International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

AI Agents as Intelligent Planning Co-Pilots: Transforming Demand Forecasting in Large-Scale Retail Enterprises

Mazdul Hasan Choudhury*[1]

Abstract: These instructions give you guidelines for preparing papers for IJISAE. Use this document as a template if you are using Microsoft Word 2007 or later. Otherwise, use this document as an instruction set. Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials. The abstract must be a concise yet comprehensive reflection of what is in your article. In particular, the abstract must be self-contained, without abbreviations, footnotes, or references. It should be a microcosm of the full article. The abstract must be between 150–250 words. Be sure that you adhere to these limits; otherwise, you will need to edit your abstract accordingly. The abstract must be written as one paragraph, and should not contain displayed mathematical equations or tabular material. The abstract should include three or four different keywords or phrases, as this will help readers to find it. It is important to avoid over-repetition of such phrases as this can result in a page being rejected by search engines. Ensure that your abstract reads well and is grammatically correct.

***Keywords*:** *engines, material, paragraph, grammatically*

1. I. Introduction

Walk the planning floor of any major national retailer and you will find something the last twenty years of technology investment has not fundamentally changed: demand planners spending most of their day reacting. Reacting to the exception flags the system produced overnight. Reacting to a promotional spike the model did not anticipate. Reacting to a supplier disruption that the forecast had no mechanism to detect. The dashboards have improved, the models have grown more sophisticated, and the data pipelines have expanded — but the underlying dynamic has remained stubbornly in place. Demand forecasting in enterprise retail is still, in most organizations, a periodic, manually intensive, exception-driven process. And that is a structural problem, because every inventory decision in a retail operation — what to procure, how much, where to position it, when to commit — flows from a forecast. At the scale of a national omnichannel retailer, a 15-percentage-point improvement in forecast accuracy is not a modeling achievement. It is a billion-dollar operational outcome that reverberates through the entire supply chain.

The gap between what retail forecasting technology could deliver and what it actually delivers inside planning organizations comes down to a layer that the research literature has largely passed over: the agentic workflow that surrounds the model. Forecasting is not a computation — it is a continuous process. Demand signals must be ingested as they emerge. Models must update as conditions shift. Outputs must be converted into business-language recommendations that planners can act on within their existing decision cadence. Planner feedback on those recommendations must be captured and fed back into the system to improve future outputs. None of this happens automatically in traditional architectures. It happens — inconsistently, incompletely, too slowly — through a combination of batch processes, manual planner judgment,

and organizational routines that cannot scale as SKU counts and channel complexity grow [1].

AI agent architectures offer a direct solution to this workflow problem — and that solution is distinct from deploying a better forecasting model. LLM-powered planning agents translate model outputs into plain-language recommendations, answer planner queries without requiring a data science intermediary, and orchestrate downstream analytical tools through API calls. DRL agents learn continuously from operational outcomes, refining inventory policies in real time rather than waiting for the next model retraining cycle. Transformer-based temporal fusion networks capture seasonal patterns, promotional responses, and cross-SKU demand interactions that simpler recurrent architectures fail to represent. When these capabilities are combined in a coordinated multi-agent system and deployed around a retailer's existing forecasting

infrastructure, the result is something meaningfully different from a more accurate model: it is a continuously operating planning co-pilot [2], [3].

Four contributions structure this paper: (1) a practitioner-grounded taxonomy of AI agent architectures for retail demand planning, with analytical differentiation of LLM-powered multi-agent systems, DRL agents, and transformer networks by capability and appropriate application domain; (2) a synthesis of peer-reviewed empirical evidence benchmarking agent-based systems on forecast accuracy, inventory efficiency, and planner productivity; (3) a five-stage deployment framework covering data infrastructure, system integration, and change management requirements; and (4) a human-in-the-loop collaboration model specifying the explainability and trust-calibration design requirements that separate durable adoption from eventual abandonment.

2. II. The Demand Forecasting Challenge in Large-Scale Retail

The numbers involved in large-scale retail demand forecasting are worth stating plainly because scale is the context that makes the AI-agent argument meaningful. A major national retailer may manage between 50,000 and 150,000 active SKUs across thousands of store locations, with each SKU-location pair constituting a distinct forecasting problem. In global supply chains, purchase commitments are made months before selling seasons begin. A 10% forecast error at order placement translates directly into either a stockout or a markdown at point of sale — with consequences that cascade through vendor relationships, working capital allocation, and customer experience simultaneously [4].

The demand signals that matter to retail forecasting do not behave the way statistical models assume. They respond to competitor promotions within hours, to social media moments within minutes, to supply disruptions before the disruption is formally reported. Traditional architectures, built to extrapolate from historical patterns, have no mechanism for continuous adaptive response at that speed. The lag between signal emergence and forecast adjustment — commonly measured in days to weeks — is a structural revenue leak that compounds across every category, location, and planning cycle in the operating year [1].

What sits behind all of this is what might fairly be called the planner attention problem. Even the most experienced demand planner has finite cognitive bandwidth. Faced with thousands of exception flags per planning cycle, planners default to heuristics — reviewing the highest-volume SKUs, applying rule-of-thumb adjustments, leaving lower-priority categories on autopilot. As retail enterprises grow their SKU counts, their channel complexity, and their geographic footprint, this model degrades predictably and silently [5].

Table 1. Traditional vs. AI agent-augmented demand forecasting: operational capability comparison across five dimensions

Dimension	Traditional Forecasting Systems	AI Agent-Augmented Forecasting
Signal Breadth	Historical sales; manually entered promotions; seasonal indices — updated on batch cycles	Live POS streams; IoT/RFID; social signals; competitor pricing feeds; macroeconomic indicators — continuous
Adaptability	Static parameters; manual review cycles that lag demand shifts by days to weeks	Continuous self-refinement; real-time signal ingestion; model policy updated without human trigger
Explainability	Model output numbers with minimal narrative; planner interprets without system support	Natural language recommendations; SHAP feature attribution; counterfactual scenario generation in business language
Scalability	Planner cognitive load degrades as SKU, location, and channel volume grows	Hierarchical multi-agent coordination designed to scale with — rather than against — operational complexity
Planner Effort	High: manual exception review, parameter maintenance, adjustment documentation each cycle	Materially reduced: agent handles routine tasks; planner attention directed to strategic, high-stakes decisions

3. III. AI Agent Architectures for Retail Demand Planning

Lumping all AI agents into a single category is a mistake with real consequences for how they get deployed. A reactive agent responds to current inputs with a pre-defined policy — useful for threshold-triggered alerts and standard exception routing, but limited in planning depth. A deliberative agent maintains an internal model of the planning environment and reasons prospectively about future states — more appropriate for scenario generation and forward inventory positioning. A multi-agent system orchestrates ensembles of specialized agents in coordinated parallel, enabling coverage of the full planning complexity that no single model can handle alone [5], [6].

LLM-powered planning agents have changed what a forecasting system can realistically deliver to a non-technical planning organization. The traditional interface between a forecasting model and a planner is a number. The LLM agent layer replaces that interface with a conversation: converting structured model outputs into narrative business explanations, responding to planner questions in plain language, generating demand scenario descriptions, and orchestrating downstream tools through function-calling APIs. In documented omnichannel retail deployments, LLM agents coordinating ARIMA time series models with real-time operational feeds have generated store-level and department-level KPI forecasts and delivered actionable recommendations directly into operational workflows without requiring data science interpretation [7].

DRL agents learn a planning policy through continuous interaction with the operational environment — adjusting inventory decisions in response to a multi-objective reward function that simultaneously penalizes forecast error, stockout cost, and overstock carrying cost [8]. Multi-agent DRL systems that combine transformer-based sequence modeling with IoT sensor data integration have delivered 18.2% lower forecast error and 23.5% reduced stockout rates against state-of-the-art non-agentic baselines [5]. Transformer-based temporal fusion networks provide the forecasting backbone that makes agent-level orchestration viable at scale, capturing long-range temporal dependencies that simpler recurrent models cannot represent with adequate fidelity [9].

Figure 1. Five-Stage AI Agent Architecture for Retail Demand Planning

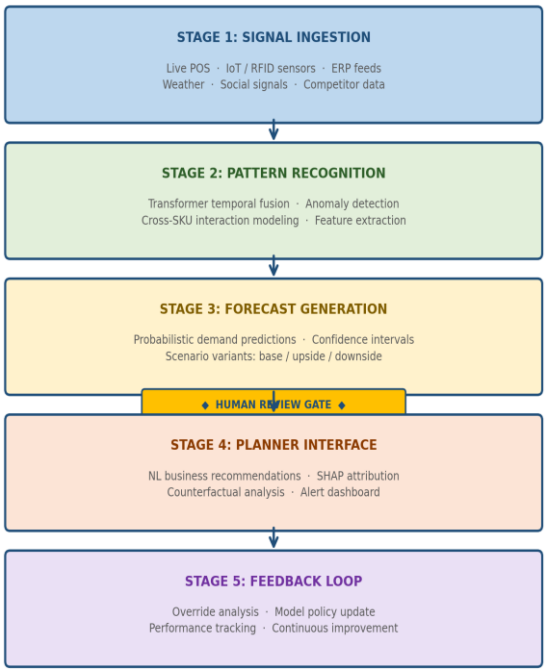

Fig. 1. Five-stage AI agent architecture for retail demand planning. Human review gates are positioned at the Stage 3/4 boundary and after Stage 4. AI agents operate continuously across all five stages; planner authority is preserved at the decision points where domain judgment creates the most value.

4. IV. Multi-Agent Coordination and Omnichannel Forecasting

The coordination challenge in omnichannel retail is not primarily a forecasting problem — it is a consensus problem. Inventory committed at store level affects distribution center availability. A sell-through acceleration in one region should trigger upstream replenishment signals before the downstream stockout becomes visible to customers. Multi-agent systems handle these interdependencies continuously, through specialized agents at each planning node reaching consensus on inventory decisions in real time. Brintrup [6] frames this as the central challenge in LLM-powered supply chain automation: enabling agents representing different organizational nodes to negotiate and converge on inventory decisions without continuous human mediation.

Validated LLM-powered multi-agent deployments in omnichannel retail are documented in peer-reviewed

literature. A deployment using four years of transactional records from a medium-sized Irish omnichannel retailer implemented an LLM agent framework coordinating ARIMA models with real-time operational data, generating KPI forecasts at store and department level — Average Transaction Value, Items per Transaction, Conversion Rate — and delivering narrative recommendations into planner workflows [7]. Multi-agent coordination architectures fall into hierarchical designs (master agent delegating to domain sub-agents) and peer-to-peer designs (agents negotiating directly). Liu et al. [10] demonstrate hierarchical multi-agent DRL outperforms centralized optimization in high-variability demand environments.

Table 2. Performance comparison across four forecasting architecture types

Architecture	MAPE (%)	Stockout vs. Baseline	Inventory Cost vs. Baseline	Planning Response
Traditional statistical (ARIMA / ETS)	28.76	Baseline (0%)	Baseline (0%)	Manual; 2-5 days lag
Single ML model (LSTM / XGBoost)	16.43	~10-12% reduction	~6-8% reduction	Semi-automated; hours
Single DRL agent (multi-objective)	~15.0	~16-18% reduction	~10% reduction	Automated; near real-time
Multi-agent DRL + transformer + IoT	~14.5	23.5% reduction	10-15% reduction	Automated; real-time

5. V. Human-AI Collaboration in Demand Planning Workflows

The consistent failure mode in enterprise AI planning deployments is not a model failure — it is an adoption failure. A technically excellent system that planners do not use, do not trust, or actively work around delivers zero operational value regardless of its accuracy. The co-pilot model proposed here positions AI agents to handle high-frequency, lower-stakes tasks — routine parameter maintenance, standard exception resolution, data quality flagging — while human planners handle contextually complex, organizationally consequential decisions that require domain knowledge, supplier relationship awareness, and strategic context that no AI system currently encodes reliably [11].

Explainability is the non-negotiable prerequisite. Research on human-AI collaboration in retail planning demonstrates clearly that planners will not systematically adopt AI recommendations they cannot interpret [11]. Planners need SHAP attribution showing which features drove the recommendation, counterfactual analysis showing how the recommendation changes if a key assumption is wrong, and recommendations expressed in business language they recognize. SHAP, LIME, and structured counterfactual generation are architectural requirements, not optional add-ons [12].

Trust calibration is the longer-run challenge. In early deployment phases, planners will override AI recommendations regularly — sometimes with genuine domain knowledge the model lacks; sometimes from habitual heuristics. The system must capture these overrides, analyze their relationship to subsequent outcomes, and update agent behavior accordingly. Organizations that skip this feedback architecture deploy systems that plateau below their potential and eventually get abandoned despite superior technical capability [6].

6. VI. Implementation Considerations and Enterprise Deployment

Data infrastructure is where ambition meets reality. The forecasting capability ceiling of any agent system is set not by the sophistication of its models but by the quality, completeness, and latency of its input data. A survey of twelve organizations transitioning to AI-driven forecasting platforms found that full-scale implementation averaged 8.7 months, with data integration consuming 42% of total project time [3]. This is not bad luck — it is predictable, and it can be managed if the data infrastructure investment is scoped honestly before technical development begins rather than discovered mid-project.

System integration is the second constraint that consistently catches organizations off guard. An AI agent that generates superior demand signals but cannot pass those signals cleanly into the ERP, WMS, and OMS that downstream teams depend on produces no operational value. End-to-end integration design —

from signal ingestion through model inference through planner interface through downstream execution trigger — must be fully scoped before technical development begins. Build-versus-buy decisions at each layer should be evaluated against existing organizational capabilities before committing to new development [13-15].

Change management determines whether technical success becomes business impact. Experienced demand planners will arrive at AI agent deployments with rational skepticism. Treating adoption as a training problem — delivering a system, running a few sessions, and hoping behavior changes — produces the highest override rates and the most rapid regression to pre-agentic practices. Organizations that make the repositioning of planner roles genuine — not just in onboarding slide decks but in how planning performance is actually measured — achieve materially higher adoption [11-15].

Table 3. Enterprise deployment framework: readiness dimensions, critical requirements, and failure modes

Readiness Dimension	Critical Requirements	Failure Mode if Absent
Data Infrastructure	Integrated, low-latency feeds: POS, ERP, WMS, IoT, supplier APIs; data quality governance	Agent recommendations unreliable; planner trust collapses before system demonstrates value
System Integration	End-to-end API connectivity: forecast output through execution systems	Forecasts remain analytical artifacts; no downstream operational impact
Explainability Design	SHAP / LIME attribution; counterfactual generation; confidence in business language	High override rates; distrust compounds; adoption plateaus below potential
Change Management	Role clarity; pilot evidence; executive sponsorship; performance metrics updated	Practitioner resistance; shadow manual processes maintained
Feedback Architecture	Override capture; planner annotation; model update loops; performance dashboards	Agent does not learn from experience; accuracy plateaus; system eventually abandoned

7. VII. Results and Measured Impact

The performance evidence for AI agent-based retail demand forecasting has moved past the stage where hedging is warranted. The most comprehensive single-study model comparison evaluates six ML models across 1,876 products spanning grocery, fashion, and electronics retail: LSTM networks achieve a MAPE of 16.43% against 28.76% for traditional statistical methods — a 42.87% improvement [3]. LLM-powered multi-agent coordination in omnichannel environments adds a further layer of KPI forecast improvement at store and department level [7]. Multi-agent DRL systems deliver 18.2% lower forecast error and 23.5% reduced stockout rates compared to state-of-the-art non-agentic baselines, with the most pronounced improvements at promotional events and seasonal transitions [5].

Inventory efficiency outcomes are equally direct in their financial implications. The 23.5% stockout rate reduction reported by Yang et al. [5] represents recovered revenue from sales events that would otherwise have been permanently lost. In best-practice AI-driven planning implementations, product unavailability reductions of up to 65% are documented, alongside inventory carrying cost reductions of 10-15% [1]. At large-scale retail enterprise values — where annual inventory carrying costs typically run at 20-30% of total inventory value — a 10-15% carrying cost reduction represents hundreds of millions of dollars in recovered working capital annually.

Planner productivity improvements may be the most strategically durable long-run outcome. In organizations where AI agents absorb routine exception resolution and standard forecast adjustment, experienced planners redirect their attention to decisions that genuinely require their expertise: new category launches, major promotional strategy, demand sensing for products without historical data.

Planners in AI-augmented environments consistently describe their work shifting from reactive exception processing to proactive strategic guidance — a change in planning function that, compounded over time, represents a capability advantage competitors cannot replicate quickly [11].

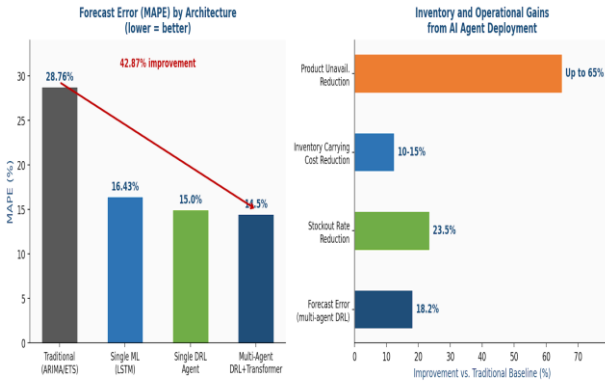

Figure 3. Measured Performance Gains from AI Agent Deployment vs. Traditional Baselines
Sources: [1] Walter et al. 2025; [3] Li et al. 2025; [5] Yang et al. 2025; [7] Nolan et al. 2025

Fig. 2. Measured performance gains from AI agent deployment in retail demand planning benchmarked against traditional statistical baselines. Left: MAPE comparison by architecture type. Right: inventory and operational gains. All figures sourced from independent peer-reviewed evaluation studies.

8. VIII. Conclusion

What this paper has argued — and what the evidence supports — is that the value of AI agents in retail demand forecasting is not primarily in the forecasting model they wrap. It is in the planning workflow they create: continuous, adaptive, explainable, and calibrated to a human-AI collaboration model that planners can actually adopt and sustain over time. A better model without that workflow infrastructure produces marginal gains in controlled evaluation and underwhelming results in production. A well-architected AI planning co-pilot changes the nature of retail planning from a reactive, periodic, manually intensive process into a continuously operating organizational capability that compounds in value as it learns.

The practical investment implications are direct. Data infrastructure and end-to-end system integration are the prerequisites that model sophistication cannot substitute for — and they must be scoped honestly before technical development begins. Explainability architecture is the adoption lever; without it, technical accuracy does not translate into planning behavior change. Change management that genuinely repositions planners as AI-augmented strategic decision-makers is the organizational design prerequisite for sustained impact. Retailers that invest in all four dimensions in parallel will generate compounding planning advantages.

Future developments in agentic retail commerce will extend AI agent responsibility beyond demand forecasting into promotional optimization, dynamic pricing coordination, autonomous replenishment, and supply-side negotiation. The data pipelines, integration architectures, explainability frameworks, and human-AI collaboration models being built in demand planning deployments today are the foundations on which those capabilities will rest. Organizations that build those foundations deliberately will hold a compounding competitive advantage as agentic commerce moves from early adoption to operational standard.

References

[1] A. Walter, K. Ahsan, and S. Rahman, "Application of artificial intelligence in demand planning for supply chains: a systematic literature review," Int. J. Logistics Manag., vol. 36, no. 3, pp. 672-719, 2025.

[2] Sunaina Sridhar et al., "A comprehensive framework for human-AI collaborative decision making in intelligent retail environments," Expert Syst. Appl., 2026.

[3] Srinivas Ankam, "AI-driven demand forecasting in enterprise retail systems," Int. J. Sci. Adv. Technol., vol. 15, no. 1, 2025.

[4] O. R. Amosu et al., "AI-driven demand forecasting: enhancing inventory management and customer satisfaction," World J. Adv. Res. Rev., vol. 23, no. 2, pp. 708-719, 2024.

[5] Y. Yang et al., "Multi-agent deep reinforcement learning for integrated demand forecasting and inventory optimization in sensor-enabled retail supply chains," Sensors, vol. 25, no. 8, p. 2428, 2025.

[6] Valeria Jannelli, "Agentic LLMs in the supply chain: towards autonomous multi-agent consensus-seeking," Int. J. Prod. Res., 2025.

[7] Darian Reyes and Pierangelo Rosati, "ARIMA forecasting with LLM-powered multi-agent coordination for omnichannel retail KPIs," in Proc. IEEE Conf., 2025.

[8] Vasanth Rajendran et al., "Automated demand forecasting in retail supply chains using deep reinforcement learning," in Proc. IEEE Conf., 2025.

[9] M. A. Shahzad et al., "Demand forecasting for retail using three-S temporal fusion (3STF) network," 2025.
[10] X. Liu et al., "Multi-agent deep reinforcement learning for multi-echelon inventory management," Prod. Oper. Manag., 2025.
[11] P. Jooss et al., "Artificial intelligence and work design: implications for frontline service employees and future research," J. Service Manag., 2025.
[12] Anmol Aggarwal, "Explainable AI for demand forecasting and price optimization: a transparent approach using tree models and SHAP," in Proc. IEEE Conf., 2025.
[13] Min Jeong An et al., "Demand forecasting in micro-fulfillment centers using association rule-based machine learning," Int. J. Prod. Econ., 2025.
[14] Zied Bahroun et al., "A systematic analysis of generative artificial intelligence for supply chain transformation," 2025.
[15] NAMEER UL HAQ QURESHI et al., "Demand forecasting for Rossmann stores using weather-enhanced deep learning model," IEEE Access, vol. 12, 2024.