Agentic AI For Dynamic E2E Sourcing

by

Diandra Riando

and

Sandeep Sharma

Submitted on **May 8, 2026** in Partial Fulfillment of the

Requirements for the Degree of Master of Applied Science in Supply Chain Management

at the Massachusetts Institute of Technology

ABSTRACT

This capstone presents a proof-of-concept (POC) agentic artificial intelligence (AI) system to transform the sourcing process for a global healthcare Sponsor. The current sourcing pipeline is constrained by fragmented data, reactive renewals, and reliance on individual judgment, leading to inefficiencies and operational risk. Against a baseline of 71 calendar days for an end-to-end sourcing cycle, the proposed solution targets a 35–40% cycle time reduction, bringing the estimated cycle time to 35–49 days. The system is a modular, multi-agent architecture comprising 12 specialized agents across two integrated parts: Part 1 focuses on generating and prioritizing sourcing opportunities before sourcing procedures begin, while Part 2 executes end-to-end Source-to-Contract (S2C) workflows aligned to the Sponsor's Desktop Procedures (DTPs), specifically DTP-01 to DTP-06, for the Information Technology (IT) Infrastructure category. The architecture includes five pre-procedure agents for opportunity generation and prioritization, and seven execution agents for S2C workflow orchestration and decision support. It combines rule-based logic, retrieval-augmented generation, lightweight machine learning (ML), and template-driven drafting to support dynamic sourcing decisions. A human-in-the-loop design ensures oversight while a shared learning-loop architecture captures reviewer edits, overrides, and validation feedback to continuously refine scoring logic, retrieval relevance, and routing decisions over time. Evaluation across simulated scenarios demonstrates an 81.9% AI Reliability Rate and 85.6% Signal Attribution Accuracy for Part 1 under clean data conditions, and a 78% First-Pass Acceptance Rate with 85% Knowledge Base Retrieval Relevance for Part 2. Expected outcomes include prioritized sourcing pipelines, faster sourcing execution, and a scalable foundation for enterprise adoption, with measurable improvements in speed, consistency, transparency, and continuous learning capability.

Capstone Advisor: Dr. Maria Jesus Saenz

Title: Executive Director, MIT Supply Chain Management (SCM) Masters Programs; Director, MIT Digital Supply Chain Transformation Lab; Research Scientist

Table of Contents

1 INTRODUCTION 3
1.1 Motivation 3
1.2 Problem Statement and Key Questions 3
1.3 Project Goals and Expected Outcomes 4
2 STATE OF THE PRACTICE 4
2.1 Sourcing as a Structured Decision Process 4
2.2 Analytics and AI in Procurement 5
2.3 Human-in-the-Loop Design 5
3 METHODOLOGY 5
3.1 Part 1: Prioritized Sourcing Opportunities using Agentic AI 6
3.1.1 Decision Modelling: Sourcing Pipeline Generation 6
3.1.2 Architecture and Design of Agents 6
3.1.3 Prioritization Framework 7
3.1.4 Model Properties and Integration of Agents 9
3.1.5 Performance Measurement Framework 9
3.2 Part 2: End-to-end Source-to-Contract (S2C) Execution 9
3.2.1 Decision Modelling: S2C Execution 10
3.2.2 Architecture and Design of Agents 10
3.2.3 Performance Measurement Framework 11
3.3 Analytical Model Design and Learning Architecture 12
3.4 Data Collection and Preparation 12
4 RESULTS AND DISCUSSION 13
4.1 Part 1: Prioritized Sourcing Opportunities 13
4.1.1 Test Cases 13
4.1.2 Test Results 14
4.2 Part 2: Source-to-Contract Execution 16
4.2.1 Test Cases 16
4.2.2 Test Results 16
4.3 Discussion 18
4.4 Limitations 18
5 RECOMMENDATIONS 18
5.1 Management Recommendations 18
5.2 Future Work 19
REFERENCES 20

1 INTRODUCTION

Our Sponsor is a multinational enterprise in the global healthcare sector, headquartered in the United States with a broad international presence across numerous markets. Publicly available information indicates that the organization employs a large global workforce and generates multi-billion-dollar annual revenues. Its portfolio spans multiple operating segments, with healthcare-related operations representing a major share of overall business activity. Over the past decade, the organization has invested heavily in innovation and digital transformation, particularly across supply chain, operations, and research functions. Recent initiatives have included the adoption of technologies such as AI, automation, robotics, and advanced analytics to improve operational efficiency and decision-making.

1.1 Motivation

Sourcing is the process of identifying, assessing, and engaging suppliers to procure goods and services (SAP, n.d.). Our Sponsor's sourcing activities face several challenges across its global operations, from fragmented data sources to high market volatility, forcing sourcing teams to spend significant manual effort to consolidate information, react to contract expirations, and evaluate supplier options. When agreements approach expiration, teams scramble to evaluate renewal options and adjust sourcing strategies, often under severe time pressure. Without a system that continuously interprets signals from contracts, demand forecasts, supplier performance, and market dynamics, the sourcing pipeline remains static, reactive, and vulnerable to disruption.

In its current state, the Sponsor's sourcing pipeline is hampered by three key pain points:

- **Fragmented data:** Data is spread across multiple systems, requiring significant manual effort to consolidate information before evaluating suppliers or preparing Request for X (RFx) packages — where X denotes a Proposal (RFP), Quote (RFQ), or Information (RFI) depending on the sourcing path.
- **Reactive renewals:** Contract renewals are often handled reactively, with teams scrambling to assess options under tight deadlines, resulting in longer cycle times and inconsistent supplier assessments.
- **Reliance on individual judgement:** The process relies heavily on the judgment and experience of individual managers, leading to variability in sourcing outcomes and loss of institutional knowledge when roles change.

These issues slow down sourcing activities, increase operational risk, and reduce the time teams can dedicate to more strategic work. This capstone evaluates whether a dynamic, agentic AI approach can address these pain points by improving cycle time, consistency, and sourcing efficiency. This creates a clear need to evaluate whether an agentic AI-enabled proof-of-concept can deliver meaningful value within the Sponsor's sourcing ecosystem.

Agentic AI refers to multi-agent systems that execute coordinated decision tasks with limited supervision. It consists of AI agents, which are ML models that mimic human decision-making to solve problems in real time (Russell & Norvig, 2020). In a multi-agent system, each agent performs a specific subtask required to reach the goal, and their efforts are coordinated through AI orchestration.

1.2 Problem Statement and Key Questions

The Sponsor's sourcing process remains largely reactive, constrained by fragmented data, long cycle times, and a reliance on manual judgment. As contract expirations approach, teams must quickly gather information, assess supplier options, and make renewal decisions often under time pressure and without consistent, data-driven insights. These challenges reduce efficiency, create variability in supplier evaluations, and limit the organization's ability to proactively plan sourcing activities.

The key questions to be answered include:

- How can sourcing pipeline generation be made more proactive and dynamic across a global procurement operation?
- How should an intelligent sourcing system be structured to support end-to-end sourcing decisions?
- What data and knowledge inputs are needed to keep sourcing recommendations timely and relevant?
- How should improvements in sourcing outcomes be measured and tracked over time, and what governance and scaling mechanisms would enable the system to expand across categories?

1.3 Project Goals and Expected Outcomes

The project's overall goal is to deliver a proof-of-concept agentic AI system that proactively generates sourcing pipelines using a dynamic data pipeline fed by contract records, category strategy documents, and supplier performance data. The system is aligned with the Sponsor's Desktop Procedures (DTPs) — a standardized set of step-by-step sourcing workflows (DTP-01 to DTP-06) that govern end-to-end procurement activities for the IT Infrastructure category. A Multi-Criteria Decision Making (MCDM) model evaluates and prioritizes sourcing options based on defined strategic and operational criteria, while a human-in-the-loop design ensures sourcing teams review, validate, and refine all outputs rather than being replaced by full automation. Their feedback provides the critical input needed to improve the agents over time — and as the system scales across additional categories and markets, this feedback loop directly shapes how scoring weights, retrieval indices, and routing logic evolve, ensuring the system grows more accurate and reliable with each sourcing cycle.

Deliverables to the Sponsor include a multi-agent AI system that generates sourcing pipelines for the IT Infrastructure category, and a comprehensive activity log capturing agent decisions, inputs, logic, and human feedback to ensure transparency and continuous improvement. Following implementation, the Sponsor is expected to gain an initial demonstration of how sourcing cycle times could be reduced, how decision-making may become more consistent, and how sourcing needs can be anticipated before contracts expire.

2 STATE OF THE PRACTICE

This section reviews how sourcing decisions are structured, how analytics and AI are currently applied in procurement, and the principles of responsible system design that inform this project's approach. Together, these insights clarify where gaps exist and what capabilities a next-generation, agentic AI-enabled sourcing system must offer.

2.1 Sourcing as a Structured Decision Process

Academic and industry research describe sourcing as a structured decision process guided by clear logic, reliable data, and coordinated activities across the sourcing cycle. Classical sourcing literature treats sourcing as a design problem that aligns requirements, market conditions, supplier capabilities, and incentives to maximize value. Seshadri (2005) frames this as a mechanism design challenge where sourcing policies must coordinate prices and performance with organizational goals. More recent procurement texts position sourcing as a continuous strategic process. Monczka et al. (2020) highlight the need for integrated category strategies, cross-functional collaboration, and ongoing supplier development rather than one-time buying. These foundations show that effective sourcing depends on well-defined decision points, consistent evaluation criteria, and a stable information environment — all essential for any intelligent sourcing system.

2.2 Analytics and AI in Procurement

Research on analytics and artificial intelligence demonstrates how digital tools can support these decisions. Wamba et al. (2018) explain how big data and ML improve forecasting, risk visibility, and operational insight. Guida et al. (2023) highlight emerging applications such as supplier scoring, risk classification, and document extraction that help structure the early stages of sourcing. Cannas et al. (2024) notes that these tools are increasingly embedded in procurement workflows, but their usefulness depends on high-quality, consistent, and accessible data. Current procurement systems also use optimization engines for award scenario analysis and rule-based workflows for compliance, but they typically operate as isolated components and do not connect upstream requirements with downstream negotiation logic.

Industry reports point to similar trends. Mittal et al. (2024) shows that leading organizations now use AI-assisted recommendations to speed category strategies, spot opportunities, and shorten cycle times. Keelvar (2025) demonstrates how modern AI assistants can draft event documents, structure evaluations, and generate award scenarios while keeping processes auditable. These developments show that AI-enabled orchestration can accelerate sourcing while preserving human decision authority. This shift is consistent with broader digital supply chain transformation frameworks, which emphasize the alignment of operations, technology, and strategy as essential enablers of sustainable procurement performance (Saenz et al., 2022).

2.3 Human-in-the-Loop Design

A recurring theme across the literature is the importance of responsible system design and human-machine collaboration. Accenture (2025) finds that high-performing organizations rely on algorithms for consistent, repeatable analysis but reserve judgment calls, trade-offs, and approvals for humans. Effective human-AI collaboration requires transparency, feedback loops, and learning mechanisms — principles that directly shape this project's human-in-the-loop design. Research on human-machine teaming further emphasizes that AI systems deliver the most durable value when humans remain active participants in decision validation, contributing domain expertise that algorithms cannot replicate (Kiron et al., 2020). Broader digital supply chain transformation frameworks reinforce this, positioning the alignment of operations, technology, and strategy as a prerequisite for sustainable procurement performance (Saenz et al., 2022). In procurement specifically, AI tools are increasingly reshaping supplier negotiations by surfacing data-driven insights, but negotiators who understand how to complement rather than defer to algorithmic recommendations achieve consistently better outcomes (Revilla & Saenz, 2025).

3 METHODOLOGY

The methodology for this project is designed to develop and evaluate a proof-of-concept (POC) intelligent sourcing system for the Sponsor’s procurement Category teams, starting with the IT Infrastructure category. The approach is grounded in four upstream data signals — contract expiration timelines, spend data, category strategy, and supplier risk profiles — which together inform prioritization before execution through DTP-01 to DTP-06. These procedures define the sequencing, responsibilities, and compliance requirements for end-to-end sourcing activities, and serve as the structural backbone for modelling decision logic, designing the agentic system, and determining the data and evaluation needs of the POC.

The system is organized into two integrated parts: Part 1 generates and prioritizes sourcing opportunities from upstream data signals, and Part 2 executes those opportunities through DTP-01 to DTP-06. Both parts share a common data and memory infrastructure and are orchestrated by the same Supervisor Agent. This modular architecture reflects the engineering systems principle that complex sociotechnical challenges are best addressed through interconnected, function-specific subsystems.

3.1 Part 1: Prioritized Sourcing Opportunities using Agentic AI

Before any formal sourcing procedure is triggered, the system monitors four upstream signals: contract expirations, spend data, category strategy, and supplier risk. Pre-DTP agents process these inputs to identify opportunities, assess urgency and strategic fit, and generate a prioritized pipeline for S2C execution. This approach addresses fragmented, reactive pipeline management by enabling proactive, data-driven sourcing decisions.

3.1.1 Decision Modelling: Sourcing Pipeline Generation

The design begins with four key signals, each treated as a decision input. For each, the system defines its role, constraints, and the appropriate analytical approach—rule-based logic, retrieval, or scoring. Table 1 summarizes the core decision, required inputs, constraints, and appropriate mechanism for each of the four upstream signals.

Phase	Core Decision	Required Inputs	Constraints	Appropriate Mechanism
Contract Expiration	Detect upcoming sourcing events	Contract expiration timelines	Data completeness, renewal policies	Rule-based trigger logic with alert prioritization
Spend Data	Identify category opportunity	IT Infrastructure spend data	Budget limits, aggregation rules	Spend analytics with trend detection
Category Strategy	Align with category direction	Category strategy from Category Management	Strategic sourcing guidelines	Strategy mapping with rule validation
Supplier Risk Profile	Assess supplier exposure	Supplier risk profile	Risk thresholds, compliance flags	Risk scoring model with threshold alerts

Table 1: *Core Decision, Required Inputs, Constraints, and Appropriate Mechanisms (Part 1)*

3.1.2 Architecture and Design of Agents

Five specialized agents carry out the pipeline generation function. Each agent is responsible for one signal domain and passes its output to the Sourcing Signal Agent, which synthesizes all signals into a unified, prioritized case list. The five specialized agents include:

- **Spend Pattern Agent:** Analyzes category spend data to identify sourcing opportunities, aggregation risks, and concentration exposure using trend detection, thresholds, and clustering.
- **Contract Renewal Agent:** Tracks contract expirations and renewal windows to trigger proactive sourcing using date logic, thresholds, and alert prioritization.
- **Category Strategy Agent:** Extracts supplier preferences, sourcing posture, and policy constraints using structured retrieval, rule mapping, and large language model (LLM) based summarization.
- **Supplier Risk Agent:** Interprets supplier risk to guide sourcing decisions using risk thresholds, compliance checks, and normalized risk data.
- **Sourcing Signal Agent:** Consolidates signals (spend, expirations, risk, strategy) to trigger sourcing cases using signal fusion, prioritization, and rule-based event logic.

Figure 1 provides an overview of the agentic AI architecture for Part 1, illustrating how the five specialized agents connect to the Sourcing Signal Agent.

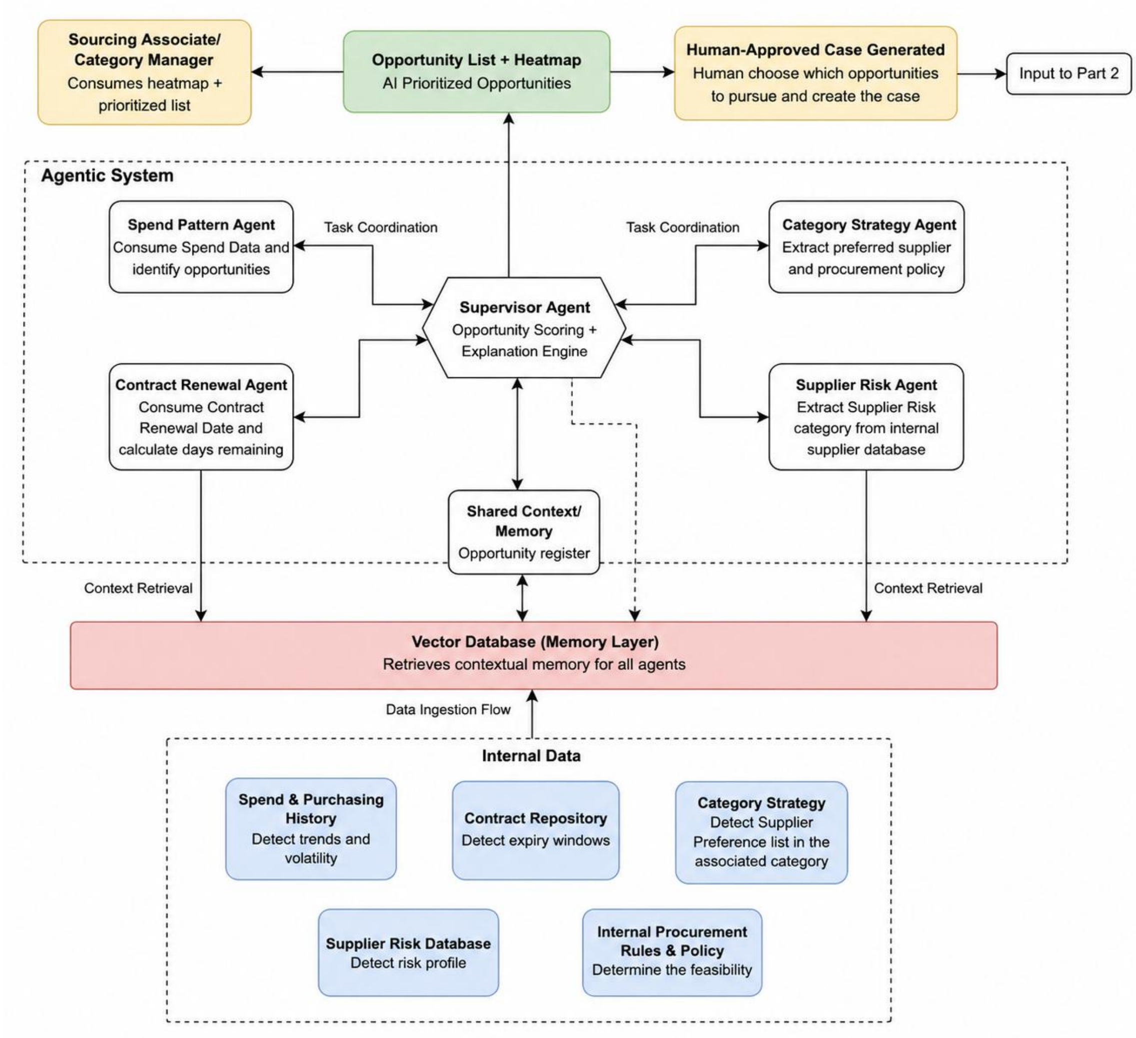

Figure 1: *Overview of the Agentic AI Architecture for Part 1*

3.1.3 Prioritization Framework

Before initiating any sourcing event, the system ranks competing opportunities to prioritize the highest-value, highest-risk, and most time-critical cases. Spreitzenbarth et al. (2024) identify spend analytics, contract monitoring, and supplier risk scoring as key AI/ ML use cases in procurement, while Mittal et al. (2024) highlight the role of AI in continuously surfacing these signals. Choudhary et al. (2023) establish weighted composite scoring as the dominant approach for supply chain prioritization. This framework operationalizes these insights into two Priority Score (PS) models: one for existing contracts and one for new sourcing requests.

Scoring Model for Existing Contracts

For contracts approaching expiration, five component scores capture timing risk, financial importance, operational risk, supplier dependency, and strategic alignment. Each is normalized to a 0–10 scale. The composite Priority Score is calculated using a weighted linear aggregation model (Choudhary et al., 2023). Weights are fully configurable by category, allowing procurement teams to tailor prioritization logic to the strategic priorities of each category. The general formulation is:

$$\mathbf{PS_{contract} = \alpha(EUS) + \beta(FIS) + \gamma(SRS) + \delta(SCS) + \varepsilon(SAS)}$$

Table 2 details each component score, description, and its default weight for IT Infrastructure category.

Component Score	Description	Default Weight (IT Infrastructure)
Expiry Urgency Score (EUS)	Captures timing risk using a step function based on months to expiry: 0–3 → 10; 3–6 → 8; 6–12 → 6; 12–18 → 4; >18 → 1. This reflects the non-linear rise in urgency as deadlines approach (Spreitzenbarth et al., 2024). EUS is the primary trigger for sourcing action.	α = 0.30
Financial Impact Score (FIS)	Normalizes contract value relative to the category maximum: *(Value / Max Value) × 10*. This preserves relative scale across contracts and enables comparability without recalibration (Choudhary et al., 2023). FIS ensures high-value contracts are prioritized alongside urgency.	β = 0.25
Supplier Risk Score (SRS)	Derived from internal supplier risk classifications and normalized to 0–10. It acts as an independent escalation factor, ensuring high-risk suppliers increase priority regardless of size or timing.	γ = 0.20
Spend Concentration Score (SCS)	Measures supplier dependency using share of category spend: >30% → 10; 20–30% → 7; 10–20% → 5; <10% → 2. High concentration increases switching risk and negotiation exposure. This dimension is monitored for threshold-based alerts.	δ = 0.15
Strategic Alignment Score (SAS)	Based on category strategy guidance and consistency with portfolio-based supplier segmentation frameworks (Kraljic, 1983): Preferred → 10; Allowed → 7; Non-preferred → 3. SAS serves as a secondary modifier.	ε = 0.10

Table 2: *Five component scores to derive the composite Priority Score for existing contracts*

Scoring Model for New Sourcing Requests

New sourcing requests lack an incumbent supplier, making SRS and SCS non-computable. The scoring model therefore substitutes implementation urgency and estimated spend as the primary drivers, supported by category importance and strategic alignment. As with the contract model, weights are fully configurable by category. The general formulation is:

$$\mathbf{PS_{new} = \alpha(IUS) + \beta(ES) + \gamma(CSIS) + \delta(SAS)}$$

Table 3 details each component score, description, and its default weight for IT Infrastructure category.

Component Score	Description	Default Weight (IT Infrastructure)
Implementation Urgency Score (IUS)	Mirrors EUS but uses the requested implementation timeline: <3 months → 10; 3–6 → 8; 6–12 → 6; >12 → 3. IUS reflects time-to-delivery as the primary triage factor (Spreitzenbarth et al., 2024).	α = 0.30
Estimated Spend Score (ES)	Normalizes estimated spend vs. pipeline maximum: (Estimated Spend / Max Spend) × 10. ES pairs with IUS as a co-equal driver, jointly accounting for 60% of the score without incumbent risk data (Choudhary et al., 2023).	β = 0.30

Category Spend Importance Score (CSIS)	Contextualizes the request within enterprise spend: (Category Spend / Max Category Spend) × 10. CSIS ensures requests in high-value categories are prioritized over similar requests in less critical areas.	$\gamma = 0.25$
Strategic Alignment Score (SAS)	Retains the same definition as in the contract model, acting as a qualifying modifier across both opportunity types.	$\delta = 0.15$

Table 3: *Four component scores to derive the composite Priority Score for new sourcing requests*

3.1.4 Model Properties and Integration of Agents

Both models share key properties: component scores are bounded [0,10], weights sum to 1.0, and the composite PS is also bounded [0,10] and increases monotonically, enabling direct ranking of opportunities without further transformation. Each score is mapped to a specific pre-DTP agent — for example, EUS to the Contract Renewal Agent and SRS to the Supplier Risk Agent. The Sourcing Signal Agent combines these scores to generate a prioritized pipeline for S2C execution. Scores are continuously updated as new data arrives, ensuring an always-current priority list — critical for sustaining trust in AI-driven procurement systems (Spreitzenbarth et al., 2024).

3.1.5 Performance Measurement Framework

System performance is assessed using two types of measures. Key Performance Indicators (KPIs) measure how well the system performs its core tasks. Key Learning Indicators (KLIs) measure how well the system improves over time based on reviewer feedback. Table 4 presents the KPIs and KLIs defined for Part 1, including descriptions and calculation methods.

Name	Description	How to Calculate
KPI: AI Reliability Rate (Prioritization Acceptance Rate)	The proportion of AI-generated sourcing opportunity rankings accepted by the sourcing team without changing the order. A high rate confirms the system is surfacing the right opportunities in the right sequence.	Opportunities accepted without re-ordering ÷ Total opportunities presented × 100. Measured each review cycle.
KPI: Signal Coverage Rate	The share of contracts and new requests for which all four required data inputs — contract expiry, spend, category strategy, and supplier risk — are available before scoring begins. Incomplete data prevents accurate prioritization.	Cases with all four signals present ÷ Total active cases × 100. Measured at each pipeline run.
KLI: Signal Attribution Accuracy	The share of prioritization decisions where the team confirms the system correctly explained why a case was ranked where it was. Supports trust and accountability in the AI's recommendations.	Cases where the stated ranking reason is confirmed correct ÷ Total cases reviewed × 100. Captured through structured reviewer feedback.

Table 4: *Performance Measures for Part 1: Sourcing Pipeline Generation*

3.2 Part 2: End-to-end Source-to-Contract (S2C) Execution

Once initiated in Part 1, a sourcing case enters the structured S2C workflow governed by DTP-01 through DTP-06. The system executes pathway selection, supplier identification and scoring, RFx preparation, negotiation support, award justification, contracting, and implementation reporting. Execution is coordinated by a Supervisor Agent, which enforces sequencing, validates inputs, and routes tasks to the appropriate agents.

3.2.1 Decision Modelling: S2C Execution

Each DTP phase involves explicit decisions supported analytically. Instead of treating procedures as narrative guidance, the project models them as structured decision nodes, defining required inputs, constraints, and the appropriate approach—deterministic logic, retrieval, scoring, or structured text generation. Table 5 summarizes the core decision, required inputs, constraints, and appropriate mechanism for each DTP phase.

Phase	Core Decision	Required Inputs	Constraints	Appropriate Mechanism
DTP 01	Select sourcing pathway	Spend, expiry, scope changes, preferred suppliers	Competitive bid policy, renewal criteria	Deterministic rule logic + LLM classification
DTP 02	Identify supplier pool and define evaluation structure	Supplier data, capabilities, performance, category strategy, KPIs	Risk flags, approved vendors, mandatory dimensions	Retrieval-augmented search + rule validation + LLM support
DTP 03	Score suppliers and generate RFx package	Service Level Agreement (SLA) history, incidents, risk indicators, templates	Cutoff thresholds, template completeness	ML-based scoring + rule validation + template-driven LLM
DTP 04	Recommend negotiation approach	Bids, benchmarks, risk profile	Category policy on negotiation types	Heuristic rules + ML-based anomaly detection
DTP 05	Prepare award justification	Scores, bids, scenario comparison	Compliance documentation needs	Rule-based comparison + LLM rationale generation
DTP 06	Produce contracting and reporting summaries	Award terms, SLAs, pricing	Required contracting and reporting fields	LLM extraction + rule validation

Table 5: *Core decision, required inputs, constraints, and appropriate mechanisms (Part 2)*

3.2.2 Architecture and Design of Agents

Seven agents execute the S2C workflow, each aligned to specific DTP phases. The Supervisor Agent coordinates all of them, enforcing sequencing and compliance throughout.

- **Supervisor Agent (DTP-01–06):** Orchestrates the S2C workflow, validates inputs, selects pathways, and coordinates agents using routing rules, context memory, and procedural retrieval.
- **Sourcing Signal Agent (DTP-01–02):** Consolidates upstream signals (spend, contract expirations, risk, and strategy) to trigger sourcing or renewal cases using signal fusion, prioritization, and rule-based event logic.
- **Supplier Scoring Agent (DTP-02–03):** Translates evaluation criteria into scores using historical data, deterministic logic, normalization, and eligibility checks.
- **RFx Draft Agent (DTP-03):** Builds RFx drafts using templates, past examples, and structured generation with completeness checks.
- **Negotiation Support Agent (DTP-04):** Analyzes bids, highlights differences, and identifies negotiation levers using comparisons, anomaly detection, and benchmarks.

- **Contract Support Agent (DTP-04–05):** Extracts key terms and prepares structured inputs using template-based extraction and validation rules.
- **Implementation Agent (DTP-05–06):** Generates rollout plans and post-award indicators using calculations, playbooks, and structured reporting.

Figure 2 provides an overview of the agentic AI architecture for Part 2, showing how the seven agents are aligned across the S2C workflow.

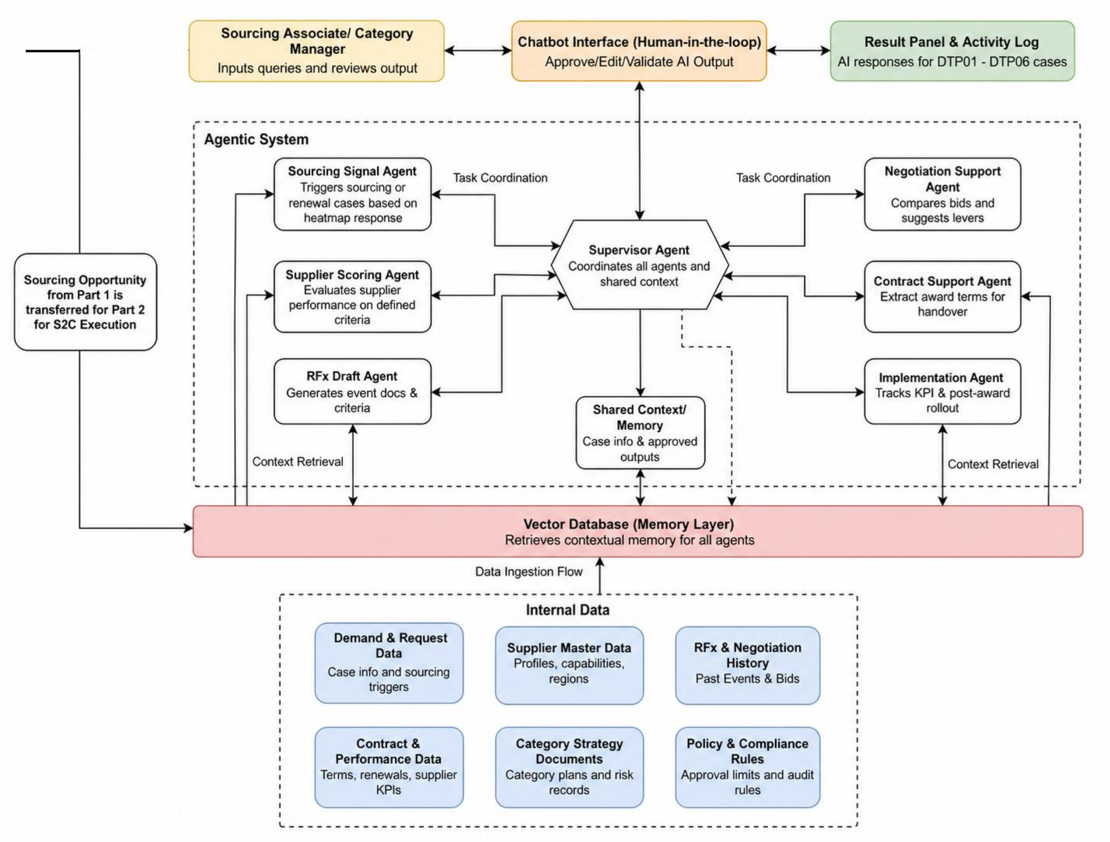

Figure 2: *Overview of the Agentic AI Architecture for Part 2*

All agents operate on two shared layers: a vector-based memory layer for contextual retrieval and a structured data layer powered by an Extract, Transform, and Load (ETL) pipeline. A human-in-the-loop interface enables sourcing professionals to review, edit, or approve all outputs before finalization.

3.2.3 Performance Measurement Framework

System performance is assessed using KPIs and KLIs, consistent with the framework defined in Part 1. Table 6 presents the KPIs and KLIs defined for Part 2, including descriptions and calculation methods.

Name	Description	How to Calculate
KPI: First-Pass Acceptance Rate	Share of AI-generated outputs (shortlists, RFx drafts, evaluations) accepted with little or no editing, indicating output quality across the S2C workflow.	Outputs accepted with no or minor edits ÷ Total outputs reviewed × 100. Measured per sourcing stage.
KPI: Stage-Level Cycle Time Reduction	Reduction in time to complete each S2C stage with AI support compared to manual baseline, indicating operational impact.	Average time with AI − Average baseline time without AI, per stage. Baseline drawn from historical sourcing records.
KLI: Edit Density by Type	Volume and type of edits (factual, structural, wording) made to AI outputs, indicating whether issues are fundamental or surface-level.	Total edits ÷ Total outputs reviewed, split by type: factual, structural, wording. Logged through the reviewer interface each cycle.
KLI: Knowledge Base Retrieval Relevance Rate	Share of retrieved documents and data considered relevant by reviewers, indicating knowledge base quality and coverage.	Retrieved items rated relevant ÷ Total retrieved items × 100. Collected via a relevance rating in the reviewer interface, per stage.

Table 6. *Performance Measures for Part 2: Source-to-Contract Execution*

3.3 Analytical Model Design and Learning Architecture

Analytical models in both parts use four core decision types consistently across agents and data contexts.

- **Heuristic Rule Logic:** In Part 1, rules govern contract expiry, spend alerts, and risk flags. In Part 2, they guide routing, evaluation, and compliance. Overrides highlight where rules need refinement.
- **Retrieval and Knowledge-Grounded Reasoning:** In Part 1, retrieval supports policy and risk insights. In Part 2, it enables access to supplier data, benchmarks, RFx content, and category strategies. User edits help improve retrieval relevance.
- **Analytical Scoring Logic:** In Part 1, scoring ranks opportunities by urgency, spend, risk, and strategy. In Part 2, scoring applies normalized metrics and price anomaly detection, standardizing evaluation.
- **Structured Language Generation:** Primarily in Part 2, agents generate RFx drafts, summaries, contract inputs, and reports using templates and retrieved data. Outputs are reviewed to refine retrieval.

Across both parts, a shared learning loop captures human validation signals—edits, overrides, and flags—to drive rule updates, retrieval improvements, scoring recalibration, and prompt tuning. KPIs and KLIs govern performance while identifying areas needing stronger control or human intervention.

3.4 Data Collection and Preparation

Data collection supports both Part 1 (Opportunity Generation) and Part 2 (S2C Execution) using anonymized Sponsor data from historical sourcing events. The proof-of-concept operates fully offline, without integration to live systems. Data collection and preparation include the following steps:

- **Data preparation using the following data sources:**
    - **Part 1:** Contract metadata, IT Infrastructure spend records, category strategy documents, and supplier risk profiles to identify sourcing triggers.
    - **Part 2:** Purchase requests, supplier master data, RFx and negotiation history, bid records, SLA performance data, and market intelligence summaries to support end-to-end execution.

- **Data processing:** All data are cleansed, de-identified, and normalized into a unified schema to ensure consistency across both parts.
- **Retrieval and knowledge access:** Documents are embedded using text models and stored in a vector database, enabling agents to retrieve DTP rules, supplier histories, category strategies, and risk data through semantic search.
- **Validation and completeness:** Rule-based checks ensure data quality (e.g., supplier completeness, scoring consistency, valid ranges). Missing data is supplemented with controlled synthetic values to maintain logical integrity.

The structured foundation ensures consistent inputs, reliable agent decisions, and scalability for future integration and expansion.

4 RESULTS AND DISCUSSION

This section presents simulated evaluation results for Part 1 (ten test cases) and Part 2 (one representative S2C case). Each sub-section summarizes the test design and corresponding KPI/KLI results based on simulated reviewer interactions using the datasets from Section 3.

4.1 Part 1: Prioritized Sourcing Opportunities

Part 1 was evaluated using ten test cases on a 100-row simulated dataset with ~70% contract renewals and ~30% new sourcing requests. Two project members acted as sourcing analysts across two testing rounds, logging reviewer actions for each output. Before Round 2, numeric validation and mandatory field pre-screening were added to reduce false-pass rates on defective inputs.

4.1.1 Test Cases

Seven cases used clean data, while three introduced invalid inputs to test resilience. The rounds evaluated outputs and reviewer responses, not agent training. In live deployment, reviewer feedback would continuously refine the system. Table 7 summarizes the ten test cases.

ID	Title	Type	Scenario Description	Expected Output and Reviewer Action
TC-01	Standard Renewal Pipeline	Business	A mix of contract renewals and new sourcing requests with complete and valid data.	Opportunities are ranked by composite score, with higher urgency prioritized. Outputs are accepted without modification.
TC-02	New Requests Only	Business	A set of new sourcing requests without existing contracts or historical references.	Opportunities are ranked based on urgency and estimated value. Reviewers may adjust rankings based on contextual priorities.
TC-03	Urgency vs. Value Trade-off	Business	Cases present trade-offs between urgency and financial importance.	More urgent cases are prioritized over higher-value but less time-sensitive ones. Rankings are validated or adjusted by reviewers.
TC-04	Non-Preferred Suppliers Only	Business	All cases involve suppliers that are not aligned with preferred sourcing strategy.	Opportunities receive lower scores with compliance considerations flagged. Reviewers request alternative sourcing options.
TC-05	High Value, Low Urgency	Business	Cases involve high financial importance but limited time sensitivity.	Lower prioritization due to reduced urgency. Reviewers may flag cases for early planning.

ID	Title	Type	Scenario Description	Expected Output and Reviewer Action
TC-06	Multi-Category Mix	Business	Opportunities span multiple categories with varying characteristics.	Ranking is based on composite scoring across categories. Outputs are validated for consistency.
TC-07	Mixed Supplier and Urgency Conditions	Business	Cases include a mix of supplier types and varying urgency levels.	Higher priority is assigned to urgent and strategically aligned cases. Reviewers approve or flag for further review.
TC-08	Invalid Data Format	Technical	Certain inputs contain improperly formatted or inconsistent values.	Affected cases are flagged and deprioritized. Reviewers reject outputs and request data correction.
TC-09	Missing Mandatory Data	Technical	Some records lack required input fields necessary for evaluation.	Incomplete cases result in unreliable rankings. Reviewers request missing data before re-evaluation.
TC-10	Invalid or Inconsistent Values	Technical	Inputs include logically inconsistent or unrealistic values.	Rankings are adversely impacted. Reviewers reject outputs and provide corrective feedback.

Table 7: *Part 1 Test Cases*

4.1.2 Test Results

Part 1 measures two KPIs — AI Reliability Rate and Signal Coverage Rate — and one KLI, Signal Attribution Accuracy, across all ten scenarios. Table 8 summarizes the results.

ID	Scenario	Type	AI Reliability Rate (KPI)	Signal Coverage Rate (KPI)	Signal Attribution Accuracy (KLI)
TC-01	Standard Renewal Pipeline	Business	92%	100%	94%
TC-02	New Requests Only	Business	78%	100%	82%
TC-03	Urgency Outweighs Spend	Business	85%	100%	89%
TC-04	Non-Preferred Suppliers Only	Business	71%	100%	76%
TC-05	High Spend, Distant Expiry	Business	83%	100%	87%
TC-06	Low-Spend Multi-Category Mix	Business	76%	100%	80%
TC-07	Mixed Supplier Status and Urgency	Business	88%	100%	91%
TC-08	Text in Numeric Fields	Technical	41%	96%	48%
TC-09	Missing Mandatory Fields	Technical	38%	96%	44%
TC-10	Invalid / Impossible Values	Technical	35%	100%	42%
Overall Average			**68.7%**	**99.2%**	**73.3%**

Table 8. *Part 1 KPI and KLI Results by Test Scenario*

Figure 3 visualizes the KPI and KLI results across all ten scenarios, with business and technical scenarios shown separately against the 80% target threshold.

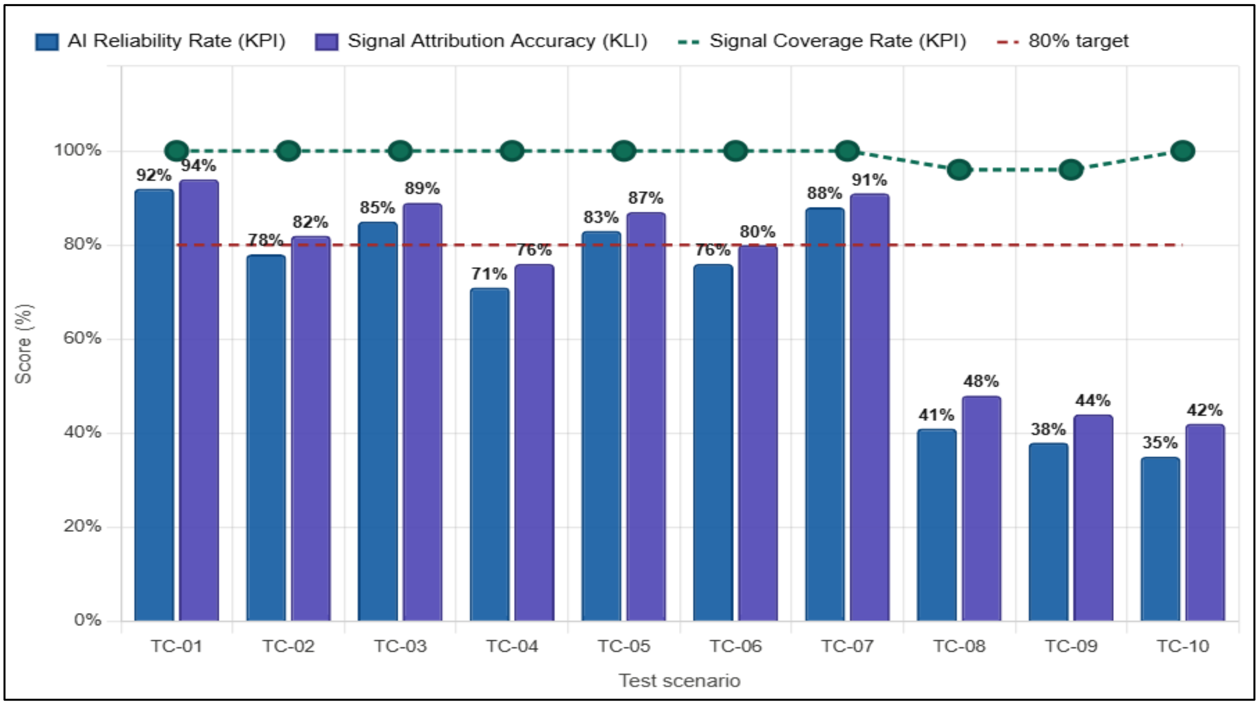

Figure 3: *Part 1 KPI and KLI results across all ten scenarios. AI Reliability / Prioritization Acceptance Rate and Signal Attribution Accuracy are shown side by side, with Signal Coverage Rate (dashed green line) overlaid. Business scenarios meet or exceed the 80% target (dashed red line), while technical scenarios fall below, highlighting the role of the human-in-the-loop mechanism in identifying defective outputs.*

AI Reliability Rate (Prioritization Acceptance Rate). Business scenarios averaged 81.9%, meeting the 80% target under clean data conditions. Lower results appeared in cases requiring contextual judgment. Technical scenarios averaged 38.0%, reflecting performance degradation due to invalid inputs; reviewers consistently identified and rejected these outputs. The overall average was 68.7%, reflecting sensitivity to data quality rather than model limitations.

Signal Coverage Rate. Business scenarios achieved full coverage, confirming complete datasets under normal conditions. Technical scenarios showed minor reductions due to missing or malformed inputs, while some inconsistencies remained undetected, highlighting the need for stronger validation. The overall average was 99.2%.

Signal Attribution Accuracy. Business scenarios averaged 85.6% showing strong alignment between system explanations and reviewer expectations, while technical scenarios averaged 44.7% due to defective inputs causing incorrect signal attribution. The overall average was 73.3%. Lower results in TC-08, TC-09, and TC-10 reflect expected degradation under poor data conditions rather than model failure. Section 5.1 outlines mitigation steps, including mandatory fields, input validation, and upstream pre-processing checks for data type, completeness, and range before scoring.

4.2 Part 2: Source-to-Contract Execution

Part 2 is evaluated using a single representative test case (TC-P2-02) that exercises all six S2C execution agents across DTP-01 to DTP-06.

4.2.1 Test Case

The test case scenario reflects a full competitive sourcing pathway without a preferred supplier, requiring sequential generation of a supplier shortlist, RFx package, evaluation, negotiation summary, award justification, and contract draft. Table 9 describes the test case used for Part 2 evaluation, covering inputs, expected outputs, and reviewer actions.

Field	Detail
Test Case ID	TC-P2-02
Title	Competitive RFx with Multi-Supplier Evaluation
Type	Business
Sourcing Pathway	Full competitive bid process covering DTP-01 to DTP-06
Scenario	A new sourcing request does not qualify for direct renewal or a preferred supplier waiver. The system identifies eligible suppliers, generates an RFx package, evaluates responses, produces a negotiation summary, drafts an award justification, and prepares a contract summary. This scenario covers the full S2C pathway within the POC scope.
Data Inputs	Sourcing request details (spend, subcategory, timeline), supplier master data, category strategy, historical RFx templates, prior bid and SLA data, and contract templates for summary generation.
Expected System Output	A complete and sequenced set of S2C outputs: (1) a supplier shortlist with eligibility rationale; (2) a structured RFx package with evaluation matrix; (3) a scored supplier comparison table; (4) a negotiation summary with bid deltas and pricing anomalies flagged. All outputs are produced in the correct DTP sequence with no missing stages.
Reviewer Actions	Reviewer approves shortlist and RFx with minor edits, adjusts evaluation weightings, validates negotiation insights, and finalizes award justification and contract summary.
Key Test Conditions	All inputs are complete and valid. No preferred supplier is available, requiring a competitive process. Evaluation criteria are retrieved from category strategy, with no data quality defects introduced.

Table 9. *Part 2 Test Case: TC-P2-02 — Competitive RFx with Multi-Supplier Evaluation*

4.2.2 Test Results

Part 2 results cover the four rate-based measures: First-Pass Acceptance Rate (KPI), Stage-Level Cycle Time Reduction (KPI), Knowledge Base Retrieval Relevance Rate (KLI), and Edit Density by Type (KLI). Table 10 presents the rate-based KPI and KLI results for TC-P2-02.

Measure	Type	Result
First-Pass Acceptance Rate	KPI	78%

Measure	Type	Result
Stage-Level Cycle Time Reduction	KPI	See note [1]
Knowledge Base Retrieval Relevance Rate	KLI	85%

Table 10. *Part 2 KPI and KLI Rate Results — TC-P2-02*

[1] Based on historical sourcing data provided by the Sponsor, the average end-to-end sourcing cycle time for DTP-01 to DTP-06 is 71 calendar days. With the agent-based platform, DTP-01 to DTP-04 (pathway selection through negotiation prep) is expected to be completed in 28–35 days. DTP-05 and DTP-06 (negotiation and contract award) may extend the overall timeline by an additional 7–14 days assuming straightforward negotiations.

Table 11 presents the edit density breakdown by type for TC-P2-02.

Measure	Type	Factual Corrections	Structural Changes	Wording Refinements
Edit Density by Type	KLI	11%	32%	27%

Table 11. *Part 2 Edit Density by Type — TC-P2-02*

First-Pass Acceptance Rate. TC-P2-02 achieved a 78% First-Pass Acceptance Rate, slightly below the 80% target. Most outputs were accepted with minor edits, while the evaluation matrix required structural adjustments and the award justification needed light refinement. The result reflects the complexity of a full competitive sourcing pathway rather than a fundamental issue in output quality.

Stage-Level Cycle Time Reduction. Against a baseline of 71 calendar days for a full end-to-end sourcing cycle (DTP-01 to DTP-06), the agent-based system is estimated to deliver a 35–50% overall cycle time reduction, bringing the total to approximately 35–49 days. This estimate is grounded in stage-level assumptions applied consistently across the sourcing workflow:

- **DTP-01 to DTP-04 (pathway selection, supplier identification, RFx preparation, and negotiation support):** AI-assisted routing, automated signal synthesis, and template-driven drafting reduce these stages from an estimated 50 days to 28–35 days, as agents consolidate contract, spend, risk, and strategy inputs that previously required significant manual effort.
- **DTP-05 to DTP-06 (negotiation and contract award):** Assuming straightforward negotiations, these stages are estimated at 7–14 days — significantly shorter than traditional timelines when suppliers are well-prepared and evaluation outputs are AI-generated and pre-validated.

The combined estimate yields a total cycle time of 35–49 days, representing a **35–50% reduction** against the 71-day baseline. This will be validated and refined against actual sourcing records in a live pilot.

Edit Density by Type. Structural edits were the largest share (32%), driven by changes in evaluation weighting. Wording edits (27%) were concentrated in RFx and justification sections, while factual corrections (11%) were minimal, indicating strong accuracy but a need for structural refinement.

Knowledge Base Retrieval Relevance Rate. The relevance rate was 85%, indicating effective retrieval overall. Irrelevant results were concentrated in the evaluation stage due to partial indexing of category strategy documents, highlighting the need for targeted knowledge base improvements.

4.3 Discussion

Three consistent findings emerge across both parts of the system.

- **The system performs reliably under clean business conditions.** Part 1 scenarios achieved an AI Reliability Rate (Prioritization Acceptance Rate) of 81.9% and Signal Attribution Accuracy of 85.6%, both meeting or exceeding the 80% target, while TC-P2-02 reached 78%, slightly below threshold. Reviewer edits were primarily structural and wording-related, indicating that the scoring logic, retrieval architecture, and generation pipeline are fundamentally sound.
- **Data quality is the primary risk factor, driving performance variability more than model capability.** Technical scenarios in Part 1 showed significant degradation across all KPIs and the KLI due to malformed or missing inputs, consistent with Cannas et al. (2024). In Part 2, retrieval gaps and structural edits were driven by incomplete knowledge base indexing rather than model limitations.
- **The human-in-the-loop mechanism serves as both a safeguard and a learning engine.** Reviewers consistently identified and rejected anomalous outputs in technical scenarios, preventing defective prioritizations from propagating into execution. Importantly, reviewer flags on missing fields, incorrect values, or implausible scores are routed back to data owners as structured correction signals — dynamically improving the underlying data quality over time, not just compensating for it.

This proof-of-concept answers all four of its core questions. Sourcing pipelines become proactive through real-time monitoring of contract expiry, spend, category strategy, and supplier risk — ranked by five specialized agents. The system is structured as twelve agents across the full sourcing cycle, with a Supervisor Agent enforcing sequencing and humans validating every output. It runs on contract records, spend data, supplier risk profiles, category strategies, and RFx history — held in a shared knowledge layer. Progress is measured through KPIs (81.9% AI Reliability, 99.2% Signal Coverage, 78% First-Pass Acceptance) and KLIs, with the architecture designed to scale beyond IT Infrastructure through weight recalibration, knowledge base expansion, and category-specific configuration.

4.4 Limitations

The POC operates in a simulated, offline environment using anonymized historical data. KPI and KLI values are based on structured test scenarios rather than live evaluations with sourcing professionals. The Part 1 scoring model is calibrated for the IT Infrastructure category, and Part 2 agents are configured for a single competitive scenario; extension to other categories and pathways requires recalibration, knowledge base updates, and additional testing. These constraints are consistent with a proof-of-concept and limit generalizability beyond controlled scenarios.

5 RECOMMENDATIONS

The proof-of-concept indicates potential improvements in sourcing efficiency, contingent on data quality and governance maturity. However, performance is highly dependent on data quality, validation rigor, and knowledge coverage.

5.1 Management Recommendations

The following recommendations support the transition from POC to pilot while strengthening data quality, governance, and evaluation to enable scalable and reliable deployment.

- **Advance to Pilot with Data Governance:** System performance is highly dependent on data quality. Enforce mandatory fields, apply input validation at entry, and assign ownership for key data sources. Initiate a pilot in selected high-quality categories using POC KPIs as benchmarks.
- **Strengthen Preferred Supplier Data:** Gaps in supplier classification increased review effort. Conduct structured data cleanup and maintain a machine-readable repository aligned with category strategy.
- **Implement Pre-Processing Validation:** Shift validation upstream by enforcing checks on data type, completeness, and range prior to scoring. Route invalid records back to data owners for correction.
- **Re-index Knowledge Base:** Address retrieval gaps by improving indexing of category strategies, supplier data, and supporting templates to enhance relevance.
- **Embed Evaluation Framework:** Integrate KPIs and KLIs from the outset. Establish baselines, track reviewer edits, and implement governance processes to monitor performance and refine model parameters over time.

5.2 Future Work

Future work should extend the proof-of-concept along four dimensions:

- **Live system integration:** The system should be connected to live contract, spend, and risk data sources through secure application programming interfaces (APIs). It includes implementing data refresh mechanisms and conflict resolution protocols to support ongoing review cycles.
- **Category and pathway expansion:** Future work should develop a structured onboarding framework to extend the system across additional procurement categories and sourcing pathways, including direct renewals, sole-source justifications, and framework-based sourcing.
- **Adaptive scoring:** Further research should explore the application of lightweight reinforcement learning techniques to dynamically adjust scoring weights based on accumulated reviewer feedback, thereby reducing reliance on manual model tuning.
- **Agent training and feedback-driven improvement:** A critical distinction exists between the test-based evaluation conducted in this POC and the training of agents in a production setting. The current evaluation assesses output quality through structured reviewer interactions — it does not retrain the agents. In a live deployment, reviewer actions such as overrides, edits, and flags would serve as continuous training signals, enabling scoring weights to be recalibrated, knowledge bases to be re-indexed, and generation prompts to be refined with each sourcing cycle. Establishing a governed feedback pipeline — defining how signals are captured, validated, and used to update agent behavior — is a critical next step before enterprise rollout.
- **Governance and change management:** A formal governance framework should be established to support adoption, including mechanisms to build user trust, define escalation protocols for human-only decisions, and ensure regular audit of system activity logs in alignment with responsible AI principles. Critically, change management efforts should specifically address algorithm aversion — the tendency of decision-makers to distrust or override algorithmic recommendations even when they are well-calibrated. Training programs should help sourcing professionals understand how to add value to the system through quality feedback rather than wholesale rejection of AI outputs, reinforcing a collaborative human-AI dynamic rather than a substitution narrative.

REFERENCES

Accenture. (2025). *Learning, Reinvented: Accelerating Human-AI Collaboration*. Retrieved from https://www.accenture.com/us-en/insights/consulting/learning-reinvented-accelerating-human-ai-collaboration

Cannas, V. G., Ciano, M. P., Saltalamacchia, M., & Secchi, R. (2024). Artificial intelligence in supply chain and operations management: A multiple case study research. *International Journal of Production Research*, 62(9), 3333–3360. https://doi.org/10.1080/00207543.2023.2232050

Choudhary, N. A., Singh, S., Schoenherr, T., & Ramkumar, M. (2023). Risk assessment in supply chains: A state-of-the-art review of methodologies and their applications. *Annals of Operations Research*, 322(2), 565–607. https://doi.org/10.1007/s10479-022-04700-9

Guida, M., Caniato, F., Moretto, A., & Ronchi, S. (2023). The role of artificial intelligence in the procurement process: State of the art and research agenda. *Journal of Purchasing and Supply Management*, 29(2), 100823. doi:10.1016/j.pursup.2022.100823

Keelvar. (2025). *The Rise of AI Agents in Enterprise Procurement*. Retrieved from https://www.keelvar.com/documents/the-rise-of-ai-agents-in-enterprise-procurement

Kiron, D., Ransbotham, S., Reeves, M., & Gerbert, P. (2020). Designing AI systems with human-machine teams. *MIT Sloan Management Review*. Retrieved from https://sloanreview.mit.edu/article/designing-ai-systems-with-human-machine-teams/

Kraljic, P. (1983). Purchasing must become supply management. *Harvard Business Review*, 61(5), 109–117.

Mittal, A., Cocoual, C., Erriquez, M., & Liakopoulou, T. (2024, June 13). Revolutionizing procurement: Leveraging data and AI for strategic advantage. *McKinsey & Company*. Retrieved from https://www.mckinsey.com/capabilities/operations/our-insights/revolutionizing-procurement-leveraging-data-and-ai-for-strategic-advantage

Monczka, R. M., Handfield, R. B., Giunipero, L. C., & Patterson, J. L. (2020). *Purchasing and Supply Chain Management* (7th ed.). Boston: Cengage Learning.

Revilla, E., & Saenz, M. J. (2025, July 24). How AI is reshaping supplier negotiations. *Harvard Business Review*. Retrieved from https://hbr.org/2025/07/how-ai-is-reshaping-supplier-negotiations

Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Hoboken, NJ: Pearson.

Saenz, M. J., Borrella, I., & Revilla, E. (2022, March/April). Digital supply chain transformation: Aligning operations and strategy. *Supply Chain Management Review*.

SAP. (n.d.). What is sourcing? Retrieved from https://www.sap.com/resources/what-is-sourcing

Seshadri, S. (2005). *Sourcing Strategy: Principles, Policy and Designs*. New York: Springer.

Spreitzenbarth, J. M., Bode, C., & Stuckenschmidt, H. (2024). Artificial intelligence and machine learning in purchasing and supply management: A mixed-methods review of the state-of-the-art in literature and practice. *Journal of Purchasing and Supply Management*, 30(1), 100896. https://doi.org/10.1016/j.pursup.2024.100896

Wamba, S. F., Gunasekaran, A., Dubey, R., & Ngai, E. W. T. (2018). Big data analytics in operations and supply chain management. *Annals of Operations Research*, 270(1), 1–4. https://doi.org/10.1007/s10479-018-3024-7