Check for updates

Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 5, pp. 432-446, October 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.5.432
© 2026 KIIE

로컬 LLM기반 RAG를 활용한 완제품 안전재고 산출 프레임워크 연구

조형우 · 유우식†

인천대학교 산업경영공학과

A Framework for Finished Goods Safety Stock Estimation Using Local LLM-Based Retrieval-Augmented Generation (RAG)

Hyoungwoo Cho · Woo-sik Yoo

Department of Industrial and Management Engineering, Incheon National University

This study proposes a local LLM-RAG-based framework for estimating finished-goods safety stock in high-mix low-volume manufacturing environments. The proposed framework maintains a statistical safety stock calculation as a quantitative anchor and complements it with context-aware reasoning using a local Large Language Model (LLM) and Retrieval-Augmented Generation (RAG). The framework first preprocesses ERP/MES data through SQL-based filtering and calculates dynamic indicators such as risk lead time, seasonality, urgent-order volatility, demand trends, and open orders. These indicators are then injected into the local LLM as structured RAG context to generate explainable safety stock adjustment rationales. The framework was implemented using empirical data from an automotive parts supplier and applied to 180 finished-goods items. The developed program generated safety stock output logs over 22 ASOF points from January to May 2026. Because safety stock is a policy-level reference value linked to production planning and MRP, the quantitative evaluation was conducted using five monthly closing points. The results showed positive penalty improvement in all monthly closing periods, while the delivery response proxy increased from 80.7% to 83.0%. These findings indicate that the proposed framework can support context-aware safety stock adjustment by jointly considering ABC grade, seasonality, demand trends, urgent-order volatility, and open orders, while avoiding excessive weekly fluctuations in policy-level safety stock values.

Keywords: Safety Stock, Large Language Model (LLM), Retrieval-Augmented Generation (RAG), Inventory Optimization

1. 서 론

1.1 연구의 배경 및 필요성

최근 글로벌 제조 산업은 다품종 소량생산 체제로 급변하고 있으며, 수요 변동성과 내부 제조 리드타임의 불확실성이 충돌하는 복잡한 환경에 직면해 있다. 이러한 불확실성 속에서 결품과 과잉 재고를 동시에 방어하기 위한 안전재고 산출은 제조 경쟁력을 좌우하는 가장 어려운 과제 중 하나이다.

기존의 공급망 관리 현장에서는 주로 과거 데이터에 기반하여 정규분포를 가정한 고정된 수식을 활용하는 전통적 수리 모델에 크게 의존해 왔다. 그러나 이러한 접근 방식은 계산 과정이 정형화되어 정적인 결과만을 도출할 뿐, 고객 요구 리드타임의 변동이나 긴급 수주의 발생과 같은 현장의 역동성과 비즈니스 맥락을 반영하지 못하는 수리적 경직성의 한계를 지닌다.

† 연락저자 : 유우식 교수, 22012 인천광역시 연수구 아카데미로119 인천대학교 산업경영공학과 8호관, Tel : 032) 835-8488, Fax : 032)835-0777, E-mail : wsyoo@inu.ac.kr
2026년 5월 7일 접수; 2026년 6월 17일, 2026년 7월 5일 수정본 접수; 2026년 7월 6일 게재 확정.

전통적 수리 모델의 경직성을 보완하기 위해, 최근 학계와 산업계에서는 심층강화학습을 비롯한 신경망 및 에이전트 기반의 데이터 주도형 AI 모델을 재고관리와 공급망 의사결정에 적용하려는 연구가 활발히 이루어지고 있다. 이러한 접근법은 동적 환경에서의 정책 탐색과 복잡한 의사결정 문제 처리에 장점을 가지지만, 실제 제조 현장에 적용하기 위해서는 충분히 정제된 학습 데이터, 안정적인 상태 정의, 그리고 결과 해석 가능성이 함께 확보되어야 한다.

그러나 다품종 소량생산 제조 현장에서는 수기 입력 지연, 인적 오류, 긴급 수주, 미납수주 처리 방식의 차이 등으로 인해 입력 데이터의 불확실성이 지속적으로 발생한다. 이러한 불확실성이 안전재고 산출 과정에 직접 반영될 경우, 단기적인 데이터 변동이 장기적인 정책성 재고 기준값의 변동으로 전이되어 생산계획과 MRP의 안정성을 저해할 수 있다. 따라서 본 연구에서는 특정 DRL 모델과의 직접 성능 비교보다는, 통계 기반 안전재고 산출값을 정량 앵커로 유지하면서 현장 맥락을 설명 가능한 방식으로 보완하는 안전재고 산출 프레임워크에 초점을 둔다.

본 연구는 <Figure 1>과 같이, 수리 모델의 정형성은 유지하되 현장 맥락을 반영하기 어려운 한계를 보완하기 위해 로컬 LLM과 RAG를 결합한 하이브리드 안전재고 산출 구조를 제안한다. 제안 구조는 SQL 기반 전처리를 통해 입력 데이터의 정합성을 확보하고, 통계 기반 안전재고 산출값을 LLM 추론의 정량 앵커로 사용한다. 이후 ABC 등급, 계절성, 수요 추세, 긴급오더 변동성, 미납수주 정보를 RAG 컨텍스트로 주입하여 품목별 안전재고 조정 근거를 생성한다. 이를 통해 본 연구는 특정 DRL 모델과의 직접 성능 우열 비교보다는, 실제 제조 데이터 환경에서 적용 가능한 설명 가능한 안전재고 산출 절차와 의사결정 근거의 투명성을 제시하는 데 초점을 둔다.

1.2 기존 연구의 동향 및 한계점

공급망 관리(SCM)와 재고관리 분야에서는 불확실한 수요와 리드타임을 고려하여 안전재고를 배치하고 보충정책을 최적화하려는 연구가 지속적으로 수행되어 왔다. Graves and Willems(2000)는 공급망 네트워크에서 수요 불확실성을 고려한 전략적 안전재고 배치 문제를 다루었으며, Bertsimas and Thiele(2006)는 수요분포에 대한 완전한 사전 지식이 없더라도 불확실성을 반영할 수 있는 강건최적화 기반 재고관리 접근법을 제시하였다. 이러한 연구들은 안전재고가 단순한 평균 수요의 보정값이 아니라, 수요 변동성, 리드타임, 서비스 수준, 공급망 구조와 연계된 정책적 의사결정 변수임을 보여준다.

최근에는 기계학습과 인공지능을 활용하여 동적 공급망 환경에서 재고 및 보충정책을 조정하려는 연구도 확대되고 있다. Priore *et al.*(2019)은 환경 변화가 빠른 공급망에서 기계학습, 특히 귀납학습을 활용하여 시점별 환경 조건에 따라 적절한 보충정책을 동적으로 선택하는 접근법을 제시하였다. Baryannis *et al.*(2019)은 공급망 리스크 관리 분야의 AI 연구를 의사결정, 예측, 학습 역량과 식별·평가·대응 과제 관점에서 체계적으로 분류하였다. 또한 Toorajipour *et al.*(2021)은 공급망 관리 전반에서 AI 기법의 적용 영역, 적용 가능성이 높은 세부 과제, 향후 연구 격차를 분석하였다. 이러한 선행연구는 AI가 공급망 의사결정을 보완할 수 있음을 보여주지만, 실제 제조 현장의 수기 입력 지연, 미납수주 처리, 긴급수주, 품목별 운영정책과 같은 맥락 변수를 안전재고 산출 과정에 직접 결합하는 문제는 여전히 추가적인 연구가 필요한 영역이다. 심층강화학습(Deep Reinforcement Learning, DRL) 역시 동적 재고관리와 공급망 의사결정 문제를 해결하기 위한 주요 접근법으로 연구되어 왔다. Oroojlooyjadid *et al.*(2022)은 Beer Game 환경에서 Deep Q-Network를 활용한 재고 최적화 가능성을 제시하였으며, Gijsbrechts *et al.*(2022)은 lost sales, dual sourcing, multi-echelon inventory 문제에서 DRL이 기존 휴리스틱 및 근사 동적계획법과 경쟁 가능한 성능을 보일 수 있음을 실증하였다. 그러나 이러한 접근법은 상태 정의, 학습 데이터 구성, 튜닝 비용, 결과 해석 가능성 측면에서 실무 적용 시 고려해야 할 요소가 존재한다. 따라서 본 연구는 DRL 모델과의 직접적인 성능 우열 비교를 수행하기보다는, 통계 기반 안전재고 산출값을 정량 앵커로 유지하고, 제조 현장의 운영 맥락을 설명 가능한 방식으로 보완하는 로컬 LLM-RAG 기반 안전재고 산출 절차를 제안하는 데 초점을 둔다.

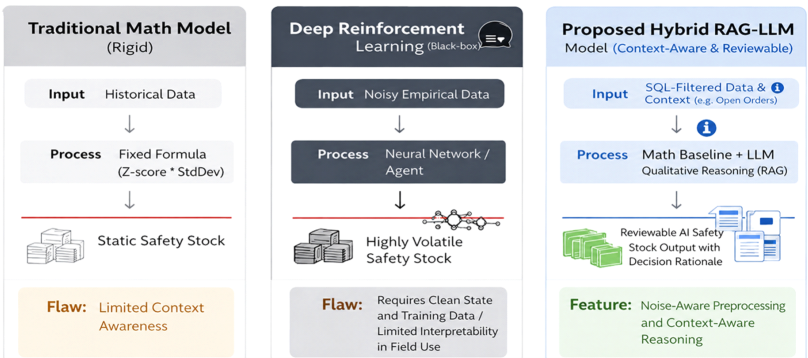


Figure 1. Conceptual Positioning of the Proposed Local LLM-RAG Framework for Explainable Safety Stock Estimation

한편, 최근 생성형 AI와 대형언어모델(LLM)의 발전은 공급망 및 운영관리 분야에서 새로운 의사결정 지원 가능성을 제시하고 있다. Srivastava *et al.*(2024)은 LLM이 공급망 관리에서 활용될 수 있는 잠재적 주제와 적용 영역을 분석하였고, Jackson *et al.*(2024)은 생성형 AI가 수요예측, 재고관리, 공급망 설계, 리스크 관리 등 다양한 SCOM 의사결정 영역에 적용될 수 있음을 제시하였다. 또한 Kosasih *et al.*(2024)은 공급망 분야에서 설명 가능한 AI와 신경기호적 접근이 실무 수용성을 높이는 데 중요하다고 보았다. 최근 Jackson *et al.*(2026)은 RAG를 활용하여 비정형 공개 데이터를 공급망 그래프로 구조화하는 방법을 제시함으로써, LLM과 RAG가 공급망 맥락정보를 구조화하고 재사용하는 데 활용될 수 있음을 보였다. 그러나 기존 LLM 및 RAG 기반 공급망 연구는 주로 공급망 지식 탐색, 공급망 매핑, 의사결정 지원, 비정형 정보의 구조화에 초점을 두고 있으며, ERP/MES 기반의 수치형 운영지표를 통계적 안전재고 산출식과 결합하여 완제품 안전재고 기준값을 산출하고, 그 조정 근거를 품목별로 설명하는 연구는 아직 제한적이다. 따라서 본 연구에서는 실데이터 기반의 통계적 안전재고 산출값을 정량 앵커로 유지하면서, ABC 등급, 계절성 지수, 수요 추세, 긴급오더 변동성, 미납수주 정보를 RAG 컨텍스트로 주입하여 품목별 안전재고 조정 근거를 생성하는 로컬 LLM-RAG 기반 하이브리드 안전재고 산출 프레임워크를 제안한다.

1.3 연구의 목적 및 방법

본 연구의 목적은 다품종 소량생산 제조 환경에서 통계 기반 안전재고 산출값을 정량 앵커로 유지하면서, 로컬 LLM과 RAG를 활용하여 품목별 운영 맥락을 설명 가능한 방식으로 반영하는 완제품 안전재고 산출 프레임워크를 제안하는 데 있다. 이를 위해 본 연구는 ERP/MES 데이터의 SQL 기반 전처리, 통계적 안전재고 산출, 구조화된 RAG 컨텍스트 구성, 로컬 LLM 기반 판단 근거 생성, 그리고 HITL 검토 및 산출 로그 축적 단계로 구성된 파이프라인을 설계하였다. 전체 절차는 <Figure 2>에 제시된 바와 같이 4개의 구역(Zone)과 A부터 J까지 이어지는 단계로 구성된다.

첫째, 구역 1. 원천 데이터 수집 단계에서는 데이터의 무결성을 확보하기 위한 전처리가 수행된다. [A. 마스터 및 전처리] 과정에서는 NULL값을 0으로 치환하는 자가치유를 수행하고, 품목의 ABC 등급 및 수주 테이블 스캔 옵션을 결정하는 중앙 정책(Policy)을 동적 로드한다. 이어지는 [B. 생산/수주/수불 빅데이터 스캔]에서는 PDBox(생산실적), BizOrder(수주), Stock(수불) 데이터를 스캔하여 고객요구 리드타임(CNT)과 긴급오더율(UVI)을 추출한다. 또한 [C. 미납 수주 연산] 단계에서 '수주량 - 출고량 = 미납 잔량(Open Order)'을 계산함으로써 향후 안전재고의 이중 카운팅을 원천 방지하며, [D. 시계열 피처 추출]을 통해 과거 36개월 데이터를 바탕으로 계절성 지수(SI)와 최근 12개월(12M) 트렌드 배열을 생성하여 동적 변동성 지표를 완성한다.

둘째, 수집 및 정제된 원천 데이터는 구역 2. '통계 연산 및 1차 DB 적재' 단계로 이관된다. [E. 통계적 안전재고 연산]에서는 앞서 추출된 Z-Score, 표준편차(StdDev), Risk_LT, UVI, SI를 공식(Eq.5)에 대입하여 1차 통계재고(SS_stat) 베이스라인을 도출한다. 이후 대용량 데이터 처리 시 발생할 수 있는 메모리 초과 오류를 방지하기 위해, [F. Staging DB 1차 일괄 적재] 과정에서 도출된 데이터를 AI_Item_SafetyStock_Log 테이블에 Bulk Insert 방식으로 일괄 적재하며 구역 2를 종료한다.

셋째, 데이터 적재가 완료되면 구역 3. 'AI 추론(Agentic LLM Reasoning)' 단계에서 수리적 한계를 보완하는 지능형 의사결정이 이루어지며, DB 부하를 최소화하기 위해 메모리 적

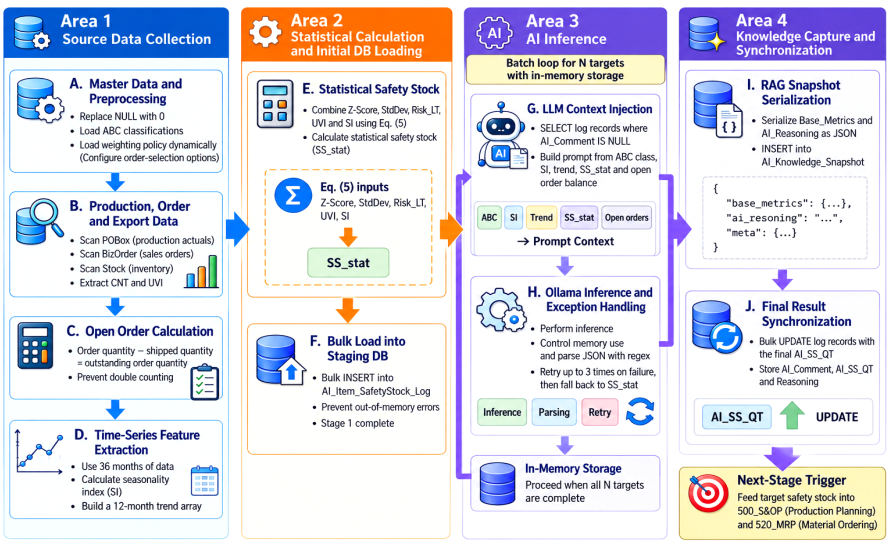


Figure 2. Proposed Local LLM_Based Context-Aware Hybrid Safety Stock Optimization Framework

재(In-Memory) 방식이 활용된다. [G. LLM Context 주입]을 위해 Log 테이블에서 AI_Comment IS NULL인 타겟만 SELECT한 후, 해당 품목의 ABC등급, SI, 트렌드, SS_stat, 미납 잔량 정보를 RAG 프롬프트 컨텍스트로 구성한다. 이를 기반으로 [H. Ollama 추론 및 예외처리] 단계에서 Target N건을 반복 추론(Batch Loop)한다. 이 과정에서 Ollama 엔진이 정성적 판단을 수행하며, 도출된 텍스트는 마크다운 제거 및 JSON 정규식 파싱을 거쳐 시스템 언어로 정제된다. 특히 시스템 안정성을 위해 파싱 실패 시 3회 재시도 로직을 수행하며, 최종 실패 시에는 1차 통계량(SS_stat)으로 대체(Fallback)하여 무중단 운영을 보장한다. Target N건의 추론 결과는 메모리에 순차적으로 적재되며, 루프가 완료된 후 다음 단계로 일괄 이관된다.

넷째, 구역 4. 'RAG 로그 축적 및 결과 동기화(RAG Persistence and Output Synchronization)' 단계에서는 LLM 추론 결과와 판단 근거를 구조화된 로그로 저장하고, 안전재고 산출 결과를 운영 시스템과 연계한다. [I. RAG 스냅샷 직렬화 적재]에서는 Base_Metrics와 AI_Reasoning 결과를 JSON 형태로 직렬화하여 AI_Knowledge_Snapshot 테이블에 저장한다. 이 데이터는 모델 파라미터를 직접 갱신하는 학습 데이터가 아니라, 이후 추론 주기에서 참조 가능한 RAG 컨텍스트 로그로 활용된다. 이어서 [J. 최종 결과 동기화] 단계에서는 산출된 최종 수량(AI_SS_QT), AI_Comment, Reasoning을 Log 테이블에 일괄 UPDATE하여 안전재고 산출 결과와 판단 근거를 함께 관리한다. 동기화된 안전재고 데이터는 후속 S&OP 및 MRP 모듈에서 참조 가능한 기준값으로 제공되며, 실제 반영 여부와 시점은 월말 확정 정책 및 현장 검토 절차에 따라 결정된다.

2. 수리적 통계 기반의 1차 안전재고 베이스라인 구축

본 장에서는 다품종 소량생산 환경의 극심한 수요 및 리드타임 변동성을 제어하기 위해, 현장 실증 데이터를 바탕으로 <Figure 2>의 구역 2와 같이 1차 통계적 안전재고를 도출한다.

2.1 현장 실데이터 기반 수요 및 리드타임 통계량 산출

본 연구에서 사용한 주요 수식 기호와 변수의 정의는 <Table 1>과 같다. 각 변수는 ERP/MES에서 추출된 수주, 납품, 출고, 수불 및 품목 마스터 데이터를 기반으로 산출되며, 이후 통계 기반 안전재고 산출과 LLM-RAG 추론 컨텍스트의 입력값으로 활용된다. 본 연구의 1차 안전재고 베이스라인은 완전히 새로운 확률모형을 제안하는 것이 아니라, 전통적인 정규분포 기반 안전재고 산출 구조를 기본으로 하되, ERP/MES에서 산출 가능한 고객 요구 리드타임(CNT), 순수 위험 리드타임(RiskLT), 긴급 수주 변동성(UVI), 계절성 지수(SIm)를 결합한 현장 적용형 확장 휴리스틱 기준값이다. 즉, 본 베이스라인은 이후 LLM이 임의의 안전재고 수량을 생성하지 않도록 하는 정량 앵커로 사용되며, 품목별 운영 맥락을 반영하기 전의 비교 기준값으로 기능한다.

$$CNT = \frac{1}{N}\sum_{i=1}^{N}(DeliveryDate_i - PODate_i) \tag{1}$$

고객 요구 리드타임(CNT)은 B2B 납품 환경에서 고객사가

Table 1. Definition of Mathematical Symbols and Variables Used in Safety Stock Estimation

Symbol	Definition	Unit / Type	Description
i	Order index	Index	Index representing an individual customer order or delivery transaction
N	Number of observations	Count	Number of observed order and delivery transactions used to calculate CNT
$PODate_i$	Purchase order date of order (i)	Date	Customer purchase order placement date
$DeliveryDate_i$	Delivery date of order (i)	Date	Actual delivery date or requested delivery date for order (i)
CNT	Customer net tolerance time	Day	Average duration between purchase order date and delivery date, representing customer-acceptable waiting time
$AvgLT$	Average manufacturing lead time	Day	Average manufacturing or supply lead time by item
$RiskLT$	Net risk lead time	Day	Net risk lead time calculated as ($AvgLT - CNT$)constrained to a minimum value of 0.01
Q_{urgent}	Urgent order quantity	Quantity	Order quantity classified as short lead-time or urgent orders
Q_{total}	Total order quantity	Quantity	Total order quantity during the same observation period
UVI	Urgent-order volatility index	%	Ratio of urgent order quantity relative to total order quantity
m	Target month	Month	Target month evaluated for seasonality index calculation

Table 1. Definition of Mathematical Symbols and Variables Used in Safety Stock Estimation (Continued)

Symbol	Definition	Unit / Type	Description
SI_m	Seasonality index of month (m)	Ratio	Seasonality weight calculated by dividing the 3-year historical average demand of the target month by the overall monthly average demand
D	Demand	Quantity	Shipment or demand quantity by item
σ_D	Standard deviation of demand	Quantity	Standard deviation representing demand volatility by item
Z	Service-level coefficient	Ratio	Normal distribution-based safety factor reflecting the target service level
SS_{stat}	Statistical safety stock	Quantity	Primary baseline safety stock calculated from mathematical statistics
$OpenOrder$	Open order quantity	Quantity	Confirmed unfulfilled order quantity calculated by deducting shipped quantity from total order quantity
AI_SS_QT	AI safety stock output field	Quantity / DB field	Database field name where the finalized AI-1T safety stock quantity is stored
$Trend_{12M}$	Twelve-month demand trend	Array / Direction	Time-series array or trend direction representing demand patterns over the past 12 months
ABC	ABC inventory grade	Category	Inventory classification grade indicating item criticality or operational priority
$AI-1T$	AI 1-Tier RAG safety stock output	Quantity	Final AI-driven safety stock generated by feeding 1-Tier RAG context into a local LLM, utilizing SSstat as a quantitative anchor
$\max(\cdot)$	Maximum operator	Function	Operator that selects the larger of two values; used to maintain a minimum protective buffer even during severe off-peak seasons

발주(PO) 후 납품까지 기다려주는 평균 여유 기간을 수리적으로 산출한 지표다. 즉, 고객의 기다림 자체가 제조사가 확보할 수 있는 '무비용 방어 시간'이 됨을 직관적으로 반영한 것이다.

$$RiskLT = \max(0.01, AvgLT - CNT) \quad (2)$$

순수 위험 리드타임(Risk_LT)은 단순 평균 리드타임에서 CNT를 차감하여 실질적으로 결품 리스크가 노출되는 순수 위험 구간을 도출한 지표다. 예를 들어 제조에 10일이 소요되더라도 고객이 7일을 기다려준다면, 우리가 순수하게 재고를 들고 방어해야 하는 위험 기간은 단 3일에 불과하다는 현장 직관을 수식화한 것이다.

2.2 수요 변동성 지표 (긴급 수주 변동성(UVI) 및 계절성 지수(SI) 도출)

$$UVI = \left(\frac{\sum Q_{urgent}}{\sum Q_{total}} \right) \times 100 \quad (3)$$

긴급 수주 변동성 지수(UVI)는 전체 수주 물량 대비 정규 수요 예측을 벗어난 기습적 단납기 오더의 비율을 정량화한 변동성 지표다. 원청의 갑작스러운 주문 폭주가 잦은 품목일수록 이 UVI 가중치가 커져, 널뛰는 긴급 오더에 대비한 방어용 재고 버퍼를 기민하게 늘려준다.

$$SI_m = \frac{Target\ Month\ Avg\ Demand\ (Past\ 3\ Yrs)}{Total\ Monthly\ Avg\ Demand\ (Past\ 3\ Yrs)} \quad (4)$$

계절성 지수(SI)는 목표 월의 과거 3년 평균 수요를 전체 3년 월평균 수요로 나눈 값으로, 수요의 비선형적 패턴을 포착하는 핵심 가중치이다. 특정 월의 지수가 1.5라면 평소보다 50% 더 팔리는 성수기를, 0.7이면 30% 덜 팔리는 비수기를 의미하여 월별 수요 요동에 완벽히 동기화된다. 여름에 쏟아지는 주문을 1년 평균으로 뭉뚱그려 계산하면 성수기엔 결품이 나고 겨울엔 악성 재고가 쌓이는 '평균의 함정'에 빠지게 된다. <Figure 3>에서 시각화한 바와 같이, 계절성 지수(SI)를 동적으로 반영한 베이스라인은 실제 수요 곡선에 선제적으로 반응하여 결품과 과잉 재고 리스크를 동시에 상쇄하는 강력한 유연성을 제공한다.

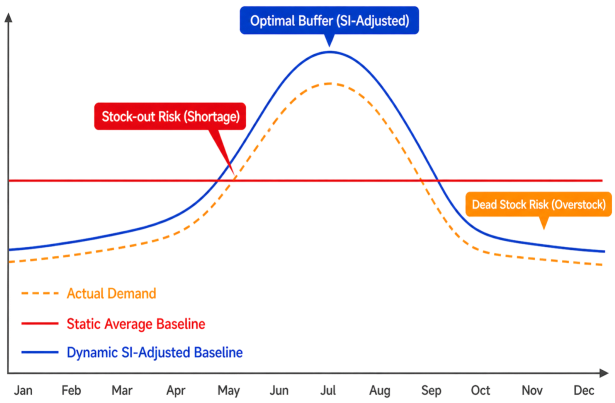


Figure 3. Comparison of Static Average Baseline vs. Dynamic SI-Adjusted Baseline overcoming the Average Trap

2.3 1차 통계적 안전재고(Mathematical Baseline) 산출

$$SS_{stat} = Z \times \sigma_D \times \sqrt{\frac{RiskLT}{30}} \times \left(1 + \frac{UVI}{100}\right) \times \max(0.5, SI_m) \quad (5)$$

1차 통계적 안전재고(SS_{stat})는 고전적 정규분포 공식에 실질 위험 리드타임(Risk_LT), 긴급오더 폭주율(UVI), 계절성 지수(SI)를 가중 곱셈으로 결합한 확장형 휴리스틱 모델이다. 정적 공식에 현장의 리스크 (기다려주는 시간, 갑작스러운 오더, 계절성)를 재고 수치로 도출해낸다. 특히 수식(5)에서 계절성 지수에서 $\max(0.5, SI_m)$ 와 같이 0.5 하한선을 설정한 것은 다품종 소량생산(HMLC) 현장의 불확실성을 제어하기 위한 수리적 안전망이다. B2B 제조 환경에서는 비수기라 할지라도 A/S 물량이나 원청의 돌발적인 긴급 수주가 언제든 발생할 수 있다. 만약 극단적인 비수기 지수(예: SI < 0.2)를 그대로 곱할 경우 방어 버퍼가 0에 가깝게 과도하게 삭감되므로, 최악의 수요에서도 통계적 안전재고의 최소 50%를 '최소 방어(Minimum Buffer)'로 유지하도록 통제한 것이다.

이렇게 산출된 1차 수리적 안전재고는 다품종 소량생산의 불확실성을 정량적으로 훌륭히 제어하지만, 아직 ABC 재고 등급이나 미납 수주 같은 비즈니스 맥락을 완벽히 이해하지는 못한다. 따라서 이 수치는 향후 3장에서 진행될 LLM 추론의 정량적 앵커(Anchor)로 활용되어, 정성적 맥락과 동적으로 융합되는 핵심 기반이 된다.

3. 로컬 LLM 및 RAG 기반 하이브리드 추론 엔진 설계

3.1 로컬 LLM 구축 및 결정론적 정형 RAG 컨텍스트 주입

제조 현장의 수불 및 발주 데이터는 외부로 유출할 수 없는 핵심 산업 보안 데이터이므로, 본 프레임워크는 망분리 환경에서 독립적으로 구동 가능한 경량화 로컬 LLM(예: Qwen2.5 14B)을 추론의 코어로 채택하였다. 정량적 의사결정의 일관성을 위해 모델의 Temperature 파라미터는 최소화(0.2 수준)하여 환각(Hallucination) 현상을 극도로 억제하였다. 특히 기존 언어모델들이 수치 데이터 연산에 취약했던 한계를 보완하고자, 비정형 검색의 한계를 극복하기 위한 결정론적 정형 검색(Deterministic Tabular Retrieval) 방식을 도입하였다.

2장에서 연산된 대상 품목의 핵심 지표(시계열 트렌드 배열, ABC 등급, 계절성 지수(SI), 긴급오더율(UVI), 1차 통계적 안전재고(SS_{stat}))와 미납 확정 수주 정보를 RAG를 통해 고속으로 추출하여 동적 컨텍스트로 LLM 프롬프트에 주입한다. 이러한 다차원 컨텍스트가 주입되는 시스템 및 사용자 프롬프트의 논리적 템플릿 구조는 <Table 2>와 같이 구조화되어, 환각을 억제하고 정량적 정확도를 극대화한다.

3.2 SCM 전문가 휴리스틱 기반 추론 및 강건한 자동회귀 설계

단순한 데이터 나열을 넘어, 프레임워크 내부에는 LLM 응답을 안정적으로 제어하기 위한 SCM 전문가 휴리스틱이 주입된다. 이 휴리스틱은 안전재고와 기 확정 생산계획(MPS) 물량을 구분하고, 미납수주가 안전재고 버퍼에 중복 반영되지 않도록 하는 운영 규칙을 포함한다. 또한 LLM은 ABC 등급, 계절성 지수, 수요 추세, 긴급오더 변동성, 미납수주 정보를 함께 검토하여 품목별 안전재고 조정 방향과 그 판단 근거를 생성한다. 본 연구에서 LLM은 통계 기반 안전재고 수량을 임의로 대체하는 수치 최적화기가 아니라, SS_{stat}을 정량 앵커로 유지한 상태에서 품목별 운영 맥락을 통합하고 판단 근거를 생성하는 의사결정 지원 계층으로 사용된다. 성수기 상향, 비수기 하향, 안정 유지, 미납수주 분리와 같은 기본 판단 유형은 결정론적 규칙엔진으로도 일부 표현할 수 있다. 그러나 실제 제조 현장에서는 ABC 등급, 계절성 지수, 수요 추세, 긴급오더 변동

Table 2. Structure of the System and User Prompt for RAG Context Injection

Prompt Role	Component	Injected Content & Instructions
System Prompt	Role & Goal	- You are an SCM and inventory management expert with 30 years of experience. - Analyze multidimensional indicators to determine the optimal safety stock(AI_SafetyStockQT)and return the rationale(AI_Comment)in JSON format.
System Prompt	Expert Guidelines	- [Rule 1] Prioritize stockout prevention for Grade A, and prioritize the mitigation of obsolete inventory for Grade B. - [Rule 2] Do not double-count confirmed open orders(Open Order)into the safety stock. - [Rule 3] Use the seasonality index (*SI*) as a primary contextual signal to guide adjustment direction, while comprehensively evaluating ABC classification, demand trend, *UVI*, and open orders to finalize adjustments.
User Prompt	Item Context (RAG)	- Item: [Item Code], Grade: [ABC Grade], SS_stat: [Primary statistical safety stock quantity] - SI: [Seasonality index], UVI: [Urgent-order volatility index], Risk_LT: [Net risk lead time] - Trend_12M: [Past 12-month monthly shipment array], Open_Order: [Confirmed open order quantity]

성, 미납수주, 생산계획 반영 여부가 동시에 작용하며, 이들 신호가 항상 동일한 방향을 가리키지는 않는다. 예를 들어 계절성 지수는 낮지만 UVI가 높거나, A등급 품목이지만 확정 미납수주가 이미 생산계획에 반영된 경우처럼 상충되는 맥락이 발생할 수 있다. 본 연구에서 LLM의 한계효용은 이러한 다차원 운영 맥락을 하나의 고정 임계값 규칙으로 환원하지 않고, 정량 앵커와 전문가 규칙의 경계 안에서 조정 방향과 판단 근거를 자연어로 구조화하여 현장 실무자가 검토할 수 있게 하는 데 있다. <Figure 4>의 사례는 A등급 품목이 성수기와 우상향 수요 추세를 보이는 경우, 통계 기반 안전재고 산출값을 기준으로 상향 조정 근거가 어떻게 생성되는지를 보여준다.

도출된 AI의 판단 근거와 최종 산출 수량은 생산계획 시스템과의 연계를 위해 지정된 JSON 포맷으로 정규화된다. 또한 현장 데이터의 이상치, 로컬 서버 부하, LLM 응답 지연, JSON 파싱 오류 등에 대비하기 위해 재시도 및 fallback 절차를 포함하였다. 파싱 실패 시 최대 3회의 재시도(Max Retries = 3)를 수행하며, 최종 실패 시에는 AI 산출값을 사용하지 않고 2장에서 산출한 통계 기반 안전재고로 대체한다. <Figure 5>는 이러한 예외 상황과 품목별 운영 맥락에 따라 제안 엔진이 어떤 판단 경로를 따르는지를 요약한 시나리오 매트릭스다. 이 구조는

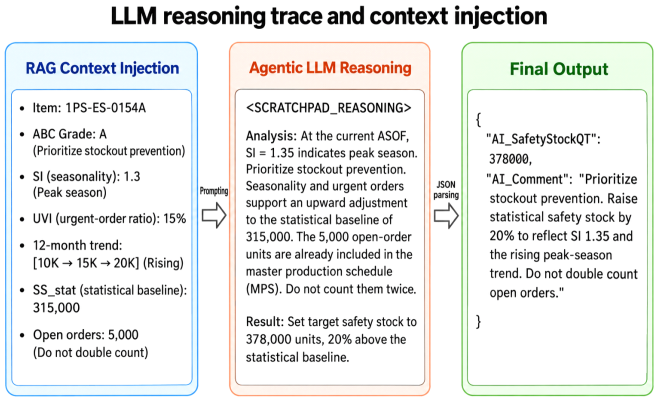


Figure 4. LLM Reasoning Trace and Context Injection Process for Dynamic Safety Stock Estimation

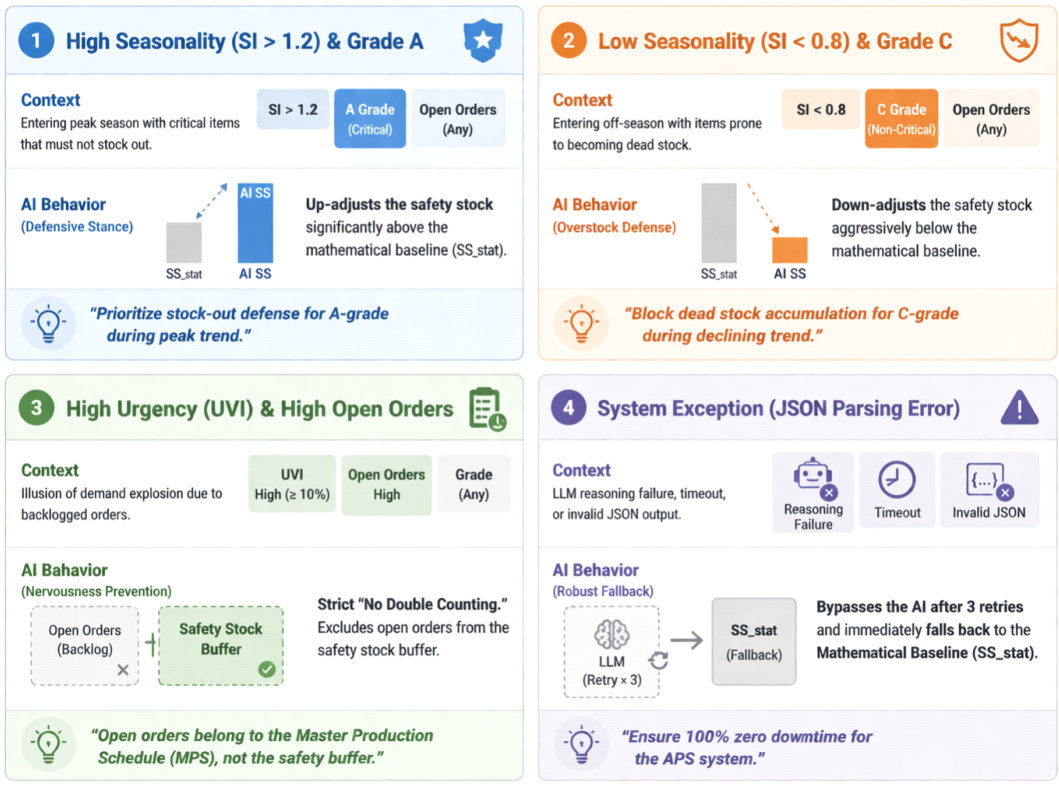


Figure 5. Contextual Scenario & AI Behavior Matrix based on SCM Expert Guidelines

LLM 응답 오류가 안전재고 산출 프로세스 전체로 전이되는 위험을 줄이고, 설명 가능한 판단 근거와 fallback 절차를 함께 제공하기 위한 것이다.

첫째, '성수기 대응(Scenario 1)'이다. 로컬 LLM은 A등급 품목이 성수기 구간(SI > 1.2)에 진입하고 수요 추세가 상승하는 경우, 통계 기반 안전재고 산출값을 기준으로 상향 조정이 필요한지 검토한다. 이때 조정 근거는 ABC 등급, 계절성 지수, 수요 추세, 긴급오더 변동성, 미납수주 정보를 함께 고려하여 생성된다.

둘째, '비수기 재고 조정(Scenario 2)'이다. B등급 품목이 비수기 하락 국면(SI < 0.8)에 진입하고 수요 추세가 감소하는 경우, LLM은 과잉재고 가능성을 고려하여 통계 기반 안전재고 대비 하향 조정 가능성을 검토한다. 이는 단순히 계절성 지수 하나만을 적용하는 것이 아니라, 품목 등급과 수요 추세를 함께 반영하는 방식이다.

셋째, '미납수주 분리 처리(Scenario 3)'이다. 긴급오더 변동성이 높거나 기확정된 미납수주가 존재하는 경우에도, 미납수주는 생산계획 또는 출고 대응 대상 수요로 구분하고 안전재고 버퍼와 별도로 처리한다. 이를 통해 동일 수요가 생산계획과 안전재고에 중복 반영되는 가능성을 줄이고, 안전재고 기준값의 불필요한 팽창을 완화한다.

넷째, '예외 상황 Fallback(Scenario 4)'이다. LLM 응답 지연, JSON 파싱 오류, 지정 포맷 이탈 등 시스템 예외가 발생하면 최대 3회의 재시도를 수행하고, 최종 실패 시에는 통계 기반 안전재고 산출값을 사용한다. 이 절차는 LLM 응답 오류가 안전재고 산출 결과에 직접 반영되는 것을 방지하고, 운영 시스템이 안정적인 기준값을 유지하도록 보완하는 역할을 한다.

4. 사례 적용 및 실험 결과

본 장에서는 제안된 로컬 LLM-RAG 기반 안전재고 산출 프레임워크의 실무 적용 가능성을 확인하기 위해, 국내 자동차 부품 2차 벤더 기업의 ERP/MES 데이터와 연동되는 안전재고 산출 프로그램을 개발하고 그 산출 결과를 분석하였다. 개발 프로그램은 2026년 1월 1일부터 2026년 5월 28일까지 총 22개 ASOF 시점에서, ABC 매출 기여도 분석에 따라 누적 매출의 약 90%를 구성하는 자동차 프레스부품 완제품 180개 핵심 품목을 대상으로 안전재고 산출 로그를 생성하였다. 다만 안전재고는 생산계획 및 MRP의 기준값으로 연동되는 정책성 재고이므로, 주차별 산출값을 모두 확정 기준으로 사용할 경우 생산계획의 변동성을 증폭시킬 수 있다. 이에 본 연구에서는 전체 ASOF 산출 로그를 검토 대상으로 유지하되, 정량 성과 비교와 추세 분석은 월말 확정 기준에 해당하는 5개 ASOF 시점을 중심으로 수행하였다. 입력 데이터로는 과거 36개월 수주・수불・출고 이력과 미납수주 정보를 활용하였다.

4.1 실험 환경 구축 및 병행성 제어(Concurrency Control) 아키텍처 구현

외부 인터넷망과 분리된 제조 현장의 엄격한 보안 요건을 충족하고자, 사내 로컬 서버 환경에 경량화된 오픈소스 대형 언어모델(Qwen2.5)과 RAG 파이프라인을 구축하였으며, 상세한 실험 환경은 <Table 3>에 명시된 바와 같다.

다품종 소량생산 현장에서는 복잡한 품목별 라우팅(Routing) 데이터의 무결성이 1차 수리적 통계 산출의 신뢰성을 좌우한다. 이를 해결하기 위해<Figure 6>과 같이 개별 품목과 설비가 매핑된 작업 준비 시간(Setup Time) 및 실적 사이클 타임(C/T)을 동적으로 동기화하는 '설비/공정 마스터 통합 관리 시스템'을 자체 개발하여 연동함으로써 입력 데이터의 정합성을 극대화 하였다. 나아가, 대규모 제조 데이터가 실시간으로 교환되는 ERP/MES 환경에서 필연적으로 발생하는 데이터 풀스캔(Full-Scan) 병목을 원천 차단하기 위해, 공통 테이블 표현식(CTE) 튜닝과 교착 상태(Deadlock) 방어 로직을 결합한 병행성 제어(Concurrency Control) 및 무중단 배포(Zero-down-time Deployment) 아키텍처를 구현하였다. 이를 통해 다수의 실무자와 AI 백그라운드 엔진이 지식 스냅샷 DB에 동시 접근하더라도 락(Lock) 발생 없이 고유 식별자를 통한 핀포인트 업데이트가 가능해져, 대규모 트랜잭션 환경에서도 실제 공장에 즉시 적용 가능한 수준의 시스템 안정성을 확보하였다

Table 3. Experimental Environment Setup

Category	Specification
Hardware	On-Premise Local Server (GPU NVIDIA RTX 4090)
Database	Microsoft SQL Server (Integrated ERP/MES view)
AI Model (LLM)	Qwen 2.5 (Local deployment via Ollama), Temperature0.2
Framework	Python 3.10, Pandas (high-performance vectorized processing), LangChain (RAG pipeline)
Target Data	180 key finished automotive press parts accounting for ~90% of cumulative revenue based on ABC sales contribution analysis /Safety stock calculation logs across 22 As-Of (ASOF) evaluation timepoints (2026.01.01. – 2026.05.28)/Quantitative performance comparison conducted focusing on 5 month-end finalized ASOF timepoints /36 months of historical orders, inventory movements (receipts/issues), shipments, and confirmed open order data

4.2 하이브리드 추론 엔진의 산출 결과 및 Human-in-the-Loop (HITL) 기반 RAG 지식 재사용

앞서 설계된 파이프라인에 따라 각 품목의 1차 통계 기반 안전재고를 산출하고, 이를 로컬 LLM에 RAG 컨텍스트로 주입하여 AI-1T 안전재고 산출 결과와 판단 근거를 생성하였다. 본 절에서는 AI-1T 산출 사례와 함께, 현장 검토 결과가 이후 추론 주기에서 RAG 컨텍스트로 재사용되는 HITL 기반 운영 구조를 설명한다. <Table 4>는 전체 ASOF 산출 로그 중 AI-1T 판단 유형을 설명하기 위해 선정한 사례이다. 사례는 성수기 상향 조정, 비수기 하향 조정, 안정 품목 유지, 미납수주 분리 처리의 네 가지 판단 유형을 보여주도록 구성하였다.

<Table 4>는 AI-1T 산출 결과가 단일 계절성 규칙에만 의존하지 않고, 품목 등급, 계절성 지수, 수요 추세, 긴급오더 변동성, 미납수주 정보를 함께 고려하여 상향, 하향, 유지 또는 분리 처리 판단을 생성하는 과정을 보여준다. A등급 성수기 품목은

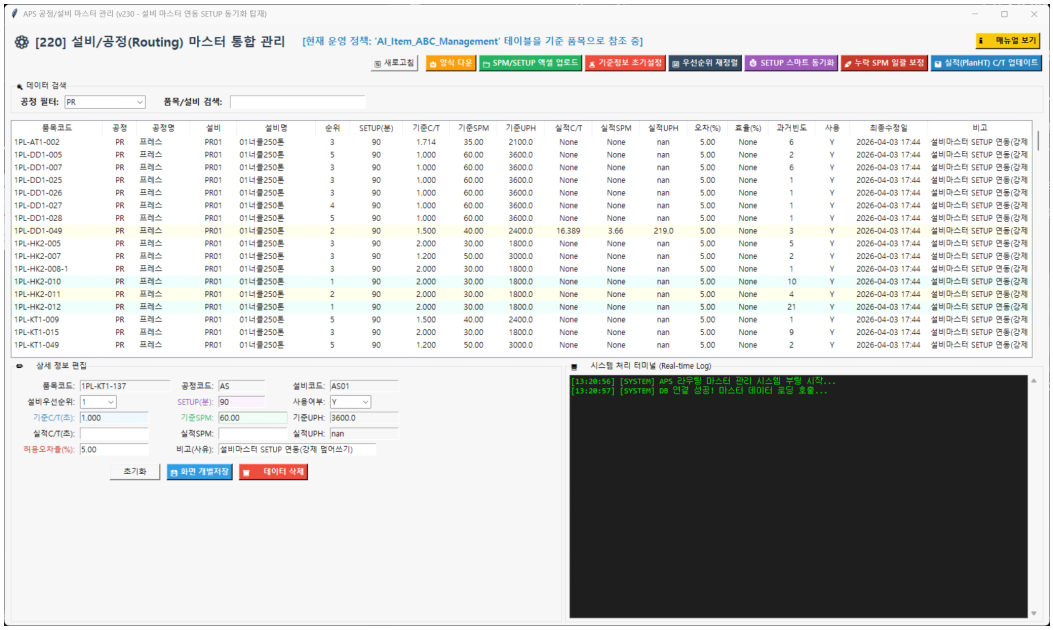


Figure 6. Conceptual Architecture of Integrated Routing Master Management and Synchronization for Data Integrity in HMLV (High-Mix Low-Volume) Production Environments

Table 4. Examples of AI-1T Safety Stock Decision Types Based on RAG Context

Item Code	Grade	SIm / UVI	Trend (Past 3 Months)	Statistical Safety Stock (SSstat)	AI Final Safety Stock	AI Decision Rationale (Reasoning Comment)
1PS-ES-0154A	A	1.35 / 15%	[Oct: 10K, Nov: 15K, Dec: 20K (Upward)]	315,000	378,000	Identified as a Grade A item with an SI of 1.35 and an upward demand trend; justified an upward adjustment compared to the statistical safety stock.
1PS-EK-040BA	B	0.70 / 45%	[Oct: 50K, Nov: 30K, Dec: 15K (Downward)]	218,500	100,000	Identified as a Grade B item entering an off-peak period (SI 0.70) with a downward demand trend; justified a downward adjustment considering excess inventory risks.
1MF-DD1-001	B	1.05 / 2%	[Oct: 40K, Nov: 41K, Dec: 40K (Stable)]	240,150	240,000	Maintained at a level similar to statistical safety stock to minimize unnecessary policy volatility, given low demand volatility and a stable recent trend.
1PS-ES-0608A	A	1.10 / 50%	[Oct: 5K, Nov: 8K, Dec: 10K (Upward)]	12,500	12,500	Although UVI is high and unfulfilled orders exist, confirmed open orders are categorized into production planning demand and are not double-counted into the safety stock buffer.

상향 조정 근거가 생성되었고, B등급 비수기 품목은 하향 조정 근거가 생성되었다. 안정 품목은 불필요한 정책값 변동을 줄이기 위해 통계 기반 안전재고와 유사한 수준을 유지하였으며, 미납수주가 존재하는 품목은 확정 수요와 안전재고 버퍼를 구분하여 중복 반영 가능성을 줄였다. 다만 <Table 4>의 사례는 LLM이 새로운 수학적 최적화 규칙을 자동으로 발견하였음을 의미하는 것이 아니라, 통계 기반 안전재고와 현장 운영 맥락을 결합하여 검토 가능한 판단 근거를 생성하는 과정을 보여주기 위한 것이다. 즉, 본 연구의 AI-1T 산출 구조는 규칙엔진을 대체하기보다, 규칙엔진이 처리하기 어려운 복합 맥락의 설명 가능성과 현장 검토 가능성을 보완하는 역할을 한다. 특히 동일한 SI 조건이라도 ABC 등급, 최근 수요 추세, UVI, 미납수주 반영 여부에 따라 상향, 하향, 유지 또는 분리 처리 판단이 달라질 수 있으므로, LLM은 이러한 판단 조합을 자연어 근거와 함께 기록하여 HITL 검토 및 이후 RAG 컨텍스트 재사용이 가능하도록 한다. 이러한 산출 및 로깅 과정은 <Figure 7>의 로컬 LLM 기반 추론 엔진 구동 화면에서 확인할 수 있다.

본 프레임워크는 AI-1T 안전재고 산출 결과와 판단 근거를 현장 실무자가 검토할 수 있도록 Human-in-the-Loop (HITL) 절차를 포함한다. <Figure 8>은 현장 검토, 피드백 저장, RAG 컨텍스트 재사용으로 이어지는 HITL 기반 지식 재사용 구조를 나타낸다.

현장 실무자가 대시보드를 통해 AI의 산출량과 결정 근거(AI Comment)를 검토하고 필요시 수량 보정 또는 의견을 남기면, 해당 피드백은 지식 스냅샷 DB에 저장된다. 이 데이터는 모델 파라미터를 직접 갱신하는 fine-tuning 데이터가 아니라, 이후 추론 주기에서 참조 가능한 RAG 컨텍스트로 재사용된다. 따라서 본 구조는 자가 학습 구조라기보다, 현장 판단 근거를 축적하고 다음 추론에서 재활용하는 HITL 기반 RAG 지식 재사용 구조로 정의할 수 있다.

4.3 월말 확정 기준 안전재고 산출 결과 및 정량 성과 분석

제안된 로컬 LLM-RAG 기반 안전재고 산출 프로그램은 2026년 1월부터 5월까지 총 22개 ASOF 시점에서 자동차 프레스부품 완제품 180개 품목에 대한 안전재고 산출 로그를 생성

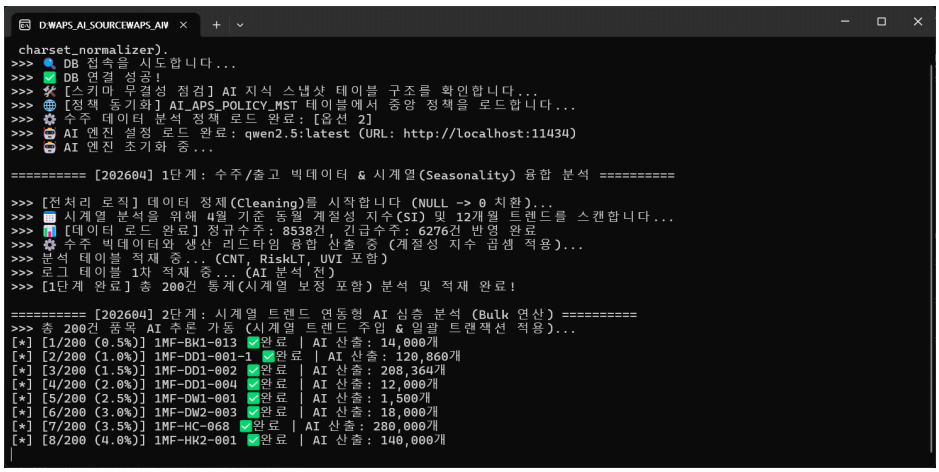


Figure 7. Interface of the Local LLM-Based Inference Engine and Decision-Rationale Logging

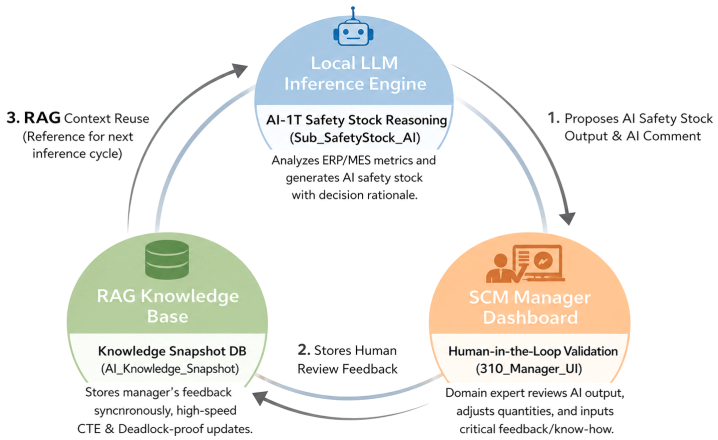


Figure 8. Human-in-the-Loop (HITL) Feedback Architecture for RAG Context Reuse

하였다. 안전재고는 생산계획 및 MRP의 기준값으로 연동되는 정책성 재고이므로, 본 연구에서는 주차별 산출값을 모두 확정 기준으로 사용하지 않고 월말 확정 기준에 해당하는 5개 시점을 중심으로 정량 성과를 분석하였다. 본 연구의 비교 기준(benchmark)은 2장에서 산출한 통계 기반 안전재고($SSstat$)로 설정하였다. $SSstat$은 동일 ERP/MES 데이터 조건에서 모든 대상 품목에 일관되게 산출 가능한 기준값이며, AI-1T 산출 결과가 기존 통계 기반 안전재고 기준 대비 과소재고 및 과잉재고 위험을 얼마나 완화하는지를 비교하기 위한 운영상 benchmark로 사용하였다. 실험 절차는 원천 데이터 전처리, 통계 기반 안전재고 산출, RAG 컨텍스트 구성, 로컬 LLM 기반 AI-1T 산출, 월말 확정 기준 성과 비교의 순서로 수행하였다.

본 연구의 정량 성과지표는 통계 기반 안전재고와 AI-1T 산출 결과를 동일한 위험 평가 기준에서 비교하기 위해 정의하였다. 여기서 통계 기반 안전재고 $SSstat$은 2장에서 산출한 수리적 베이스라인이며, 본 연구의 benchmark로 사용된다. AI-1T는 $SSstat$을 정량 앵커로 사용하고, 1-Tier RAG 컨텍스트를 로컬 LLM에 주입하여 산출된 AI 기반 최종 안전재고를 의미한다. 월말 기준 시점 m에서 품목 i의 안전재고 산출값을 S_i,m이라고 할 때, 본 연구의 penalty 함수는 과소재고 위험과 과잉재고 위험을 함께 반영하는 운영상 위험 점수로 정의하였다. 과소재고 위험은 해당 품목의 월말 기준 위험대응 필요버퍼보다 안전재고가 부족한 경우 발생하는 위험이며, 과잉재고 위험은 필요버퍼를 초과하여 불필요한 재고 부담이 발생하는 위험을 의미한다. 따라서 본 연구의 penalty 는 실제 회계 비용을 직접 추정한 값이 아니라, 동일한 기준에서 $SSstat$과 AI-1T 산출값을 비교하기 위한 normalized operational risk score로 사용하였다.

$$P_m(S)=\frac{\sum_{i=1}^{N_m}A_{i,m}\left[\frac{WU_{i,m}\max(0,B_{i,m}-S_{i,m})}{B^*_{i,m}}+\frac{WO_{i,m}\max(0,S_{i,m}-B_{i,m})}{B^*_{i,m}}\right]}{\sum_{i=1}^{N_m}A_{i,m}} \quad (6)$$

$$B^*_{i,m}=max(B_{i,m},\epsilon)$$

$$PI_m=\frac{P_m(SS_{stat,m})-P_m(AI\text{-}1T_m)}{P_m(SS_{stat,m})}\times 100 \quad (7)$$

$$DRP_m=\frac{\sum_{i=1}^{N_m}A_{i,m}min(AI\text{-}1T_{i,m},DB_{i,m})}{\sum_{i=1}^{N_m}A_{i,m}DB_{i,m}}\times 100 \quad (8)$$

수식 (6)에서 B_i,m은 월말 기준 시점에서 품목별 수요 변동성, 순수 위험 리드타임, 긴급 수주 변동성, 계절성 지수, 품목 등급 및 미납수주 분리 여부를 고려하여 정의한 위험대응 필요버퍼를 의미한다. WU_i,m과 WO_i,m은 각각 과소재고 위험과 과잉재고 위험의 상대적 중요도를 반영하는 가중치이다.

수식 (7)의 penalty 개선율은 동일한 penalty 함수에 대해 통계 기반 안전재고 $SSstat$을 적용했을 때의 위험 점수와 AI-1T 산출값을 적용했을 때의 위험 점수를 비교한 상대 개선율이다. 따라서 본 지표는 AI-1T가 통계 베이스라인 대비 과소재고 및 과잉재고 위험을 얼마나 완화하였는지를 나타낸다.

Table 5. Definition of Symbols Used in Performance Evaluation Metrics

Symbol	Definition	Unit / Type	Description
m	Monthly closing evaluation point	Month	Performance evaluation timepoint based on monthly closing
i	Item index	Index	Index representing evaluated finished product items
N_m	Number of evaluated items at month (m)	Count	Number of finished goods items evaluated at month-end closing point (m)
$S_{i,m}$	Safety stock quantity of item (i) at month (m)	Quantity	Safety stock quantity of item (i) evaluated at the month-end closing point (m)
$SS_{stat,m}$	Statistical baseline safety stock vector at month (m)	Quantity vector	Baseline safety stock vector calculated via the statistical formula at month-end, serving as the benchmark in this study
$AI\text{-}1T_m$	AI-1T safety stock vector at month (m)	Quantity vector	AI-driven final safety stock vector generated by feeding 1-Tier RAG context into the local LLM, utilizing SSstat,m as a quantitative anchor
$AI\text{-}1T_{i,m}$	AI-1T safety stock quantity of item (i) at month (m)	Quantity	Final safety stock quantity generated by the AI-1T model for item (i) at the month-end closing point (m)
$B_{i,m}$	Required risk-response buffer of item (i) at month (m)	Quantity	Required risk-mitigation buffer accounting for demand volatility, net risk lead time, urgent order volatility, seasonality index, item grade, and separation of open orders
$B^*_{i,m}$	Adjusted required buffer of item (i) at month (m)	Quantity	Adjusted required buffer modified to ensure numerical denominator stability when *Bi,m* approaches zero
ϵ	Small positive constant	Constant	An infinitesimal positive constant introduced to prevent division-by-zero errors

Table 5. Definition of Symbols Used in Performance Evaluation Metrics (Continued)

Symbol	Definition	Unit / Type	Description
$A_{i,m}$	Item importance weight of item (i) at month (m)	Weight	Evaluation weight reflecting item criticality, ABC classification grade, or operational priority
$WU_{i,m}$	Under-stock risk weight of item (i) at month (m)	Weight	Weight reflecting the relative penalty and operational severity of under-stock (stockout) risk
$WO_{i,m}$	Over-stock risk weight of item (i) at month (m)	Weight	Weight reflecting the relative penalty and holding/obsolescence severity of over-stock (excess inventory) risk
$UnderRisk_{i,m}(S)$	Normalized under-stock risk	Ratio	Normalized under-stock risk measuring the extent to which the evaluated safety stock falls short of the required risk-mitigation buffer
$OverRisk_{i,m}(S)$	Normalized over-stock risk	Ratio	Normalized over-stock risk measuring the extent to which the evaluated safety stock exceeds the required risk-mitigation buffer
$P_{m(S)}$	Operational risk penalty at month (m)	Score	Normalized operational risk penalty score synthesizing both under-stock and over-stock risks for safety stock vector (S) at month-end closing point (m)
$P_m(SS_{stat,m})$	penalty of statistical baseline	Score	Operational risk penalty score incurred when adopting the statistical baseline safety stock (SSstat,m)
$P_m(AI1T_m)$	penalty of AI-1T output	Score	Operational risk penalty score incurred when adopting the AI-1T safety stock output (AI-1Tm)
PI_m	penalty improvement at month (m)	%	Percentage reduction (improvement rate) in operational risk penalty achieved by AI-1T relative to the statistical baseline
$DB_{i,m}$	Required delivery-response buffer of item (i) at month (m)	Quantity	Baseline required buffer necessary to absorb delivery lead-time risks for item (i) at the month-end closing point
DRP_m	delivery response proxy at month (m)	%	Operational proxy metric quantifying how effectively the AI-1T safety stock covers the required delivery-risk response buffer
min(·)	Minimum operator	Function	Operator selecting the minimum of two values; prevents overestimating the coverage metric when AI-1T safety stock exceeds the required buffer
max(·)	Maximum operator	Function	Operator selecting the maximum of two values; ensures non-negativity when under-stock or over-stock risks are absent

수식 (8)의 delivery response proxy는 실제 납품률 또는 납기 준수율을 직접 대체하는 지표가 아니라, 월말 확정 기준 안전재고 조정값이 납품 리스크 대응 방향으로 얼마나 기여할 수 있는지를 평가하기 위한 운영상 보조지표이다. 본 연구에서는 이 두 지표를 이용하여 월말 확정 기준 5개 시점에서 통계 베이스라인 대비 AI-1T 산출 결과의 개선 방향을 비교하였다.

성과 비교는 2장에서 산출한 통계 기반 안전재고 SSstat을 benchmark로 설정하고, 동일한 penalty 함수에 대해 AI-1T 안전재고 산출 결과를 비교하는 방식으로 수행하였다. 비교 지표는 수식 (7)의 penalty 개선율과 수식 (8)의 delivery response proxy로 구성하였다. penalty 개선율은 통계 기반 안전재고 대비 AI-1T 산출값이 과소재고 및 과잉재고 위험을 얼마나 완화하였는지를 나타내며, delivery response proxy는 월말 확정 기준 안전재고 조정값이 납품 리스크 대응 방향으로 얼마나 기여할 수 있는지를 나타내는 운영상 보조지표로 사용하였다.

<Figure 9>는 월말 확정 기준 5개 시점에서 AI-1T 안전재고 산출 결과의 penalty 개선율과 delivery response proxy의 변화를 나타낸다. Penalty 개선율은 1월 2.59%, 2월 1.96%, 3월 2.62%, 4월 2.74%, 5월 5.08%로 나타났다. 2월에는 개선폭이 일시적으로 낮아졌으나, 모든 월말 기준 시점에서 통계 기반 안전재고 대비 양의 개선율을 보였다. Delivery response proxy는 1월 80.7%, 2월 81.5%, 3월 82.1%, 4월 82.6%, 5월 83.0%로 나타났다.

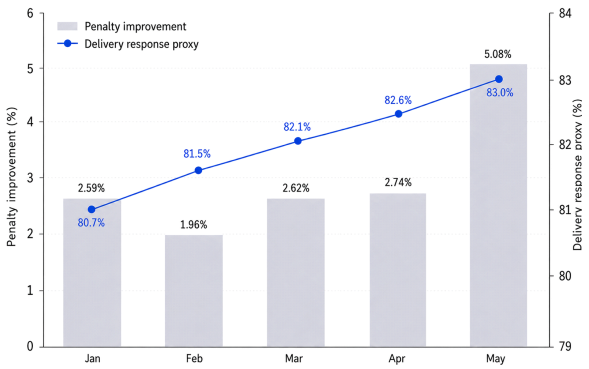


Figure 9. Monthly Closing-Based Performance Trend of AI-1T Safety Stock Outputs Compared with the Statistical Baseline

<Table 6>은 <Figure 9>의 월말 기준 분석 결과를 수치로 요

Table 6. Monthly Closing-Based Performance Summary of AI-1T Safety Stock Outputs

Month	penalty improvement (%)	delivery response proxy (%)	Interpretation
Jan	2.59	80.7	Positive penalty improvement at the first monthly closing point after cold-start screening
Feb	1.96	81.5	penalty improvement remained positive despite a temporary decrease
Mar	2.62	82.1	penalty improvement remained positive
Apr	2.74	82.6	Positive improvement was maintained under monthly closing evaluation
May	5.08	83.0	Highest penalty improvement among the five monthly closing points

Note. Penalty improvement was calculated as the relative reduction in the normalized operational risk penalty of AI-1T compared with the statistical baseline SSstat. The delivery response proxy is an operational indicator that reflects the extent to which monthly confirmed safety stock values support delivery-risk responsiveness; it does not directly represent actual delivery performance.

약한 것이다. AI-1T 산출 결과는 월별 변동을 보였으나, 모든 월말 기준 시점에서 통계 기반 안전재고 대비 양의 penalty 개선율을 보였다. 특히 5월에는 penalty 개선율이 5.08%로 가장 높게 나타났으며, delivery response proxy는 83.0%로 관측되었다. 본 연구에서 cold-start screening은 시스템 초기 구동 직후 RAG 로그와 안전재고 산출 로그가 충분히 축적되지 않은 초기 주차의 변동성을 월말 확정 기준 성과 비교에서 제외하는 절차를 의미한다. 즉, 원천 ASOF 로그 전체는 검토 대상으로 유지하되, 정량 성과 평가는 각 월의 정책 확정 기준에 해당하는 월말 ASOF 값을 중심으로 수행하였다.

이상의 결과는 제안 프로그램이 단순히 재고 총량을 증감시키는 방식이 아니라, 품목별 ABC 등급, 계절성, 수요 추세, 긴급오더 변동성, 미납수주 정보를 함께 고려하여 과소재고와 과잉재고 위험을 균형 있게 조정하는 방향으로 작동했음을 보여준다. 또한 주차별 ASOF 산출 로그와 월말 확정값을 분리함으로써, 안전재고 정책값의 과도한 변동이 생산계획으로 전이되는 문제를 완화할 수 있다.

5. 결 론

5.1 연구의 요약 및 실무적 기여

본 연구는 다품종 소량생산 환경에서 통계 기반 안전재고 산출값에 로컬 LLM-RAG 기반 현장 맥락 추론을 결합하는 완제품 안전재고 산출 프레임워크를 제안하였다. 개발 프로그램은 2026년 1월부터 5월까지 총 22개 ASOF 시점에서 자동차 프레스부품 완제품 180개 품목에 대한 안전재고 산출 로그를 생성하였으며, 안전재고가 생산계획 및 MRP와 연동되는 정책성 기준값이라는 점을 고려하여 월말 확정 기준 5개 시점을 중심으로 정량 성과를 분석하였다. 분석 결과, AI-1T 안전재고 산출 결과는 모든 월말 기준 시점에서 통계 기반 안전재고 대비 양의 penalty 개선율을 보였고, delivery response proxy는 1월 80.7%에서 5월 83.0%로 높아졌다. 이는 제안 프레임워크가 품목별 ABC 등급, 계절성, 수요 추세, 긴급오더 변동성, 미납수주 정보를 함께 고려하여 과소재고와 과잉재고 위험을 균형 있게 조정하는 방향으로 작동할 수 있음을 보여준다.

특히 실무적 관점에서 본 연구는, 보안 제약과 자본력이 부족한 중소/중견 제조 기업(SME)에 망분리 환경에서도 즉시 도입 가능한 저비용 · 고효율의 AI 기반 안전재고 의사결정 지원 가이드라인을 제시했다는 점에 가장 큰 의미가 있다. 고가의 클라우드 인프라나 대규모 데이터 파이프라인 없이도, 현장 데이터의 보안을 완벽히 유지하며 기존 시스템(APS)과 즉각적으로 연동할 수 있는 실증적이고 현실적인 대안을 제시한 것이다.

5.2 학술적 의의

학술적 측면에서 본 연구는 공급망 관리(SCM) 분야의 AI 적용 패러다임을 단순한 기계적 수치 최적화에서 '상황 인지형 추론(Context-Aware Reasoning)'으로 전환했다는 점에서 다음과 같은 세 가지 학술적 의의를 지닌다.

첫째, 통계 기반 안전재고 산출값과 RAG 기반 현장 맥락 추론을 결합한 설명 가능한 안전재고 산출 구조를 제시하였다. 본 연구는 수기 입력 지연, 인적 오류, 긴급 수주, 미납수주 등 제조 현장에서 발생하는 데이터 불확실성을 SQL 기반 전처리와 통계적 안전재고 산출 과정을 통해 1차적으로 정리하고, 그 결과를 로컬 LLM 추론의 정량 앵커로 활용하였다. 이후 ABC 등급, 계절성, 수요 추세, 긴급오더 변동성, 미납수주 정보를 RAG 컨텍스트로 주입함으로써, 통계 모델만으로는 반영하기 어려운 품목별 운영 맥락을 안전재고 조정 근거에 결합하였다.

둘째, 미납수주와 안전재고 버퍼의 역할을 분리하는 운영 규칙을 안전재고 산출 과정에 반영하였다. 본 연구에서는 확정된 미납수주를 미래 불확실성에 대비하기 위한 안전재고 버퍼와 구분하여 처리함으로써, 동일 수요가 생산계획과 안전재고에 중복 반영될 가능성을 줄이고자 하였다. 또한 주차별 ASOF 산출 로그를 모두 정책 확정값으로 사용하지 않고 월말 확정 기준으로 성과를 평가함으로써, 단기 수요 변동 관찰과 안전재고 정책값 확정을 분리하는 절차를 제시하였다. 이는 안전재고 기준값의 과도한 변동이 생산계획과 MRP로 전이되는 문제를 완화하기 위한 운영적 안정화 장치로 해석할 수 있다.

셋째, 안전재고 산출 결과와 판단 근거를 함께 제시하는 설명 가능한 의사결정 구조를 구현하였다. 이때 LLM의 역할은

고정된 임계값 규칙을 단순히 반복 적용하는 것이 아니라, 통계 기반 안전재고를 기준으로 여러 운영 신호를 함께 검토하고, 그 결과를 현장 실무자가 이해할 수 있는 판단 근거로 변환하는 데 있다. 따라서 본 연구의 학술적 의의는 LLM을 독립적인 최적화 모델로 제시한 데 있는 것이 아니라, 수리적 기준값, 운영 휴리스틱, RAG 기반 맥락정보, HITL 검토 절차를 결합한 설명 가능한 안전재고 의사결정 구조를 제시한 데 있다. 제안 프레임워크는 최종 안전재고 수량만을 출력하는 방식이 아니라, ABC 등급, 계절성, 수요 추세, 긴급오더 변동성, 미납수주 처리 여부 등 주요 판단 요소를 함께 기록한다. 이러한 RAG 기반 맥락정보와 AI-1T 판단근거의 대표 사례는 <Table 7>에 제시하였다. <Table 7>은 AI-1T가 단일 계절성 임계값만을 적용한 것이 아니라, 품목 등급, 계절성 상태, 수요 추세, 위험 태그 및 guardrail 적용 여부 등 복수의 운영 신호를 함께 참조하여 판단근거를 구성하였음을 보여준다. 이를 통해 실무자는 AI-1T 산출 결과가 어떤 운영 맥락을 근거로 상향, 하향 또는

Table 7. Representative 1-Tier RAG Context Signals and AI-1T Decision Rationales

Case Type	Item Code	ABC	RAG Context Signals	AI-1T Direction	RAG-based Decision Rationale
Peak-season A-grade increase	1OM-DD1-015	A	SI: Peak season Trend: Mixed / Stable Risk: LOW_COVERAGE	Upward adjustment	Grade A, peak season, and LOW_COVERAGE signals were committed to REF_MEMORY, logged as rationale to prioritize the current feature vector for upward adjustment in subsequent inferences.
Conflicting signal: off-season but rising demand	1MF-BK1-013	A	SI: Off-peak season Trend: Recent increase Risk: LOW_COVERAGE	Upward adjustment	Despite the off-peak signal, Grade A, recent upward trend, and LOW_COVERAGE were co-identified, illustrating multi-contextual reasoning rather than rigid adherence to a single SI rule.
Off-season B-grade decrease	1OM-KT1-004	B	SI: Off-peak season Trend: Recent decrease Risk: GUARDRAIL_APPLIED	Downward adjustment	Grade B, off-peak season, recent downward trend, and active guardrail signals were co-recorded, serving as rationale for downward adjustment to mitigate excess inventory.
Fallback / statistical baseline maintained	1MF-HK2-001-1	A	SI: Normal season Trend: Insufficient history Risk: SHORTAGE	Maintain baseline	Even with data sparsity (insufficient history) and a SHORTAGE signal, the model falls back to maintaining the statistical baseline rather than executing volatile AI adjustments.
Guardrail-based reduction	1MF-KT1-002	B	SI: Normal season Trend: Recent decrease Risk: GUARDRAIL_APPLIED	Downward adjustment	Rationales suppressing excessive safety stock inflation based on recent downward demand trends and guardrail triggers were committed to REF_MEMORY.
Peak-season B-grade increase	1MF-SP-020200	B	SI: Peak season Trend: Mixed / Stable Risk: LOW_COVERAGE	Upward adjustment	Demonstrates that upward adjustment can be justified even for Grade B items when peak season and LOW_COVERAGE signals co-occur.
Conflicting signal: A-grade but decreasing trend	1OM-DD1-003	A	SI: Normal season Trend: Recent decrease Risk: LOW_COVERAGE	Upward adjustment	A conflicting context case where Grade A criticality and LOW_COVERAGE take precedence over a recent downward trend, justifying defensive upward adjustment.
B-grade with rising demand	1MF-DW1-001	B	SI: Normal season Trend: Recent increase Risk: LOW_COVERAGE	Upward adjustment	Illustrates that co-occurring recent demand increases and LOW_COVERAGE signals serve as legitimate rationale for upward adjustment even for Grade B items.
Stable context with low-coverage defense	1MF-DD1-002	A	SI: Normal season Trend: Mixed / Stable Risk: LOW_COVERAGE	Upward adjustment	Under normal seasonality and stable demand, Grade A and LOW_COVERAGE signals jointly warrant upward adjustment from a stockout-prevention perspective.

유지되었는지 검토할 수 있으며, 필요시 Human-in-the-Loop 절차를 통해 현장 판단을 보완할 수 있다. 이러한 구조는 제조 현장에서 AI 기반 안전재고 산출 결과의 해석 가능성과 검토 가능성을 높이는 데 기여한다.

5.3 한계점 및 향후 연구 과제

본 연구는 로컬 LLM-RAG 기반 안전재고 산출 프레임워크를 실제 제조 데이터와 연동하여 구현하고, 통계 기반 안전재고 대비 AI-1T 산출 결과를 월말 확정 기준으로 비교하였다는 점에서 의의가 있다. 그러나 본 연구의 정량 평가는 단일 자동차 부품 벤더의 완제품 안전재고 산출 사례를 중심으로 수행되었으며, 평가 대상은 ABC 매출 기여도 분석에 따라 누적 매출의 약 90%를 구성하는 180개 핵심 품목으로 한정하였다. 이러한 품목 선정은 안전재고 정책 변화가 운영성과에 미치는 영향이 큰 주요 품목군을 우선 검증하기 위한 실무적 범위 설정이라는 점에서 타당성이 있으나, 저매출 품목, 간헐수요 품목, 장기 꼬리(long-tail) 품목의 수요 특성까지 충분히 대표한다고 보기는 어렵다. 또한 본 연구는 DRL 등 다른 AI 기반 재고 최적화 모델과의 직접 성능 비교를 포함하지 않았다. 따라서 향후 연구에서는 동일한 데이터 조건에서 통계 기반 모델, 규칙 기반 휴리스틱, 규칙엔진, DRL 또는 기타 AI 기반 모델을 함께 비교하여 각 접근법의 적용 범위와 성능 차이를 보다 체계적으로 검토할 필요가 있다. 특히 SI 임계값, ABC 등급, 수요 추세, UVI, 미납수주 분리 규칙으로 구성된 결정론적 규칙 엔진과 AI-1T 산출 구조를 비교함으로써, LLM 도입이 정량 성과뿐 아니라 판단근거의 검토 가능성, 예외상황 처리, HITL 피드백 재사용 측면에서 갖는 한계효용을 추가로 검증할 필요가 있다. 또한 본 연구에서는 안전재고가 생산계획 및 MRP와 연동되는 정책성 기준값이라는 점을 고려하여, 주차별 산출 로그 전체를 확정값으로 사용하지 않고 월말 확정 기준 5개 시점을 중심으로 성과를 분석하였다. 분석 결과, AI-1T 산출 결과는 모든 월말 기준 시점에서 통계 기반 안전재고 대비 양의 penalty 개선율을 보였고, delivery response proxy도 1월 80.7%에서 5월 83.0%로 높아졌다. 향후에는 1년 이상 장기 운영 데이터를 축적하여 월말 기준 안전재고 조정이 실제 납품 이력, 생산계획 실행 결과, 재고회전율 및 과잉재고 비용에 미치는 영향을 후속 연구에서 검증할 예정이다.

마지막으로, 본 연구의 범위는 완제품 안전재고 산출 모듈에 한정되어 있다. 향후 연구에서는 안전재고 산출 결과를 S&OP, MPS, MRP, CRP 및 생산 스케줄링 모듈과 연계하여, 안전재고 정책값 변화가 생산계획과 자원부하에 미치는 영향을 통합적으로 분석할 필요가 있다. 또한 HITL 피드백은 모델 파라미터를 직접 갱신하는 fine-tuning이 아니라 RAG 컨텍스트로 재사용되는 지식 축적 구조이므로, 향후에는 피드백 데이터의 품질 관리, 재사용 주기, 의사결정 로그의 검증 절차를 고도화하는 방향으로 확장할 수 있다.

참고문헌

Baryannis, G., Validi, S., Dani, S., and Antoniou, G. (2019), Supply Chain Risk Management and Artificial Intelligence: State of the Art and Future Research Directions, *International Journal of Production Research*, **57**(7), 2179-2202.

Bertsimas, D. and Thiele, A. (2006), A Robust Optimization Approach to Inventory Theory, *Operations Research*, **54**(1), 150-168.

Gijsbrechts, J., Boute, R. N., Van Mieghem, J. A., and Zhang, D. J. (2022), 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.

Graves, S. C. and Willems, S. P. (2000), Optimizing Strategic Safety Stock Placement in Supply Chains, *Manufacturing & Service Operations Management*, **2**(1), 68-83.

Jackson, I., Ivanov, D., Dolgui, A., and Namdar, J. (2024), Generative Artificial Intelligence in Supply Chain and Operations Management: A Capability-Based Framework for Analysis and Implementation, *International Journal of Production Research*, **62**(17), 6120-6145.

Jackson, I., Saenz, M. J., Ivanov, D., and Ma, B. J. (2026), Supply Chain Mapping through Retrieval-Augmented Generation: Applications to the Electronics Industry, *Journal of the Operational Research Society*, **77**(7), 1-21.

Kosasih, E. E., Papadakis, E., Baryannis, G., and Brintrup, A. (2024), A Review of Explainable Artificial Intelligence in Supply Chain Management Using Neurosymbolic Approaches, *International Journal of Production Research*, **62**(4), 1510-1540.

Oroojlooyjadid, A., Nazari, M., Snyder, L. V., and Takac, M. (2022), A Deep Q-Network for the Beer Game: Deep Reinforcement Learning for Inventory Optimization, *Manufacturing & Service Operations Management*, **24**(1), 285-304.

Priore, P., Ponte, B., Rosillo, R., and de la Fuente, D. (2019), Applying Machine Learning to the Dynamic Selection of Replenishment Policies in Fast-Changing Supply Chain Environments, *International Journal of Production Research*, **57**(11), 3663-3677.

Srivastava, S. K., Routray, S., Bag, S., Gupta, S., and Zhang, J. Z. (2024), Exploring the Potential of Large Language Models in Supply Chain Management: A Study Using Big Data, *Journal of Global Information Management*, **32**(1).

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., and Fischl, M. (2021), Artificial Intelligence in Supply Chain Management: A Systematic Literature Review, *Journal of Business Research*, **122**, 502-517.

저자소개

조형우: 인천대학교 산업경영공학과에서 2023년 석사학위를 취득하고 인천대학교에서 산업경영공학과 박사과정에 재학 중이다. 연구 분야는 스마트제조, APS + AI, 제조 데이터 최적화(Optimization) 이다.

유우식: 1986년 서울대학교에서 학사학위를 취득하고 한국과학기술원에서 1988년 석사, 1992년 박사학위를 취득하고 1996년부터 인천대학교 산업경영공학과에서 교수로 재직 중이다. 연구 분야는 스마트 제조, CAD/CAM이다.