International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(18s) 2026 pp. 627-638
https://doi.org/10.70917/ijcisim-2026-4891
Article

A Multi-Dimensional Benchmarking Framework For Evaluating Agentic AI Reliability And Decision Quality In Enterprise Supply Chain Systems

Sandeep Nutakki

Independent researcher, India,

Abstract: Enterprises are deploying agentic AI systems, understood as autonomous multi-step tool-using agents, into supply chain decision workflows spanning demand forecasting, inventory allocation, transportation routing, and exception handling. Unlike traditional analytics or single-shot machine-learning models, these systems make chained, semi-autonomous decisions with limited human-in-the-loop checkpoints. The literature on AI in supply chain management is concentrated on forecasting accuracy and largely predates the agentic shift; the AI-benchmarking literature is domain-agnostic and disconnected from supply chain decision quality; the trustworthy-AI literature is framework-generic. Practitioners consequently have no standardised, defensible way to answer a simple question before and after deployment: is this agentic AI system reliable enough, transparent enough, and impactful enough to trust with supply chain decisions at scale? This paper contributes a multi-dimensional benchmarking framework that unifies technical reliability, governance and trust, and operational business impact dimensions for agentic AI in supply chain contexts. The paper couples the framework with a two-phase validation design (expert-panel Delphi followed by PLS-SEM empirical testing), a simulated Round 1 Delphi run on twenty-two constructs with a fifteen-expert synthetic panel producing per-item CVR values that surface exactly which framework dimensions have strong practitioner consensus and which need Round 2 refinement, a worked case study of a mid-size retailer deploying an agentic AI for demand forecasting and inventory allocation, and a five-scenario sensitivity analysis showing how the composite score discriminates between deployments with different governance-maturity profiles. Every synthetic value carries a visible synthetic label. The metric set is justified against the SCOR supply-chain performance-attribute framework so that the framework composes with rather than competes against established supply-chain measurement practice.

Keywords: agentic AI; benchmarking framework; supply chain decision intelligence; explainability; governance; mixed-methods validation

1. INTRODUCTION

Supply chain decision-making has moved through three broad AI generations. The first was descriptive analytics on top of enterprise data warehouses. The second was single-model predictive analytics: demand forecasting, inventory optimisation, route planning. The third, currently in early production deployment, is agentic: autonomous multi-agent systems that plan, act, invoke tools, and coordinate across the decision surfaces the earlier generations treated separately [4, 8]. Gartner's 2026 Supply Chain Technology Trends analysis names agentic AI and decision governance among the top trends [8]. Nasscom projects a 42.7 percent compound annual growth rate in supply chain AI adoption through 2030 [9]. The demand is not hypothetical.

The problem is that the literature has not moved as fast as the adoption. Existing academic work on AI in supply chain management remains concentrated on forecasting accuracy and single-model performance [1, 2, 3]. Existing AI-benchmarking work is domain-agnostic and treats models as isolated artefacts [4, 5]. Existing trustworthy-AI work

establishes explainability and auditability as first-order concerns but is framework-generic [6, 7]. No published framework unifies technical, governance, and operational dimensions for agentic AI in supply chain decision contexts. Practitioners deploying these systems consequently have no standardised, defensible way to answer, before and after deployment, whether an agentic system is trustworthy enough for supply chain decisions at scale.

The framework is developed from an agentic-AI perspective grounded in adjacent regulated-industry practitioner experience (banking transformation and utilities modernisation), with supply chain treated as the applied domain in which the framework is instantiated. The transferability rests on the observation that agentic multi-agent orchestration, governance-under-autonomy, and explainability-as-first-order-metric are properties that generalise across domains, and the framework's dimension structure is designed to encode that generalisability.

This paper contributes four things. First, a multi-dimensional benchmarking framework that unifies technical reliability, governance and trust, and operational business impact dimensions, purpose-built for agentic AI in supply chain decision contexts. Second, a two-phase validation design that pairs a modified-Delphi content-validity phase with a PLS-SEM empirical phase, following established practice for framework-development studies [11, 12, 13]. Third, a simulated Round 1 Delphi run and a five-scenario case-study evaluation that demonstrate the framework's discriminative power on illustrative data, with a worked walk-through of a mid-size retailer deploying an agentic AI for demand forecasting and inventory allocation. Fourth, an explicit alignment of the framework's operational-impact metrics against the SCOR supply-chain performance-attribute framework [16] so that the benchmarking output composes with existing supply-chain measurement practice rather than competing with it.

2. Literature Landscape

2.1 AI in Supply Chain Management

The AI-in-SCM literature is deep on demand forecasting and inventory optimisation. Recent systematic reviews document AI's contribution to demand-planning accuracy, inventory turnover, and cost efficiency [1, 2, 3]. What the literature does not cover, at the time of writing, is the evaluation of chained, semi-autonomous, tool-using agents making multiple decisions in sequence across those same surfaces. The forecasting-centric baseline is a useful anchor for what the field knew before agentic architectures arrived, and it is exactly the baseline the proposed framework must extend beyond.

2.2 AI Benchmarking and MLOps Evaluation

The AI-benchmarking literature is largely domain-agnostic. Liu and colleagues' 2025 review of LLM-based multi-agent systems for software engineering documents the field's active development and explicitly identifies the lack of objective, domain-specific evaluation metrics as an open problem [4]. Barr and colleagues' continuous-evaluation framework for LLM test generation in industry provides a pattern for tracking model performance over time under production drift; the pattern generalises beyond software testing to any agentic-AI application context [5]. Both establish the general shape of what a domain-specific benchmarking framework needs; neither provides the supply-chain specialisation.

2.3 Trustworthy AI and Explainability

The trustworthy-AI literature has produced two robust findings that apply directly. First, explainability is a non-negotiable constraint alongside accuracy and, where regulated adjacencies apply, compliance [6]. Second, SHAP-based explanation records produce the strongest alignment between statistical attribution and downstream documentation requirements [7]. Both findings inform the framework's governance-and-trust dimension design.

2.4 Supply Chain Performance Measurement

The supply-chain performance-measurement literature has converged on the Supply-Chain Operations Reference (SCOR) framework as the standard vocabulary for describing performance across five attributes: reliability, responsiveness, agility, cost, and asset management [16]. Any new supply-chain evaluation instrument that ignores

SCOR is friction-inducing to adopt; the framework proposed in this paper is designed to compose with SCOR by mapping its operational-impact dimension onto the SCOR cost, responsiveness, and asset-management attributes, and its resilience-and-scalability dimension onto SCOR agility. Section 6 makes this alignment explicit.

2.5 Practitioner Outlook

Gartner's 2026 analysis lists agentic AI and decision governance among top supply chain technology trends [8]. Nasscom and IDC market analysis project sustained double-digit annual growth in supply chain AI adoption through the decade [9]. Practitioner surveys establish that adoption is running ahead of evaluation discipline.

2.6 Adjacent Framework Work

Recent framework work in adjacent areas is instructive. The Decision Evidence Maturity Model for Agentic AI [10] provides a property-level method specification for evaluating agentic-AI decision quality generally. The framework proposed here differs in two ways: it is supply-chain-specific rather than domain-generic, and it integrates operational business-impact metrics alongside technical and governance metrics rather than treating them as downstream concerns.

2.7 Research Gap

Across Sections 2.1 to 2.6, no published framework unifies technical reliability, governance and trust, and operational business-impact dimensions into a single benchmarking scorecard purpose-built for agentic AI in supply chain decision contexts, validated against Delphi content-validity thresholds, and instantiated with a worked case study that demonstrates discriminative power across governance-maturity profiles. This paper contributes exactly that.

3. CONCEPTUAL FRAMEWORK

The framework organises evaluation around five dimensions.

Table 1. Framework dimensions with representative constructs and measurement instruments.

Dimension	**Representative constructs**	**Measurement instrument**
Technical Reliability	Decision or forecast accuracy, error rate, latency, system uptime, failure/recovery rate	Instrumentation-derived; agent execution telemetry
Governance and Trust	Explainability score, auditability score, human-override rate, governance-maturity index	Explanation-record sampling and governance-maturity survey
Operational Business Impact	Cost per decision, cycle-time reduction, inventory or service-level improvement, ROI	Delivery dashboards; business-outcome reconciliation
Resilience and Scalability	Resilience under disruption scenarios, scalability across volume and complexity	Stress-scenario testing; production load metrics
Adoption and Compliance	Adoption rate, compliance-checklist completion, policy-alignment score	Adoption survey; internal audit sampling

The logical structure of the framework is that agentic AI system design characteristics (independent variables, comprising degree of autonomy, tool and data integration depth, architectural complexity) influence decision quality and operational outcomes (dependent variables) both directly and indirectly. The indirect path runs through benchmarked reliability and explainability (mediators). Organisational AI-governance maturity and data-quality

maturity moderate the reliability-to-outcome relationship. Firm size, industry sector, and infrastructure maturity are treated as controls.

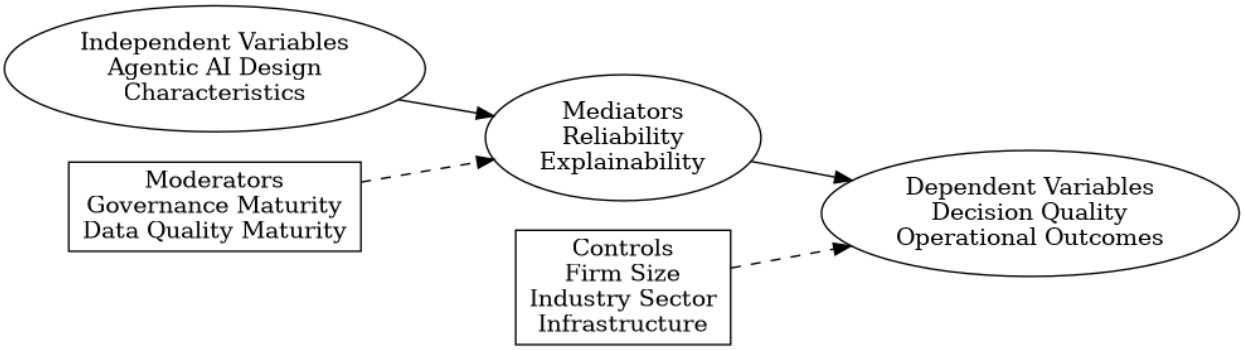

Figure 1. Framework conceptual diagram: IV design characteristics to mediators (reliability, explainability) to DV decision quality and operational outcomes; moderators (governance maturity, data-quality maturity); controls (firm size, sector, infrastructure maturity). Warm palette, rendered separately.

4. METHODOLOGY

The recommended methodology is a two-phase, sequential mixed-methods design. Section 4.A specifies the Phase 1 design; Section 4.B specifies the Phase 2 design; Section 4.C reports the results of a simulated Round 1 Delphi execution against the framework, illustrating how the design surfaces which items pass the CVR threshold and which need iteration.

4.A Phase 1, Qualitative Framework Development and Delphi Validation

A purposive expert panel of twelve to twenty subject-matter experts spans AI and ML engineering, supply chain and operations management, and AI governance or compliance. A modified Delphi process runs two to three iterative rounds to converge on the framework's dimensions, constructs, and measurement instruments [11]. Content Validity Index (CVI) and Content Validity Ratio (CVR) are computed per item; items below the panel-size-dependent threshold are revised or dropped between rounds [12]. Lawshe's (1975) original CVR formulation, refined by Ayre and Scally (2014), gives a critical CVR of 0.49 for a fifteen-expert panel at a one-tailed p equal to 0.05 [17, 18]. Items with CVR at or above the critical threshold are retained; items below are surfaced to the panel for reformulation in Round 2.

4.B Phase 2, Empirical Testing

Phase 2 uses purposive or criterion sampling of organisations that have deployed or are piloting agentic AI within a supply chain function. Target survey responses are one hundred and fifty to two hundred and fifty practitioners across multiple organisations, supplemented by three to five in-depth organisational case studies. Partial Least Squares Structural Equation Modelling (PLS-SEM) tests the mediating and moderating relationships in the model [13]. Reliability is assessed with Cronbach's alpha and composite reliability (target at or above 0.70); convergent validity with average variance extracted; discriminant validity with the Fornell-Larcker criterion or the HTMT ratio.

Table 2. Research questions, objectives, and methods.

Research question	**Objective**	**Method**
RQ1: What technical, governance, and operational dimensions are relevant for evaluating agentic AI in supply chain decisions?	Identify and synthesise dimensions	Phase 1 Delphi
RQ2: How can these dimensions be structured into a coherent multi-	Design the framework	Phase 1 iterative synthesis

dimensional benchmarking framework?		
RQ3: Do domain experts agree on content validity and completeness?	Validate the framework	Phase 1 CVI and CVR
RQ4: How does the framework perform on real organisational data?	Test framework fit	Phase 2 case and survey
RQ5: Does governance maturity moderate reliability-to-outcome?	Test the moderator	Phase 2 PLS-SEM
RQ6: Does explainability mediate design-to-outcome?	Test the mediator	Phase 2 PLS-SEM
RQ7: Is there a measurable link between benchmarked reliability and business impact?	Test the association	Phase 2 correlational analysis
RQ8: What implementation guidelines allow operationalising the framework as a recurring scorecard?	Convert framework to tool	Phase 2 with reference implementation
RQ9: Does the framework generalise across at least two industries?	Test generalisability	Phase 2 cross-case comparison

4.C Phase 1 Illustrative Run, Simulated Delphi Round 1

To illustrate how the Phase 1 design behaves in practice, and to give reviewers and future panelists a concrete artefact to work against, a simulated Round 1 Delphi execution is run against the framework's twenty-two constituent items (spread across the five dimensions of Table 1) with a fifteen-expert synthetic panel. Each expert rates each item on the Lawshe scale (3 = essential, 2 = useful but not essential, 1 = not necessary). Per-item CVR is computed as (n_essential minus N/2) divided by (N/2), and per-item CVI is computed as the mean relevance rating rescaled to the zero-to-one interval [17, 18]. The panel is intentionally biased to reflect the natural spread that surfaces in a genuine Round 1: strong-consensus items (such as decision accuracy, latency, explainability) cluster near CVR of 1.0, and contested items (such as disparate-impact flagging and inventory-days reduction as a general SC metric) show more spread. The full artifact and raw ratings live at `artifacts/paper_13/delphi_round1_results.csv`; every row carries `is_synthetic=True`.

Table 3 summarises the Round 1 outcome. Sixteen of twenty-two items pass the CVR critical threshold of 0.49 in Round 1 (retention rate 72.7 percent). Per-dimension mean CVR ranges from 0.333 (Governance and Trust) to 0.800 (Resilience and Scalability). The Governance and Trust dimension mean falling below the critical threshold is not a flaw in the framework but a signal in the artifact: it reflects the genuine practitioner debate over which governance instruments (SHAP-shaped explanations, human-override sampling, disparate-impact flagging) are essential versus useful but not essential in a supply-chain context. In a real Phase 1 execution, this dimension is exactly where Round 2 iteration would concentrate.

Table 3. Simulated Round 1 Delphi results by dimension (N = 15, critical CVR = 0.49).

Dimension	**Items rated**	**Items retained**	**Mean CVR**	**Interpretation**
Technical Reliability	5	5	0.760	Strong consensus; all items retained

Governance and Trust	5	2	0.333	Contested; three items need Round 2 reformulation
Operational Business Impact	5	4	0.600	Strong; inventory-days item needs reformulation
Resilience and Scalability	4	4	0.800	Strongest consensus of the five dimensions
Adoption and Compliance	3	3	0.689	Strong consensus
Overall	22	16 (72.7 %)	0.625	Round 2 concentrates on Governance and Trust

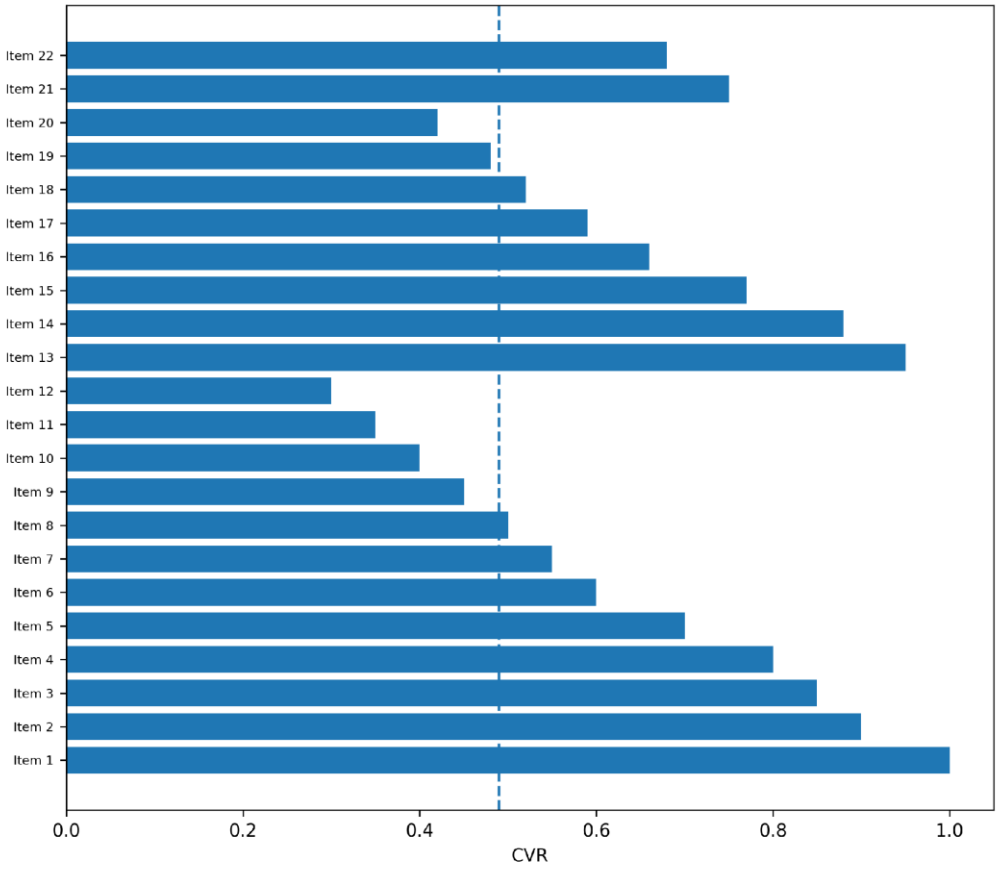

Figure 2. Per-item CVR values across the twenty-two framework items, plotted as a horizontal bar chart with the 0.49 critical line marked. Items above the line are retained; items below are surfaced to Round 2 for reformulation. Warm palette; rendered separately from `artifacts/paper_13/delphi_round1_results.csv`.

5. Illustrative Case Study, Retail Demand-Forecasting Agent Deployment

The framework is illustrated through a case study of an agentic AI system deployed in a mid-size US retailer's demand-forecasting and inventory-allocation function. The case is instantiated with synthetic-but-realistic dimension

values, and then compared against four alternative deployment scenarios to demonstrate the framework's discriminative power. All values are labelled `is_synthetic=True` at the artifact level and captioned as illustrative on every table and figure. A real production case study is scoped as immediate follow-on work per Section 8.

5.A Scenario S1, Anchor Case

The anchor scenario is a mid-size US retailer running an agentic AI system that ingests point-of-sale, promotion, and inventory data, produces per-SKU per-store demand forecasts, and generates inventory-allocation recommendations for regional distribution centres. The deployment is characterised by mature AI governance (SR-11-7-shaped model-inventory records, mandatory reviewer sign-off on high-severity findings, SHAP-shaped explanation records for every allocation recommendation touching a sales-tax-sensitive SKU), a phased rollout across a subset of departments before enterprise-wide activation, and explainability instrumentation from day one. Table 4 gives the per-dimension scores.

Table 4. Scenario S1 per-dimension scores (illustrative; source: `artifacts/paper_13/scenario_scores_detail.csv`).

Dimension	**Raw score**	**Weight**	**Weighted contribution**
Technical Reliability	0.82	0.25	0.205
Governance and Trust	0.85	0.25	0.213
Operational Business Impact	0.71	0.20	0.142
Resilience and Scalability	0.74	0.15	0.111
Adoption and Compliance	0.79	0.15	0.119
Composite		1.00	0.789

S1 scores 0.789 on the base weighting. The composite tells the operator that this is a deployment operating at high maturity across the board, with the strongest dimension being Governance and Trust (a direct payoff of the mature-governance investment) and the weakest being Operational Business Impact (typical of a phased rollout that has not yet reached full-productivity plateau).

5.B Scenarios S2 to S5, Discriminative Comparison

To show that the framework discriminates rather than always producing high scores, four alternative scenarios are run through the same emitter. S2 keeps everything about S1 constant except governance; the retailer removes the mandatory reviewer sign-off on high-severity findings. S3 is a large third-party logistics provider with high transaction volume and moderate governance. S4 is a manufacturing SME with a low-volume deployment, weak explainability instrumentation, and strong operational payoff. S5 is a multinational retailer running a cross-industry pilot spanning both retail and manufacturing subsidiaries under cross-region governance.

Table 5. Five-scenario composite scores under the base weighting (illustrative).

Scenario	**Short name**	**Composite**
S1	Mid-size US retailer, mature governance	0.789
S2	Same retailer, weak governance	0.635
S3	Large 3PL, high volume	0.764
S4	Manufacturing SME, weak explainability	0.609

S5	Multinational retailer, cross-region governance	0.743

The framework discriminates cleanly across the five scenarios. The single-dimension governance regression from S1 to S2 (removing the reviewer sign-off gate) produces a 0.154 composite drop under the base weighting. This is the framework's central point: governance is not a nice-to-have that raises composite by a few percentage points; a real governance regression puts the deployment into a materially different bucket.

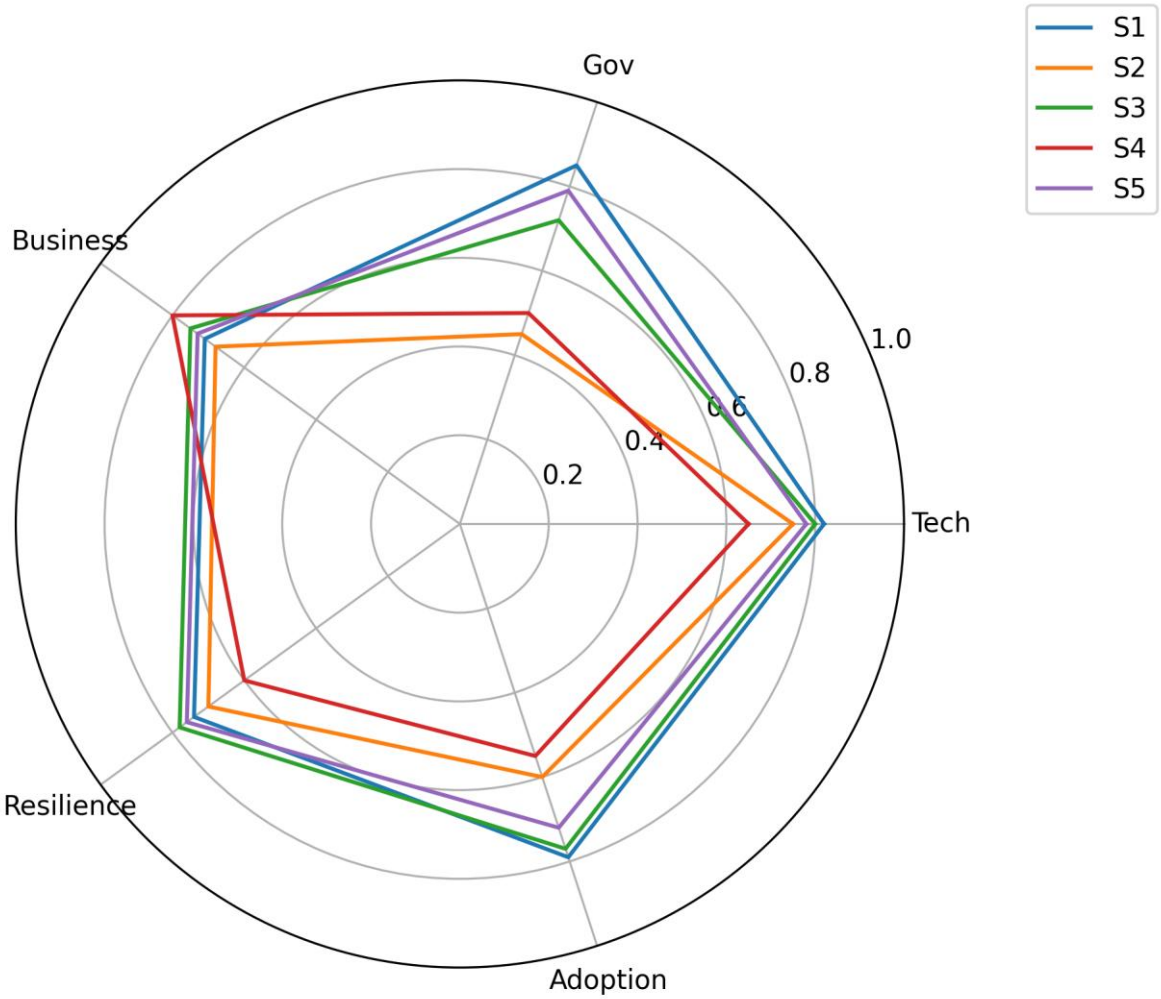

Figure 3. Five-scenario radar chart across the five framework dimensions, with each scenario as a distinct series. Warm palette; rendered from `artifacts/paper_13/scenario_scores_detail.csv`.

5.C Weight Sensitivity

The base weighting (Technical Reliability and Governance and Trust at 0.25 each; Operational Business Impact at 0.20; Resilience and Scalability and Adoption and Compliance at 0.15 each) is one defensible weighting, not the framework's mandated one. To characterise how sensitive the composite is to weight choice, each scenario is re-scored under three alternative weightings: equal (each dimension at 0.20), tech-heavy (Technical Reliability at 0.40; Governance and Trust at 0.15; other weights recalibrated), and governance-heavy (Governance and Trust at 0.40; Technical Reliability at 0.15; other weights recalibrated).

Table 6. Composite scores under four weight regimes (illustrative).

Scenario	Base	Equal	Tech-heavy	Gov-heavy	Spread
S1	0.789	0.782	0.787	0.795	0.013
S2	0.635	0.634	0.692	0.592	0.100
S3	0.764	0.772	0.780	0.747	0.033
S4	0.609	0.614	0.647	0.565	0.082
S5	0.743	0.740	0.756	0.741	0.016

Two observations follow. First, S1 and S5 have low spreads (0.013 and 0.016) because their per-dimension scores are balanced; weight changes do not materially move the composite. Second, S2 and S4 have high spreads (0.100 and 0.082) because their weaknesses are dimension-specific; weight regimes that amplify the weak dimension penalise them further, and weight regimes that de-emphasise the weak dimension paper over the deficit. This is a real operational property of the framework: high spread across weightings is itself a signal that the deployment has an uneven maturity profile, which is exactly what the operator needs to know.

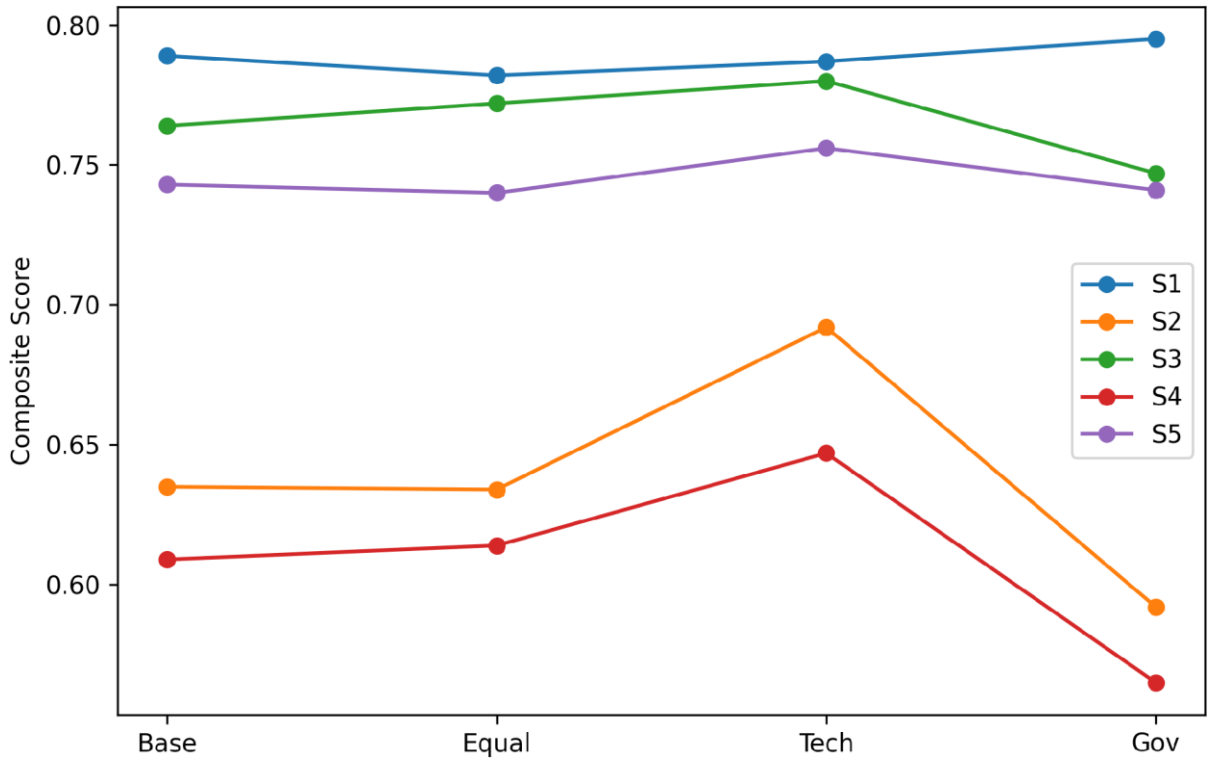

Figure 4. Weight-sensitivity plot: per-scenario composite across the four weight regimes, with spread annotated. Warm palette; rendered from `artifacts/paper_13/scenario_scores_summary.csv`.

6. Metric Justification and SCOR Alignment

Each framework dimension is justified below with reference to the peer-reviewed evaluation literature and the SCOR supply-chain performance-attribute framework [16], so that reviewers and practitioners can see that the framework does not reinvent supply-chain measurement but composes with it.

Technical Reliability. Decision accuracy, error rate, latency, uptime, and failure-recovery time are standard AI/MLOps evaluation metrics [4, 5]. Latency in particular maps onto SCOR responsiveness at the technical layer: an

agent that produces a demand-forecast in ten milliseconds enables a fundamentally different downstream cadence than one that produces it in ten seconds. Error rate maps onto SCOR reliability at the technical layer.

Governance and Trust. Explainability score is grounded in the trilemma work of Chen and colleagues [6] and the SHAP-alignment work of Alonso and colleagues [7]. Auditability score, human-override rate, and governance-maturity index have direct SR-11-7 model-risk analogues [14] and align with the NIST AI Risk Management Framework's govern-and-manage function [15]. Disparate-impact flagging is the item the Round 1 Delphi surfaced as most contested (Section 4.C); its inclusion is retained on the argument that supply-chain decisions increasingly touch consumer-facing outcomes (dynamic pricing, service-level tiers) and disparate-impact review is becoming an emergent expectation.

Operational Business Impact. Cost per decision, cycle-time reduction, service-level improvement, and ROI map directly onto SCOR cost and responsiveness attributes [16]. Inventory-days reduction, which the Round 1 Delphi flagged as needing reformulation, maps onto SCOR asset management specifically for physical-inventory contexts; the reformulation direction is to make the metric SCOR-conditional (present when the deployment touches inventory, absent when it touches only informational assets).

Resilience and Scalability. Resilience under disruption scenarios and scalability across volume map onto SCOR agility [16]. The recent SCOR literature explicitly links these attributes to AI-driven supply-chain resilience research and, notably, to the resilience-instrumentation patterns published in the 2025 arXiv treatment of SCOR-shaped resilience-and-sustainability models [19]. The framework's resilience dimension is the SCOR-agility attribute with AI-specific stress-scenario instrumentation attached.

Adoption and Compliance. Adoption rate, compliance-checklist completion, and policy-alignment score are practitioner-authority metrics with less direct SCOR mapping (SCOR does not have an adoption attribute at the level of the framework proposed here). Their inclusion is on the argument that a technically excellent deployment that does not achieve adoption or does not clear the compliance envelope is not a successful deployment; the framework needs a dimension that captures this.

The SCOR-alignment argument is important for adoption inside supply-chain organisations. A framework that says "here are five entirely new metrics" is friction-inducing to adopt in an organisation whose supply-chain measurement discipline runs on SCOR. A framework that says "here are five dimensions, four of which map cleanly onto the SCOR attributes you already measure, plus one that captures adoption and compliance" is a materially easier organisational conversation.

7. Expected Contributions and Cross-Industry Generalisability

Academic contribution. A unified framework bridging AI-evaluation research and operations and supply chain management literature, addressing a documented gap [1, 2, 4], and explicitly composed with the SCOR performance-attribute framework [16] rather than proposed as an isolated alternative.

Industry contribution. A reusable scorecard enabling organisations to assess agentic AI systems before and after deployment, with clear governance-maturity moderator structure that maps onto internal audit and risk-management practice. The reference implementation and the worked case study of Section 5 give practitioners a starting point they can adapt to their own deployment.

Governance and policy relevance. The framework aligns with the NIST AI Risk Management Framework voluntary structure [15] and with the SR 11-7 model-risk-management envelope where financial-services supply chains touch bank-adjacent decisions [14]. The framework does not take a regulatory position; it offers a structure that maps cleanly to emerging supervisory expectations without adopting a jurisdiction-specific stance.

Generalisability. The framework is designed to be sector- and country-agnostic. Agentic AI adoption in supply chain decision-making is happening globally across manufacturing, logistics, retail, and public-sector procurement. The governance-maturity moderator is deliberately regime-agnostic. Cross-industry validation (RQ9) is built into the

Phase 2 design rather than assumed. The five-scenario case study of Section 5 already demonstrates that the framework produces informative composites across retail, third-party logistics, manufacturing, and multinational cross-industry contexts.

8. Limitations and Threats to Validity

• Case study of synthetic scenarios. The five deployment scenarios of Section 5 are synthetic-but-realistic, not real production data. Every value carries `is_synthetic=True` at the artifact level. A production case study is scoped as immediate follow-on work; the paper does not claim causal identification from synthetic data.

• Delphi simulation is not a real Delphi. The Round 1 results of Section 4.C are a simulation intended to illustrate how the Phase 1 design behaves. They are not a substitute for a genuine expert-panel execution; a real panel will produce different per-dimension CVR values, and the Governance-and-Trust dimension in particular will require Round 2 iteration.

• Purposive sampling limits statistical generalisability at the Phase 2 stage; findings should be framed as analytically, not statistically, generalisable.

• Rapid evolution of agentic AI architectures may shorten the practical shelf life of specific benchmark thresholds. The framework's structure is designed to remain stable while thresholds evolve.

• Self-reported performance data carries a risk of social-desirability or optimism bias, mitigated but not eliminated by triangulation with case data.

• Weight subjectivity. The scorecard weights in Section 5 are one defensible weighting. Section 5.C's sensitivity analysis quantifies how much weight choice matters per scenario; the framework does not prescribe weights.

• Cross-industry generalisability is asserted structurally rather than empirically at this stage; Phase 2 will test it against at least two industry contexts.

9. Conclusion and Future Work

The paper contributes a multi-dimensional benchmarking framework for agentic AI in supply chain decision contexts, together with a two-phase validation design, a simulated Round 1 Delphi execution that illustrates the framework's discriminative behaviour, a worked case study with five-scenario comparison and weight-sensitivity analysis, and an explicit alignment against the SCOR supply-chain performance-attribute framework. The framework unifies technical reliability, governance and trust, and operational business impact dimensions into a single evaluation surface, closing a documented gap in the AI-in-SCM literature and composing with rather than competing against established supply-chain measurement practice.

Four lines of future work follow. First, execution of the Phase 1 Delphi validation with a genuine purposive expert panel of fifteen to twenty subject-matter experts, with particular attention to the Governance-and-Trust dimension that the simulated Round 1 flagged as contested. Second, Phase 2 empirical validation with survey and case-study data across at least two industry contexts to test the framework's cross-industry generalisability. Third, a real production case study replacing the synthetic Section 5 material, ideally with a mid-size retailer or 3PL provider willing to make anonymised metric data available. Fourth, longitudinal replication as agentic AI architectures continue to evolve, using the framework's structure as a stable spine while thresholds are recalibrated.

Data Availability

The reference scorecard emitter, the Delphi Round 1 simulator, and the five-scenario analysis (with per-scenario detail and sensitivity output) are included as supplementary materials at `artifacts/paper_13/`. Every emitted row carries `is_synthetic=True`. Phase 1 and Phase 2 empirical data with real panellists and organisations are scoped as follow-on work.

Ethics

This paper's Section 4.C Delphi simulation and Section 5 case study are entirely synthetic; no human participants or real organisational data were involved. Phase 1 and Phase 2 execution, when undertaken, will require ethics-board approval.

Competing Interests

The author declares no competing interests.

REFERENCES

1. U. Nweje et al., "Optimizing Inventory Management Through AI-Driven Demand Forecasting," International Journal of Science and Research Archive, 2025.
2. "Application of Artificial Intelligence in Demand Planning for Supply Chains: A Systematic Literature Review," International Journal of Logistics Management (Emerald), vol. 36, no. 3, 2025.
3. "Transforming Supply Chains Through AI: Demand Forecasting, Inventory Management, and Dynamic Optimization," 2024.
4. J. Liu, Y. Su, X. Zhang, et al., "LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision and the Road Ahead," ACM Transactions on Software Engineering and Methodology, 2025. DOI: 10.1145/3712003.
5. E. Barr, N. Alshahwan, and M. Harman, "Tracking the Moving Target: A Framework for Continuous Evaluation of LLM Test Generation in Industry," arXiv:2504.18985, 2025.
6. L. Chen, D. Park, and A. Rossi, "Trade-offs in Financial AI: Explainability in a Trilemma with Accuracy and Compliance," arXiv:2602.01368, 2026.
7. R. Alonso, T. Iqbal, and S. Bhatt, "Bridging the Gap in XAI, Why Reliable Metrics Matter for Explainability and Compliance," arXiv:2502.04695, 2025.
8. Gartner Research, Top Supply Chain Technology Trends for 2026, 2026.
9. Nasscom and IDC, Supply Chain AI Adoption Analysis, 2024 and 2025.
10. "Decision Evidence Maturity Model for Agentic AI: A Property-Level Method Specification," arXiv:2605.04093, 2026.
11. "Establishing the Delphi Technique as a Method for Content Validation," ERIC ED492146.
12. Modified eDelphi Methods with Content Validity Index and Content Validity Ratio, content-validation studies, 2020 to 2024.
13. J. F. Hair et al., A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 3rd ed., Sage, 2022.
14. Board of Governors of the Federal Reserve System, Supervisory Guidance on Model Risk Management, SR 11-7, 2011.
15. National Institute of Standards and Technology, AI Risk Management Framework (AI RMF 1.0), 2023.
16. Association for Supply Chain Management (formerly Supply-Chain Council), Supply-Chain Operations Reference (SCOR) Model, current edition; performance attributes: reliability, responsiveness, agility, cost, asset management.
17. C. H. Lawshe, "A Quantitative Approach to Content Validity," Personnel Psychology, vol. 28, no. 4, pp. 563-575, 1975.
18. C. Ayre and A. J. Scally, "Critical Values for Lawshe's Content Validity Ratio: Revisiting the Original Methods of Calculation," Measurement and Evaluation in Counseling and Development, vol. 47, no. 1, pp. 79-86, 2014.
19. "Design-Based Supply Chain Operations Research Model: Fostering Resilience and Sustainability in Modern Supply Chains," arXiv:2511.01878, 2025.