LLM decisions. Together, these results provide evidence of human-AI complementarity in inventory control: LLMs improve human decisions, and human judgment adds value beyond simply following an LLM.

1 Introduction

[Tianyi: Should we call it OR algorithms?] Operations research (OR) algorithms have been successfully deployed in large-scale systems for inventory and supply chain decisions. A standard workflow is predict-then-optimize: estimate uncertain inputs from historical data, then compute an optimal or near-optimal decision under a specified model. While effective in many stable settings, these approaches are often rigid. They rely on fixed modeling assumptions and on a training distribution that may not match the environment at decision time, and they typically do not incorporate contextual knowledge that is difficult to encode in a formal model. In practice, these limitations motivate a human-in-the-loop process: OR tools serve as decision support, and human experts make the final call, using contextual judgment to accept, adjust, or override algorithmic recommendations.

Recent advances in large language models (LLMs) suggest a new opportunity to improve such pipelines. Unlike traditional OR algorithms, LLMs can incorporate natural-language context and draw on broad world knowledge to form reasonable judgments even when the formal model is

incomplete or the environment shifts in unexpected ways. At the same time, LLMs are a nascent technology for operational decision-making, and it remains unclear when they help, how they should be deployed alongside existing OR heuristics, and how they interact with human decision makers.

This paper takes a step toward understanding the role of LLMs in inventory decision-making. We focus on a multi-period inventory control setting where the objective is to maximize profit. Classic OR baselines perform well when demand and lead times are stationary and well modeled, but many realistic settings deviate from these assumptions. The environment can experience non-stationary demand, ranging from temporal shocks, sudden change points, or seasonal trends. There can also be uncertainty from the supply side such as unpredictable lead times. These features create instances that are not fully specified by a fixed parametric model and that reward context-sensitive reasoning and rapid adjustment.

We ask three main research questions.

1. **Can LLMs improve upon a standard OR inventory algorithms on challenging instances?** We consider instances designed to stress rigid methods through temporal shocks, seasonality, and supply uncertainty. We evaluate whether an LLM can use the observable history and contextual descriptions of the environment to recommend better actions than a traditional baseline.

2. **How should LLMs be deployed in an algorithmic decision pipeline?** There are multiple plausible architectures to incorporate LLMs. An LLM may directly output an action, it may modify an OR recommendation, or it may provide the input parameters to an OR algorithm. We test several deployment choices that represent realistic ways practitioners might use LLMs as an additional layer on top of existing OR tools.

3. **How do LLMs interact with human decision makers, and do they create complementarity?** Since humans commonly remain responsible for final decisions in practice, we ask whether LLM decision support improves human performance relative to traditional OR decision support, and whether humans add value beyond simply following an LLM recommendation. Evidence on both margins would support *human-AI complementarity*.

We tackle these questions using an experiment with two parts.

Benchmark design and algorithmic experiment. First, we construct a benchmark suite of inventory instances where a simple OR baseline is expected to struggle. The benchmark includes the following stressors: (i) an ad hoc demand change point, (ii) demand trends, and (iii) unpredictable lead times. We then compare the baseline OR heuristic to LLM-based policies that attempt to incorporate context and adapt more quickly to regime shifts.

[Jackie: Include more details about the environment and maybe more discussion about each stressor. Also enumerate all the LLM deployment variants.]

Human-in-the-loop experiment. Second, we study human-in-the-loop decision making in a classroom experiment. Participants make inventory decisions in the same environment under three different decision-support modes. In one mode, participants receive a traditional OR recommendation; in another, they receive an LLM recommendation (for example, an LLM that interprets the history and possibly the OR output); and in a third, they receive the LLM recommendation but retain full discretion to deviate from it.

[Jackie: State more details about the experiment.]

Main findings. We summarize three main findings.

First, on the benchmark instances, LLM-based decision policies improve performance relative to the standard OR baseline. These gains are largest in settings with temporal shocks and other forms of non-stationarity where rigid history-based heuristics adapt slowly. [Jackie: summarize main quantitative results, which stressors contribute most.]

Second, in the classroom experiment, humans make better decisions when equipped with LLM recommendations than when equipped with a traditional OR recommendation. This result suggests that LLMs can add value even when the final decision is made by a human, and not only in fully automated settings. [Jackie: Add details of main results]

Third, human decision makers outperform fully automated decisions that simply follow the LLM recommendation. This indicates that human judgment remains valuable even when LLM support is available, and that the best performance arises from combining the two. Together, these findings provide evidence of human-AI complementarity in inventory control: LLMs improve human decisions, and human judgment adds value beyond simply following the LLM.

2 Game Description and Benchmark Instances

We study a multi-period inventory control game in which a decision maker manages a single product over a finite horizon of T periods. Each period $t = 1, 2, \ldots, T$ proceeds as follows:

1. **Observe state.** The decision maker observes current on-hand inventory $I_t^{\text{on-hand}}$ and past realized demand.
2. **Place an order.** The decision maker chooses an order quantity $q_t \geq 0$.
3. **Receive arrivals.** Shipments from earlier orders may arrive and are added to on-hand inventory.
4. **Realize demand and sell.** Demand D_t is realized; sales equal $\min\{D_t, I_t^{\text{on-hand}}\}$.
5. **Carry over inventory.** Any leftover inventory carries over to period $t{+}1$.

Lead times. Orders have a *promised* lead time of L periods, but actual lead times may deviate from the promise, and orders may be lost.

Reward. The per-period reward equals sales profit minus holding cost:

$$R_t \;=\; P \cdot (\text{units sold at } t) \;-\; H \cdot (\text{units carried over after } t).$$

The total performance of a policy is the sum of rewards over the horizon, $\sum_{t=1}^{T} R_t$.

Initialization. Each play-through begins with five realized demand values revealed as "historical data". The initial on-hand and in-transit inventories are zero, and the decision maker is told the profit P, the holding cost H, and the promised lead time L.

2.1 Benchmark Instances

An *instance* is a single play-through of the inventory game with a fixed horizon and a fixed realization of all stochastic quantities (e.g., demands and order arrivals), which we hold deterministic within an instance for reproducibility and fair comparisons across decision policies.

Across all instances, we fix the profit-to-holding-cost ratio to be $P/H = 4$. We consider two lead-time settings: (i) a fixed promised lead time $L = 2$, and (ii) a stochastic lead time where each order's realized lead time is drawn uniformly from $\{1, 2, 3, \infty\}$, where ∞ indicates the shipment never arrives.

Synthetic instances. We construct four synthetic demand environments, each with a horizon of $T = 50$ periods:

- **Instance 1 (IID demand).** For all periods t, demand is $\mathcal{N}(100, 25)$.
- **Instance 2 (Sudden demand shift).** For periods $t = 1, \ldots, 15$, demand is $\mathcal{N}(100, 25)$; for periods $t = 16, \ldots, 50$, demand is $\mathcal{N}(200, 25\sqrt{2})$.
- **Instance 3 (Always-increasing demand).** In period t, demand is $\mathcal{N}(100t, 25\sqrt{t})$.
- **Instance 4 (Variance increase).** For periods $t = 1, \ldots, 15$, demand is $\mathcal{N}(100, 25)$; for periods $t = 16, \ldots, 50$, demand is Uniform$[0, 200]$.

These instances are designed to stress simple history-averaging heuristics via non-stationarity (change points and trends) and distributional shifts (variance changes).

Real-data instances. We also construct instances grounded in real demand trajectories from the H&M Personalized Fashion Recommendations dataset hosted on Kaggle. The dataset contains daily sales at the product level. We select six popular products and aggregate their sales to the weekly level over a year, yielding a demand trajectory of 52 periods per product. The first five demand points are provided as historical data, hence the game lasts $T = 47$ periods. [Jackie: Say that LLM and human are provided with item description.]

3 Part 1: Algorithmic Experiment

3.1 Algorithms

[Jackie: Describe the different algorithms (e.g., OR, OR to LLM, etc.)]

3.2 Results

4 Part 2: Human-in-the-loop Experiment

4.1 Experiment Setup

We conducted the experiment as part of regular classroom instruction in three different courses [Jackie: specify courses]. Participants played the multi-period inventory game described in Section 2. Each participant completed three runs of the game on three real-data instances and three different collaboration modes. [Jackie: Describe instances more?]

Collaboration modes (treatments). We studied three collaboration modes that vary how an OR heuristic, an LLM agent, and the human participant share decision authority:

1. **Mode A (OR → Human decision).** A formula-based OR tool recommends an order quantity each period, and the participant makes the final order decision.

2. **Mode B (OR → LLM → Human decision).** The OR recommendation is passed to an LLM agent, which provides a recommendation and natural-language reasoning. The participant observes the OR recommendation, the LLM recommendation, and the LLM reasoning, and then makes the final order decision.

3. **Mode C (OR → LLM with human guidance).** The LLM agent makes the order decisions, and the participant provides high-level guidance every four periods to steer future actions. [Jackie: Describe this in more detail.]

The three real-world instances and the order in which they were played are fixed across participants. Each participant experiences each of the three collaboration modes exactly once, with participants randomized uniformly across the six possible mode orders. The game was not a mandatory activity and neither the participation nor their game performance had an impact on the students' grades.

The primary outcome is final performance in each instance, measured by the cumulative reward. We also record anonymous interaction logs, including text prompts and responses (when applicable), decisions, and timestamps. After the activity, participants completed a brief post-activity survey about their experience with the decision-support tools. No personally identifiable information is collected or stored as part of the research dataset.

4.2 Results

[Jackie: add introductory paragraph.] The hypotheses and the fixed-sequence testing procedure in this section were pre-registered prior to conducting the classroom experiment (see the pre-registration at `<PREREG_LINK>`).

4.2.1 LLM vs. OR decision-support (Mode B vs. Mode A)

We first test whether adding the LLM layer improves outcomes relative to providing only the OR recommendation. Concretely, we compare Mode B (OR + LLM + human decision) to Mode A (OR + human decision) using the pooled OLS specification with subject fixed effects and instance fixed effects:

$$Y_{ij} \;=\; \alpha_i \;+\; \beta_j \;+\; \tau_{c(i,j)} \;+\; \varepsilon_{ij},$$

where Y_{ij} is the total score obtained by subject i on instance j, α_i captures subject-level heterogeneity, β_j captures instance difficulty, and $\tau_{c(i,j)}$ is the treatment effect for the decision-support mode used in that subject-instance pair. Standard errors are clustered at the subject level.

Across 187 subject-instance observations from 69 subjects, Mode B yields a statistically significant improvement over Mode A. The estimated treatment effect is $\widehat{\tau}_B = 1867$ (SE = 673), which is positive and significant ($t = 2.77$; one-sided $p = 0.0036$, two-sided $p = 0.0072$).

4.2.2 LLMs as decision-support vs. LLMs as decision-makers (Mode B vs. Mode C)

We next test whether delegating decisions to the LLM with periodic human guidance (Mode C) performs differently from keeping the human as the final decision maker with LLM support (Mode

B). Following the fixed-sequence procedure, we test

$$H_2 : \tau_C - \tau_B \neq 0$$

using a two-sided test at level $\alpha = 0.025$ (conditional on rejecting H_1).

We find a statistically significant difference between the two modes. The estimated contrast is $\widehat{\tau}_C - \widehat{\tau}_B = -2083$ (SE = 692); this difference is significant ($t = -3.01$, two-sided $p = 0.0037$), and remains significant under the fixed-sequence threshold ($p < 0.025$). Thus, conditional on H_1, we reject H_2's null of no difference and conclude that Mode B achieves higher performance than Mode C in our classroom experiment.

[Jackie: Comment that this result depends on how good the LLM is.]

4.2.3 TODO: Mechanisms

[Jackie: I think we need a better understanding of the potential mechanisms for "human-AI complementarity". For example, it could be that when an LLM provides reasoning to override the OR algorithm, the human uses this reasoning to do a further override that improves outcomes. We should look into the traces to see if there are any interesting patterns of override/adherence behavior.]

4.2.4 Some other analysis

[Jackie: Here is a possible criticism to the "complementarity" insight: what if the following was happening: the population can be divided into two groups who (1) follow the OR / LLM recommendation blindly, and (2) achieve a high performance which does not depend on whether they were in mode A or B. In this case, there is no "human-AI complementarity", it's just that performance improves with the LLM due to the first group who blindly follows the recommendation (since LLM recommendation in general is better than the OR recommendation).]

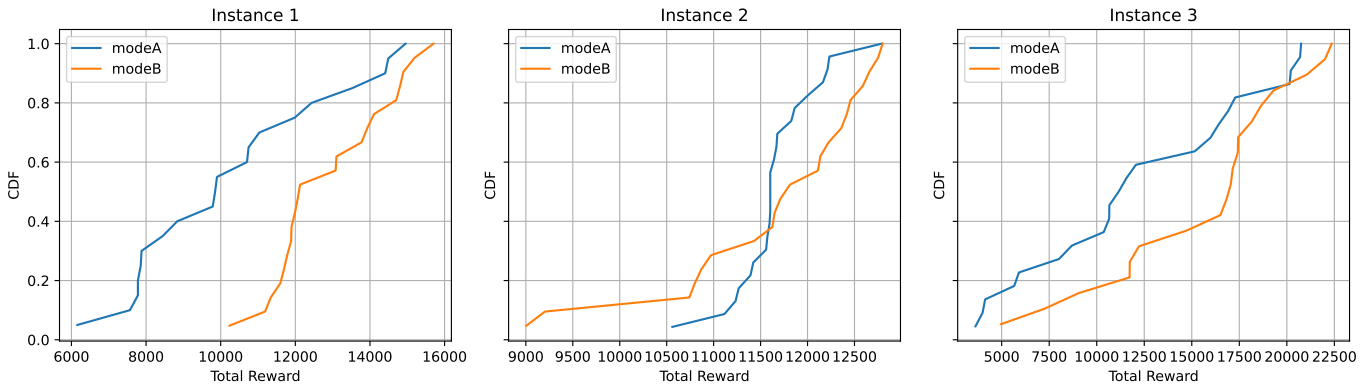

Figure 1: CDF of the rewards by instance and mode.

Let Y_{ij} be the reward for subject j on instance i, and let $\bar{Y}_i$ be the average reward for instance i. Define $R_{ij} = Y_{ij} - \bar{Y}_i$ as the improvement from the average for that instance. For subject j, let $R_j(m)$ be their improvement for the instance where they played mode $m \in \{A, B, C\}$. Then, define $\Delta_j = R_j(B) - R_j(A)$ be the difference in their improvement when they played under mode B compared to mode A. fig. 2 plots the distribution of Δ_j across participants. We see that $\Delta_j > 0$

for 67.8% of participants. This means that for most users, they performed (relatively) better on mode B than in mode A.

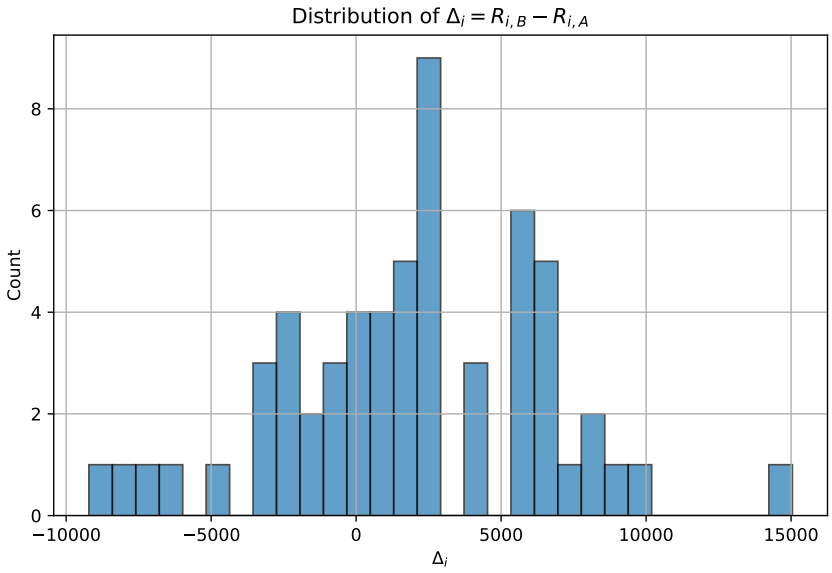

Figure 2: CDF of the rewards by instance and mode.

References