BEYOND THE ORDER QUANTITY: BENCHMARKING BEHAVIORAL BIASES OF LLM INVENTORY AGENTS UNDER UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can solve quantitative problems, but inventory agents must act under uncertain demand, lead times, and partner behavior repeatedly. Existing evaluations, however, rarely distinguish limited task competence from systematic behavioral bias or move beyond a single inventory setting. We introduce AIM-Bench, a benchmark that evaluates LLM inventory agents across five environments, from a single-period newsvendor problem to multi-period and multi-echelon supply chains. AIM-Bench varies the source and structure of uncertainty and evaluates agents at three levels: operational outcomes, distance from an ex-post optimal action, and behavioral signatures such as mean anchoring, demand chasing, and the bullwhip effect. Experiments with eight state-of-the-art LLMs reveal that strong reasoning ability does not ensure reliable replenishment, because even the strongest models retain systematic behavioral biases: some models still anchor orders toward mean demand, and all evaluated models amplify demand variability in the Beer Game. These biases can remain hidden when models are compared only by aggregate cost or stockout rate. We further test two lightweight interventions. Cognitive-reflection prompting reduces anchoring for several models, information sharing substantially weakens the bullwhip effect of some models. None of these remedies is uniformly effective. Our results show that LLM inventory agents should be evaluated not only by what outcomes they achieve, but also by how their decisions respond to uncertainty.

1 INTRODUCTION

Large language models (LLMs) are increasingly embedded in agents that observe an environment, retain history, reason about goals, and execute actions. ReAct interleaves reasoning with interaction, while AgentBench and AgentBoard evaluate LLMs through multi-round tasks (Yao et al., 2023; Liu et al., 2024; Ma et al., 2024). As such agents move into operational use, inventory replenishment stands out as a demanding application: orders must be chosen before demand and delivery conditions are known, ordering too little creates shortages, and ordering too much increases holding cost. Multi-period systems add outstanding orders and random lead times, and multi-echelon systems add interactions among decision makers. This setting also changes what evaluation must measure. A model that answers a quantitative question correctly may still fail when actions alter later states, feedback is incomplete, and errors accumulate over time, so sequential performance cannot be inferred from static reasoning accuracy alone. Classical models formalize the underlying trade-offs (Arrow et al., 1951; Kaplan, 1970), and learning-based methods optimize them from data or simulation (Qi et al., 2022; Gijsbrechts et al., 2022; Liu et al., 2025). A general-purpose LLM agents must infer a replenishment policy from instructions and interaction history without task-specific training, yet they often fail to maintain it once their earlier decisions change the state.

A growing body of work builds what we call *LLM supply-chain agents*: systems in which an LLM repeatedly maps an observed inventory state to an operational action, possibly alongside other agents. Within this class, existing studies differ mainly in scope and mechanism. InvAgent uses LLMs as autonomous decision makers in a multi-agent inventory system (Quan & Liu, 2025); Wang et al. develop a domain-oriented LLM for supply-chain knowledge and strategic games (Wang et al., 2025); Yoshizato et al. introduce memory retrieval for adaptive multi-echelon ordering (Yoshizato

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2026); and RetailBench evaluates agents that jointly manage replenishment, pricing, suppliers, and cash flow (Zhang et al., 2026). These studies establish the potential of LLM-based operational agents, but they mainly report system performance or the benefit of one specific mechanism. The problem with this emphasis is that it cannot tell a designer *why* an agent fails or *which* remedy to apply: a poor outcome may reflect a misunderstood objective, misestimated demand, ignored goods in transit, or overreaction to recent observations, while a reasonable final cost can conceal a fragile policy that alternates between excessive orders and shortages. Two agents can also obtain similar service levels through very different order trajectories. We argue that this diagnostic gap is a missing piece in current evaluations: final reward, cost, or success rate cannot separate these cases or guide a targeted correction, and inventory evaluation should therefore connect outcomes to the decisions that produced them. AgentBoard makes a related observation in general agent tasks, showing that trajectory analysis reveals failures hidden by final success (Ma et al., 2024); we bring this principle into inventory management, where normative references make the diagnosis precise.

Behavioral operations provides a useful language for this diagnosis. Human decision makers anchor newsvendor orders toward mean demand, chase recent demand, and respond to alternative frames (Schweitzer & Cachon, 2000; Bostian et al., 2008; Bolton & Katok, 2008; Schultz et al., 2018). Local overreaction can also propagate upstream as the bullwhip effect (Lee et al., 1997; Croson & Donohue, 2006). These patterns are measurable action signatures, not only explanations for high cost. LLM studies report related framing, risk, and judgment biases (Jia et al., 2024; Chen et al., 2025; Cui et al., 2025), but one-shot evidence does not establish how a stateful inventory agent behaves. Figure 1 illustrates this difference.

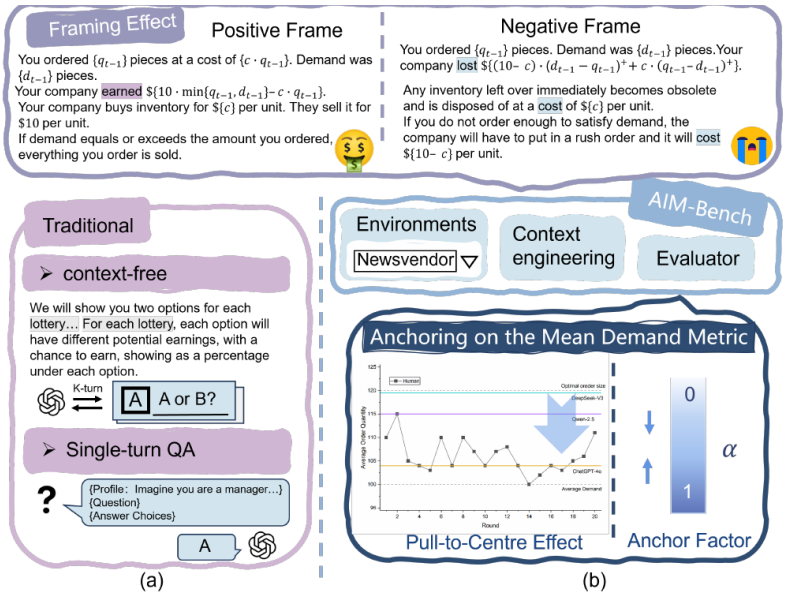

Figure 1: From one-shot behavioral testing to agent evaluation. (a) Traditional newsvendor studies evaluate isolated choices. (b) AIM-Bench evaluates decisions through repeated interactions with an inventory environment.

We address three questions. First, how does an LLM agent respond when uncertainty changes from demand to random lead time and partner behavior? Second, can action-level and behavioral metrics explain differences hidden by aggregate outcomes? Third, can a diagnosed bias guide a targeted intervention? Answering these questions requires controlled variation in uncertainty and metrics for both consequences and trajectories, so that task difficulty can be separated from a recurring decision heuristic. This requirement directly shapes our benchmark: we introduce AIM-Bench, whose five environments span single-period, multi-period, and multi-echelon decisions and together vary uncertainty in demand, vendor lead time, and partner behavior. The environments form a progression: the newsvendor problem isolates stochastic demand; multi-period replenishment adds delayed and stochastic arrivals; the Beer Game, two-level warehouse network, and supply-chain

network add distributed decisions and heterogeneous supply routes. This structure tests whether a behavior remains local, disappears with feedback, or compounds through interaction.

The benchmark framework we proposed evaluates agents at three complementary levels. Operational metrics measure cost, service reliability, and inventory exposure. A process metric compares multi-period trajectories with ex-post reference actions (Qi et al., 2022; Liu et al., 2026). Behavioral diagnostics test mean anchoring, demand chasing, framing sensitivity, and demand amplification. This protocol connects what happened with how the model decided, and the diagnosis then motivates targeted interventions: cognitive reflection for insufficient adjustment and information sharing for local overreaction. Applying this protocol to eight state-of-the-art LLMs, we find a consistent gap between general reasoning capability and operational reliability; we defer the detailed findings to the experiments and conclusions.

In this paper, we make the following contributions:

- We introduce AIM-Bench, a unified evaluation suite for LLM inventory agents under demand, lead-time, and partner-behavior uncertainty. Its five environments connect single-step behavioral tests with sequential and multi-agent replenishment.
- We propose a layered protocol that combines operational outcomes, action-level optimality, and behavioral diagnostics. This protocol distinguishes poor outcomes from systematic decision bias.
- We evaluate eight open- and closed-source LLMs, including reasoning-oriented models released in 2026, and identify recurring failure patterns such as mean anchoring and demand amplification. We further show that cognitive reflection, information sharing, and extended thinking can mitigate targeted biases, but do not provide a universal fix.

2 Related Work

2.1 Inventory Control, Learning, and Simulation

The newsvendor problem captures the single-period trade-off between excess inventory and unmet demand (Arrow et al., 1951); multi-period formulations add carry-over inventory and stochastic lead times (Kaplan, 1970). As state, sourcing, and network complexity increase, data-driven methods learn policies from records or simulation. Qi et al. derive ex-post actions for end-to-end replenishment under nonstationary demand and random lead time (Qi et al., 2022), while deep reinforcement learning addresses lost-sales, dual-source, and multi-echelon systems (Gijsbrechts et al., 2022; Oroojlooyjadid et al., 2022; Liu et al., 2025). MABIM provides a multi-agent, multi-echelon inventory simulator for comparing operations-research and learning policies (Yang et al., 2023). InventoryGPT instead uses Transformer-based imitation learning for large-scale order timing and quantity (Liu et al., 2026). These methods learn specialized policies and mainly evaluate aggregate outcomes, they do not test whether general-purpose LLMs can reliably act across diverse inventory settings, nor do they examine the process and behavioral signatures of their decisions. AIM-Bench addresses this gap by evaluating general-purpose LLMs at the outcome, process, and behavior levels.

2.2 LLMs and Agents in Supply-Chain Management

Early work on LLM inventory decisions focuses on either behavioral choices or a specific supply-chain setting. Su et al. and Chen et al. examine LLM decisions and human-like biases in newsvendor and managerial scenarios (Su et al., 2023; Chen et al., 2025). Moving from single decisions to agents, InvAgent introduces an LLM-based multi-agent system for inventory management and evaluates its zero-shot decisions in a Beer Game-style environment (Quan & Liu, 2025). Wang et al. combine retrieval-augmented generation with supply-chain knowledge and use the resulting model in examinations and strategic supply-chain games (Wang et al., 2025). Yoshizato et al. show that structured strategy prompts can work in restricted cases and introduce memory retrieval to improve adaptation across multi-echelon scenarios (Yoshizato et al., 2026). RetailBench expands the scope to long-horizon supermarket operation, where agents jointly manage replenishment, pricing, supplier selection, assortment, and cash flow (Zhang et al., 2026). This emerging literature studies increasingly capable systems. AIM-Bench is complementary: it holds the agent interface fixed, varies the source of uncertainty, and diagnoses decision bias rather than proposing a new policy architecture.

2.3 Evaluation of Interactive LLM Agents

LLM agent evaluation has moved from static questions toward interaction with tools and environments. ReAct interleaves reasoning and action (Yao et al., 2023). AgentBench covers operating-system, database, game, household, and web tasks and identifies long-term reasoning as a common limitation (Liu et al., 2024). AgentBoard adds fine-grained progress measures for multi-turn, partially observable environments (Ma et al., 2024). These benchmarks motivate trajectory analysis but do not model the operational consequences of repeated quantitative decisions. AIM-Bench brings this principle into inventory management: its process metrics use a normative inventory reference, and its behavioral metrics diagnose structured departures from that reference.

2.4 Behavioral Operations and LLM Decision Bias

Behavioral operations documents predictable deviations from normative inventory policies. In the newsvendor problem, decision makers often pull orders toward mean demand (Schweitzer & Cachon, 2000; Bostian et al., 2008), chase recent demand (Bolton & Katok, 2008), and respond to how equivalent outcomes are framed (Tversky & Kahneman, 1981; Schultz et al., 2018). Prospect theory explains how reference points and asymmetric treatment of gains and losses can shape risky choice (Kahneman & Tversky, 1979). In multi-echelon systems, coordination risk, local feedback, and limited information can amplify orders upstream (Croson et al., 2014; Croson & Donohue, 2003; 2006). Parallel studies show that LLMs can reproduce human-like risk preferences and judgment biases (Dillion et al., 2023; Jia et al., 2024; Cui et al., 2025; Chen et al., 2025). However, these tendencies are often measured with one-shot prompts. We study whether they appear in repeated inventory actions, how they interact with operational uncertainty, and whether a behavior-specific intervention changes the resulting policy.

3 AIM-Bench: Benchmark Design

3.1 Design Principle and Task Formulation

AIM-Bench is designed around one principle: the source of uncertainty and the source of failure should both be identifiable. We therefore vary uncertainty along three axes—demand, lead time, and partner behavior—and increase decision structure from a single-period choice to sequential and multi-agent control. This follows the environment-based evaluation of LLM agents (Liu et al., 2024; Ma et al., 2024), but grounds every state transition and metric in an inventory process.

Across all environments, the LLM receives a textual task description and the currently observable inventory state. It returns an order action in a constrained text format. The environment validates and executes the action, samples the uncertain events, and provides the next observation. Reasoning and acting are therefore separated from environment dynamics in the same broad sense as interactive language-agent frameworks (Yao et al., 2023). We keep this interface consistent across environments. The comparison then changes the inventory problem and available information, not the basic mechanism used to query the model.

We formulate the environments as a Partially Observable Markov Game (POMG) (Littman, 1994). The formal definition and the interaction memory are provided in Appendix A.

3.2 Environment Suite

The five environments form a progression rather than a collection of unrelated tasks. Figure 2 summarizes their structures and uncertainty sources. NVP isolates stochastic demand in a single-period choice; MPR adds carry-over inventory and stochastic lead times; BG introduces four serial echelons and partner-behavior uncertainty; TWN tests source selection under heterogeneous routes; and SCN combines multi-agent interaction with dual sourcing. This progression lets us ask whether a behavior that appears harmless in NVP persists, disappears, or compounds in sequential and multi-agent settings. Detailed mechanics, state transitions, cost parameters, and uncertainty sources are provided in Appendix B.

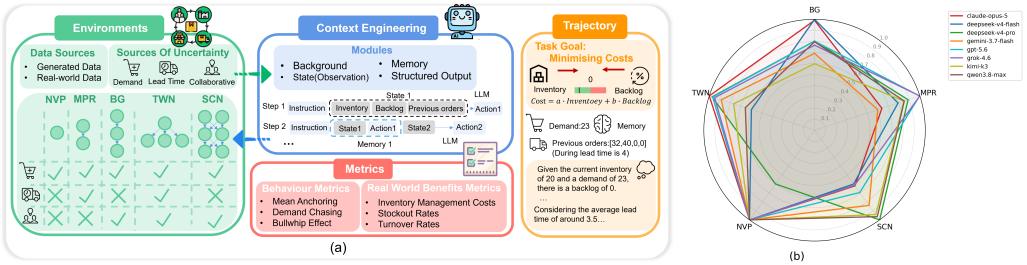

Figure 2: Overview of AIM-Bench. (a) Five environments organize uncertainty by demand, lead time, and partner behavior. (b) Management scores summarize model performance across environments (normalized per environment; larger is better).

4 Layered Evaluation Protocol

No single score can distinguish all relevant failure modes. AIM-Bench therefore evaluates agents at three levels. Outcome metrics measure operational consequences. Process metrics compare actions with an ex-post reference policy. Behavioral diagnostics measure structured deviations such as anchoring and demand amplification. The levels are complementary. Outcome metrics establish practical effect, process metrics localize deviations along a trajectory, and behavioral metrics test whether those deviations follow a recurring pattern. This design parallels the move from final success to analytical trajectory evaluation in AgentBoard (Ma et al., 2024), while using domain-specific references rather than generic subgoal completion.

4.1 Newsvendor Decisions and Behavioral Diagnostics

In NVP, the agent chooses an order quantity q before observing stochastic demand D with cumulative distribution F. Each unit costs c and sells for r. Unmet demand is lost, and leftover inventory has zero salvage value. The underage and overage costs are therefore $c_u = r - c$ and $c_o = c$. The critical ratio and the optimal order follow from these costs. Each agent completes 20 rounds. To test framing sensitivity, we present mathematically equivalent objectives in two forms: a positive frame emphasizes profit, while a negative frame emphasizes loss relative to perfect information. This design follows the broader framing literature (Tversky & Kahneman, 1981; Kahneman & Tversky, 1979) and its application to inventory decisions (Schultz et al., 2018). A reliable inventory policy should not reverse its risk preference only because the wording changes.

NVP also provides two natural anchors: mean demand and the most recent demand realization. Human newsvendor experiments repeatedly observe insufficient adjustment away from these values (Schweitzer & Cachon, 2000; Bostian et al., 2008; Bolton & Katok, 2008). We use the same behavioral structure to measure whether the agent relies on an accessible value instead of adjusting toward q^*.

Anchoring on the Mean Demand. Following prior behavioral work (Bolton & Katok, 2008), we model the observed order as a partial adjustment toward the optimum, parameterized by an anchoring factor α. A value near zero indicates a full adjustment toward the optimum, while a value near one indicates strong anchoring on mean demand.

Demand Chasing Heuristic. Demand chasing occurs when the agent reacts to the latest realization even though demand is independent across periods. We quantify it as $\rho(q_t, d_{t-1})$ (Bolton & Katok, 2008). Under independent and identically distributed demand, a positive correlation indicates that the previous demand is inappropriately driving the current order.

Exact formulas for the critical ratio, the optimal order, the anchoring model, and the demand-chasing measure are provided in Appendix C.1.

4.2 Multi-Period Action Quality

MPR considers one item over periods $t = 1, \ldots, T$, following the periodic-review setting with stochastic lead times (Kaplan, 1970; Qi et al., 2022). The starting inventory I_t may be positive or negative, where a negative value denotes backorders. Orders are placed at fixed review times t_m, for $m = 1, \ldots, M$. Order m has quantity a_m, random lead time L_m, and arrival time $v_m = t_m + L_m$. We assume that orders do not cross in transit. After arrivals and demand D_t, inventory evolves with demand and random arrivals, and the period cost applies holding cost h to positive inventory and backorder cost b to negative inventory. Outcome metrics summarize the cost of the full trajectory, but they do not show whether individual orders were well calibrated. We therefore compute an ex-post reference order using realized demand and lead times, following the decomposition of Qi et al. (Qi et al., 2022). Ex-post actions have also been used as imitation targets for Transformer-based inventory control (Liu et al., 2026); here, we use them only for evaluation. We report the trajectory-level action distance, which is the Euclidean distance between the agent's orders and the ex-post reference. This metric distinguishes policies that obtain similar final outcomes through very different order trajectories.

Exact equations for the inventory evolution, the period cost, the ex-post reference order, and the action distance are provided in Appendix C.2.

4.3 Multi-Echelon Dynamics and Demand Amplification

BG uses a fixed lead time L. Let a_t^i be the quantity shipped by echelon i to its downstream partner during period t. The sellable stock of echelon i combines its current inventory I_t^i and the shipment a_{t-L}^{i+1} that has just arrived from upstream. Its inventory evolves with shipments and demand, and the system cost is the sum of holding and backorder costs across the M echelons.

The bullwhip effect occurs when order variability grows as demand signals move upstream (Forrester, 1958; Lee et al., 1997). It may arise from rational supply-chain dynamics, but controlled studies also identify behavioral causes (Croson & Donohue, 2002; 2003; 2006). Let $D_t^{(1)}$ be the order signal at a downstream echelon and $D_t^{(2)}$ the signal observed by its immediate upstream partner. We measure amplification as the ratio of their standard deviations. Values above one indicate amplification; values below one indicate smoothing. Because information sharing can reduce behavioral amplification (Zhao & Zhao, 2015; Sarkar & Kumar, 2015), we compare local observations with a setting that exposes partner information to each agent.

Exact formulas for the multi-echelon dynamics, the system cost, and the bullwhip ratio are provided in Appendix C.3.

4.4 Operational Outcome Metrics

Across sequential environments, we report average inventory cost, stockout rate, and an inventory ratio, consistent with the need to evaluate both cost and service in data-driven inventory control (Qi et al., 2022; Gijsbrechts et al., 2022). These metrics capture different trade-offs: cost combines penalties, stockout rate measures service reliability, and the inventory ratio reveals whether a low stockout rate is achieved through excessive stock.

Average Inventory Management Cost. For a horizon of T periods, average cost is the mean period cost over the horizon.

Inventory Ratio. We normalize inventory by demand, yielding an inventory-to-demand ratio (not the conventional inventory turnover rate, whose numerator and denominator are reversed). Lower IR indicates a leaner inventory position when service levels are comparable.

Stockout Rate. Let T_s be the number of periods with a stockout. The stockout rate is the fraction of such periods.

Exact formulas for all three metrics are provided in Appendix C.4.

5 EXPERIMENTS

5.1 MODELS AND SETUP

We evaluate eight open- and closed-source models: GPT-5.6 (OpenAI, 2026), Claude-Opus-5 (Anthropic, 2026), Gemini-3.7-Flash (Google DeepMind, 2026), DeepSeek-V4-Pro and DeepSeek-V4-Flash (DeepSeek-AI, 2026), Qwen3.8-Max (Alibaba Cloud, 2026), Kimi-K3 (Moonshot AI, 2026), and Grok-4.6 (xAI, 2026). All models use the same environment interface and receive the same task information within each condition. NVP contains 20 decisions per model and frame; the MPR analysis contains five completed seeds per model except Gemini-3.7-Flash, for which four are available; and the BG comparison contains one completed trajectory per model and information condition. Detailed operational outcomes across all four environments and all evaluated models are provided in Appendix D.

5.2 RESULTS AND FINDINGS

Finding 1: Most models achieve the normative optimum and show no measurable anchoring or demand chasing, and cognitive reflection provides little benefit. Figure 3 shows that seven of the eight models produce anchoring factors and demand-chasing correlations of exactly zero under both the positive frame and the negative frame, without and with cognitive reflection, with orders fixed at the theoretical optimum $q^* = 120$. As a result, the cognitive-reflection prompt produces no measurable change for them. DeepSeek-V4-Pro is the sole exception, exhibiting a small anchoring factor. Cognitive reflection reduces the anchoring factor from 0.108 to 0.035 under negative frame and the demand chasing correlation from 0.327 to 0.192. It indicates that the NVP task is now saturated for most frontier LLMs, cognitive reflection cannot improve performance on these tasks.

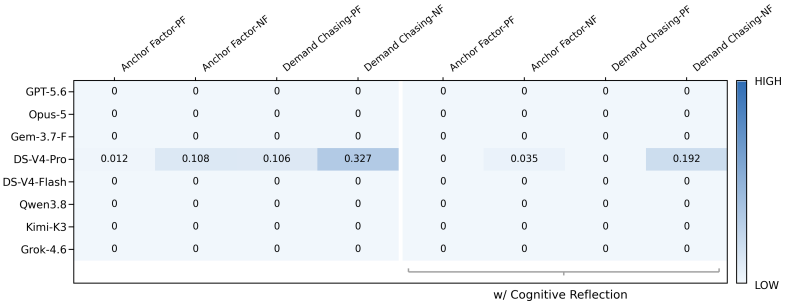

Figure 3: Anchoring and demand-chasing diagnostics for the high-margin NVP condition with $c = 3$, under the positive frame (PF) and negative frame (NF), without and with cognitive reflection.

Finding 2: Behavioral effects do not transfer reliably across contexts. Prior work reports framing-induced risk reversal for LLMs in the disease-outbreak problem (Zhang et al., 2023), while behavioral research shows that framing effects depend on the decision context and reference point (Tversky & Kahneman, 1981; Schultz et al., 2018). We do not observe the same reversal in NVP. The two frames change wording but do not reliably move orders across the risk-neutral solution for any of the eight models. This result cautions against treating a behavioral tendency measured in one prompt domain as a stable property of an LLM agent, consistent with broader evidence that LLM decision behavior varies across uncertain contexts (Jia et al., 2024).

Finding 3: Frontier LLMs converge to the same rational solution and, unlike humans, are not swayed by recent demand. The NVP results reveal a striking convergence: despite different vendors, architectures, and training corpora, seven of eight models independently produce the same optimal order ($q^* = 120$) in every round, under both frames and both reflection conditions. LLMs encodes a common normative solution to the newsvendor problem, and that the models' decision behavior is not driven by idiosyncratic training differences. In contrast to human newsvendors, who reliably exhibit insufficient adjustment and demand chasing (Schweitzer & Cachon, 2000; Bostian et al., 2008), the decisions made by LLMs are not influenced by the most recent demand. The sole exception, DeepSeek-V4-Pro, still shows only minor residual bias, which means that LLMs show less bias in NVP problems than humans.

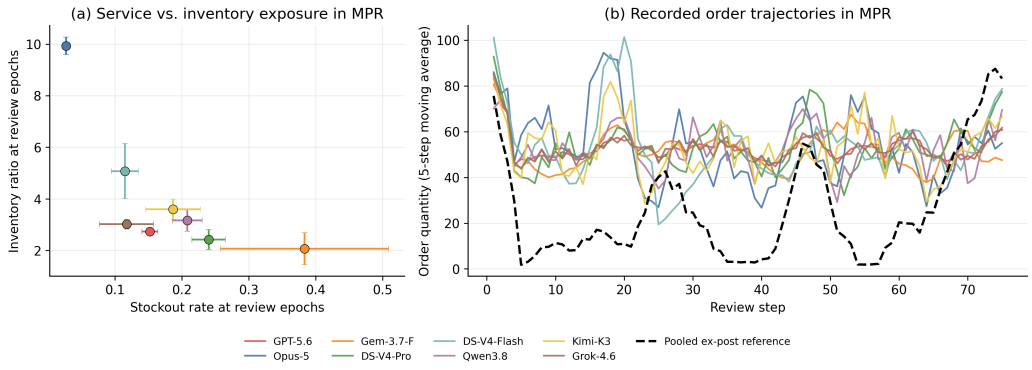

Figure 4: MPR behavior. (a) Service and inventory exposure show the trade-off between stockout rate and inventory ratio across different models. (b) Order trajectories: most of the models place persistently large orders, consistent with overestimating demand uncertainty.

Finding 4: All models amplify demand, but information sharing has model-dependent effects. On the trajectory observed in our experiments, all evaluated models exhibit the bullwhip effect in BG (Figure 5). Without information sharing, L_1/L_0 ranges from 1.44 to 2.69 and L_2/L_0 ranges from 2.01 to 2.79. DeepSeek-V4-Pro produces the largest ratios in this condition. Information sharing reduces L_1/L_0 for three of the eight models and reduces L_2/L_0 for five. Claude-Opus-5 shows the largest reduction in L_2/L_0, from 2.01 to 1.35. In contrast, GPT-5.6, Qwen3.8-Max, and Grok-4.6 exhibit increases at both measured echelons. For Grok-4.6, L_1/L_0 increases from 1.97 to 2.13 and L_2/L_0 increases from 2.25 to 2.31. Information sharing therefore changes demand amplification, but it is not uniformly beneficial across models.

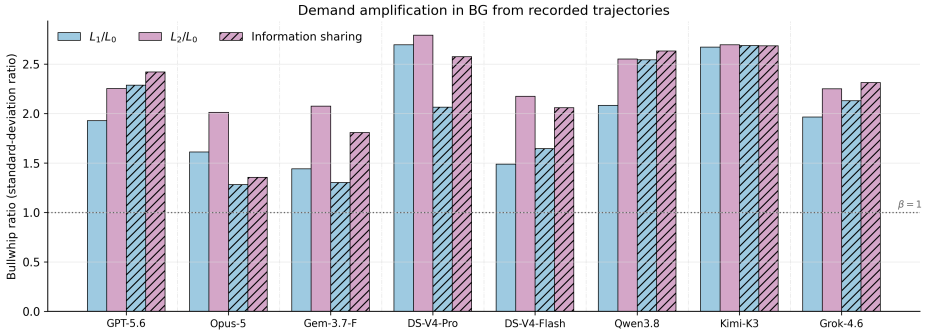

Figure 5: Demand amplification in BG. L_1/L_0 and L_2/L_0 denote the order-variability ratios of the first and second upstream echelons relative to customer demand. Information sharing (hatched) reduces the bullwhip effect for some of the models.

Finding 5: Process metrics separate policies that average cost alone can merge. MPR average daily cost and action distance capture different aspects of policy quality (Figure 6). DeepSeek-V4-Pro and Kimi-K3 obtain similar average daily costs, 134.65 and 136.83, respectively, but their mean action distances differ substantially, 450.97 versus 616.15. GPT-5.6 and Grok-4.6 form a second cluster: their average daily costs are 89.99 and 89.64, and their action distances are 386.96 and 389.55. Their review-epoch service and inventory profiles still differ, with stockout rates of 0.152 and 0.117 and inventory ratios of 2.73 and 3.02, respectively. These results show that average cost, service exposure, and distance from the ex-post reference provide complementary policy information.

Finding 6: Static optimality does not provide a useful ranking of dynamic inventory competence. Seven models tie at a 100% exact-optimal rate in NVP, yet occupy very different positions in MPR. GPT-5.6 and Grok-4.6 have the lowest mean daily costs and action distances. Claude-

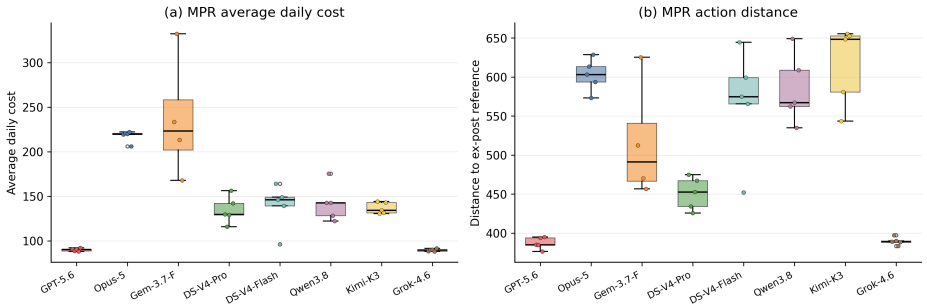

Figure 6: MPR outcome and process metrics across five seeds. Distance from the ex-post reference order provides additional separation when aggregate costs or stockout rates are similar.

Opus-5, despite a low review-epoch stockout rate of 0.027, has the highest inventory ratio (9.93) and a mean action distance of 602.38. Gemini-3.7-Flash has the lowest inventory ratio (2.07) but the highest stockout rate (0.383) and cost (236.79). Thus, neither one-shot correctness nor any single operational metric supports a general ranking. Recent inventory-agent benchmarks similarly find that reliability changes with scenario and horizon (Yoshizato et al., 2026; Zhang et al., 2026); our results identify the transition from a static formula to stateful control as one concrete source of that instability.

Generational comparison. Because AIM-Bench holds the interface and environments fixed, we can compare the 2026 lineup with the previous generation that we evaluated under identical conditions: Qwen-2.5-72B, DeepSeek-V3, GPT-4o, GPT-4.1, and Gemini-2.5-Flash-Lite (Yang et al., 2025; DeepSeek-AI et al., 2025; OpenAI, 2024; Comanici et al., 2025). The best 2026 model improves the best previous MPR cost from 332 (GPT-4.1) to 224 (Grok-4.6) and the median anchoring factor drops from about 0.7 to almost 0. The bias pattern, however, is unchanged: all models amplify demand in BG, and rankings still reverse across uncertainty structures. One year of rapid progress has therefore attenuated, but not eliminated, the behavioral failure modes that AIM-Bench measures.

6 Conclusions

AIM-Bench studies a gap that outcome-only evaluation misses: an LLM can produce plausible inventory results while following a systematically biased policy. Across five environments and eight state-of-the-art models, we find demand amplification, and strong changes in model ranking as uncertainty shifts from demand to lead time and partner behavior. Action-level distance explains differences that cost and stockout rate sometimes hide. The mitigation experiments reinforce the same point. Cognitive reflection is useful when anchoring is the diagnosed failure, information sharing is useful when local decisions amplify upstream demand for some of the models. None of these remedies is universal. Reliable deployment therefore requires matching the evaluation and the intervention to the uncertainty and failure mechanism of the target inventory setting.

7 Limitations

This study uses simulated environments and a limited set of model versions, so the results are not estimates of performance in a specific firm. In most environments, each model–condition pair was evaluated using only a single trajectory. As a result, we cannot estimate run-to-run variance or construct confidence intervals, and the reported results should be interpreted as descriptive observations from the sampled trajectories rather than stable estimates of average performance or effect size. Future work should therefore replicate each condition with multiple independent trajectories before drawing general conclusions. The mitigation study covers prompt-based cognitive reflection and observation-level information sharing. However, fine-tuning, reinforcement learning, and tool-supported optimization may behave differently. Finally, behavioral metrics identify consistent action patterns but not their internal cause, and closed training procedures prevent causal attribution.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and code debugging. All AI-assisted outputs were thoroughly reviewed, verified, and tested by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. The experimental setup, including the versions of models is described in detail in section 5.1. We have also provided a full description of the evaluation metrics in section 4 to assist others in reproducing our experiments.

REFERENCES

Alibaba Cloud. Alibaba unveils Qwen3.8-Max. https://www.alibabacloud.com/en/press-room/alibaba-unveils-qwen3-8-max, 2026.

Anthropic. Introducing Claude Opus 5. https://www.anthropic.com/news/claude-opus-5, 2026.

Kenneth Arrow, Theodore Harris, and Jacob Marschak. Optimal inventory policy. *Econometrica*, 19(3):250–272, 1951.

Gary E. Bolton and Elena Katok. Learning-by-doing in the newsvendor problem: A laboratory investigation. *Manufacturing & Service Operations Management*, 10(3):519–538, 2008.

AJ A. Bostian, Charles A. Holt, and Angela M. Smith. Newsvendor "pull-to-center" effect: Adaptive learning in a laboratory experiment. *Manufacturing & Service Operations Management*, 10(4): 590–608, 2008.

Yang Chen, Samuel N. Kirshner, Anton Ovchinnikov, Meena Andiappan, and Tracy Jenkin. A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do? *Manufacturing & Service Operations Management*, 27(2):354–368, 2025.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Rachel Croson and Karen Donohue. Experimental economics and supply-chain management. *Interfaces*, 32(5):74–82, 2002.

Rachel Croson and Karen Donohue. Impact of POS data sharing on supply chain management: An experimental study. *Production and Operations Management*, 12(1):1–11, 2003.

Rachel Croson and Karen Donohue. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science*, 52(3):323–336, 2006.

Rachel Croson, Karen Donohue, Elena Katok, and John Sterman. Order stability in supply chains: Coordination risk and the role of coordination stock. *Production and Operations Management*, 23(2):176–196, 2014.

Ziyan Cui, Ning Li, and Huaikang Zhou. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5: 627–634, 2025.

DeepSeek-AI. Deepseek-v4 technical report. https://api-docs.deepseek.com/news/news260424/, 2026.

DeepSeek-AI, Aixin Liu, et al. Deepseek-v3 technical report, 2025.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

Jay Forrester. Industrial dynamics: A major breakthrough for decision makers. *Harvard Business Review*, 36(4):37–66, 1958.

Joren Gijsbrechts, Robert N. Boute, Jan A. Van Mieghem, and Dennis J. Zhang. Can deep reinforcement learning improve inventory management? Performance on lost sales, dual-sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management*, 24(3):1349–1368, 2022.

Google DeepMind. Introducing Gemini 3.7 Flash. `https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/`, 2026.

Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. Decision-making behavior evaluation framework for LLMs under uncertain context. *Advances in Neural Information Processing Systems*, 37:113360–113382, 2024.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Robert S. Kaplan. A dynamic inventory model with stochastic lead times. *Management Science*, 16 (7):491–507, 1970.

Hau L. Lee, V. Padmanabhan, and Seungjin Whang. The bullwhip effect in supply chains. *Sloan Management Review*, 38:93–102, 1997.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 157–163. Elsevier, 1994.

Meng Liu, Yang Bai, Meng Qi, and Zuo-Jun Max Shen. Inventory management with transformer: Automated decision making for order timing and quantity. *Service Science*, 2026. Articles in Advance.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024.

Xiaotian Liu, Ming Hu, Yijie Peng, and Yaozu Yang. Multi-agent deep reinforcement learning for multi-echelon inventory management. *Production and Operations Management*, 0(0), 2025.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Moonshot AI. Kimi K3 model card. `https://www.kimi.com/en/blog/kimi-k3`, 2026.

OpenAI. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

OpenAI. Introducing GPT-5.6. `https://openai.com/index/gpt-5-6/`, 2026.

Afshin Oroojlooyjadid, MohammadReza Nazari, Lawrence V. Snyder, and Martin Takáč. A deep q-network for the beer game: Deep reinforcement learning for inventory optimization. *Manufacturing & Service Operations Management*, 24(1):285–304, 2022.

Meng Qi, Yuanyuan Shi, Yongzhi Qi, Chenxin Ma, Rong Yuan, Di Wu, and Zuo-Jun Max Shen. A practical end-to-end inventory management model with deep learning. *Management Science*, 69 (2):759–773, 2022.

Yinzhu Quan and Zefang Liu. Invagent: A large language model based multi-agent system for inventory management in supply chains, 2025.

Sourish Sarkar and Sanjay Kumar. A behavioral experiment on inventory management with supply chain disruption. *International Journal of Production Economics*, 169:169–178, 2015.

Kenneth L. Schultz, Lawrence W. Robinson, Lawrence J. Thomas, John Schultz, and John O. McClain. The use of framing in inventory decisions. *Production and Operations Management*, 27 (1):49–57, 2018.

Maurice E. Schweitzer and Gérard P. Cachon. Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science*, 46(3):404–420, 2000.

Jiaqi Su, Yilin Lang, and Kay-Yut Chen. Can AI solve newsvendor problem without making biased decisions? A behavioral experimental study. *SSRN Electronic Journal*, 2023.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

Haoyu Wang, Jianzhe Jiang, L. Jeff Hong, and Guangxin Jiang. LLMs for supply chain management, 2025.

xAI. Grok 4.6. `https://x.ai/news/grok-4-6`, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report, 2025.

Xianliang Yang, Zhihao Liu, Wei Jiang, Chuheng Zhang, Li Zhao, Lei Song, and Jiang Bian. A versatile multi-agent reinforcement learning benchmark for inventory management, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Kenta Yoshizato, Kazuki Shimizu, Ryota Higa, and Takanobu Otsuka. AI agent systems for supply chains: Structured decision prompts and memory retrieval, 2026. Full version of an extended abstract accepted at AAMAS 2026.

Lei Zhang, Jiayi Wang, Jun Wu, and Zhe Zhang. Retailbench: Benchmarking long-horizon reasoning and coherent decision making of LLM agents in realistic retail environments, 2026.

Tianjun Zhang, Dandan Ma, and Michael A. Saunders. Is ChatGPT humanly irrational? Research Square Preprint, 2023.

Yingshuai Zhao and Xiaobo Zhao. On human decision behavior in multi-echelon inventory management. *International Journal of Production Economics*, 161:116–128, 2015.

A Formal Problem Formulation

We formulate the environments as a Partially Observable Markov Game (POMG) (Littman, 1994):

$$\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{\mathcal{O}_i\}_{i\in\mathcal{N}}, \mathcal{P}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, \gamma), \tag{1}$$

where $\mathcal{N}$ is the agent set, $\mathcal{S}$ is the state space, and $\mathcal{A}_i$ and $\mathcal{O}_i$ are the action and observation spaces of agent i. The transition function $\mathcal{P}$ captures demand, order arrival, and inter-agent dynamics; $\mathcal{R}_i$ captures the corresponding inventory objective. At period t, the policy also receives a bounded interaction memory

$$m_{i,t} = \{(o_{i,\tau}, a_{i,\tau})\}_{\tau=t-k}^{t-1}. \tag{2}$$

This memory turns the model from a one-shot solver into a sequential decision maker. It contains only information made available by the environment and does not reveal latent simulator states. Partial observability matters because outstanding deliveries, local partner states, and future demand are not always known. It is also central to analytical agent evaluation, where success depends on maintaining a coherent belief across multiple interactions (Ma et al., 2024).

B Environment Suite Details

Detailed information about the five environments are listed below.

Newsvendor Problem (NVP). The agent chooses one order quantity before observing stochastic demand (Arrow et al., 1951). NVP isolates demand uncertainty and supports controlled tests of anchoring, demand chasing, and framing. Because a closed-form critical-fractile solution is available, the environment also provides a clear normative reference.

Multi-Period Replenishment (MPR). Inventory carries across periods and orders arrive after stochastic vendor lead times (Kaplan, 1970; Qi et al., 2022). MPR tests whether the agent can coordinate current inventory, outstanding orders, future demand, and uncertain arrival time. It separates uncertainty in *when* inventory arrives from uncertainty in *how much* demand occurs.

Beer Game (BG). Four agents represent the retailer, wholesaler, distributor, and manufacturer in a serial supply chain (Croson et al., 2014). Each agent observes only local information and must replenish from its upstream partner. BG exposes uncertainty about partner behavior and allows us to measure demand amplification. It has also served as a testbed for reinforcement-learning and LLM-based inventory agents (Oroojlooyjadid et al., 2022; Quan & Liu, 2025).

Two-Level Warehouse Network (TWN). A central warehouse supplies several downstream mini-warehouses. A mini-warehouse can also order from the manufacturer through a slower direct channel. TWN tests source selection and allocation under heterogeneous routes and lead times.

Supply Chain Network (SCN). Each downstream node can use a normal source or a faster, more expensive source. SCN combines multi-agent interaction with dual sourcing and tests whether an agent trades delivery speed against procurement cost.

Table 1: Detailed parameters for five environments.

Environment	Horizon	Cost Parameters	Demand Parameters
NVP	20	$c = 3, r = 10$	Uniform on $[50, 150]$
BG	50	$H = 1$, $B = 1$, $L = 4$, $S_I = 10$, $S_O = 10$, $F_C = 5$	Pre-generated stochastic demand
MPR	75	$H = 1, b = 9$	Empirical demand series
TWN	75	$W_H = 0.5$, $w_H = 2$, $W_{\text{lost}} = 2$, $w_{\text{lost}} = 2$, $W_{\max} = 60000$, $w_{\max} = 30000$	Empirical demand series
SCN	200	$H = 1.0$, $B = 1.0$, $C = (0.1, 0.2)$, $S_I = 2$, $S_O = 2$, $L = 4$	Uniform on $\{0, 1, 2, 3, 4\}$

Notes. Horizon denotes the number of periods per episode. **NVP**: c is unit cost, r is selling price; demand is uniform on $[50, 150]$. **BG**: H and B are holding and backlog cost coefficients; L is the fixed lead time; S_I and S_O are initial inventory and initial order quantity; F_C is the fixed ordering cost. Demand is generated stochastically before the episode. **MPR**: H and B are holding and backlog cost coefficients. **TWN**: W_H and w_H are holding cost coefficients for the central and forward warehouses; W_{lost} and w_{lost} are lost-sales penalty coefficients; $W_{\max}$ and $w_{\max}$ are capacity limits. **SCN**: H and B are per-node holding and backlog cost coefficients; $C = (0.1, 0.2)$ are the regular and rush ordering cost coefficients; S_I and S_O are initial inventory and initial order quantity; L is the lead time. Demand conforms to discrete uniform on $\{0, 1, 2, 3, 4\}$.

C Evaluation Metrics and Behavioral Diagnostics

C.1 Newsvendor Decisions and Behavioral Diagnostics

The underage and overage costs are $c_u = r - c$ and $c_o = c$. The critical ratio and the optimal order are

$$\eta = \frac{c_u}{c_u + c_o}, \qquad q^* = F^{-1}(\eta). \tag{3}$$

Let μ denote mean demand. Following prior behavioral work (Bolton & Katok, 2008), we write the observed order as

$$q = \mu + (1 - \alpha)(q^* - \mu), \tag{4}$$

where α is the anchoring factor. A value near zero indicates a full adjustment toward the optimum, while a value near one indicates strong anchoring on mean demand.

Demand chasing occurs when the agent reacts to the latest realization even though demand is independent across periods. We quantify it as $\rho(q_t, d_{t-1})$ (Bolton & Katok, 2008). Under independent and identically distributed demand, a positive correlation indicates that the previous demand is inappropriately driving the current order.

C.2 Multi-Period Action quality

After arrivals and demand D_t, inventory evolves as

$$I_{t+1} = I_t - D_t + \sum_{m=1}^{M} a_m \mathbf{1}\{t = v_m\}. \tag{5}$$

The period cost applies holding cost h to positive inventory and backorder cost b to negative inventory:

$$\begin{aligned} C_t = h \left[I_t - D_t + \sum_{m=1}^{M} a_m \mathbf{1}\{t = t_m + L_m\} \right]^+ \\ + b \left[-I_t + D_t - \sum_{m=1}^{M} a_m \mathbf{1}\{t = t_m + L_m\} \right]^+, \end{aligned} \tag{6}$$

where $[x]^+ = \max\{x, 0\}$.

We compute an ex-post reference order using realized demand and lead times, following the decomposition of Qi et al. (Qi et al., 2022):

$$a_m^* = \max\left(d_{[v_m, s_m^*]} - I_{v_m}, 0\right), \tag{7}$$

where $d_{[u,v]} = \sum_{t=u}^{v} d_t$ is realized cumulative demand and $s_m^* = \lfloor h(v_{m+1} - v_m)/(h + b) \rfloor + v_m$ is the corresponding replenishment endpoint. We report the trajectory-level action distance

$$d(a, a^*) = \sqrt{\sum_{m=1}^{M} (a_m - a_m^*)^2}. \tag{8}$$

C.3 Multi-Echelon Dynamics and System Demand Amplification

BG uses a fixed lead time L. The inventory of echelon i evolves as

$$I_{t+1}^i = I_t^i + a_{t-L}^{i+1} - a_t^i. \tag{9}$$

The system cost is the sum of holding and backorder costs across the M echelons:

$$\begin{aligned} C_t^{\text{total}} = \sum_{i=1}^{M} \Big[h_i \left[I_t^i + a_{t-L}^{i+1} - a_t^i \right]^+ \\ + b_i \left[-I_t^i - a_{t-L}^{i+1} + a_t^i \right]^+ \Big]. \end{aligned} \tag{10}$$

Let $D_t^{(1)}$ be the order signal at a downstream echelon and $D_t^{(2)}$ the signal observed by its immediate upstream partner. We measure amplification as

$$\beta = \frac{\sigma_2}{\sigma_1}, \tag{11}$$

where $\sigma_j = \text{std}(D_t^{(j)})$. Values above one indicate amplification; values below one indicate smoothing.

C.4 Operational Outcome Metrics

Average inventory management cost:

$$\mathrm{AC} = \frac{1}{T}\sum_{t=1}^{T} C_t. \tag{12}$$

Inventory ratio:

$$\mathrm{IR} = \frac{\sum_{t=1}^{T} I_t}{\sum_{t=1}^{T} D_t}. \tag{13}$$

This quantity is an inventory-to-demand ratio, not the conventional inventory turnover rate, whose numerator and denominator are reversed. Lower IR indicates a leaner inventory position when service levels are comparable. Stockout rate:

$$\mathrm{SR} = \frac{T_s}{T}. \tag{14}$$

D Operational Results across AIM-Bench Environments

	BG			MPR			TWN			SCN		
Model	Avg. Cost	IR	SR	Avg. Cost	IR	SR	Avg. Cost	IR	SR	Avg. Cost	IR	SR
gpt-5.6	112.22	0.44	0.68	90.04	2.75	0.15	442.85	0.59	1.00	22.94	**0.24**	0.81
claude-opus-5	58.22	**0.24**	0.74	218.28	9.98	**0.03**	**341.23**	0.69	1.00	25.49	0.41	0.77
gemini-3.7-flash	143.82	0.30	0.70	232.35	**2.36**	0.29	699.60	0.0142	1.00	18.00	0.35	0.80
deepseek-v4-pro	111.98	0.85	0.59	129.22	3.05	0.21	425.28	**0.0035**	1.00	**12.42**	0.63	**0.56**
deepseek-v4-flash	**55.82**	0.92	**0.37**	163.32	6.34	0.11	1607.30	13.37	1.00	26.22	0.69	0.65
qwen3.8-max	113.82	0.68	0.62	142.15	3.04	0.23	1425.58	11.20	1.00	13.66	0.29	0.79
kimi-k3	170.56	0.31	0.82	143.61	2.98	0.24	1072.73	8.67	0.95	14.46	0.33	0.74
grok-4.6	122.68	0.36	0.71	**89.29**	3.31	0.07	501.85	6.57	**0.40**	25.71	0.63	0.61