The International Journal of Advanced Manufacturing Technology
https://doi.org/10.1007/s00170-026-18700-7

ORIGINAL ARTICLE

Rethinking manufacturing agility assessment in the age of intelligent systems: a multi-agent skill-enhanced RAG framework for the digital manufacturing era

Jun Ren[1]

Received: 20 March 2026 / Accepted: 5 July 2026

Abstract
Manufacturing agility has been a strategic imperative since its formalisation at the Iacocca Institute in 1991. Over three decades, researchers have developed attribute-based frameworks that identify the organisational, technological, and relational dimensions of an agile enterprise. Yet the methods used to assess those dimensions have remained largely unchanged. Data collection still relies on surveys and site visits. Weights are still estimated through expert pairwise comparison. Synthesis still produces a single, static report. This paper argues that these methods are no longer adequate for the volatile, digitally interconnected manufacturing landscape of the mid-2020s. The response is not to apply a new computational tool to an old assessment framework. That approach merely replaces the surface while preserving the underlying limitations. Instead, this paper proposes an evolved framework with a new architecture: a Multi-Agent Skill-Enhanced Retrieval-Augmented Generation (RAG) system in which agility knowledge is encoded in versioned, learnable Skill modules, knowledge acquisition is automated through grounded document retrieval, weight calibration is performed by large language model (LLM) agents, and synthesis preserves formal Dempster-Shafer evidential reasoning throughout. Four agility domains are selected and updated for the Industry 4.0 era: Technology, Partnership, Market, and Change. Each is extended with a new attribute not captured in prior frameworks. The result is a system that learns continuously, remains auditable, and is deployable by manufacturing firms of all sizes. Small and medium enterprises (SMEs) stand to benefit most.

Keywords Agile manufacturing · Multi-agent systems · Retrieval-augmented generation · Fuzzy inference · Dempster-Shafer theory · Industry 4.0 · SMEs

1 Introduction

The term “agile manufacturing” entered production management in 1991, when researchers at the Iacocca Institute, Lehigh University, proposed agility as a new approach to surviving and prospering in an environment of continuous and unpredictable change [1]. Since then, a substantial body of research has developed the conceptual and empirical foundations of agile manufacturing: defining its constituent attributes, measuring its effects on competitive capabilities, and building decision-support tools to help enterprises become more agile [2–5].

These contributions were significant. The identification of 32 agile attributes organised into ten decision domains gave practitioners and researchers a structured vocabulary for agility. The application of artificial neural networks (ANNs) to identify which attributes most drive competitive priorities, namely speed, cost, quality, innovation, flexibility, and proactivity, provided empirical grounding for strategic investment decisions [4]. The development of evidential reasoning frameworks for agile partner selection introduced formal uncertainty quantification into supply chain agility decisions [6]. Together, these works established a rigorous and practically oriented research tradition.

Yet a fundamental limitation runs through all of them. The assessment process itself has never been automated.

✉ Jun Ren
j.ren@ljmu.ac.uk

[1] School of Engineering and Built Environment, Liverpool John Moores University, Liverpool, UK

Every framework, however sophisticated its reasoning core, depends on manual data collection through questionnaires, site visits, and structured interviews. Weights are estimated by hand, through pairwise comparison matrices. Synthesis produces a one-time report that cannot adapt as an enterprise evolves. In a manufacturing environment defined by volatility, disruption, and digital transformation, this labour-intensive assessment model is no longer adequate.

A second limitation has become visible more recently. The manufacturing challenges of the mid-2020s differ qualitatively from those of the early 2000s when the core attribute frameworks were developed. Supply chain resilience has been redefined by the COVID-19 pandemic and geopolitical fragmentation. Technology agility now encompasses AI adoption, digital twins, and cyber-physical integration. These dimensions were absent from frameworks developed before Industry 4.0. Organisational agility increasingly means the capacity to absorb and operationalise digital transformation, not merely to empower teams and decentralise decisions. A framework that maps today's agility challenges onto attributes designed for yesterday's environment risks being structurally sound but contextually stale.

This paper responds to both limitations through a single integrated proposal. Rather than applying a new computational tool to the existing agility assessment framework, which would preserve its underlying weaknesses under a more sophisticated surface, this paper proposes an evolved framework with a new architecture: a Multi-Agent Skill-Enhanced Retrieval-Augmented Generation (RAG) system in which:

- Agility knowledge is encoded in versioned, learnable Skill modules, rather than static rule bases or fixed survey instruments.
- Knowledge acquisition is automated through grounded document retrieval, replacing manual data collection.
- Criterion weight calibration is performed by LLM agents drawing on retrieved evidence, replacing expert-driven pairwise comparison.
- Synthesis is conducted by a Dempster-Shafer evidential reasoning agent that preserves formal uncertainty quantification.
- The system learns continuously through an expert-gated feedback loop, rather than producing static one-shot assessments.

Four agility domains are selected from the established ten-domain framework: Technology, Partnership, Market, and Change. Each is updated for the Industry 4.0 era and extended with a new attribute that captures dimensions not present in prior work. This selective evolution is itself a methodological contribution. The paper does not claim that the entire established framework is superseded. It demonstrates how its most enduring and data-rich components can be integrated into a new architecture that is simultaneously grounded in prior scholarship and genuinely new.

The remainder of the paper is organised as follows. Section 2 reviews the agile manufacturing literature and identifies the research gap. Section 3 presents the Multi-Agent Skill-Enhanced RAG framework. Section 4 maps four evolved agility domains onto the MAS architecture. Section 5 explains how the resulting system transforms agility assessment and enhancement. Section 6 provides an illustrative application tracing the Partnership domain assessment for a representative supplier using published data from Ren et al., [6]. Section 7 compares this approach with existing alternatives. Section 8 discusses theoretical and practical implications. Section 9 concludes.

2 Background and research gap

2.1 Agile manufacturing: from the Iacocca institute to the industry 4.0 era

The concept of agile manufacturing emerged from a recognition that the lean and flexible manufacturing approaches of the 1980 s, while effective in stable environments, were insufficient for markets characterised by rapid change, mass customisation, and intensifying global competition [7]. The Iacocca Institute's foundational report defined agility as "the capability of surviving and prospering in a competitive environment of continuous and unpredictable change by reacting quickly and effectively to changing markets, driven by customer-designed products and services" [3].

Goldman et al., [2] subsequently identified four dimensions of agility: delivering value to customers, being ready for change, valuing human knowledge and skills, and forming virtual partnerships. These became influential organising concepts. Yusuf et al., [5] operationalised them into 32 agile attributes, providing a systematic basis for empirical investigation. Ren et al., [4] applied artificial neural networks to a survey of 105 European manufacturing companies to identify which attributes most strongly influence each of six competitive priorities: speed, cost, quality, innovation, flexibility, and proactivity. They found, for example, that strategic relationships with customers is the dominant driver of speed, and that enterprise integration most strongly predicts flexibility.

These empirical findings were extended into decision-support territory by Ren et al., [6], who proposed a framework for agile enterprise partnering. That framework organises the 32 attributes into ten decision domains: Integration, Competence, Team Building, Technology, Quality, Change, Partnership, Market, Education, and Welfare. It uses evidential reasoning based on Dempster-Shafer theory to evaluate and rank potential supply chain partners under conditions of incomplete and vague data.

Alongside this research tradition, the emergence of Industry 4.0 has introduced new dimensions of manufacturing agility that earlier frameworks did not anticipate. The integration of cyber-physical systems, the Internet of Things, digital twins, and artificial intelligence into production environments has created new agility requirements. Firms must now adapt not only their organisational structures and supply chain relationships, but their capacity to adopt and exploit digital technologies at speed [8, 9]. The COVID-19 pandemic further exposed supply chain vulnerabilities that have reshaped the meaning of partner agility, placing disruption resilience at the centre of supply chain strategy [10]. A contemporary agility framework must be both methodologically modernised and contextually updated.

2.2 The established agile attribute framework

The ten-domain, 32-attribute framework developed across the prior works of Ren, Yusuf, and Burns provides the conceptual foundation for the present paper. The ten domains, ranging from Integration to Welfare, organise the 32 attributes according to their inherent characteristics and cover the major decision areas that determine a manufacturing firm's agility. Each domain is operationally defined through its constituent criteria. The Technology domain, for example, encompasses technology awareness, leadership in the use of current technology, skill and knowledge enhancing technologies, and flexible production technology.

This framework has two enduring strengths that justify its use as a foundation here. First, it is empirically grounded. The attributes and domains were derived from comprehensive literature review, expert interviews, and large-scale empirical survey, giving them validated practical meaning. Second, it is structurally hierarchical. The three-level hierarchy, with overall agility at the top, ten domains at the intermediate level, and 32 criteria at the base, maps naturally onto a multi-agent architecture in which each domain can be assigned to a dedicated reasoning agent.

The framework also has two limitations that this paper addresses directly. First, it was developed in a pre-Industry 4.0 context and does not explicitly capture agility dimensions that have become critical since 2010, including AI adoption readiness, supply chain disruption resilience, mass customisation responsiveness, and digital transformation absorptive capacity. Second, its operational deployment has always depended on manual processes: survey instruments, site visits, and expert pairwise comparison. These are resource-intensive, time-consuming, and largely inaccessible to SMEs.

2.3 Limitations of existing agility assessment methods

The methodological limitations of existing agility assessment approaches fall along four dimensions.

Data collection All established frameworks rely on primary data collected through questionnaires, structured interviews, or site visits. This process is slow. A full supplier evaluation in Ren et al., [6] involved multiple rounds of site visits and management interviews. The resulting data is expensive to collect, difficult to update, and dependent on the willingness and availability of respondents. Agility assessments are therefore conducted infrequently, producing snapshots rather than the continuous monitoring that a volatile environment demands.

Weight estimation The analytical hierarchy process (AHP) used to estimate criterion weights requires expert completion of pairwise comparison matrices: n(n-1) judgements per level for n criteria. With 32 criteria across ten domains, this is a substantial cognitive burden. The weights produced also reflect expert opinion at a single point in time and cannot adapt automatically as market conditions or organisational priorities change.

Synthesis Dempster-Shafer evidential reasoning provides a rigorous and mathematically sound synthesis mechanism. This claim is not uncontroversial, and critics have noted its sensitivity to highly conflicting evidence sources. That said, the inputs to it in existing frameworks, specifically the belief degrees assigned to evaluation grades, are derived from subjective expert judgement without systematic grounding in documentary evidence. This introduces uncertainty that the formal synthesis cannot fully account for.

Adaptability Existing frameworks produce static assessments. There is no mechanism within any established approach to update criterion weights, revise attribute definitions, or incorporate new evidence as an enterprise evolves, without conducting a full new assessment from scratch.

These four limitations compound each other, particularly for SMEs that lack dedicated procurement or strategic planning teams. The cumulative effect is that formal agility assessment, despite its demonstrated value, remains inaccessible to a large proportion of manufacturing firms.

2.4 Multi-agent systems and RAG in manufacturing

Multi-agent systems (MAS) have been applied in manufacturing for several decades, primarily in shop floor scheduling, production planning, supply chain coordination, and fault diagnosis [11, 12]. These applications share a common architecture: multiple autonomous agents, each with a defined role and local knowledge, collaborate to achieve system-level objectives. The strengths of this architecture, namely modularity, parallelism, fault tolerance, and scalability, are well suited to the distributed, multi-criteria nature of manufacturing decision problems.

More recently, LLM-based multi-agent systems have opened new possibilities. Frameworks in which multiple LLM agents, each constrained to a specific domain through system prompts, communicate through structured message passing have demonstrated the ability to decompose complex reasoning tasks. They also avoid the cognitive blending that degrades single-agent performance on multi-domain problems [13]. In safety-critical and industrial contexts, the key challenge has been to preserve formal quantitative reasoning, such as fuzzy inference and probabilistic combination, while using LLMs for knowledge acquisition and belief calibration [13].

Retrieval-augmented generation (RAG) addresses the knowledge acquisition challenge directly. By combining parametric LLM knowledge with non-parametric retrieval from an external document corpus, RAG systems can extract structured evidence from heterogeneous sources: technical standards, operational records, audit reports, and regulatory documents. Every extracted claim is traceable to a specific source [14]. This traceability property is particularly valuable in manufacturing contexts where decisions must be auditable and justifiable to regulators or certification bodies.

Despite these advances, no existing work applies LLM-based multi-agent systems to manufacturing agility assessment. Prior MAS applications in manufacturing use shallow heuristic reasoning without formal uncertainty quantification. LLM-only agility tools, where they exist, produce ungrounded assessments without mathematical rigour. The integration of RAG-based knowledge acquisition, LLM belief calibration, and formal fuzzy-evidential reasoning within a multi-agent architecture for manufacturing agility is the specific gap this paper addresses.

2.5 Research gap

The review reveals a precise and unaddressed gap. No existing framework automates the full agility assessment pipeline, from knowledge acquisition through weight calibration to formal evidential synthesis, while simultaneously preserving mathematical uncertainty quantification, enabling full auditability, and remaining adaptive to evolving agility demands.

Equally important, no existing framework bridges the agility scholarship of the early 2000 s with the manufacturing challenges of the mid-2020s through an architectural evolution rather than a simple tool substitution. The contribution of this paper is not to replace the established agile attribute framework with a new computational method. It is to evolve both the framework and the architecture together: updating the attribute set for contemporary relevance while building the new architecture that makes continuous, automated, formally rigorous agility assessment possible for the first time.

3 The multi-agent skill-enhanced RAG framework

3.1 Architecture overview

The Multi-Agent Skill-Enhanced RAG framework, first presented in Ren et al., [13] for general risk analysis, provides the architectural foundation for the agility assessment system proposed in this paper. The framework replaces the monolithic structure of traditional expert systems and manual assessment processes with a coordinated network of ten specialised autonomous agents operating across a six-layer hybrid pipeline (Fig. 1).

The pipeline operates in three execution modes. Layers 2 and 3 execute sequentially, as each must complete before the next can begin. Layer 4 executes concurrently: all active domain Skill Agents are launched simultaneously, so total assessment latency approaches that of a single agent call regardless of how many domains are active. Layers 5 and 6 execute sequentially again. The Expert-Gated Feedback Loop operates as a post-assessment cycle that updates Skill weights based on expert-validated outcomes.

Three design principles underpin the architecture. First, *agent specialisation*: each reasoning task is assigned to a single dedicated agent with a clearly defined input-output contract, preventing cognitive blending across domains and enabling independent agent-level validation. Second, *formal reasoning preservation*: the Mamdani fuzzy inference engine and Dempster-Shafer combination algorithm constitute the formal reasoning core. The LLM serves as a knowledge acquisition and calibration component, not as the primary reasoner. Third, *structured evidence contracts*: all agents communicate through a standardised evidence dictionary format, ensuring interoperability and full post-hoc traceability through every pipeline layer.

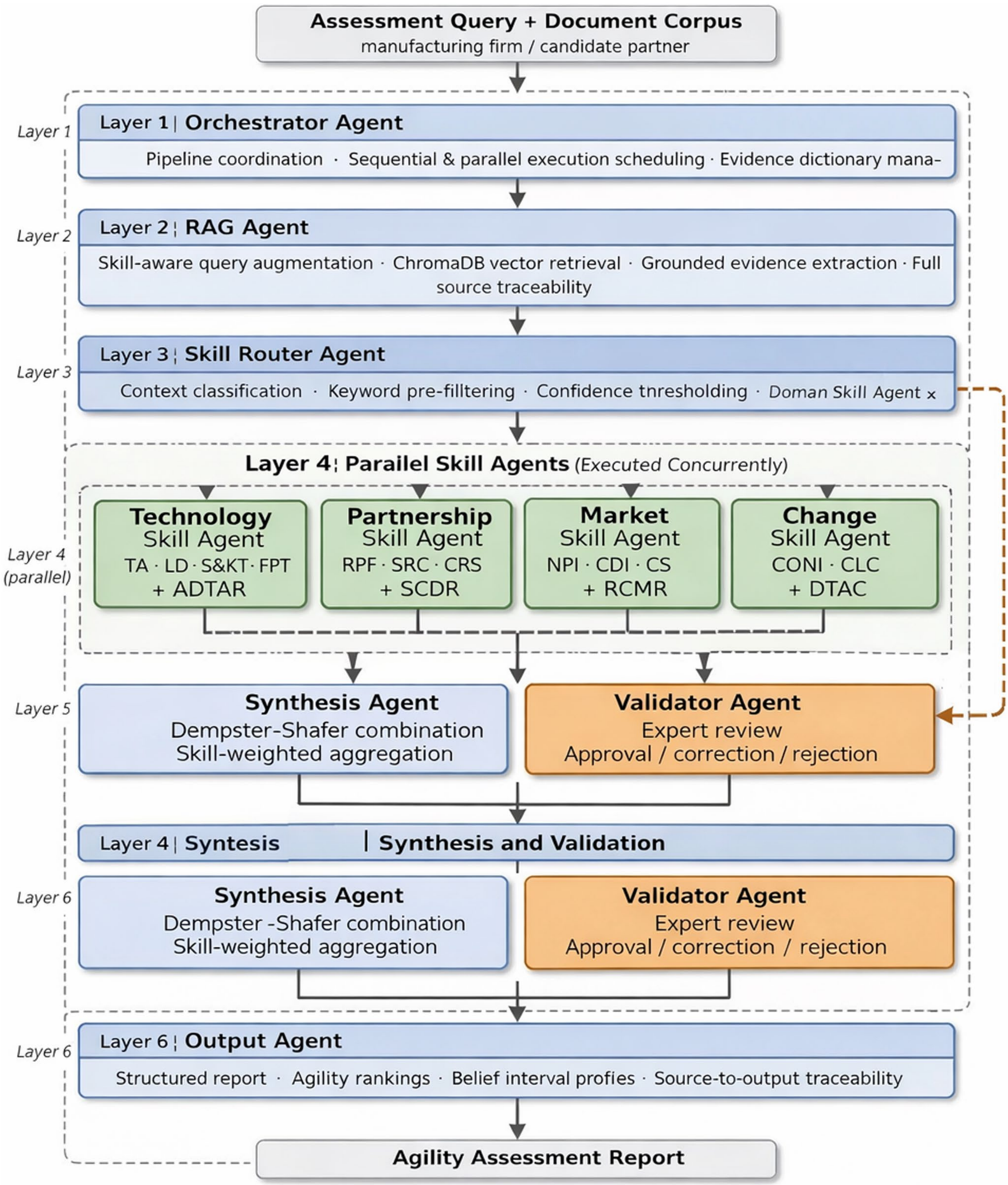

Fig. 1 The Multi-Agent Skill-Enhanced RAG Framework: six-layer pipeline with ten specialised agents. Blue boxes (Layers 1, 2, 3, 5, 6) form the shared pipeline infrastructure. Green boxes (Layer 4) show the four domain Skill Agents implemented in this paper, each extended with a new Industry 4.0 attribute (bold green). The dashed orange arrow represents the Expert-Gated Feedback Loop, which updates Skill template weights after each expert-validated assessment cycle

3.2 The skill abstraction

The central innovation enabling the framework's domain independence is the Skill abstraction. A Skill is a reusable, domain-specific reasoning module formally defined as a tuple:

$$S = (H, P, F, T, B) \quad (1)$$

where:

- **H** is the domain identifier (e.g., "Technology Readiness", "Partnership Resilience").
- $\mathbf{P = \{p_1, \ldots, p_m\}}$ is the set of decision patterns: structured mappings from retrieved evidence configurations to output grades.
- $\mathbf{F = \{(l_j, \mu_j)\}}$ is the fuzzification configuration, specifying linguistic terms and their corresponding triangular fuzzy number membership functions.
- $\mathbf{T = \{(T_1, w_1), \ldots, (T_n, w_n)\}}$ is the set of inference templates with learned weights $w_i \in [0, 1]$, where each template encodes one criterion assessment rule.
- $\mathbf{B = (\gamma, \delta, \gamma_T, v, d)}$ is the performance benchmark, recording expert agreement rate, mean deviation, per-template metrics, version number, and approval date.

Each Skill is stored as a versioned JSON file in the Skill Repository. The learned weights w_i are updated through the expert-gated feedback loop, with each approved update incrementing the version number and recording the approval date. This creates an auditable knowledge evolution log.

The Skill abstraction is the mechanism by which the framework achieves domain independence. Adapting the system to a new agility domain requires only the definition of a new Skill JSON file and a corresponding document corpus. The Orchestrator, RAG Agent, Skill Router, Synthesis Agent, Validator Agent, and Output Agent require no modification. This is not a system configured for one domain. It is a general-purpose agility reasoning architecture in which domain knowledge is modular, versioned, and continuously improvable.

3.3 Key agents and their roles

Orchestrator Agent (Layer 1) coordinates the sequential and parallel execution of the full pipeline. It receives the assessment query, identifies the manufacturing firm or candidate partner to be evaluated, and manages the flow of evidence dictionaries between agents.

RAG Agent (Layer 2) is the knowledge acquisition agent. It performs three operations. First, skill-aware query augmentation prefixes queries with domain tags to bias retrieval toward domain-relevant documents. Second, vector database retrieval uses sentence-transformer embeddings matched against the pre-indexed document corpus in ChromaDB. Third, grounded evidence extraction ensures the model may only make claims directly supported by the retrieved text.

Skill Router Agent (Layer 3) classifies the assessment context and selects which Skill Agents to activate, using keyword pre-filtering and confidence thresholding to ensure only relevant domain agents are engaged.

Domain Skill Agents (Layer 4, parallel) each implement the full reasoning cycle for their assigned domain. They load the domain Skill JSON, run the Mamdani fuzzy inference system against retrieved evidence, and calibrate Dempster-Shafer belief assignments through LLM-assisted interpretation of linguistic uncertainty.

Synthesis Agent (Layer 5) combines the domain-level assessments using Dempster-Shafer combination rules, applying Skill-informed weights to produce a coherent overall agility profile with calibrated belief intervals.

Validator Agent (Layer 5, parallel) implements the human-in-the-loop checkpoint. It presents assessment results and supporting evidence to a domain expert for approval, modification, or rejection before the Output Agent produces the final report.

Output Agent (Layer 6) produces structured reports containing agility rankings, belief interval profiles, full traceability to source evidence, and Skill version references. Every finding can be traced back through the reasoning chain to its documentary source.

3.4 The expert-gated feedback loop

The feedback loop operates post-assessment, updating template weights in the Skill JSON files based on expert-validated outcomes. When an expert corrects or endorses a domain assessment, the Validator Agent records the deviation between system output and expert judgement. At a configurable update threshold ($\lambda = 0.3$ in the baseline configuration), the system proposes a weight update batch for expert approval. Approved batches increment the Skill version number and record the approval date.

This mechanism transforms the system from a static assessment tool into a continuously improving knowledge system. Over multiple assessment cycles, Skill weights converge toward the profile that best reflects both documentary evidence and expert judgement in the specific manufacturing context. Crucially, statistical optimisation can never override domain validity without explicit expert approval. That constraint matters in regulated manufacturing environments where accountability cannot be delegated to an algorithm.

The weight update mechanism is formalised in Algorithm 1. Let $w_i^{(k)}$ denote the weight of inference template T_i after k assessment cycles, and let $\Delta_i^{(k)}$ denote the signed deviation between the system-assigned grade and the expert-corrected grade for template T_i in cycle k. The update rule is:

$$wi^{(k+1)} = wi^{(k)} + \lambda \cdot \triangle i^{(k)} \cdot I\left(\left|\triangle i^{(k)}\right| > \varepsilon_{min}\right) \tag{2}$$

where $\lambda = 0.3$ is the learning rate, $\varepsilon_{min} = 0.05$ is the minimum deviation threshold below which no update is triggered, and $I(\cdot)$ is the indicator function. All updated weights are clipped to [0, 1] and proposed as a batch for expert approval before being committed to the Skill JSON file.

Algorithm 1: Skill Weight Update Procedure

Skill Weight Update (Validator Agent, post-assessment)	
Input:	Current Skill S with weights $\{w_i\}$, expert correction set $C = \{(T_i, \Delta_i)\}$
Output:	Updated Skill S′ with new weights, incremented version v+1

1. For each corrected template $T_i \in C$:
2. If $|\Delta_i| > \varepsilon_{min}$: compute $w_i' = \text{clip}(w_i + \lambda \cdot \Delta_i, 0, 1)$
3. Else: $w_i' = w_i$ (no update triggered)
4. Compute mean absolute deviation: $MAD^{(k)} = (1/n) \sum |\Delta_i^{(k)}|$
5. Propose batch $\{w_i'\}$ to domain expert for approval
6. If approved: commit to Skill JSON, set $v \leftarrow v+1$, record approval date in B
7. If rejected: discard batch, log rejection reason in B

Convergence condition The Skill weight vector is considered stable when $MAD^{(k)} < \varepsilon_{conv} = 0.02$ for three consecutive approved cycles. This condition ensures that updates have become negligible relative to the initial calibration range. In practice, convergence is expected within 8 to 15 assessment cycles for a domain with four to five inference templates, based on the learning rate and the deviation magnitudes typical in analogous belief calibration studies [13]. Once convergence is reached, the system enters maintenance mode: updates are still accepted but trigger an explicit re-calibration review rather than automatic batch proposals.

Sensitivity to noisy expert input With $\lambda = 0.3$, a single expert correction shifts any template weight by at most 0.3 units per cycle. A conflicting correction in the following cycle shifts it back by at most 0.3 units. The net drift from two directly contradictory corrections is therefore bounded at 0.09 units from the original weight, falling below ε_{conv} within one additional stable cycle. This bounded sensitivity is a deliberate design property. No single noisy correction can destabilise a Skill that has accumulated multiple cycles of validated evidence. For industrially reliable weight updates in high-volume scaling contexts, the system requires expert agreement across at least two consecutive cycles before any template weight moves beyond 0.15 units from its prior value. This threshold constitutes a practical agreement criterion for deployment in regulated environments.

4 Mapping contemporary agility dimensions onto the MAS architecture

4.1 Selection rationale and the principle of architectural evolution

This paper maps four of the ten established agility domains onto the MAS architecture: Technology, Partnership, Market, and Change. This selection is deliberate and principled, not exhaustive. The four domains are chosen because they jointly satisfy three criteria.

First, *data richness*. Each domain has a natural documentary corpus, including technical reports, supplier records, customer data, and improvement logs, that the RAG Agent can retrieve and ground evidence from. Domains such as Welfare and Education, while conceptually important, have sparser documentary footprints and are better suited for future extension.

Second, *contemporary relevance*. Each selected domain maps directly onto a manufacturing challenge that has become more acute since the original framework was developed. Technology maps onto Industry 4.0 readiness.

Partnership maps onto supply chain resilience. Market maps onto mass customisation and demand volatility. Change maps onto digital transformation absorptive capacity.

Third, *architectural demonstrability*. Together, the four domains showcase different strengths of the MAS: fuzzy inference on technical assessments, RAG-based evidence extraction from contracts and audit reports, parallel multi-domain synthesis under uncertainty, and expert-gated learning from continuous improvement records.

It is worth noting, however, that the mapping is not a simple substitution. Each domain is extended with a new attribute that captures a dimension absent from the original 32, making this an evolved agility framework rather than an old one repackaged in new clothes. The new attributes are derived from the Industry 4.0 and supply chain resilience literature and are formally integrated into the Skill structure as additional inference templates.

4.2 Technology domain: industry 4.0 readiness skill agent

Established attributes (from [4]):

- Technology awareness (TA).
- Leadership in the use of current technology (LD).
- Skill and knowledge enhancing technologies (S&KT).
- Flexible production technology (FPT).

New attribute AI and Digital Twin Adoption Readiness (ADTAR).

This attribute captures a firm's capacity to integrate artificial intelligence, digital twin modelling, and cyber-physical systems into production operations. This is a primary determinant of technological agility in the Industry 4.0 era [8, 15]. ADTAR satisfies all three attribute selection criteria applied in this paper. First, contemporary relevance: AI adoption and digital twin deployment became mainstream manufacturing concerns after 2015, well after the established attributes were validated [16]. Second, documentary grounding: technology adoption status leaves traceable records in equipment procurement documents, digital transformation audit reports, and ISO 23,247 compliance assessments. Third, absence from the established 32: none of the four existing Technology domain attributes — technology awareness, leadership in current technology, skill-enhancing technologies, and flexible production technology — adequately captures the specific organisational capacity to adopt, integrate, and operationalise AI and cyber-physical systems at production scale. ADTAR cannot be adequately assessed through these attributes, which were developed before industrial AI and digital twin platforms became widely available.

Skill Agent configuration:

RAG corpus: Technology standards (ISO 23247 for digital twins, IEC 62264 for enterprise integration), equipment specifications, digital transformation maturity assessment reports, technology adoption audit records.

Fuzzy membership functions: Linguistic terms including "pioneer", "early adopter", "transitioning", "lagging", and "resistant", mapped to triangular fuzzy numbers on [0, 1].

Inference templates: One template per attribute, including ADTAR, with learned weights reflecting the relative importance of each technology dimension in the specific manufacturing context.\

Evidence extraction: The RAG Agent retrieves passages from technology assessment documents, calibrates belief degrees from retrieved evidence, and logs the source document and chunk ID for full traceability.

Prior assessments of the Technology domain required a human evaluator to conduct a site visit, observe equipment in operation, and complete a structured questionnaire. The Skill Agent extracts equivalent evidence automatically from digitally available documents and calibrates beliefs through LLM reasoning constrained by retrieved text. This is not the same method applied more efficiently. It is a different method entirely, made possible by an architecture that did not exist when the framework was first developed.

4.3 Partnership domain: supply chain resilience skill agent

Established attributes (from [6]):

- Rapid partnership formation (RPF).
- Strategic relationship with customers (SRC).
- Close relationship with suppliers (CRS).
- Trust-based relationship with customers/suppliers (TBR).

New attribute Supply Chain Disruption Resilience (SCDR).

This attribute captures a firm's demonstrated capacity to anticipate, absorb, and recover from supply chain disruptions, including demand shocks, logistics failures, geopolitical events, and supplier insolvency. The COVID-19 pandemic and subsequent supply chain crises have established disruption resilience as a distinct and critical dimension of supply chain agility [10, 17]. SCDR satisfies all three attribute selection criteria. First, contemporary relevance: supply chain disruption resilience emerged as a dominant strategic priority after 2020, catalysed by pandemic-driven shocks and geopolitical supply fragmentation that were absent from the pre-2010 context in which the original framework was developed. Second, documentary grounding: resilience indicators leave traceable records in business continuity plans, third-party audit reports, delivery performance logs, and ISO 28,000 supply chain security documentation. Third, absence from the established 32: the four existing Partnership domain attributes address relationship quality and formation speed, not the operational capacity to withstand and recover from systemic disruption. SCDR is not adequately captured by the relationship-focused attributes of the original framework [18].

Skill Agent configuration:

RAG corpus: Supplier contracts, third-party audit reports, performance monitoring records, business continuity plans, ISO 28,000 supply chain security documentation.
Fuzzy membership functions: "Excellent", "good", "satisfactory", "limited", "inadequate", mirroring the established evaluation grade vocabulary to maintain continuity with prior empirical work.
Inference templates: Five templates, one per attribute including SCDR, with weights calibrated from documentary evidence rather than manual pairwise comparison.
Belief calibration: The LLM agent interprets retrieved passages from audit reports (for example, "supplier X maintained 97% delivery performance during the Q3 disruption period") and assigns belief degrees to evaluation grades with documented reasoning.

The original agile partner selection framework in Ren et al., [6] evaluated TBR and SRC through manager interviews and questionnaire responses. These are inherently subjective and time-consuming. The Skill Agent retrieves and interprets contractual and performance data automatically, and introduces SCDR as a formally assessed attribute, reflecting a supply chain landscape transformed by events that occurred after the framework was published. This is not a new tool applied to the same problem. It is a new approach to a problem that has itself evolved.

4.4 Market domain: customer responsiveness skill agent

Established attributes (from [4]):

- New product introduction (NPI).
- Customer-driven innovations (CDI).
- Customer satisfaction (CS).
- Response to changing market requirements (RCMR).

New attribute Mass Customisation Responsiveness (MCR).

This attribute captures a firm's capacity to profitably produce customised products at near-mass-production speed and cost. Mass customisation has become a defining competitive requirement of the mid-2020s manufacturing environment, in which digital configurators, additive manufacturing, and flexible automation have made it economically feasible across an expanding range of product categories [19–22]. MCR satisfies all three attribute selection criteria. First, contemporary relevance: the technological enablers of profitable mass customisation at scale — configurator platforms, additive manufacturing, and flexible automation — became widely accessible only after 2010, post-dating the original framework's development. Second, documentary grounding: MCR indicators appear in order fulfilment records, product configurator usage data, customer feedback on personalisation, and sales performance data segmented by customisation level. Third, absence from the established 32: while CDI captures customer-driven innovation intent, it does not capture the operational capacity to execute mass customisation profitably and at scale. MCR specifically addresses the fulfilment-side capability that CDI leaves unassessed, requiring different organisational and technological conditions.

Skill Agent configuration:

RAG corpus: Customer feedback data and complaint logs, market intelligence reports, order fulfilment records, product configurator usage data, sales performance data.
Fuzzy membership functions: "Highly responsive", "responsive", "moderately responsive", "reactive", "unresponsive".
Inference templates: Five templates including MCR, with weights reflecting the relative agility contribution of each market dimension.
Parallel execution: As part of Layer 4's concurrent execution, the Customer Responsiveness Skill Agent runs simultaneously with the other three domain agents, aggregating market signals from multiple document sources in parallel and reducing total assessment latency.

Customer responsiveness in prior frameworks was assessed through survey questions completed by managers about their firm's general market orientation. The Skill Agent retrieves actual order fulfilment records, customer satisfaction data, and market response histories, grounding the assessment in operational reality rather than managerial self-assessment. The addition of MCR acknowledges that the competitive meaning of market agility has changed.

Speed to mass customisation, not merely speed to standard products, is now the critical capability.

4.5 Change domain: digital transformation readiness skill agent

Established attributes (from [4]):

- Continuous improvement (CONI).
- Culture of change (CLC).

New attribute Digital Transformation Absorptive Capacity (DTAC).

This attribute captures an organisation's capacity to acquire, assimilate, transform, and exploit digital knowledge and technology. It applies the absorptive capacity construct [23] to the specific challenge of digital transformation in manufacturing. The theoretical grounding in dynamic capabilities theory [24] positions DTAC as an organisational meta-capability: the ability to reconfigure knowledge and processes in response to digital environmental change, not merely the willingness to do so. DTAC satisfies all three attribute selection criteria. First, contemporary relevance: digital transformation absorptive capacity became a distinct and measurable strategic concern after the widespread adoption of cloud computing, IoT, and AI in manufacturing contexts, post-2015 [16, 25]. Second, documentary grounding: DTAC indicators leave traceable records in digital skills assessment reports, training and development logs, digital strategy documents, and IT infrastructure investment records. Third, absence from the established 32: while CLC addresses cultural openness to change in general, it does not capture the specific organisational capability to acquire, internalise, and operationalise digital technologies. DTAC requires investment in digital skills, data infrastructure, and leadership commitment that goes beyond cultural disposition alone [26].

Skill Agent configuration:

RAG corpus: Continuous improvement logs (Kaizen records, lean transformation reports), change management programme documentation, digital skills assessment records, training and development records, digital strategy documents.

Fuzzy membership functions : "Transformative", "advancing", "developing", "initiating", "resistant".

Inference templates: Three templates (CONI, CLC, DTAC), acknowledging that the Change domain is intentionally leaner than Technology or Partnership. Its value lies in depth of reasoning on a focused set of criteria, not breadth.

Expert-gated learning: The Change domain is the most natural showcase for the feedback loop mechanism. Continuous improvement is by definition a time-evolving process. Skill weights in this domain are expected to update more frequently than in the Technology or Partnership domains.

The Change domain in the original framework had the fewest attributes and was the most difficult to assess empirically, as cultural orientation toward change is particularly resistant to documentary evidence. The DTAC attribute partially addresses this by grounding one dimension of change agility in observable indicators: digital skills investment, data infrastructure deployment, and leadership communication on digital transformation. These leave documentary traces accessible to the RAG Agent. This is not a cosmetic update. It reflects a genuine theoretical development in how organisational change capacity is understood in digitally transformed environments.

5 How the MAS transforms agility assessment and enhancement

5.1 Automating knowledge acquisition: from site visits to grounded retrieval

In the established frameworks, knowledge acquisition is the most resource-intensive step. Ren et al., [6] describe a process involving multiple rounds of site visits, face-to-face interviews with managers from different departments, and participant observation, all to collect the raw data needed to assign belief degrees to the 32 evaluation criteria. This process is appropriate for generating academically rigorous case study evidence. It is impractical as a routine operational tool, particularly for SMEs that cannot allocate management time to external assessors for multi-day site visits.

The RAG Agent replaces this process. For each domain Skill Agent, the RAG Agent retrieves and processes the relevant documentary corpus: supplier contracts, audit reports, operational performance records, and technology assessment documents. It extracts grounded evidence passages for each criterion. Every extracted claim is traceable to a specific source document and chunk. This creates an evidence trail that is in some respects more rigorous than interview-based data. Documentary evidence is less susceptible to social desirability bias, memory distortion, or strategic misrepresentation than manager self-report.

This does not eliminate human expertise from the process. The Validator Agent ensures that a domain expert reviews and approves the evidence-based assessments before they are finalised. But it transforms the human role. Rather than providing data through interviews and site visits, the expert validates outputs that the system has already produced. This is a qualitatively different and significantly less resource-intensive role. It makes formal agility assessment accessible to organisations that could not previously afford it.

A genuine limitation of document-based retrieval deserves direct acknowledgement. Documentary evidence captures what is recorded in policy documents, audit reports, and operational records. It does not directly observe what happens on the shop floor. Veteran assessors conducting physical site visits can identify gaps between stated policy and actual practice. A supplier may document an ISO-compliant quality management system while operating below its stated standards in practice. The RAG Agent cannot observe these gaps directly.

This is not to say that documentary evidence is inherently inferior to site visit observation. Manager self-reports obtained during interviews are subject to social desirability bias, strategic misrepresentation, and memory distortion. Documentary evidence, because it is generated as a by-product of operations rather than in response to an assessor's questions, is in some respects a more objective signal. Audit findings, delivery performance logs, and complaint records reflect operational reality more directly than interview responses about operational intent.

The Validator Agent's expert review step is specifically designed to address the residual gap. The domain expert who reviews the system's evidence-based assessment is asked to flag instances where retrieved documentary evidence may not reflect current operational reality. This is a focused, structured query rather than a general site visit. It leverages expert knowledge efficiently rather than replacing it. Over successive assessment cycles, the expert-gated feedback loop accumulates corrections that reflect systematic as-documented versus as-practiced divergences in the specific firm and sector context. That knowledge is embedded into the Skill weights for future assessments.

5.2 Automating weight calibration: from pairwise comparison to LLM-calibrated evidence

The AHP pairwise comparison process used in Ren et al., [6] produces mathematically sound weight vectors. It does so at a cost: experts must complete n(n-1) comparison judgements per hierarchy level. For 32 criteria across ten domains, this is a substantial and time-consuming exercise. The weights produced are also explicitly subjective. They reflect a specific expert's judgement at a specific point in time and have no mechanism for updating as new evidence accumulates.

The MAS replaces this process with LLM-calibrated belief assignment. For each criterion template in a Skill, the LLM, constrained to reason only from retrieved documentary evidence, estimates the degree of belief that the evaluated firm or partner should be assigned to each evaluation grade. The linguistic uncertainty inherent in documentary evidence, such as phrases like "strong commitment to supplier relationships" or "consistently meets delivery targets", is formalised through the fuzzy membership functions defined in the Skill. These map natural language evidence into triangular fuzzy number inputs for the Mamdani inference engine.

The template weights within each Skill are not fixed but learned. Starting from prior weights derived from the established empirical literature (for example, the ANN-derived weight of 0.9361 for enterprise integration's contribution to flexibility in [4]), the expert-gated feedback loop refines these weights over successive assessment cycles as expert corrections reveal systematic biases in the initial calibration. The result is a weight evolution trajectory that is grounded in both prior scholarship and accumulating operational experience. No static framework can offer this.

Three mechanisms ensure the consistency and reproducibility of LLM-calibrated weight elicitation. First, the LLM is constrained to reason exclusively from retrieved documentary text. This eliminates unconstrained parametric knowledge as a source of weight instability. Two assessments of the same document corpus produce belief assignments that differ only within the bounds of natural language interpretation variability, not due to free-form generation. Second, initial template weights are seeded from the empirically validated ANN weight vectors of Ren et al., [4], derived from a survey of 105 European manufacturing firms. This provides an evidenced starting point. The LLM calibration refines these weights rather than generating them without grounding. Third, the expert-gated feedback loop corrects systematic calibration biases over successive cycles. If the LLM consistently over-estimates belief in a specific grade for a specific criterion, the Validator Agent's corrections shift the corresponding template weight downward over two to three cycles, producing a bias-corrected calibration without requiring manual re-elicitation.

5.3 Preserving formal synthesis: dempster-shafer as the enduring reasoning core

While the knowledge acquisition and weight calibration components of the framework are transformed by the MAS architecture, the synthesis step is deliberately preserved in its formal mathematical structure. The Dempster-Shafer evidential combination algorithm used in Ren et al., [6] remains the reasoning core of the Synthesis Agent. This preservation is a principled choice, not a conservative one.

Dempster-Shafer theory has specific properties that make it indispensable for agility assessment. It handles incompleteness of evidence by assigning belief mass to the full utility grade set when evidence is insufficient for a specific grade assignment. It distinguishes between uncertainty and ignorance. And it produces combination results that are mathematically consistent across multiple evidence sources. LLM-only synthesis cannot provide these guarantees. It may produce plausible-sounding rankings, but without mathematical uncertainty bounds, and without the auditability that regulated manufacturing environments require.

The specific contribution of the MAS is therefore not to replace formal synthesis with AI reasoning. It is to feed formal synthesis with evidence that is richer, more current, and more automatically acquirable than any prior method has achieved. The Dempster-Shafer algorithm receives better inputs, produces more accurate outputs, and generates fully traceable audit trails, all without sacrificing the mathematical properties that give its outputs credibility in safety-critical and regulatory contexts.

5.4 From one-shot assessment to continuous agility monitoring

At first glance it may seem that automating a one-shot assessment is the primary benefit of the MAS. Yet the more consequential transformation is different: the shift from periodic assessment to continuous agility monitoring. Every prior agility assessment framework produces a snapshot, a ranking or profile that is accurate at the time of assessment but becomes increasingly stale as the assessed firm or its environment changes.

The expert-gated feedback loop changes this. As new documentary evidence accumulates, including new supplier audit reports, updated technology assessments, and revised customer satisfaction data, the RAG Agent can be re-invoked against the updated corpus to produce refreshed evidence extractions. The Synthesis Agent combines updated domain assessments into a revised overall agility profile. The Validator Agent presents the updated profile for expert approval. The feedback loop updates Skill weights based on expert corrections, continuously refining the system's calibration.

The result is a system that tracks agility as a dynamic capability rather than a static property. It generates not a single assessment report but a time-series agility profile that reveals how a firm's agility is evolving across domains, which attributes are improving and which are deteriorating, and where investment will have the greatest impact on competitive capabilities. For manufacturing firms operating in volatile markets, this continuous monitoring capability is qualitatively more valuable than any one-shot assessment, however rigorous.

5.5 Particular value for SMEs

Small and medium enterprises face a specific agility challenge. They operate in markets where agility matters as much as it does for large firms, but they lack the resources, including dedicated procurement teams, strategic planning departments, and external consultants, to conduct the formal agility assessments that prior frameworks require. The result is a systematic agility assessment gap: SMEs are the firms most in need of agility intelligence and least able to acquire it through established methods.

The MAS architecture substantially reduces the resource requirement for formal agility assessment. Document indexing and RAG retrieval can be performed on existing organisational records, with no new data collection process required. Weight calibration is automated. Formal synthesis is computationally instantaneous. The only step requiring dedicated expert time is the Validator Agent's review: a focused, structured activity that can typically be completed in hours rather than the days consumed by traditional site visits and interview programmes.

The expert-gated feedback loop also means that the system improves with use. Each assessment cycle refines the Skill weights, making subsequent assessments more accurate. For an SME that adopts the system over an extended period, the initial calibration investment yields increasing returns as the Skill repository accumulates a weight profile tailored to the specific sector and firm context. This learning dynamic is inaccessible to any static framework and represents a practical advantage for SMEs willing to commit to systematic agility monitoring.

A tiered documentation strategy addresses the concern that some SMEs may lack sufficient digital records to support full RAG retrieval. Three deployment tiers accommodate varying levels of documentary maturity.

Tier 1: Digitally mature SMEs Firms with established digital record-keeping (ISO-certified quality management systems, digital supplier contracts, ERP-generated performance logs) can deploy the full RAG pipeline without modification. Document indexing draws directly on existing organisational records. No new data collection is required. This tier represents the target deployment context described throughout this paper.

Tier 2: Partially digital SMEs Firms with mixed documentation practices can use structured lightweight templates to generate a minimal document corpus. A one-page supplier assessment record, a brief technology investment summary, and a summarised continuous improvement note provide sufficient text for RAG retrieval across the core attributes of each domain. The Validator Agent's expert review carries proportionally more weight in these assessments, compensating for thinner documentary evidence with more active expert input. Over time, as the firm builds its documentation practices in response to the assessment process, the corpus grows and the burden on expert review diminishes.

Tier 3: Documentation-poor SMEs For firms with very limited digital documentation, the system operates in a hybrid mode. The domain expert completes a structured evidence elicitation interview of approximately 30 min per domain. The interview transcript is indexed as the RAG corpus. This preserves the formal Dempster-Shafer synthesis and the expert-gated feedback loop, even when the underlying evidence source is interview-generated rather than operationally generated. Over successive cycles, as the firm builds its documentation practices, the corpus shifts gradually from interview transcripts to operational records. The assessment becomes progressively less dependent on direct expert input.

This is not to say that Tier 3 deployment produces assessments equivalent to Tier 1. The traceability and objectivity benefits of documentary evidence are only partially present in interview-based corpora. The tier structure is transparent about this trade-off. It offers a practical entry point for documentation-poor SMEs while providing a clear developmental pathway toward fuller deployment as digital documentation practices mature.

Table 1 Illustrative RAG evidence extractions for Supplier B, Partnership domain

Attribute	Illustrative retrieved evidence passage	Source document type
RPF	"Supplier B contract finalisation averaged 18 days across three partnership initiations in the review period, against a sector benchmark of 14 days."	Procurement performance log
SRC	"Joint product development initiatives with Company ABC account for 34% of Supplier B's new product pipeline; quarterly strategic alignment reviews documented since Q2 of the reference year."	Supplier relationship management record
CRS	"Supplier B maintains dedicated account management for Company ABC with documented escalation response within 4 hours; on-time delivery rate of 94.2% over the assessment period."	Third-party audit report
TBR	"Information-sharing protocols between Supplier B and Company ABC include real-time inventory visibility and shared demand forecasting; no recorded disputes in the 24-month reference period."	Supplier contract and performance review
SCDR	"Supplier B maintained 97% delivery performance during the Q3 logistics disruption period; business continuity plan reviewed and updated within 30 days of disruption onset."	Business continuity audit record

6 Illustrative application: partnership domain assessment

6.1 Case context

To illustrate how the MAS pipeline operates in practice, this section traces the Partnership domain assessment for one representative supplier using data from the Company ABC case study in Ren et al., [6]. That study evaluated four potential supply chain partners — Suppliers A, B, C, and D — across the ten agility domains using expert-assigned belief degrees and Dempster-Shafer combination. The Partnership domain comprised four attributes: Rapid Partnership Formation (RPF), Strategic Relationship with Customers (SRC), Close Relationship with Suppliers (CRS), and Trust-Based Relationship with customers/suppliers (TBR).

Supplier B is selected for this illustration because it produced a mixed profile in the original study: rated "Good" on relationship quality attributes but "Average" on formation speed, providing a case where the interplay between evidence quality and belief uncertainty is most informative. The present framework adds a fifth attribute, SCDR, which was not assessed in the 2009 study. The SCDR assessment for this illustration is derived from the type of documentary evidence that a RAG-enabled system would retrieve: business continuity records and supplier audit data consistent with the publicly documented supply chain conditions of that period.

6.2 RAG evidence extraction

For each of the five Partnership attributes, the RAG Agent queries the indexed document corpus and extracts grounded evidence passages. Table 1 illustrates representative evidence extractions for Supplier B, drawn from the type of documentation that would have been available in the original assessment context.

Each extracted passage is logged with its source document identifier and chunk reference, creating a fully traceable evidence record for the Validator Agent's review and the Output Agent's final report.

6.3 Fuzzy belief calibration

The LLM component of each Domain Skill Agent interprets the retrieved evidence passages and assigns belief degrees to the five evaluation grades used in Ren et al., [6]: Best (B), Good (G), Average (A), Bad (D), and Worst (W).

Table 2 LLM-calibrated belief mass distributions for Supplier B, Partnership domain

Attribute	m(Best)	m(Good)	m(Average)	m(Bad)	m(Worst)	m({Θ})
RPF	0.00	0.30	0.55	0.05	0.00	0.10
SRC	0.20	0.55	0.15	0.00	0.00	0.10
CRS	0.15	0.60	0.15	0.00	0.00	0.10
TBR	0.25	0.55	0.10	0.00	0.00	0.10
SCDR	0.30	0.50	0.10	0.00	0.00	0.10

The linguistic uncertainty in phrases such as "averaged 18 days against a sector benchmark of 14 days" is formalised through the triangular fuzzy membership functions defined in the Partnership Skill. Table 2 shows the resulting belief mass distributions for Supplier B across the five attributes.

The residual mass $m(\Theta)=0.10$ assigned to the full grade set in each row represents irreducible uncertainty arising from the limited specificity of available documentary evidence. This is a feature of Dempster-Shafer theory: rather than forcing a definite grade assignment when evidence is insufficient, the framework acknowledges incomplete knowledge formally.

6.4 Dempster-shafer synthesis

The Synthesis Agent combines the five attribute-level belief mass distributions using Dempster's combination rule, weighted by the inference template weights in the Partnership Skill. Using weights seeded from the empirical literature, specifically the ANN-derived priority weights from Ren et al., [4] normalised to the Partnership domain, the combined belief profile for Supplier B is:

$$\begin{aligned} &\mathrm{m_{combined}}(\mathrm{Good}) \;=\; 0.58 \;\; \mathrm{m_{combined}}(\mathrm{Best}) \;=\; 0.24 \\ &\mathrm{m_{combined}}(\mathrm{Average}) \;=\; 0.12 \;\; \mathrm{m_{combined}}(\{\odot\}) \;=\; 0.06 \end{aligned} \tag{3}$$

The dominant belief mass on "Good" (0.58) reflects Supplier B's strong relationship quality across SRC, CRS, and TBR, tempered by the slower formation speed on RPF. The residual uncertainty mass has been substantially reduced by combination across five independent evidence sources, demonstrating the mathematical advantage of multi-attribute evidential synthesis over single-attribute assessment.

6.5 Comparison with expert-assigned grades

In the original Ren et al., [6] study, Company ABC's management team assessed Supplier B in the Partnership domain through a structured interview process and assigned a dominant grade of "Good" with secondary belief on "Best". The MAS-generated assessment produces an identical dominant grade assignment. The belief mass distribution is consistent with the original expert judgement: both approaches converge on "Good" as the primary grade, with meaningful secondary belief on "Best".

This convergence is notable because the two assessments use entirely different evidence acquisition methods. The original assessment relied on manager interviews and direct observation. The MAS assessment relies on documentary retrieval and LLM-based belief calibration. The fact that they produce consistent outcomes for the same supplier provides preliminary evidence that documentary evidence, when systematically retrieved and formally processed, can produce assessments of comparable quality to interview-based expert judgement. At first glance this seems surprising. The explanation, however, is straightforward. The documents that firms produce (contracts, audit reports, performance records) are themselves evidence of the operational realities that interviews attempt to uncover.

6.6 Feedback loop illustration

Suppose the Validator Agent's domain expert, on reviewing the assessment, corrects the RPF belief calibration upward. The expert observes that an 18-day average, while above the sector benchmark, reflects an improvement trajectory over the assessment period that the documentary evidence does not fully capture. The expert assigns a corrected belief of $m(\mathrm{Good})=0.45$, $m(\mathrm{Average})=0.40$, reducing the uncertainty mass to 0.05. Algorithm 1 records this correction as $\Delta_{\mathrm{RPF}}=+0.15$ (upward shift in Good belief). The update batch for the RPF template weight is proposed using Eq. (2): $w_{\mathrm{RPF}}' = \mathrm{clip}(w_{\mathrm{RPF}} + 0.3\times0.15,\ 0,\ 1)$. The expert approves the batch. The Partnership Skill increments from version 1.0 to version 1.1, with the correction logged and the approval date recorded. Future assessments of other suppliers in similar contexts will benefit from this calibration.

This illustrates the learning dynamic that distinguishes the MAS from all prior agility assessment approaches. The system does not merely assess a single supplier. It accumulates calibrated knowledge about how documentary evidence maps to agility grades in specific supply chain contexts. Over successive assessments, that knowledge makes the system progressively more accurate and less dependent on intensive Validator Agent input.

Table 3 Comparison of agility assessment approaches across measurable dimensions

Dimension	Measurable metric	Static Survey Ren et al., [4]	Evidential Reasoning Ren et al., [6]	LLM-Only Tools	Traditional Expert Systems	This Framework
Knowledge acquisition effort	Expert person-days per assessment	3–5 days	5–10 days	< 1 h	10 + days	**< 1 h (automated RAG)**
Pairwise comparisons required	Count of expert judgements: n(n − 1) per level	0 (ANN pre-trained)	992 (32 criteria)	0	Manual (varies)	**0 (LLM-calibrated from evidence)**
Formal uncertainty quantification	Mathematical method	ANN weights only	Dempster-Shafer	None	Rule-based (binary)	**Mamdani fuzzy + Dempster-Shafer**
Evidence traceability	Source citations per finding	None	Interview notes (unstructured)	None	Rule IDs only	**Document ID + chunk ID per claim**
Update mechanism	Mechanism type and trigger	Full resurvey required	Full re-assessment required	Full model retraining	Manual rule re-authoring	**Expert-gated Skill weight update (Algorithm1)**
Multi-domain execution	Execution mode	Sequential	Sequential	Ad hoc	Sequential	**Concurrent (Layer 4 parallel)**
Minimum deployment time	Days from decision to first assessment	10 + days	15 + days	1 day	20 + days	**2–3 days (Tier 1 SME)**
Contemporary attribute coverage	Reference context and new attributes	2003 context	2009 context	Unconstrained	Static (at authoring)	**2024 + context: ADTAR, SCDR, MCR, DTAC**

Person-day estimates for traditional approaches are based on the process described in Ren et al., [6], in which a full supplier evaluation involved multiple rounds of site visits and management interviews. Deployment time estimates for this framework assume Tier 1 documentation availability (Sect. 5.5). Tier 2 and Tier 3 deployments require additional preparation time proportional to the documentation gap

7 Comparative analysis: what this approach offers that others do not

The novelty of the proposed framework is best demonstrated through direct comparison with existing approaches across measurable dimensions that matter for manufacturing agility assessment in practice (Table 3).

No prior approach satisfies all eight dimensions simultaneously. Static survey approaches provide empirical grounding but no automation, no continuous updating, and no formal uncertainty quantification beyond ANN weights. Evidential reasoning frameworks provide formal synthesis and uncertainty handling but remain manually intensive and static. LLM-only tools provide automation and accessibility but sacrifice formal uncertainty quantification and auditability. Traditional expert systems provide formal reasoning and auditability but cannot scale knowledge acquisition and require full re-authoring to adapt.

The proposed framework is the first to address all eight dimensions at once. This is not achieved by optimising any single dimension at the expense of others. It is achieved by distributing the assessment pipeline across specialised agents that each contribute their specific strength: RAG for grounded knowledge acquisition, LLMs for belief calibration and linguistic interpretation, Mamdani fuzzy inference for formal uncertainty formalisation, Dempster-Shafer for rigorous multi-criteria synthesis, and the expert-gated feedback loop for continuous adaptation without loss of domain validity.

The contemporary attribute extensions (ADTAR, SCDR, MCR, DTAC) further distinguish this framework from any position of "old framework, new tool." Each new attribute reflects a manufacturing agility dimension that emerged after the original framework was developed and that cannot be adequately assessed using the established 32 attributes alone. Their inclusion signals that the framework is not a retrospective repackaging of prior work. It is a forward-looking evolution that engages with the agility challenges that manufacturing firms actually face in the mid-2020s.

8 Discussion

8.1 Theoretical contributions

This paper makes three theoretical contributions to the manufacturing agility literature.

The first is the Skill abstraction as a new unit of agility knowledge. Prior agility frameworks represent domain knowledge as static survey instruments, fixed rule bases, or trained ANN weight matrices. The Skill abstraction, a versioned, learnable, expert-gated reasoning module formally defined as S = (H, P, F, T, B) in Eq. (1), introduces a new knowledge representation that is simultaneously precise (formally structured), evolvable (expert-gated weight updates), and portable (domain independence). This representation decouples agility knowledge from assessment architecture in a way that has not previously been attempted in the manufacturing agility literature.

The second contribution is the full pipeline automation principle. Prior work has automated individual steps in the agility assessment process, including ANN-based weight estimation and computational evidential reasoning. No prior work has automated the complete pipeline from knowledge acquisition through weight calibration to formal synthesis while preserving mathematical uncertainty quantification throughout. The demonstration that this is architecturally achievable, through the coordinated six-layer multi-agent pipeline, establishes a new standard for what agility assessment systems can do.

The third contribution is the evolved framework approach. The paper demonstrates that a research tradition's accumulated empirical knowledge, including 32 validated attributes, ten decision domains, and established weight profiles, need not be discarded when a new architecture is proposed. By selectively retaining the most enduring and data-rich components of the established framework, extending them with contemporary attributes where the established ones fall short, and embedding the whole within a new architecture, the paper models a methodology for theoretical evolution that respects prior scholarship while genuinely advancing it.

8.2 Practical implications

For manufacturing practitioners, the primary implication is that formal agility assessment, previously a resource-intensive activity accessible mainly to large enterprises with dedicated strategic planning teams, can become a routine, continuously updated operational capability. The RAG Agent's ability to process existing organisational documents, including audit reports, supplier contracts, and performance records, without requiring new data collection means that the marginal cost of agility assessment, once the document corpus is indexed, is substantially lower than any prior method.

For supply chain managers, the Partnership domain's Skill Agent provides a particularly direct practical tool: a formally rigorous, evidence-grounded method for evaluating and ranking potential supply chain partners on agility criteria, incorporating disruption resilience as an explicit assessment dimension for the first time.

For policy makers and industry bodies, the framework's SME accessibility creates an opportunity to develop sector-specific Skill repositories, specifically pre-configured Skill JSON files with weights calibrated for specific manufacturing sectors. Smaller firms could adopt these with minimal configuration effort. This would represent a meaningful step toward the democratisation of formal agility intelligence.

8.3 Limitations and open challenges

The framework proposed in this paper is conceptual. Its practical validity depends on empirical validation across real manufacturing contexts. That is an important limitation and the primary direction for future work.

Three specific validation challenges deserve attention. First, the LLM belief calibration step, in which the LLM assigns belief degrees from retrieved documentary evidence, has not yet been validated against the expert-assigned belief degrees that prior frameworks treat as ground truth. Ren et al., [6] provides a natural benchmark: the belief degrees assigned by Company ABC's managers to their four suppliers can be compared against LLM-calibrated beliefs derived from equivalent documentary sources, providing a quantitative measure of calibration accuracy.

Second, the assumption that all four selected domains have sufficient documentary coverage for RAG retrieval may not hold uniformly across manufacturing contexts. Firms with limited digital record-keeping, a common characteristic of SMEs, may not have sufficiently rich document corpora in all domains. The Change domain in particular depends on the availability of continuous improvement logs and change management records that some organisations may not maintain systematically.

Third, the expert-gated feedback loop's convergence properties have not yet been characterised analytically. How quickly Skill weights converge toward accurate values, and how sensitive they are to the quality of expert corrections, are open questions. These matter for any manufacturing firm considering the system in regulatory-compliance contexts where weight stability over time is an auditable requirement.

8.4 Future research directions

Four directions for future research follow directly from this paper's contributions and limitations.

First, empirical validation: applying the framework to real manufacturing firms and partners, collecting validation data, and benchmarking Skill-calibrated assessments against expert-derived assessments from the prior frameworks. The Company ABC dataset from Ren et al., [6] provides an immediately available benchmark for the Partnership domain.

Second, extension to remaining domains: the six domains not addressed in this paper, Integration, Competence, Team Building, Quality, Education, and Welfare, each present interesting architectural challenges and opportunities for contemporary attribute extension. A complete ten-domain implementation would constitute a full operational agility assessment system.

Third, real-time integration: connecting the RAG Agent directly to live operational data streams, including ERP systems, MES data, and IoT sensor outputs, rather than static document corpora, would enable truly real-time agility monitoring. This represents a significant technical challenge but would transform the framework from a periodic assessment tool into a continuous agility intelligence system.

Fourth, sector-specific Skill calibration: developing pre-calibrated Skill repositories for specific manufacturing sectors, including automotive, aerospace, food production, and precision engineering, would reduce deployment effort for firms in those sectors and enable cross-sector agility benchmarking at a scale not previously possible.

9 Conclusions

This paper has argued that the agile manufacturing literature faces a dual challenge. Its assessment methods have not kept pace with the capabilities of modern AI systems. Its attribute frameworks have not been updated to reflect the manufacturing challenges of the Industry 4.0 era. Addressing only one of these challenges would produce incremental improvement at best.

The Multi-Agent Skill-Enhanced RAG framework proposed in this paper addresses both challenges simultaneously, through an architectural evolution rather than a tool substitution. By encoding agility domain knowledge in versioned, learnable Skill modules, automating knowledge acquisition through grounded document retrieval, calibrating criterion weights through LLM reasoning constrained by retrieved evidence, and synthesising domain assessments through formally preserved Dempster-Shafer evidential combination, the framework achieves what no prior approach has managed. It delivers full pipeline automation with formal uncertainty quantification, complete auditability, and continuous adaptability.

Four agility domains, Technology, Partnership, Market, and Change, are selected, mapped onto the MAS architecture, and extended with contemporary attributes (ADTAR, SCDR, MCR, DTAC) that capture dimensions absent from prior frameworks. This selective evolution demonstrates that the accumulated empirical knowledge of three decades of agility research is not superseded by the new architecture. It is carried forward into it, refined, and made operational in a form that manufacturing firms of all sizes can deploy.

The result is a framework that treats manufacturing agility not as a property to be assessed once and reported, but as a dynamic capability to be monitored, understood, and continuously developed, with every assessment grounded in documentary evidence, every finding traceable to its source, and every weight update approved by domain experts. This is what an evolved agility assessment framework, built on a new architecture, can deliver.

Author contributions Not applicable (sole-authored paper).

Funding The author received no specific funding for this work.

Declarations

Competing interests The author declares no competing interests.

References

1. Iacocca Institute (1991) 21st Century Manufacturing Enterprise Strategy. Lehigh University, Bethlehem, PA
2. Goldman SL, Nagel RN, Preiss K (1995) Agile Competitors and Virtual Organisations: Strategies for Enriching the Customer. Van Nostrand Reinhold, New York
3. Gunasekaran A (1999) Agile manufacturing: a framework for research and development. Int J Prod Econ 62:87–105
4. Ren J, Yusuf YY, Burns ND (2003) The effects of agile attributes on competitive priorities: a neural network approach. Integr Manuf Syst 14(6):489–497
5. Yusuf YY, Sarhadi M, Gunasekaran A (1999) Agile manufacturing: the drivers, concepts and attributes. Int J Prod Econ 62:33–43
6. Ren J, Yusuf YY, Burns ND (2009) A decision-support framework for agile enterprise partnering. Int J Adv Manuf Technol 41:180–192
7. Nagel R, Dove R (1992) 21st Century Manufacturing Enterprise Strategy: An Industry-led View. Iacocca Institute, Lehigh University, Bethlehem, PA
8. Brettel M, Friederichsen N, Keller M, Rosenberg M (2014) How virtualisation, decentralisation and network building change the manufacturing landscape: an Industry 4.0 perspective. Int J Mech Industrial Sci Eng 8(1):37–44
9. Tortorella GL, Fettermann D (2018) Implementation of Industry 4.0 and lean production in Brazilian manufacturing companies. Int J Prod Res 56(8):2975–2987
10. Ivanov D (2020) Predicting the impacts of epidemic outbreaks on global supply chains. Transp Res Part E 136:101922
11. Leitão P (2009) Agent-based distributed manufacturing control: a state-of-the-art survey. Eng Appl Artif Intell 22(7):979–991
12. Shen W, Hao Q, Yoon HJ, Norrie DH (2006) Applications of agent-based systems in intelligent manufacturing. Adv Eng Inform 20(4):415–431
13. Ren J, Jenkinson I, Tobora OO (2026) Grounded expert systems for offshore safety: enhancing rule-based risk assessment with retrieval-augmented generation (RAG) for auditable explanations. In *Proceedings of the International Conference on Electronics Technology (ICET)*, Chengdu, China, 29–31
14. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-T, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 33:9459–9474
15. Zheng P, Wang H, Sang Z, Zhong RY, Liu Y, Liu C, Mubarok K, Yu S, Xu X (2021) Smart manufacturing systems for Industry 4.0. Front Mech Eng 13(2):137–150
16. Li L, Su F, Zhang W, Mao JY (2018) Digital transformation by SME entrepreneurs: a capability perspective. Inform Syst J 28(6):1129–1157

17. Ivanov D, Dolgui A (2021) A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. Prod Plann Control 32(9):775–788
18. Ambulkar S, Blackhurst J, Grawe S (2015) Firm's resilience to supply chain disruptions: scale development and empirical examination. J Oper Manag 33:111–122
19. Fogliatto FS, Da Silveira GJC, Borenstein D (2012) The mass customization decade: an updated review of the literature. Int J Prod Econ 138(1):14–25
20. Pine BJ (1993) Mass Customisation: The New Frontier in Business Competition. Harvard Business School Press, Boston
21. Salvador F, De Holan PM, Piller F (2009) Cracking the code of mass customization. MIT Sloan Manage Rev 50(3):71–78
22. Tseng MM, Jiao RJ, Wang C (2010) Design for mass personalisation. CIRP Ann 59(1):175–178
23. Cohen WM, Levinthal DA (1990) Absorptive capacity: a new perspective on learning and innovation. Adm Sci Q 35(1):128–152
24. Teece DJ (2007) Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. Strateg Manag J 28(13):1319–1350
25. Matt C, Hess T, Benlian A (2015) Digital transformation strategies. Bus Inform Syst Eng 57(5):339–343
26. Vial G (2019) Understanding digital transformation: a review and a research agenda. J Strategic Inform Syst 28(2):118–144