ExplainableMRP via Trace Visibility: Faithful Natural-Language Explanations for Deterministic Planning

Anonymous ACL submission

Abstract

Material Requirements Planning (MRP) is the cornerstone of manufacturing, yet it often operates as an opaque black box. When shortages occur, planners are left manually sifting through thousands of transactional rows to diagnose the root cause, as current systems provide numerical outcomes without the "why." While Large Language Models (LLMs) promise to bridge this gap, applying LLMs to this deterministic domain risks misinterpretation and hallucination. We argue that reliable explainability in MRP cannot be deduced post hoc. It must be traced. We introduce EXPLAINABLEMRP, a framework that transforms the LLM from an independent reasoner into a faithful narrator of the execution trace. By combining schema normalization, explicit trace exposure, and evidence-constrained generation, we demonstrate that our approach resolves input ambiguity and suppresses hallucination to negligible levels. Code will be released upon publication.

1 Introduction

Material Requirements Planning (MRP) is a foundational mechanism in manufacturing and supply-chain operations (Ptak, 2011). Although MRP computes material availability through a fully deterministic netting process with a single correct outcome, existing systems typically expose only final results, offering little insight into how shortages arise (Fransoo et al., 2010). As a result, the literature lacks a clear and operational account of how deterministic planning computations should be explained to human planners (Madathil et al., 2025).

Recent work has explored Large Language Models (LLMs) as a natural interface for explaining complex system behavior (Zhao et al., 2024). However, applying LLMs to industrial MRP data introduces a fundamental mismatch: MRP is governed by exact arithmetic rules, whereas LLMs rely on probabilistic inference. When explanation is treated as a purely linguistic task detached from the underlying computation, deterministic facts are reduced to stochastic guesses, leading to input misinterpretation and post hoc rationalization (Turpin et al., 2023).

We argue that explainability in deterministic planning systems must be grounded in *trace visibility*. Because MRP outcomes are uniquely determined by a sequence of inventory updates and netting operations, faithful explanations must align with this execution trace; retrospectively inferred explanations are therefore unverifiable and potentially misleading (Jacovi and Goldberg, 2020). From this perspective, explanation is not free-form language generation but a constrained narration problem in which every claim must correspond to a verifiable computational step.

Based on this view, we propose EXPLAINABLEMRP, a trace-centered framework for generating faithful natural-language explanations of MRP outcomes. The framework structures explanation around schema normalization of heterogeneous inputs, explicit exposure of the deterministic MRP trace, and evidence-constrained reasoning.

We evaluate EXPLAINABLEMRP through three research questions on input interpretation, trace-aligned explanation quality, and hallucination control. Our contributions are:

- We formalize explainability in deterministic planning systems as *trace visibility*, reframing explanation as alignment with deterministic computation rather than post hoc language generation.
- We propose and empirically validate EXPLAINABLEMRP, a trace-centered framework that combines schema normalization, deterministic trace exposure, and evidence-constrained reasoning to generate faithful explanations and prevent hallucination.

2 Related Work

2.1 MRP and Material Planning Systems

MRP is one of the oldest and most standardized planning mechanisms in manufacturing and supply chain operations (Orlicky, 1974; Mabert, 2007). Introduced in the 1960s, MRP follows a deterministic netting–offset procedure in which projected inventory, incoming supply, and planned orders are combined to compute material availability (Jacobs et al., 2011). Because the logic is entirely rule-based, the output is uniquely determined once the inputs are fixed (Hopp and Spearman, 2011).

Despite widespread use, modern studies highlight structural limitations in traditional MRP. The method assumes stable lead times, fixed usage coefficients, and batch recalculation cycles, making the system brittle under volatility (Hopp and Spearman, 2011). When disruptions occur, such as demand surges, supplier delays, or unplanned consumption, the deterministic offset mechanism can amplify small deviations, causing shortages to emerge abruptly (Steele, 1975).

Various extensions, including Demand-Driven MRP (DDMRP), Advanced Planning and Scheduling (APS) systems, and hybrid optimization frameworks, have been proposed to improve responsiveness (Ptak and Smith, 2016; Stadtler, 2005). However, these approaches typically preserve the core MRP netting logic. Commercial platforms such as SAP or Oracle provide pegging reports that link demand to supply orders, but they do not expose the temporal evolution of inventory levels or the causal sequence of netting events that lead to a shortage (Elbahri et al., 2019). As a result, the internal decision process of MRP remains opaque, limiting planners' ability to diagnose why shortages occur.

2.2 AI Models Used in Supply Chain Management

Recent advances in Supply Chain Management (SCM) have led to extensive adoption of machine learning and AI across forecasting, inventory control, scheduling, and network optimization (Ferreira and Reis, 2023). In particular, deep learning and reinforcement learning approaches have shown strong performance under uncertainty in tasks such as demand forecasting and adaptive replenishment (Salinas et al., 2020; Oroojlooyjadid et al., 2022).

However, these AI-driven methods remain largely peripheral to the core MRP engine. They are typically applied to upstream or downstream decision layers, such as forecasting or policy optimization, while treating MRP as a fixed, rule-based backend (Gyulai et al., 2017).

As a result, material-level deterministic planning remains largely unexplored: existing approaches neither integrate AI into the deterministic netting-and-offset computation nor explain the internal planning logic of MRP (Madathil et al., 2025).

2.3 Explainability of NLP and LLM

Explainability has become an essential requirement in decision-support systems, particularly in high-stakes operational environments (Gunning et al., 2019). In supply chains, prior work on explainability has primarily focused on interpreting predictive models, such as demand forecasts, risk scores, or anomaly detections, using model-agnostic techniques such as SHAP or LIME (Ferreira and Reis, 2023; Wang et al., 2025; Ribeiro et al., 2016; Lundberg and Lee, 2017).

In parallel, NLP research has emphasized faithful explanation generation, highlighting the limitations of post hoc rationalization and the risk of hallucination in LLMs (Jacovi and Goldberg, 2020; Turpin et al., 2023). Techniques such as chain-of-thought prompting and tool-augmented reasoning aim to align generated explanations with verifiable computational steps (Wei et al., 2022; Schick et al., 2023).

However, most explainable NLP research targets classification or mathematical reasoning benchmarks, rather than deterministic operational algorithms (Cobbe et al., 2021). In material planning, the explainability challenge is fundamentally different: users must understand which specific sequence of demand, supply, and lead-time events caused a shortage. This requirement demands transparency into a deterministic planning process, not interpretation of a learned statistical model. Consequently, existing explainability techniques do not directly address the needs of MRP.

2.4 LLM-based or Agent-based Reasoning

Recent work on LLM-based reasoning explores multi-agent architectures that decompose complex tasks into specialized components such as simulation, planning, explanation, and verification (Wu et al., 2024; Du et al., 2023). These approaches are particularly suitable for domains where deterministic computations must be translated into human-understandable explanations.

Despite their relevance, such architectures have not been applied to material planning. Existing LLM-based SCM studies focus on qualitative decision support or document-level reasoning, without interfacing with the MRP engine or aligning natural-language explanations with the underlying netting computation (Atalan, 2025). As a result, there is currently no framework that combines deterministic MRP logic with verifiable, evidence-grounded LLM explanations.

2.5 Gap Summary

Across these domains, two consistent gaps emerge. First, although AI techniques are widely explored in forecasting, inventory optimization, and scheduling, they have not been incorporated into the deterministic netting-and-offset mechanism of MRP. The core material planning computation remains rule-based, opaque, and isolated from modern AI tooling. Second, existing explainability methods, both in SCM and NLP, do not address the challenge of explaining deterministic planning computations. Planners require a transparent causal chain identifying which demand events, supply events, and lead-time offsets contributed to a specific shortage or recommendation, yet current approaches offer no mechanism for aligning explanations with the underlying deterministic MRP execution trace. These gaps jointly motivate the need for a framework that exposes the full deterministic MRP trace and supports verifiable, evidence-grounded natural-language explanations, as formalized in Section 3.

3 Problem Definition

MRP computes future inventory positions through a deterministic netting-and-offset procedure. Given (i) initial on-hand stock S_0, (ii) a set of dated demand events $\mathcal{D} = \{(t_i, d_i)\}$, (iii) scheduled purchase orders $\mathcal{P} = \{(t_j, p_j)\}$, and (iv) a fixed lead time L, the projected inventory at time t is defined as:

$$S(t) = S_0 + \sum_{t_j \le t} p_j - \sum_{t_i \le t} d_i \qquad (1)$$

The first shortage date is:

$$t^* = \min\{\, t \in \mathcal{H} : S(t) < 0 \,\} \qquad (2)$$

where $\mathcal{H}$ denotes the discrete planning horizon. If t^* exists, the planned replenishment arrival date is $t_{\text{arrival}} = t^* - L$. As formalized in Eq. 1 and Eq. 2, the MRP outcome is uniquely determined by the cumulative demand and supply events. However, commercial MRP systems typically expose only the endpoints $(t^*, t_{\text{arrival}})$, without revealing the intermediate inventory evolution that leads to the shortage.

3.1 Opacity of Deterministic Netting

Because intermediate netting states are not exposed, planners cannot identify which specific demand or supply event caused inventory to become negative, nor how inventory evolved prior to the shortage. As a result, diagnosis and corrective actions are limited to generic responses, such as expediting replenishment, without understanding the underlying cause.

To support causal analysis, we denote the full deterministic netting trace as:

$$\mathcal{T} = \{(t, S(t), D_t, P_t) \mid t \in \mathcal{H}\},$$

where $S(t)$ is the projected inventory at time t, $D_t = \sum_{t_i \le t} d_i$ is cumulative demand, and $P_t = \sum_{t_j \le t} p_j$ is cumulative supply up to time t. Each element of the trace records how cumulative demand and supply events incrementally update the projected inventory over time, until a shortage is first observed. The absence of this trace constitutes the **MRP Explainability Gap**. A concrete example of a deterministic MRP trace is provided in Appendix B.

3.2 Formal Problem Definition

We define the EXPLAINABLEMRP framework as a tuple $\langle X, \mathcal{T}, f_{\text{ev}}, g_\theta \rangle$, where X denotes the input tables (Bill of Materials (BOM), Plan, Stock, Open Purchase Order (OpenPO)), $\mathcal{T}$ is the deterministic MRP trace, $f_{\text{ev}} : \mathcal{T} \to E$ extracts structured evidence, and g_θ generates a natural-language explanation. The extracted evidence set E includes trace-validated facts such as the shortage date, cumulative demand and supply, inventory trajectories, and causally relevant trace events e.g., demand spikes or supply gaps.

The objective is to generate an explanation y that is both faithful to the extracted evidence and aligned with the deterministic trace. Figure 1 illustrates the overall workflow from trace generation to evidence extraction and constrained explanation generation.

While our primary focus is faithful explanation of baseline MRP outcomes, the proposed framework also supports optional what-if analy-

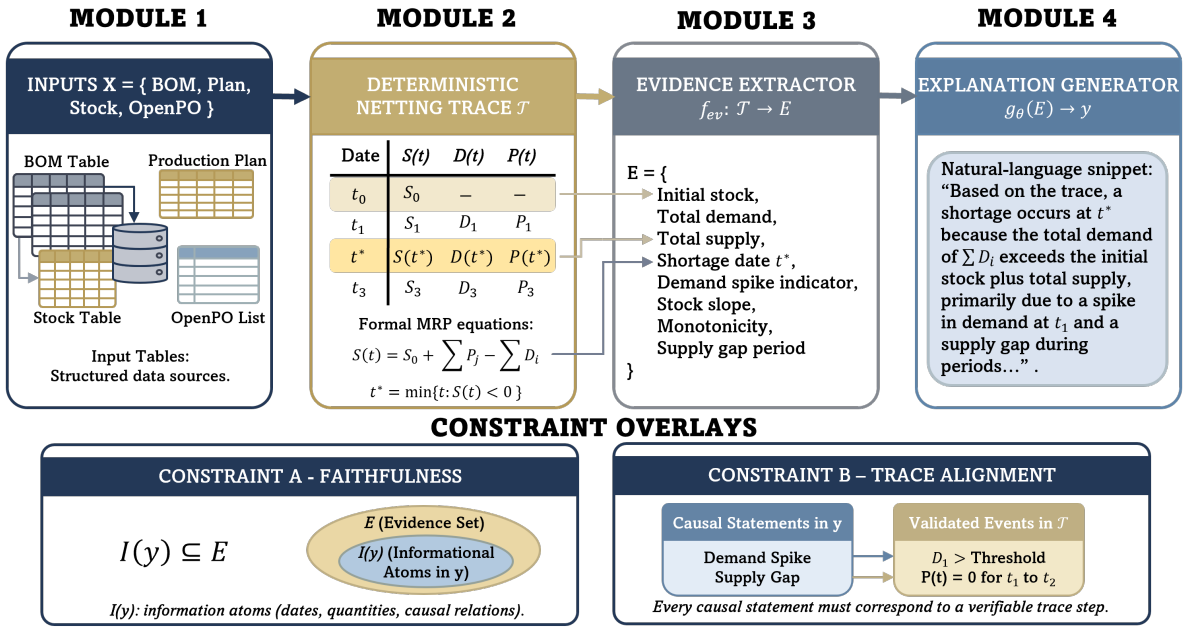

Figure 1: **Conceptual framework of ExplainableMRP.** Deterministic MRP computation is transformed into structured evidence that constrains natural-language explanations under explicit faithfulness and trace-alignment requirements.

sis by recomputing deterministic traces under user-specified input perturbations.

Faithfulness Let $\mathcal{I}(y)$ denote the set of informational atoms extracted from the explanation y, such as dates, quantities, and causal links. We require:

$$\mathcal{I}(y) \subseteq E,$$

ensuring that all assertions in y are grounded in the deterministic evidence.

Trace Alignment Every causal statement in y must correspond to a verifiable step in the deterministic trace $\mathcal{T}$, ensuring that explanations reflect the actual netting dynamics rather than post hoc rationalizations.

Trace Exposure Requirement The system must expose the full trace $\mathcal{T}$, not only the derived endpoints $(t^*, t_{\text{arrival}})$, to support human verification.

4 Method

The task formalized in Section 3 requires a system that can (1) transform heterogeneous enterprise spreadsheets into a consistent schema, (2) execute the deterministic MRP computation to obtain the full trace $\mathcal{T}_m$, (3) derive the structured evidence E_m, and (4) generate a faithful explanation y_m aligned with the trace. ExplainableMRP operationalizes this task through a modular agent-based architecture composed of three computational agents and an interactive frontend.

4.1 System Overview

Concretely, the system realizes the transformation

$$X_m \to \mathcal{T}_m \to E_m \to y_m,$$

where X_m denotes the raw data associated with material m.

Each stage of this transformation is handled by a specialized agent: DataAgent normalizes heterogeneous enterprise spreadsheets into a consistent schema, SimulationAgent executes the deterministic MRP procedure to generate the full trace $\mathcal{T}_m$ and structured evidence E_m, and ReasoningAgent generates a natural-language explanation y_m constrained by the faithfulness and trace-alignment requirements.

4.2 Definition of an Agent

An *agent* is a modular computation unit that performs a well-defined subtask in the ExplainableMRP pipeline and produces a structured output along with interpretable logs. Agents collectively implement:

$$(\mathcal{T}_m, E_m, y_m) = \text{ExplainableMRP}(X_m).$$

4.3 DataAgent: Normalizing Enterprise Spreadsheets

DataAgent normalizes heterogeneous BOM, Plan, Stock, and OpenPO spreadsheets into a consistent schema required for deterministic MRP computation, including unit standardization and temporal alignment.

4.4 SimulationAgent: Baseline and Scenario Execution

SimulationAgent supports deterministic MRP execution under both baseline and scenario-modified inputs, producing full execution traces and trace-derived evidence.

Baseline Execution. Given normalized inputs X_m, SimulationAgent executes the deterministic MRP procedure to generate the complete trace $\mathcal{T}_m$. In contrast to commercial MRP systems that expose only final outcomes, all intermediate inventory states are retained. The trace is transformed into structured evidence E_m via the deterministic evidence extractor f_{ev}, implemented as a rule-based symbolic engine that queries predefined events and quantities from $\mathcal{T}_m$ without additional language model inference.

Scenario Execution. For user-specified scenarios, SimulationAgent deterministically modifies X_m (e.g., lead-time shifts, demand scaling, or purchase order suppression) and recomputes the corresponding trace $\mathcal{T}_m^{(\text{scen})}$ and evidence $E_m^{(\text{scen})}$. Baseline and scenario outcomes are reported side by side, enabling interpretable what-if analysis within the same execution framework.

4.5 ReasoningAgent: Generating Faithful Explanations

Given the evidence E_m, ReasoningAgent produces the explanation y_m under the faithfulness constraint: $\mathcal{I}(y_m) \subseteq E_m$. To operationalize this constraint, we extract informational atoms $\mathcal{I}(y_m)$, including explicit dates, numerical quantities, and causal relations, from the generated text using pattern-based parsing and predefined linguistic templates.

This extraction process serves a dual purpose: it verifies factual assertions and computes diagnostic sets $\texttt{EvidenceUsed}(y_m)$ and $\texttt{AlignmentIssues}(y_m)$. $\texttt{EvidenceUsed}(y_m)$ is determined by mapping informational atoms to their canonical representations in E_m, while any assertion in $\mathcal{I}(y_m)$ that cannot be grounded in the deterministic evidence set is flagged as an `AlignmentIssue`. This mechanism reveals the degree of grounding and ensures that the natural-language output remains a faithful reflection of the underlying planning logic.

4.6 Frontend Integration

The frontend enables interactive editing of BOM, Plan, Stock, and OpenPO tables. Any modification triggers recomputation of $\mathcal{T}_m$, E_m, and y_m, supporting real-time what-if analysis and reproducible planning.

5 Experiments

5.1 Dataset Description

Experiments were conducted on anonymized operational data derived from an actual pharmaceutical manufacturing environment. All item identifiers and numerical values were perturbed to ensure confidentiality, while preserving the structural relationships and temporal dependencies required for deterministic MRP execution.

The dataset includes multi-level BOMs spanning multiple production steps, production plans over a 16-month horizon, lot-level inventory records, and open purchase orders. Notably, the OpenPO table contains unstructured natural-language comments, such as supplier delays or logistics issues, which are retained to evaluate the integration of deterministic MRP data with auxiliary textual signals.

5.2 Baselines

To isolate the contribution of each component in EXPLAINABLEMRP, we compare the proposed system against three baselines.

Raw-LLM. An LLM directly generates explanations from raw spreadsheets without schema normalization or preprocessing. This baseline evaluates the LLM's inherent ability to interpret heterogeneous industrial data.

No-Trace (Traditional MRP Style). This setting exposes only aggregate outcomes, namely the shortage date (t^*) and planned arrival date (t_{arrival}), without revealing the intermediate deterministic computation trace $\mathcal{T}$.

Unconstrained Trace. The full deterministic trace $\mathcal{T}$ is provided to the LLM, but explanations are generated without explicit evidence-grounding or faithfulness constraints. This baseline isolates

the effect of trace visibility from that of evidence constrained reasoning.

5.3 Evaluation Metrics

We evaluate explanation quality using four complementary metrics capturing input interpretation accuracy, causal alignment, and faithfulness. All metrics are computed via a deterministic rule-based parser (Appendix D.1) to ensure consistent and reproducible evaluation.

Input Misinterpretation Rate (IMR). IMR measures the proportion of input fields incorrectly interpreted by the LLM:

$$\text{IMR} = \frac{\text{Incorrectly Interpreted Fields}}{\text{Total Input Fields}}.$$

Trace Alignment Score (TAS). TAS is a human expert evaluation on a three-point Likert scale that assesses whether causal claims in an explanation are consistent with the deterministic netting logic and inventory trajectories in the MRP trace $\mathcal{T}$. It evaluates three dimensions: faithfulness (consistency with the deterministic trace), specificity (explicit identification of concrete causal events and quantities), and clarity (overall readability and coherence).

Evidence Coverage (EC). EC measures the fraction of extracted evidence E incorporated into the explanation y:

$$\text{EC} = \frac{|\text{EvidenceUsed}(y)|}{|E|}.$$

Hallucination Rate (HR). HR quantifies the proportion of statements in y not grounded in either the deterministic trace $\mathcal{T}$ or evidence set E:

$$\text{HR} = \frac{|\text{Hallucinated Statements}|}{|\text{Total Statements}|}.$$

5.4 RQ1: Does the DataAgent Reduce Input Misinterpretation?

This experiment evaluates whether the proposed *DataAgent* reduces the *IMR* of LLMs when processing heterogeneous industrial spreadsheets.

Experimental Setup. We compare two conditions: **Raw-LLM**, which directly processes raw CSV/Excel files with heterogeneous column names, and **ExplainableMRP**, in which all input tables are first normalized by the DataAgent into a canonical MRP schema prior to LLM interaction. Experiments are conducted on three core MRP tables, namely BOM, Stock, and OpenPO. For the Raw-LLM condition, the model is provided with explicit natural-language descriptions of each column header and instructed to map them to canonical MRP fields, following common prompt-engineering practices for schema alignment.

Table	Raw-LLM	ExplainableMRP
BOM	25.0%	0.0%
Stock	16.7%	0.0%
OpenPO	16.7%	0.0%
Average	19.5%	0.0%

Table 1: IMR comparison between Raw-LLM and EXPLAINABLEMRP.

Results. Table 1 reports the IMR for each table. When operating on raw spreadsheets, the LLM frequently misinterprets non-quantitative fields as MRP variables, resulting in an average IMR of 19.5%. After schema normalization, all columns are deterministically aligned with the canonical schema. As a result, we observed no instances of input misinterpretation in the evaluated dataset, indicating that enforcing a canonical schema effectively eliminates ambiguity in column semantics.

5.5 RQ2: Does Trace Visibility Improve Explanation Quality?

This research question examines whether access to a deterministic MRP trace improves the quality of natural-language explanations generated by LLMs. We hypothesize that trace visibility primarily enhances causal grounding and concreteness rather than surface-level readability.

Experimental Setup. For each material, we generate paired explanations under identical MRP outcomes, differing only in whether selected trace entries are provided. The **No-Trace** condition exposes only aggregate outcome summaries, whereas the **With-Trace** condition additionally provides selected trace entries capturing key inventory transitions. Experiments are conducted on eight materials, yielding sixteen paired explanations.

Human Evaluation. A blinded paired evaluation is conducted by three domain-informed evaluators. Each explanation is independently rated along three dimensions: *faithfulness*, *specificity*, and *clarity*, and evaluators indicate an overall preference between the paired explanations.

Metric	No-Trace	With-Trace
Faithfulness	2.29	**2.38**
Specificity	2.38	**2.38**
Clarity	**2.63**	2.55
Average	2.43	**2.44**

Table 2: Human evaluation results (components of TAS) comparing explanations generated with and without trace visibility. Scores are averaged across all materials and evaluators.

Results. As shown in Table 2, explanations generated with trace visibility achieve a higher average faithfulness score, while specificity remains comparable and clarity shows a slight decrease. A paired Wilcoxon signed-rank test on faithfulness scores across all material–evaluator pairs does not indicate a statistically significant difference between the trace and no-trace conditions ($p = 0.62$). As a robustness check, we additionally report paired t-test results in Table 6.

Crucially, reliance solely on aggregate statistics obscures the conditional utility of trace visibility. The lack of global significance reflects systematic heterogeneity across materials with differing planning complexity, which is masked when results are averaged.

Qualitative analysis reveals that trace visibility yields clear benefits specifically in materials with (i) long planning horizons, (ii) interacting demand and supply events, or (iii) delayed replenishment effects. In such cases (e.g., Materials 1, 3, and 7), trace access enables evaluators to verify the temporal sequence of inventory depletion, cumulative demand growth, and insufficient supply, thereby supporting explicit causal verification.

By contrast, materials with relatively linear demand–supply dynamics (e.g., Materials 6 and 8) admit fluent explanations even without trace access. In these cases, exposing the full trace provides limited additional explanatory value and may slightly reduce surface-level clarity. The complete human evaluation protocol and raw per-material scores are reported in Appendix D.

5.6 RQ3: Does Evidence-Constrained Reasoning Improve Faithfulness?

While RQ2 shows that trace visibility improves causal grounding, trace access alone does not guarantee faithful reasoning. LLMs may still introduce unsupported claims. This research question examines whether explicitly constraining explanations to extracted evidence further improves faithfulness and reduces hallucination.

Experimental Setup. We compare two explanation-generation settings with identical MRP outcomes and deterministic traces: **Unconstrained**, where the LLM freely generates explanations using the trace, and **Evidence-Constrained (ExplainableMRP)**, where the LLM is restricted to ground all claims in an explicit evidence set extracted from the trace

Automatic Evaluation Metrics. We evaluate faithfulness using EC and HR.

Metric	Unconstrained	Constrained
EC	**1.00**	0.73
HR	0.08	**0.00**

Table 3: Automatic evaluation results for RQ3. Evidence-constrained reasoning eliminates hallucination while maintaining high evidence utilization.

Results. Table 3 reports the averaged EC and HR scores. Under evidence-constrained reasoning, we observed no unsupported claims in the evaluated dataset. While EC is reduced compared to the unconstrained setting, this reduction reflects an intentional design choice that prioritizes verification and safety over narrative completeness.

In the unconstrained setting, explanations frequently combine grounded facts with extrapolated or implied causal claims, increasing apparent coverage at the cost of introducing unsupported statements. In contrast, evidence constraints function as a safety mechanism that systematically suppresses unsupported claims, resulting in zero observed hallucinations (detailed in Appendix C.4). From this perspective, the reduced EC should not be interpreted as a loss of explanatory quality, but as a deliberate trade-off that favors precision, auditability, and trustworthiness over narrative breadth.

6 Discussion

6.1 Explainability Begins at the Input Level

The results of RQ1 show that many explanation errors originate before any reasoning occurs. When given heterogeneous industrial spreadsheets, LLMs often misinterpret the semantic roles of input fields, and these errors propagate downstream to compromise subsequent reasoning. By contrast, the proposed DataAgent deterministically enforces a

canonical MRP schema, eliminating semantic ambiguity and reducing IMR to zero. These findings indicate that explainability requires explicit control over input representation, rather than post hoc interpretation alone.

6.2 Why Trace Visibility Improves but Does Not Guarantee Faithfulness

Trace visibility is not uniformly beneficial across all materials; rather, its explanatory value emerges primarily in planning scenarios where causal attribution is non-trivial due to temporal depth, interacting demand events, or delayed supply effects.

RQ2 examines whether exposing selected deterministic MRP trace entries improves the causal grounding of natural-language explanations. Human evaluation results show that trace visibility tends to improve faithfulness, while its effect on specificity is case-dependent and its impact on surface-level clarity is marginal. This pattern indicates that traces primarily support explanations by enabling causal verification grounded in concrete numerical and temporal evidence, rather than by improving linguistic fluency.

The observed variability across materials highlights a fundamental distinction between surface-level fluency and deep causal verification. In simpler planning scenarios, explanations generated without trace visibility can remain fluent and even preferred.

Crucially, the benefits of trace visibility concentrate in materials with deeper causal structures and longer planning horizons. These findings indicate that trace visibility provides *conditional utility*: it is a necessary but insufficient condition for faithful explanation in deterministic planning contexts.

6.3 Evidence Constraints as a Mechanism for Faithful Explanation

RQ3 addresses the limitation that trace visibility alone is insufficient by examining whether explicit evidence constraints are required to prevent hallucination and ensure faithful explanations. The results show a clear separation between unconstrained and evidence-constrained explanations: While unconstrained explanations often introduce unsupported claims or speculative causal narratives, evidence-constrained explanations substantially reduce HR at the cost of lower EC, reflecting a deliberate trade-off favoring verifiability and precision.

This result highlights a critical distinction between *seeing* the trace and *being forced to use* the trace. Without explicit constraints, LLMs may treat traces as optional context rather than as binding evidence. By contrast, the proposed evidence-constrained reasoning mechanism enforces alignment between the explanation and the underlying deterministic computation.

6.4 Implications for LLM-Based Decision Support Systems

Overall, the results of RQ1–RQ3 suggest that explainability in deterministic planning systems should be treated as a pipeline property rather than a model capability. Reliable explanations require (i) schema-level normalization of inputs, (ii) explicit exposure of deterministic traces, and (iii) enforcement of evidence-bounded generation.

This perspective challenges prevailing approaches to explainable AI that focus primarily on post hoc interpretation or feature attribution. In domains such as MRP, where system behavior is fully determined by transparent rules, explainability should be defined as trace visibility and faithful narration of the underlying process.

7 Conclusion

This paper revisits explainability in deterministic planning systems through the lens of MRP. We show that, in such systems, explainability should not be defined as post hoc interpretation or feature attribution, but rather as faithful exposure and narration of the underlying deterministic process.

We propose ExplainableMRP, a pipeline that enforces explainability at three critical stages: input normalization through a schema-aware DataAgent, explicit exposure of deterministic execution traces, and evidence-constrained natural-language explanation generation. Through three targeted research questions, we demonstrate that (i) schema normalization eliminates input misinterpretation, (ii) trace visibility improves factual grounding and supports specificity in complex planning scenarios, and (iii) evidence constraints reduce hallucination and improve faithfulness.

Overall, reliable explanation in deterministic planning emerges not from the language model alone, but from how information is structured, exposed, and constrained across the reasoning pipeline, enabling transparent and trustworthy industrial decision support.

Limitations

All explanations and human evaluations in this study are conducted in Korean. While English is used as the language of academic reporting, language-specific factors may influence perceived fluency and evaluation outcomes.

This work has several limitations that point to directions for future research. First, our experimental evaluation is conducted on anonymized data from a single industrial domain, namely pharmaceutical manufacturing. While the input schema and datasets are domain-specific, the underlying deterministic netting and offset logic follows standard APICS principles that are widely adopted across manufacturing sectors such as automotive and electronics. Consequently, the core components of our framework, including the SimulationAgent and ReasoningAgent, are inherently transferable to any standard MRP environment without modification.

Second, the human evaluation of explanation quality relies on a limited number of domain-informed evaluators. However, we emphasize the high complexity density of the evaluated dataset: each material instance represents a deep causal structure involving multi-level BOMs, long planning horizons (up to 16 months), and tightly coupled demand and supply events. Evaluating explanation faithfulness against such complex real-world traces provides a more rigorous test of causal grounding than large-scale evaluations on simplified or synthetic benchmarks. Although the number of evaluated materials is limited, each material constitutes a full-scale industrial MRP instance with long planning horizons and dense causal interactions across multiple projects, production steps, and time periods. As a result, the effective evaluation unit in this study is not an isolated data point, but a deeply structured planning scenario, prioritizing depth of causal complexity over the breadth of independent samples.

Third, our framework focuses on explaining deterministic MRP outcomes under fixed planning logic. It does not address stochastic planning settings, adaptive replanning policies, or optimization-based MRP variants, where uncertainty and decision-making play a larger role.

Finally, although EXPLAINABLEMRP supports optional what-if analysis by recomputing deterministic traces under perturbed inputs, we do not quantitatively evaluate downstream decision impact or usability in this work. While such user studies are an important direction for future research, the primary focus of this paper is to establish the foundational correctness and faithfulness of trace-aligned explanations, which is a prerequisite for meaningful evaluation of trust, usability, and operational decision impact in real-world deployments.

Ethics, Risks, and Use of AI Assistants

This work investigates the use of LLMs as explanation interfaces for deterministic material planning systems, rather than as autonomous decision-making agents. The proposed framework does not alter MRP decisions, forecasts, or planning logic, but generates natural-language explanations grounded in an explicit, verifiable execution trace.

Potential risks arise if generated explanations are interpreted as normative recommendations or if trace information is incomplete or incorrectly specified. To mitigate these risks, the proposed system enforces strict schema normalization, deterministic trace exposure, and evidence-constrained generation, explicitly restricting explanations to trace-derived facts.

AI assistants were used during the research process for language refinement and editing, but all technical design choices, formal definitions, experimental setups, and interpretations were developed and verified by the authors. The authors take full responsibility for the content of the paper and its claims.

References

Abdulkadir Atalan. 2025. The ChatGPT application on quality management: A comprehensive review. *Journal of Management Analytics*, 12(2):229–259.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, and Reiichiro Nakano. 2021. Training verifiers to solve math word problems. *Computing Research Repository*, arXiv:2110.14168.

Yilun Du, Shuang Li, Joshua B. Tenenbaum, and Antonio Torralba. 2023. Improving factuality and reasoning in language models through multi-agent debate. In *Proceedings of the 40th International Conference on Machine Learning*.

Faisel Mohamed Elbahri, Omar Ismael Al-Sanjary, Musab A. M. Ali, Zakiya Ali Naif, Omar Ahmed Ibrahim, and M. N. Mohammed. 2019. Difference comparison of SAP, Oracle, and Microsoft solutions based on cloud ERP systems: A review. In *2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE.

Bárbara Ferreira and João Reis. 2023. Artificial intelligence in supply chain management: A systematic literature review and guidelines for future research. In *Proceedings of the International Joint Conference on Industrial Engineering and Operations Management*, Cham. Springer.

Jan C. Fransoo, Toni Waefler, and John R. Wilson, editors. 2010. *Behavioral Operations in Planning and Scheduling*. Springer Science & Business Media, New York, NY.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120.

Dávid Gyulai, András Pfeiffer, and László Monostori. 2017. Robust production planning and control for multi-stage systems with flexible final assembly lines. *International Journal of Production Research*, 55(13):3657–3673.

Wallace J. Hopp and Mark L. Spearman. 2011. *Factory Physics*. Waveland Press, Long Grove, IL.

F. Robert Jacobs, Richard B. Chase, and Nicholas J. Aquilano. 2011. *Manufacturing Planning and Control for Supply Chain Management*. McGraw-Hill, New York, NY.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.

Vincent A. Mabert. 2007. The early road to material requirements planning. *Journal of Operations Management*, 25(2):346–356.

Abhilash Puthanveettil Madathil, Xichun Luo, Qi Liu, Charles Walker, Rajeshkumar Madarkar, and Yi Qin. 2025. A review of explainable artificial intelligence in smart manufacturing. *International Journal of Production Research*, pages 1–44.

Joseph A. Orlicky. 1974. *Material Requirements Planning: The New Way of Life in Production and Inventory Management*. McGraw-Hill, New York, NY.

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

Carol A. Ptak. 2011. *Orlicky's Material Requirements Planning*, 3 edition. McGraw-Hill, New York, NY.

Carol A. Ptak and Chad Smith. 2016. *Demand Driven Material Requirements Planning (DDMRP)*. Industrial Press, New York, NY.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

David Salinas, Valentin Flunkert, and Jan Gasthaus. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551.

Hartmut Stadtler. 2005. Supply chain management and advanced planning: Basics, overview and challenges. *European Journal of Operational Research*, 163(3):575–588.

Daniel C. Steele. 1975. The nervous MRP system: How to do battle. *Production and Inventory Management*, 16(4):83–89.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Guodong Wang, Huabing Wang, Shanshan Lv, Xinxue Kang, and Qing'an Cui. 2025. Explainable artificial intelligence for manufacturing process modelling and optimisation: An integrated ANN and LIME approach. *International Journal of Production Research*, pages 1–22.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Qingyun Wu, Gagan Bansal, Jason Wei, Yi Zhang, Yiran Wu, Qingsong Wen, Andrew M. Dai, and Denny Zhou. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the First Conference on Language Modeling*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

A Formal Definitions and Notation

This appendix defines the core objects used in EXPLAINABLEMRP to ensure clarity and reproducibility.

MRP Inputs. For a material m, the input instance is defined as $X_m = \{S_0, \mathcal{D}, \mathcal{P}, L, \mathcal{H}\}$, where S_0 is the initial on-hand stock, $\mathcal{D} = \{(t_i, d_i)\}$ is the set of dated demand events, $\mathcal{P} = \{(t_j, p_j)\}$ is the set of dated supply (purchase order) events, L is a fixed lead time, and $\mathcal{H}$ is a discrete planning horizon.

Projected Inventory and Shortage. Projected inventory at time t is computed by deterministic netting:

$$S(t) = S_0 + \sum_{t_j \le t} p_j - \sum_{t_i \le t} d_i. \tag{3}$$

The first shortage date is:

$$t^* = \min\{\, t \in \mathcal{H} : S(t) < 0 \,\}. \tag{4}$$

Deterministic MRP Trace. A deterministic trace for material m is a time-ordered record of netting states:

$$\mathcal{T}_m = \{(t, S(t), D_t, P_t) \mid t \in \mathcal{H}\}, \tag{5}$$

where $D_t = \sum_{t_i \le t} d_i$ is cumulative demand and $P_t = \sum_{t_j \le t} p_j$ is cumulative supply up to time t. Each element of the trace corresponds to an explicit instantiation of Eq. 3 at time t, and the shortage condition is detected by Eq. 4.

Evidence Extraction. We define a deterministic evidence extractor

$$f_{\text{ev}} : \mathcal{T}_m \to E_m, \tag{6}$$

where E_m is a finite set of trace-validated facts used as admissible support for explanations. In our implementation, evidence includes (non-exhaustively) the shortage date, total demand/supply, usage spike dates, and a recent inventory trend, all computed deterministically from the trace (see Appendix B for an example).

Faithfulness and Trace Alignment Requirements. Let y_m be a natural-language explanation for material m, and let $I(y_m)$ be the set of *informational atoms* extracted from y_m (e.g., dates, quantities, and causal relations). We require the faithfulness condition:

$$I(y_m) \subseteq E_m. \tag{7}$$

In addition, every causal statement in y_m must correspond to a verifiable step in $\mathcal{T}_m$ (*trace alignment*), so that explanations reflect the deterministic netting dynamics rather than post hoc rationalizations.

Faithfulness Metrics. We evaluate faithfulness using Evidence Coverage (EC) and Hallucination Rate (HR).

Let $\text{EvidenceUsed}(y_m) \subseteq E_m$ denote the subset of evidence items referenced in y_m.

$$\text{EC}(y_m) = \frac{|\text{EvidenceUsed}(y_m)|}{|E_m|}. \tag{8}$$

Let $\text{Hallucinated}(y_m)$ be the set of statements in y_m that cannot be grounded in either $\mathcal{T}_m$ or E_m.

$$\text{HR}(y_m) = \frac{|\text{Hallucinated}(y_m)|}{|\text{Statements}(y_m)|}. \tag{9}$$

Trace Alignment Score (TAS). TAS is a human evaluation score assessing whether causal claims in y_m are consistent with the deterministic trace $\mathcal{T}_m$ and its implied netting dynamics. Evaluators rate *faithfulness*, *specificity*, and *clarity* on a Likert scale; we report averaged scores across evaluators and materials in the main text.

B Example of a Deterministic MRP Trace

Table 4 shows an excerpt of the deterministic MRP trace for *Material 1*. The trace records cumulative demand D_t, cumulative applied PO P_t, and the resulting projected inventory $S(t)$ over the planning horizon. In our implementation, when the inventory first becomes negative, the system registers the shortage date and immediately applies all remaining open PO quantity in the same step (see the SimulationAgent logic).

This example illustrates two key properties. First, the shortage event is uniquely determined by the cumulative effect of dated demand and supply events (Eq. 3–4). Second, trace visibility makes the causal sequence auditable: planners can identify the exact demand increment that first violates $S(t) \ge 0$ and verify how supply actions affect subsequent inventory evolution.

C Deterministic Execution and Evidence Construction

This appendix documents the deterministic components underlying EXPLAINABLEMRP, including the raw input data format, the MRP execution algorithm, evidence extraction, and the prompt structures used for explanation generation.

Date	Demand$_t$	PO$_t$	S_{t-1}	$S(t)$ before PO	$S(t)$ after PO
2026-03-20	139.191	0	324.900	185.709	185.709
2026-03-21	14.143	0	185.709	171.566	171.566
2026-03-22	125.060	0	171.566	46.506	46.506
2026-03-23	139.203	276	46.506	-92.697	183.303
2026-03-24	14.143	0	183.303	169.160	169.160
2026-03-25	125.060	0	169.160	44.100	44.100
2026-03-26	139.263	0	44.100	-95.163	-95.163

Table 4: Excerpt of the deterministic netting trace for Material 1 around the first shortage event. Demand$_t$ denotes the daily demand increment; PO$_t$ denotes the PO quantity applied at that date. The first shortage is detected on 2026-03-23 when $S(t)$ becomes negative before PO application, after which remaining PO is applied immediately.

C.1 Raw Dataset Format (Masked Industrial Data)

Experiments are conducted on anonymized operational datasets derived from an industrial manufacturing environment. All identifiers and numerical values are masked, while preserving the structural, temporal, and causal relationships required for deterministic MRP execution.

The raw dataset consists of four enterprise spreadsheets commonly found in ERP systems:

Bill of Materials (BOM). The BOM table specifies component-to-product relationships and usage coefficients. Columns include product identifiers, component identifiers, usage quantities, and unit information. Column names and units vary across files and vendors.

Production Plan (Plan). The Plan table encodes time-phased production demand. Each column may correspond to a production date or batch, with heterogeneous naming conventions and mixed units.

Inventory Records (Stock). The Stock table provides initial on-hand inventory levels. It may include auxiliary fields such as warehouse locations, inspection status, or free-text comments that are not directly relevant to MRP computation.

Open Purchase Orders (OpenPO). The OpenPO table lists outstanding supply orders. In addition to structured fields (quantity, due date), this table often contains unstructured natural-language comments (e.g., supplier delays or logistics issues).

These spreadsheets are heterogeneous in schema, naming conventions, and units. Without explicit normalization, large language models frequently misinterpret column semantics, motivating the use of a deterministic schema-aware DataAgent.

C.2 Deterministic MRP Execution Algorithm

The *SimulationAgent* implements a deterministic, rule-based MRP execution engine. This component performs exact netting and offset computation and does not invoke any language model.

Inputs.

- Initial on-hand stock S_0
- Demand events $\mathcal{D} = \{(t_i, d_i)\}$
- Supply events $\mathcal{P} = \{(t_j, p_j)\}$
- Fixed lead time L
- Discrete planning horizon $\mathcal{H} = \{t_1, \ldots, t_T\}$

Output. A deterministic MRP trace

$$\mathcal{T} = \{(t, S(t), D_t, P_t)\}_{t \in \mathcal{H}}.$$

This algorithm explicitly instantiates the inventory update rule (Eq. 1) at every time step. Given fixed inputs, the resulting trace $\mathcal{T}$ is uniquely determined.

PO Application Convention. In our implementation, when the inventory first becomes negative, all remaining open purchase order quantities scheduled at the same time step are applied immediately. This convention is used consistently in Algorithm 1 and the trace examples reported in Appendix B.

C.3 Deterministic Evidence Extraction Function

To ensure faithful explanation, explanations are restricted to a finite set of evidence deterministically extracted from the trace.

We define an evidence extraction function

$$f_{\text{ev}} : \mathcal{T} \to E,$$

where E denotes the admissible evidence set.

Algorithm 1 Deterministic MRP Trace Computation

Require: Initial stock S_0, demand events $\mathcal{D}$, supply events $\mathcal{P}$, planning horizon $\mathcal{H}$, lead time L
Ensure: Deterministic trace $\mathcal{T}$

cumulative_demand $\leftarrow 0$
cumulative_supply $\leftarrow 0$
inventory $\leftarrow S_0$
$\mathcal{T} \leftarrow \emptyset$
shortage_detected $\leftarrow$ **False**
for each time $t \in \mathcal{H}$ (chronological order) **do**
  for all $(t_i, d_i) \in \mathcal{D}$ where $t_i = t$ **do**
    cumulative_demand $\leftarrow$ cumulative_demand $+d_i$
  end for
  for all $(t_j, p_j) \in \mathcal{P}$ where $t_j = t$ **do**
    cumulative_supply $\leftarrow$ cumulative_supply $+p_j$
  end for
  inventory $\leftarrow S_0 +$ cumulative_supply $-$ cumulative_demand
  Append (t, inventory, cumulative_demand, cumulative_supply) to $\mathcal{T}$
  if inventory < 0 **and** shortage_detected = **False then**
    shortage_date $\leftarrow t$
    arrival_date $\leftarrow t - L$
    shortage_detected $\leftarrow$ **True**
  end if
end for

Evidence Items. From a trace $\mathcal{T}$, the extractor deterministically computes:

- Initial stock S_0
- Total cumulative demand D_T
- Total cumulative supply P_T
- First shortage date $t^* = \min\{t : S(t) < 0\}$
- Inventory trajectory $\{S(t)\}_{t \in \mathcal{H}}$
- Demand spike events (large single-period increases)
- Supply gaps (intervals with increasing demand but no incoming supply)

All evidence items are computed via fixed rules without statistical inference or language model involvement.

Role in Explanation. Let y be a generated explanation and $I(y)$ its extracted informational atoms. Faithfulness is enforced by the constraint:

$$I(y) \subseteq E.$$

C.4 Prompt Templates and Evidence Injection

This section documents the prompt structures used to generate explanations.

Raw-LLM Baseline Prompt. The baseline prompt provides raw spreadsheets and requests an explanation without trace access or constraints:

> Explain why a shortage occurs for the given material based on the provided tables.

ExplainableMRP Prompt. The proposed framework uses a structured prompt consisting of:

1. Schema-normalized input tables
2. Deterministic trace-derived evidence serialized as JSON
3. An explicit instruction restricting explanations to the provided evidence

Evidence Constraint Instruction.

> Generate an explanation using only the provided evidence. If a claim cannot be supported by the evidence, state that it is unavailable. Do not speculate or introduce external information.

Output Schema and Verification. The model must return a structured JSON object containing `reasoning`, `cause`, `action`, `EvidenceUsed`, and

AlignmentIssues. Outputs are parsed deterministically to verify evidence usage and trace alignment. If validation fails, the system falls back to a deterministic template-based explanation.

Scope of Evaluation. All quantitative and qualitative evaluations reported in this paper (RQ1–RQ3) are conducted on baseline MRP executions without hypothetical scenario modifications. While the system supports interactive scenario simulation (e.g., demand scaling or supply removal), these capabilities are treated as system features and are not evaluated in the current study.

D Evaluation and Verification Details

This appendix documents the parsing and validation logic used for faithfulness verification, as well as the human evaluation protocol and raw evaluation results. These details are provided to ensure transparency and to address potential concerns regarding subjectivity or heuristic bias.

D.1 Parsing Rules and Validation Logic

To verify explanation faithfulness, all generated explanations are subjected to deterministic parsing and validation.

Informational Atom Extraction. Given an explanation y, we extract a set of *informational atoms* $I(y)$ using rule-based pattern matching. Atoms correspond to explicitly stated: (i) dates (e.g., shortage dates or arrival dates), (ii) numerical quantities (e.g., inventory levels, demand or supply amounts), and (iii) causal relations linking demand or supply events to inventory changes.

The extraction process relies on predefined lexical templates and regular expressions, rather than learned classifiers, to ensure deterministic behavior.

Unsupported Claim Detection. An informational atom $a \in I(y)$ is considered *unsupported* if $a \notin E$, where E is the trace-derived evidence set. Unsupported atoms are flagged as AlignmentIssues. This criterion is purely rule-based and does not involve human judgment or language model inference.

Evidence Usage Tracking. Atoms successfully matched to evidence items are recorded in EvidenceUsed. This mapping enables automatic computation of EC and HR.

Fallback Handling. If parsing fails due to malformed output or missing required fields, the system falls back to a deterministic template-based explanation that reports only the shortage date and aggregate demand and supply. This conservative fallback prevents unverified claims from entering the evaluation.

Fallback Usage in Evaluation. To assess whether fallback handling influenced the reported hallucination results, we manually inspected all explanations generated for RQ3 using the evaluation spreadsheet. We confirm that no explanations in the evaluated dataset were replaced by template-based fallbacks. All evaluated outputs contained trace-derived evidence items and material-specific numerical references. Accordingly, the reported zero HR reflects the effect of explicit evidence constraints and deterministic validation, rather than post-hoc output filtering or output substitution.

D.2 Human Evaluation Protocol

Human evaluation is conducted to assess whether trace visibility improves the perceived quality of explanations, particularly in terms of causal grounding.

Material Selection. The evaluated materials are selected to represent non-trivial planning cases in which deterministic netting produces meaningful causal structures. Specifically, materials are included if they (i) experience at least one shortage event within the planning horizon, (ii) involve multiple dated demand events, and (iii) have overlapping supply and demand periods. Materials with trivial configurations (e.g., a single demand event or immediate supply satisfaction) are excluded to avoid degenerate explanation cases.

Evaluators. Three domain-informed evaluators with experience in material planning and manufacturing operations participated in the study. Evaluators were not involved in system development.

Evaluation Setup. For each material, evaluators are presented with two explanations corresponding to identical MRP outcomes: one generated with trace visibility and one without. The presentation order is randomized for each evaluator and each material. Evaluators are not informed which explanation has access to the deterministic trace.

Evaluation Interface. The evaluation is implemented using a web-based survey tool. For each

explanation pair, evaluators answer the following questions:

- **Faithfulness**: Is the explanation consistent with deterministic MRP logic?
- **Specificity**: Does the explanation identify concrete dates, quantities, and causal events?
- **Clarity**: Is the explanation easy to understand and well structured?

Each dimension is rated on a three-point Likert scale.

In addition, evaluators indicate an overall preference between the two explanations.

Blinding and Randomization. All evaluations are double-blinded with respect to explanation condition. Randomization is applied independently per material to mitigate ordering effects.

D.3 Human Evaluation Instructions

This subsection provides the full instructions and questionnaire used for the human evaluation of explanation quality.

Purpose of the Study. This survey was conducted for research purposes to evaluate the quality of natural-language explanations for material shortage situations in MRP. All responses were collected anonymously and used solely for research.

Overview of Each Evaluation Task. In each question, participants were presented with the following information for a single material:

- A summary of the shortage outcome (Summary),
- A partial execution trace representing the actual deterministic computation that led to the shortage (Trace),
- Two natural-language explanations (Explanation A and Explanation B).

The execution trace serves as the ground-truth evidence indicating how the shortage occurred. Participants were asked to evaluate how well each explanation aligns with the provided trace. Both explanations were generated based on the same material and the same shortage outcome, but may differ in wording and structure.

Randomization Protocol. To mitigate ordering and presentation bias:

- The assignment of explanations to labels *A* and *B* was randomized for each question.
- The presentation order of explanation conditions (e.g., with-trace vs. no-trace) was randomized across questions.

Participants were not informed about how the explanations were generated or about the internal mechanisms of the system.

Participant Instructions. Participants were instructed to:

- Read both explanations carefully.
- Evaluate which explanation better explains the shortage situation based solely on the provided summary and trace.
- Rely on intuitive and honest judgment; no prior expertise in MRP or supply-chain planning was required.

Evaluation Criteria. Each explanation was evaluated independently using a 3-point Likert scale according to the following criteria.

Faithfulness. *Is the explanation grounded in the provided summary and trace?*

- 1: Largely inconsistent with the evidence.
- 2: Partially grounded in the evidence.
- 3: Accurately reflects key quantities and temporal dynamics.

Specificity. *Does the explanation refer to concrete quantities, dates, and changes?*

- 1: Mostly abstract or vague.
- 2: Mentions some numerical details.
- 3: Clearly specifies dates, quantities, and changes.

Clarity. *Is the explanation easy to understand for non-experts?*

- 1: Difficult to understand.
- 2: Moderately clear.
- 3: Very clear and easy to understand.

Overall Preference. After rating both explanations, participants were asked to indicate their overall preference:

- Explanation A,
- Explanation B,
- No clear preference.

Ethical Considerations. No personally identifiable information was collected. Participants were informed that their responses would be anonymized and used only for academic research purposes.

D.4 Raw Human Evaluation Scores

Table 5 reports the complete raw human evaluation scores for all materials, evaluators, and experimental conditions. Scores are reported without aggregation to allow independent inspection of per-material and per-evaluator variation.

D.5 Statistical Analysis and Robustness Checks

Using the raw faithfulness scores reported in Table 5, we conduct a paired Wilcoxon signed-rank test across all material–evaluator pairs ($n = 24$). The test yields a statistic of $W = 59.5$ with a p-value of 0.62, indicating that the aggregate difference between the trace and no-trace conditions is not statistically significant.

This result does not imply the absence of systematic benefits from trace visibility. As discussed in the main text, the effect of trace access varies substantially across materials with different structural complexity, which is not fully captured by aggregate significance tests.

Paired *t*-test as a Robustness Check. To complement the non-parametric analysis, we additionally perform a paired t-test on the same material–evaluator pairs. Table 6 summarizes the results.

Across all three dimensions, the paired t-tests similarly indicate no statistically significant differences, with small effect sizes (Cohen's $d_z \leq 0.17$). These results are consistent with the Wilcoxon test and reinforce the conclusion that trace visibility does not yield uniform improvements across all materials.

Interpretation. The statistical analyses indicate that the effect of trace visibility is not uniform across all materials. Aggregate significance tests therefore provide an incomplete characterization of its impact. As demonstrated by the qualitative anal-

Mat.	Eval.	Condition	Faith.	Spec.	Clar.
1	E1	No-Trace	2	2	3
	E1	With-Trace	3	3	2
	E2	No-Trace	2	2	3
	E2	With-Trace	3	3	2
	E3	No-Trace	2	2	2
	E3	With-Trace	2	2	3
2	E1	No-Trace	2	2	3
	E1	With-Trace	2	2	3
	E2	No-Trace	2	2	3
	E2	With-Trace	2	3	3
	E3	No-Trace	2	2	2
	E3	With-Trace	2	2	2
3	E1	No-Trace	2	2	2
	E1	With-Trace	3	3	3
	E2	No-Trace	1	2	1
	E2	With-Trace	2	3	3
	E3	No-Trace	2	2	2
	E3	With-Trace	2	2	3
4	E1	No-Trace	2	2	2
	E1	With-Trace	2	2	3
	E2	No-Trace	2	2	2
	E2	With-Trace	2	3	3
	E3	No-Trace	2	2	3
	E3	With-Trace	2	2	2
5	E1	No-Trace	2	2	2
	E1	With-Trace	3	2	3
	E2	No-Trace	2	3	2
	E2	With-Trace	2	2	3
	E3	No-Trace	1	2	3
	E3	With-Trace	2	2	2
6	E1	No-Trace	3	3	3
	E1	With-Trace	2	2	3
	E2	No-Trace	3	3	3
	E2	With-Trace	2	2	2
	E3	No-Trace	3	2	3
	E3	With-Trace	3	2	3
7	E1	No-Trace	2	2	2
	E1	With-Trace	3	3	3
	E2	No-Trace	2	3	3
	E2	With-Trace	3	3	3
	E3	No-Trace	2	2	2
	E3	With-Trace	3	3	3
8	E1	No-Trace	3	3	3
	E1	With-Trace	2	2	2
	E2	No-Trace	3	3	3
	E2	With-Trace	2	2	2
	E3	No-Trace	2	3	3
	E3	With-Trace	1	2	2

Table 5: Complete raw human evaluation scores for all materials, evaluators, and conditions. Faith., Spec., and Clar. denote faithfulness, specificity, and clarity, respectively. All scores are reported on a five-point Likert scale.

Metric	Δ Mean	$t(23)$	p
Faithfulness	+0.083	0.49	0.63
Specificity	+0.083	0.44	0.66
Clarity	+0.167	0.81	0.43

Table 6: Paired t-test results comparing With-Trace and No-Trace conditions ($n = 24$).

ysis, trace-based explanations are especially valuable in materials with deeper causal structures and longer planning horizons, where causal verification is otherwise difficult. In contrast, for materials with relatively linear demand–supply dynamics, trace access yields limited additional benefit and may not translate into higher aggregate scores. These findings suggest that the primary contribution of trace visibility lies in supporting causal verification in complex planning scenarios, rather than in consistently improving average evaluation metrics.

D.6 Qualitative Examples

To complement quantitative analysis, we include representative qualitative examples illustrating when trace visibility provides clear explanatory benefits and when its impact is limited. All examples are reproduced verbatim from the evaluated explanations.

D.7 Qualitative Failure and Correction Example

We present a representative qualitative example to illustrate when trace visibility materially improves causal verification.

Material 3 (Delayed Replenishment). In the no-trace condition, the explanation incorrectly assumed an immediate inventory recovery following a purchase order, despite a non-zero lead time. With trace access, the explanation explicitly referenced the delayed arrival date recorded in the trace, correctly attributing the shortage to cumulative demand growth during the lead-time window.

This example highlights how trace visibility prevents superficially fluent but causally incorrect explanations by exposing temporally grounded constraints that are otherwise omitted.

E Frontend and System Implementation Details

This appendix documents the interactive frontend used in our experiments and demo. The frontend is implemented in Streamlit and orchestrates the full agent pipeline, including automatic file role classification, schema normalization, deterministic MRP execution, and trace-grounded explanation generation. All screenshots shown in this appendix correspond to the actual system used in our study and are presented in Korean, reflecting the language of the underlying industrial deployment.

E.1 Frontend Overview

The frontend provides an end-to-end workflow: (i) multi-file spreadsheet upload, (ii) automatic assignment of files to MRP roles (BOM/Plan/Stock/OpenPO), (iii) in-browser table editing and lead-time specification, (iv) baseline MRP execution and optional what-if scenario simulation, and (v) inspection of deterministic traces and agent logs.

The interface is explicitly designed to expose intermediate computation artifacts (e.g., trace tables, shortage dates, and evidence summaries), rather than only final shortage endpoints typically shown in commercial MRP systems.

E.2 Agent Orchestration in the Frontend

After file upload, the frontend applies an *AutoClassifierAgent* to infer the role of each spreadsheet (BOM, Plan, Stock, OpenPO). Users may manually override the assignment before execution. The selected tables are then passed to a *DataAgent* for schema normalization (`unify`), producing canonical BOM, Stock, and OpenPO tables for downstream deterministic computation.

Auxiliary AutoClassifierAgent. In addition to the three core agents described in the main paper (DataAgent, SimulationAgent, and ReasoningAgent), the frontend includes an auxiliary AutoClassifierAgent that assists users by automatically inferring the roles of uploaded spreadsheets. This component operates exclusively at the data ingestion stage and does not participate in deterministic MRP execution, evidence extraction, or explanation generation. Accordingly, it is excluded from all experimental evaluations reported in this paper.

MRP execution is performed by a deterministic runner (`MRPRunner`), which produces per-material summaries and full simulation traces. The frontend supports two execution modes:

- **Baseline**: execute deterministic MRP on the normalized inputs.
- **Scenario**: apply user-specified perturbations

(pre-defined or free-form) and execute deterministic MRP on the modified inputs.

E.3 Scenario Specification and Application

The frontend supports both button-based and free-form natural language scenarios. Pre-defined scenarios deterministically modify one of the following: (i) BOM usage (demand scaling), (ii) per-material lead times (lead-time shifts), or (iii) open purchase order quantities (supply scaling).

For free-form scenarios, a language model is used only to parse the user's natural-language instruction into a structured JSON specification with fields `material`, `leadtime_delta`, `demand_multiplier`, and `po_multiplier`. The resulting JSON is then applied deterministically to the relevant tables. Thus, language model usage is strictly limited to scenario parsing, while all numerical MRP execution remains deterministic.

E.4 Trace and Log Exposure

To support explainability, the frontend exposes:

- Per-material key dates (shortage date, required arrival date, suggested order),
- A time-indexed simulation trace (inventory, cumulative demand, cumulative PO usage),
- Demand spike markers derived from the trace,
- Full agent logs, including AutoClassifierAgent, DataAgent, SimulationAgent, and ReasoningAgent logs, shown in expandable panels.

Figure 2 illustrates the EXPLAINABLEMRP frontend, which exposes both intermediate agent reasoning traces and the final trace-grounded explanation. By making input interpretation, schema normalization, deterministic MRP execution, and shortage reasoning explicitly visible, the interface enables users to inspect and verify every step from raw data ingestion to explanation generation.

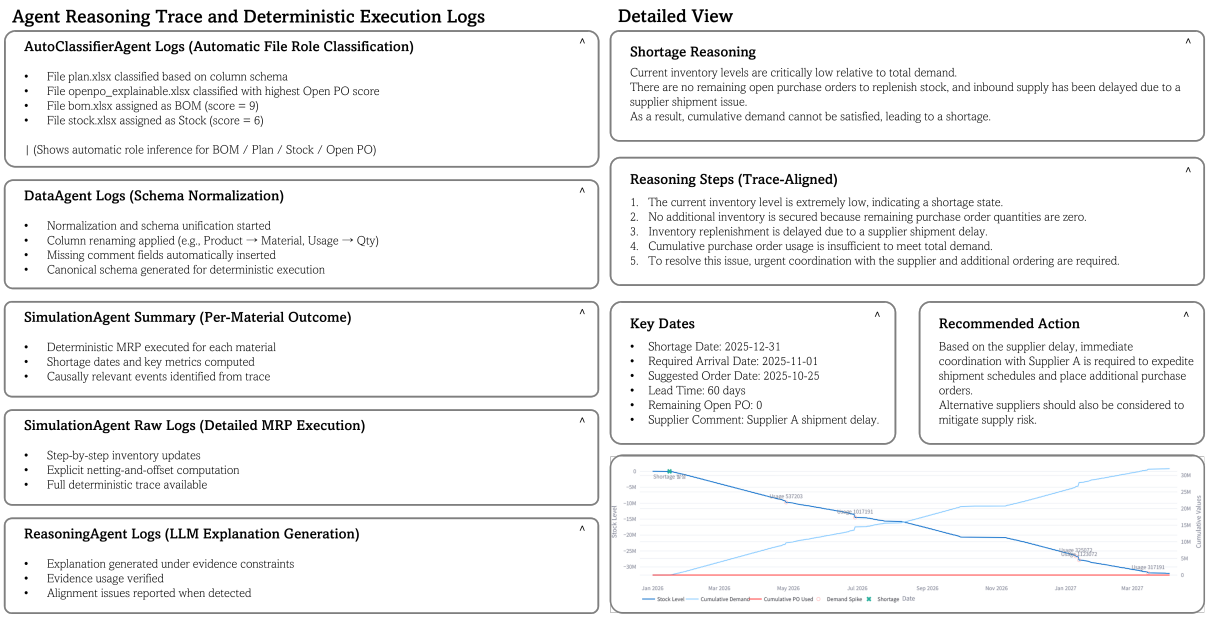

Figure 2: EXPLAINABLEMRP frontend exposing agent reasoning traces and trace-grounded explanations (Korean-language interface). The interface displays logs from the AutoClassifierAgent, DataAgent, SimulationAgent, and ReasoningAgent to make input classification, schema normalization, and deterministic MRP execution explicit and inspectable. The resulting explanation combines natural-language reasoning with numerical evidence derived directly from the deterministic trace, including shortage timing, key dates, recommended actions, and inventory–demand trajectories. The shortage event corresponds exactly to the time step at which projected inventory becomes negative in the MRP trace.