Human Factors in Robots, Drones and Unmanned Systems, Vol. 218, 2026, 18–27
https://doi.org/10.54941/ahfe1007674

Agentic LLMs for Scalable, Verifiable System Health Digital Twins

Christopher M. Norton[1], Krishna R. Pattipati[1], Jordan Thurston[1], Deepak Haste[1], Sudipto Ghoshal[1], Somnath Deb[1], and William Lawless[2]

[1]Qualtech Systems, Inc., Rocky Hill, CT, USA
[2]Paine College, Augusta, GA, USA

ABSTRACT

System Health Management (SHM) digital twins have evolved from specialized engineering tools into enterprise-wide critical systems supporting diagnostics and lifecycle decision support, yet scaling the creation, validation, and maintenance of detailed causal models remains a bottleneck due to labor-intensive, expert-driven processes that do not scale with system complexity or lifecycle evolution. This paper presents an AI-driven framework addressing this challenge through a tightly integrated neuro-symbolic architecture that combines agentic large language models (LLMs) as constrained knowledge extraction agents with a rigorous symbolic reasoning core grounded in multi-functional causal modelling, enforcing structural, semantic, and logical constraints to transform extracted knowledge into verifiable, executable diagnostic models while shifting human expertise toward validation, governance, and continuous improvement. The framework implements an end-to-end "ingest–extract–structure–verify" pipeline converting artifacts (i.e., technical manuals, schematics, FMECA data) into formal causal models compatible with TEAMS and SysML-based representations, providing a single source of truth for downstream applications including fault detection and isolation, prognostics, sensor optimization, training scenario generation, and lifecycle-informed design. Demonstrated results show up to an 80% reduction in engineering effort and rapid model generation at previously impractical scales, with aerospace and space system deployments confirming accurate, scalable operational reasoning, while an enterprise operating model treats the digital twin as a governed, evolving asset integrated across design, operations, maintenance, and training, enabling continuous adaptation from field data and offering a practical path to trustworthy, adaptive digital twins that deliver sustained enterprise-scale value.

Keywords: Agentic LLMs, System health digital twins, Model based systems engineering, Systems modelling

INTRODUCTION

System Health Management (SHM) digital twins have evolved from specialized engineering tools into enterprise-critical systems supporting diagnostics and lifecycle decision support. Despite advances in diagnostic reasoning, their scalability remains constrained by the difficulty of creating, validating, and maintaining detailed causal models that capture complex system relationships. This knowledge acquisition bottleneck—rather

Received March 30, 2026; Revised April 27, 2026; Accepted May 14, 2026; Available online July 20, 2026

than computational inference—has emerged as the primary barrier to widespread deployment.

This paper presents a neuro-symbolic framework that addresses this challenge by combining agentic large language models (LLMs) with a formally defined causal reasoning core. In this architecture, LLMs act as constrained knowledge extraction agents, transforming heterogeneous lifecycle artifacts into structured, verifiable models under strict semantic and logical constraints. The approach shifts human effort from manual model construction to validation and governance, enabling scalable, high-confidence model generation.

The framework implements an end-to-end "ingest–extract–structure–verify" pipeline that produces multi-functional causal models compatible with TEAMS and SysML-based representations, supporting applications including fault isolation, prognostics, sensor optimization, and training. By integrating this capability within an enterprise operating model, the digital twin becomes a governed, evolving asset that adapts continuously based on operational feedback. This work reframes SHM scalability as a knowledge lifecycle problem and provides a practical path toward deployable, enterprise-scale digital twins. Human considerations are integrated throughout system design, development, fielding, sustainment, and retirement. The attention to human systems integration in system development programs drove hundreds of human-centred design improvements. Efforts were concentrated to maximize total system performance through improvements in human workload, ease of maintenance, and personnel safety which resulted in a cost avoidance of billions of dollars and prevention of hundreds of fatalities and disabling injuries for the system (Booher and Minninger, 2003).

INFERENCE SCALABILITY VS. THE KNOWLEDGE BOTTLENECK

To understand the present bottleneck in scalable SHM digital twins, it is necessary to revisit what is, in fact, a solved problem: large-scale diagnostic inference. The technical foundations for reasoning about faults, tests, and maintenance actions in complex systems were laid more than three decades ago, evolving from early work on sequential diagnosis to the integrated Testability Engineering and Maintenance System (TEAMS) framework. These foundations established that, when supported by appropriate representations and algorithms, diagnostic reasoning can scale to very large systems. What remained unresolved—and what has consistently limited enterprise-scale deployment—is the scalability of knowledge acquisition, validation, and sustainment.

Sequential Diagnosis and the Scalability of Inference (1990–2000)

Early work on sequential diagnosis established rigorous information- and decision-theoretic foundations for fault isolation under uncertainty, framing diagnosis and test selection as a joint stochastic optimization problem solved via heuristic-guided search (e.g., AO*) capable of handling multiple faults, unreliable tests, and mode-dependent behaviour without combinatorial explosion. The introduction of Multi-Signal Flow Graphs (MSFGs) provided a compact, hierarchical representation aligned with system structure,

enabling efficient large-scale reasoning that was later operationalized in the TEAMS framework, which demonstrated diagnostic inference over tens of thousands to ~100,000 failure sources, even on limited 1990s-era hardware. Subsequent deployments across domains such as aerospace, rotorcraft, and medical systems consistently confirmed that model-based reasoning scales effectively in real, safety-critical environments, establishing that inference scalability has been solved for decades. However, model construction and sustainment have remained manual, labor-intensive, and difficult to scale, requiring subject-matter experts to encode and continuously update causal relationships from diverse artifacts; as systems evolve, the cost and complexity of maintaining these models dominate lifecycle effort, revealing that the true bottleneck is structural, knowledge acquisition and sustainment, vice computational inference.

NEURO-SYMBOLIC FORMALISM

A key implication of the preceding discussion is that large language models alone cannot solve the scalability challenges of SHM digital twins. While agentic LLMs excel at extracting and synthesizing information from heterogeneous, unstructured lifecycle artifacts, they do not inherently reason over dynamic system behaviour, operational modes, temporal constraints, or fault propagation. Without a precise target representation, LLM-driven extraction risks producing static, inconsistent, or non-executable outputs unsuitable for diagnosis, prognostics, or certification. The mathematical formalism introduced in the next section provides this critical structure by defining exactly what the agentic pipeline may produce. By constraining outputs to a well-defined class of multi-functional causal models—with explicit semantics for structure, failure propagation, observability, temporal behaviour, and mode dependence—free-form text synthesis becomes a disciplined structure generation problem. This formal grounding transforms agentic LLMs from speculative generators into scalable productivity multipliers, producing outputs that are verifiable and executable within an enterprise SHM digital twin lifecycle. The “Symbolic” core is thus not a static graph but a Multi-Functional Causal Model capable of dynamic state estimation, redundancy management, mode switching, and signal transformation, ensuring that the LLM-driven extraction pipeline produces actionable, rigorous, and certifiable models.

The Causal Graph Structure

We define the diagnostic model as a tuple $G = (V, E, \Sigma, A)$, representing a directed, semantically enriched graph that encodes system structure, failure behaviour, observability, and propagation dynamics, and which defines the admissible hypothesis space for agentic extraction. The node set V captures the physical, functional, and logical entities, including structural elements (systems, subsystems, components), associated failure modes, their resulting effects, and observation points such as sensors and tests; additional logic and control constructs, such as mode switches, redundancy relationships,

and functional transformations, govern system behaviour under varying operational conditions. The signal set Σ represents the finite set of system-relevant functions or signals, with the edges E denoting directed, signal-labelled causal dependencies that encode conditional and temporal propagation of faults through the system. Observability is defined through mappings between failure effects and the subset of signals detectable by each observation node, enabling formal reasoning about detection and isolation. The set of node attributes A capture domain-specific parameters, including propagation delays, test latencies, failure rates, and repair costs, grounding the model in system physics and operational constraints.

Dynamic Update of the Diagnostic Matrix and Agentic Extraction as Function Approximation

Rather than treating the Dependency Matrix (D-matrix) as a static lookup table, our framework generates it dynamically via a mapping function Ψ in the TEAMS inference engine, where $D(t) = \Psi(G, S_{mode}, \Delta_{time})$ encodes the causal graph, current operational mode, and temporal effects such as propagation delays and test latencies, enabling real-time updates as system states evolve and extending the model to handle multi-valued outcomes and test uncertainty, including false alarms and missed detections, with inference framed as identifying the most probable set of failure modes or components given the observed test results. The Agentic LLM pipeline formalizes this process as a probabilistic approximation Φ of a mapping Φ from unstructured artifacts—PDFs, diagrams, text—to structured models, producing a candidate model $\hat{M}$ that is subsequently refined through a verification function V_{verify}, which combines human oversight with logical constraints and minimizes a loss function penalizing deviations from physical reality and engineering rules, thereby constraining outputs to valid causal model syntax, mitigating hallucination risks, and ensuring provenance, traceability, and certifiability suitable for safety-critical domains. As concluded in (Rai et al.), the model supports multiple views derived from the single source of truth separating the structure (the causal graph) from inference (the algorithms operating on the graph). The elements of architecture that are key to scalability are enumerated as follows.

1. **Fault Tree Analysis (FTA):** Derived by traversing G to identify minimal cut sets for critical system failures.
2. **FMECA:** Automatically generating Failure Modes, Effects, and Criticality Analysis tables by forward-propagating failure modes $f \in F$ to system-level effects, $V_{effects}$.
3. **Sensor Optimization and Supportability Analysis:** Quantify sensor placement optimization and support costs via optimized fault isolation strategies during design.
4. **Automatic Diagnosis:** On-board health monitoring and mission impact assessment, remote diagnosis and tele-diagnosis, and embedded diagnosis based on implementation architectures that range from smart self-diagnosing and intelligent machines to centralized sensor processing in the cloud.

5. **Prognostics:** Extending the graph with time-dependent degradation functions, enabling RUL prediction, time-to-alarm (TTA) and time-to-maintenance (TTM).
6. **Make Every Technician Perform like an Expert:** This includes customer self-help and online help, multilingual call centre support (non-expert call centre agents, prompted by remote diagnostic servers) and dispatching right materials, tools, skills and knowledge relevant to the observed context. The model enables personalized training by generating interactive training content, fault scenarios, quizzes, and what-if missions. FRACAS, telemetry feedback and learning records store (LRS) data not only refine models but also suggest new training scenarios (e.g., emerging fault signatures).
7. **Autonomic Logistics:** Interface to autonomic logistics and commercial customer relationship management (CRM) software by leveraging web-based methods to interface with automatic logistics environment for proactive parts ordering and prognostics system health management.

AGENTIC LLM PIPELNE ARCHITECTURE: THE "INGEST-EXTRACT-STRUCTURE-VERIFY" LOOP

The central contribution of this work is the Agentic LLM pipeline designed to solve the mapping problem previously identified. Unlike standard "Chat with PDF" approaches, which rely on semantic search and often hallucinate relationships, our architecture employs a Multi-Agent Workflow with strict role separation and specific, well-defined tasks for each agent. The objective is to decompose the massive multi-disciplinary cognitive load of model creation into discrete, verifiable tasks: Ingestion, Extraction, Structuring, and Validation. This pipeline can be viewed as an operational realization of the constrained mapping $\hat{\Phi}$ introduced in Section 3.3, with each agent responsible for enforcing a different subset of structural, semantic, and provenance constraints.

The Agentic LLM pipeline operationalizes the mapping from unstructured artifacts to structured causal models through a multi-stage "ingest–extract–structure–verify" workflow, in which each stage enforces complementary structural, semantic, and provenance constraints. Rather than relying on free-form generation, the pipeline decomposes model construction into discrete, verifiable tasks: ingestion normalizes heterogeneous inputs and engineering artifacts; extraction identifies entities and relationships while enforcing causal and syntactic constraints; structuring assembles a candidate model and highlights ambiguities or low-confidence elements; and verification incorporates human subject-matter expert review as the final authority. To further reduce hallucination risk, the approach integrates context engineering with graph-based retrieval, ensuring that model generation is grounded in explicit relationships rather than raw text. This architecture shifts effort from manual authoring to targeted validation, enabling scalable, high-confidence model generation suitable for safety-critical applications.

Closed-Loop Digital Twin Architecture

With the bottleneck addressed by the agentic pipeline, the barrier to enterprise deployment shifts from model creation to lifecycle integration, transforming the resulting model from a static artifact into the engine of a continuous closed loop. Using TEAMS-Designer as the reference implementation, which supports RAAML-compliant SysML v2 import/export and reasoning over multi-functional causal graphs, the digital twin emerges from an ordered, ontology-driven extraction process into a computable knowledge model rather than descriptive documentation. Structural decomposition, functional roles, failure modes, mode-dependent D-matrix, propagation paths, and test designs are explicitly represented as entities within the causal graph, enabling direct execution by the Neuro-Symbolic reasoning core and allowing common technical questions—component identification, functional purpose, fault causality, observable symptoms, diagnostic procedures, and configuration compatibility—to map cleanly to specific model layers rather than ad hoc reasoning. This layered construction supports structured queries for troubleshooting, prognostics, mission impact assessment, sensor optimization, and training scenario generation over a shared enterprise asset. Importantly, the methodology is system-agnostic: while illustrated with vehicle and aerospace examples, the same modelling spine applies to aircraft, power plants, manufacturing cells, medical devices, and networked cyber-physical systems, with only domain-specific ontologies and physics differing, while the extraction workflow, verification logic, and reasoning engines remain unchanged—enabling enterprise-scale deployment without reintroducing the historical bottleneck.

The Six-Step SHM Process and Application to Training Curriculum Development

The digital twin does not exist in a vacuum; it drives a continuous enterprise process. This process transforms the digital twin from a static artifact into a living asset, with the Agentic pipeline enabling rapid updates as new field data reveals novel failure modes, such as unanticipated part failures; agents ingest service reports, propose structural graph updates, and present them to SMEs for approval, closing the loop between Operations and Design. In Breaking the Knowledge Acquisition Bottleneck (Rai et al.), the Knowledge Acquisition phase achieved a 17× acceleration in modelling tasks compared to manual baselines while maintaining 92% accuracy and 91% completeness, and the Operational Reasoning phase reduced "False Pulls" by 17% with 97% accuracy, demonstrating both efficiency and operational scalability for edge deployment and closed-loop impact. Across three case studies, the Neuro-Symbolic framework enabled model generation by up to 17× faster while scaling to demanding safety-critical environments. As detailed in recent literature on operating digital twins (Rosen et al.), the core executes a six-step closed loop process as demonstrated in Figure 1.

The SHM digital twin operates as a closed-loop system in which sensed data and observations are continuously evaluated against the causal model

to detect anomalies, isolate faults, predict future behaviour, and inform maintenance decisions. Rather than emphasizing the individual steps, the key contribution is the integration of diagnostic reasoning with enterprise actions and feedback: maintenance outcomes are not terminal events but inputs to the model lifecycle. When discrepancies arise, such as unsuccessful repairs or previously unmodeled failure behaviour, they expose gaps in the causal structure, triggering a formal Model Update Request. This request is processed by the agentic pipeline, which ingests new field evidence, proposes structural or parametric updates, and submits them for SME validation, thereby closing the loop between Operations and Design.

A recurring failure in digital transformation is treating Digital Twins as isolated technical artifacts rather than enterprise systems embedded in organizational processes. For SHM digital twins to deliver sustained value in mission-critical domains, they must be integrated with design, certification, operations, maintenance, training, and governance workflows, functioning as living, managed assets whose value accrues through continuous use, structured feedback, and controlled evolution. From an enterprise perspective, a diagnostic digital twin is akin to software, data, or intellectual property, underpinning decisions that affect safety, availability, cost, and regulatory compliance, and requiring clear ownership and stewardship by a designated authority or multi-disciplinary team serving as the "Source of Truth" for approving agentic pipeline updates. The twin follows a distinct lifecycle, evolving from initial design (v1.0) through testing (v1.1) and branching for different fleet configurations (v1.2-Block 4 vs. v1.2-Block 5), with explicit version control ensuring that diagnostic reasoning remains aligned with the physical configuration, preventing unsafe or misleading maintenance actions.

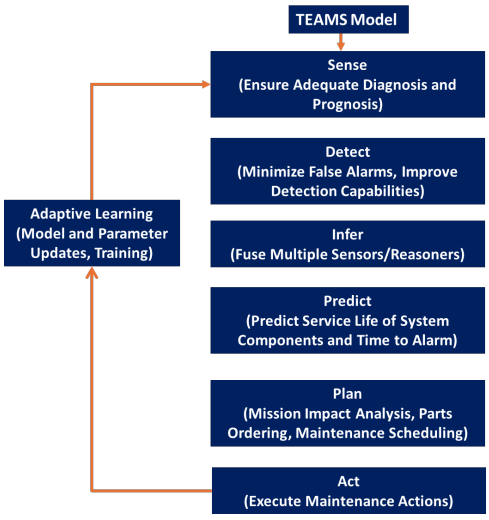

Figure 1: The intelligent reasoning engine core (sense – detect – infer – predict – plan – act) followed by a triggered model adaptation process. real-time data (error codes, sensor data) feeds the model-based reasoner, which executes the Sense-Detect-Isolate-Predict-Plan-Act loop to drive maintenance actions.

Beyond operational diagnostics, the validated causal model also supports workforce development by enabling an Intelligent Tutoring System that generates adaptive, model-driven training without manual curriculum design. Because the causal graph explicitly encodes failure modes, effects, and observables, it can be used to automatically generate realistic "what-if" fault scenarios and diagnostic exercises, allowing technicians to train on system-level reasoning rather than static procedures. Training interactions, in turn, provide feedback on both user performance and model fidelity: recurring diagnostic errors across users may indicate ambiguity or gaps in the underlying model or source documentation. This creates a secondary feedback loop in which training data informs both model refinement and content generation, further integrating operations, training, and continuous improvement within the digital twin lifecycle (see Figure 2).

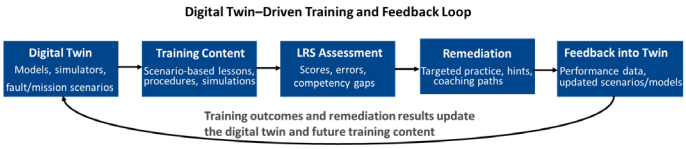

Figure 2: The enterprise feedback loop for training. operational data flows back into the digital twin to generate new training scenarios, closing the loop between operations and workforce development.

Closed-Loop Feedback

The value of an SHM digital twin emerges only when embedded in a closed-loop enterprise-wide operating model defined by a "Model–Sense–Infer–Predict–Plan–Act" cycle. In the forward loop, design artifacts such as FTA, FMECA, testability analyses, and schematics seed the initial digital twin, which is deployed to support field operations; in the feedback loop, maintenance outcomes—including fault-not-found events, newly observed failure modes, and recurring issues—serve as inputs that drive continuous model adaptation. These signals are processed by the agentic pipeline, where contradictions between observed outcomes and model predictions trigger automated model defect reports and targeted re-evaluation of causal relationships. Conceptually, this architecture functions as an enterprise-level supervisory control system, with the digital twin mediating between fast-time-scale physical processes and slower human decision-making to ensure that operational knowledge is systematically incorporated into design, certification, and training activities.

Governance, Configuration Management, and Assurance

Because the digital twin encodes domain knowledge rather than executable code alone, assurance must address model completeness, consistency, and coverage, not just software correctness, necessitating explicit governance. Diagnostic models require strict configuration management, with every change to the causal graph traceable to authorizing documents, and updates must undergo regression testing against historical "Gold Standard" scenarios to ensure fault isolation paths remain valid. Traceability is maintained through

the Neuro-Symbolic architecture, linking every node and edge to its source artifact for auditability. Security and access control are critical: role-based access differentiates diagnostic execution, model validation, and agentic pipeline administration, while data sovereignty ensures proprietary subsystem models can integrate without exposing internal logic, enabling multi-party supply chain deployment without violating IP or regulatory constraints. Clear organizational roles are also essential, with engineers, operators, maintainers, and managers interacting with the twin at appropriate abstraction levels, and the Human–AI collaboration model balancing automation for scale with human responsibility for judgment and approval. This combination of governance, security, traceability, and accountability underpins trust, certification, and enterprise-scale adoption of SHM digital twins.

Interoperability and Enterprise Integration

Enterprise-scale adoption of SHM digital twins requires interoperability across tools, vendors, and organizational boundaries, as even technically advanced twins risk becoming isolated silos without consistent semantics and exchange mechanisms. In SHM, interoperability must span model structure, semantic meaning, and operational access, and the proposed framework addresses all three by aligning the neuro-symbolic diagnostic core and agentic knowledge acquisition pipeline with emerging standards. At the structural level, the TEAMS multi-functional causal model is aligned with the SysML v2 metamodel, enabling bidirectional exchange with Model-Based Systems Engineering (MBSE) environments, where system elements, failure modes, and observability constructs map directly to SysML definitions, preserving traceability between design and diagnostic reasoning. Semantic consistency is enforced by mapping extracted entities and relationships to controlled enterprise ontologies, with alignment mechanisms reconciling terminology across heterogeneous artifacts to ensure that equivalent concepts are represented uniformly across engineering, logistics, and maintenance systems. For operational integration, the framework adopts the Asset Administration Shell (AAS) as a standardized interface layer, exposing diagnostic and prognostic capabilities through a dedicated submodel that allows external systems to query the twin for health state, fault isolation, and decision support outputs via vendor-neutral interfaces. By decoupling internal reasoning from external access and grounding all representations in shared structural and semantic standards, the framework supports scalable deployment across federated, multi-vendor environments. In contrast to existing digital twin standards that emphasize architecture but not knowledge lifecycle, this approach explicitly integrates knowledge acquisition, validation, and sustainment, providing a practical path toward interoperable, scalable, and certifiable SHM digital twins.

RESEARCH DIRECTIONS

The proposed framework enables scalable knowledge acquisition but also motivates several key research directions toward fully adaptive SHM digital twins. First, self-correcting models should autonomously update parameters

such as failure rates, test reliabilities, and operational costs from field data, while preserving structural constraints and auditability, with human oversight reserved for topology changes. Second, adaptive sensing strategies should dynamically reconfigure observability based on mission context and uncertainty, enabling real-time trade-offs between sensing cost and diagnostic risk. Third, distributed and federated reasoning architectures are required to scale toward system-of-systems applications, allowing localized inference with coordinated exchange of high-level health states. Finally, formal assurance methods and enterprise adoption models must ensure trust, validation, and sustained organizational integration, recognizing that scalability depends as much on governance and process maturity as on technical capability.

CONCLUSION

This paper shows that the primary limitation in scaling SHM digital twins is not diagnostic inference, but the acquisition, validation, and sustainment of structured knowledge. By combining agentic LLM-based extraction with a formally constrained neuro-symbolic reasoning core, the proposed framework enables rapid, scalable generation of verifiable causal models while preserving the rigor required for safety-critical applications.

Beyond model creation, the framework establishes a closed-loop enterprise architecture in which operational data continuously informs model evolution, supporting diagnostics and life-cycle decision support from a shared knowledge base. Alignment with MBSE standards, enterprise ontologies, and operational interfaces enables interoperability across tools and organizations, allowing digital twins to function as integrated enterprise assets rather than isolated artifacts.

Taken together, these contributions provide a practical foundation for trustworthy, adaptive SHM digital twins that can scale with system complexity and lifecycle demands, enabling sustained value across design, operations, and maintenance in mission-critical environments.

REFERENCES

Deb, S., Mathur, A., Willett, P.K. and Pattipati, K.R. (1998) "Decentralized Real-time Monitoring and Diagnosis," IEEE Systems, Man, and Cybernetics Conference, San Diego, CA, October.

Luo, J., Choi, K., Pattipati, K.R., Qiao, L. and Chigusa, S. (2006) "Distributed fault diagnosis for networked, embedded automotive systems," in: Proc. IEEE International Conference on SMC, Taipei, Taiwan.

Luo, J., Pattipati, K.R., Qiao, L. and Chigusa, S. (2005) "Agent-based Real-time Fault Diagnosis," IEEE Aerospace Conference, Big Sky, Montana, March.

Rai, R., Pattipati, K. R., Norton, C., Thurston, J., Haste, D., Ghoshal, S., Deb, S., and Lawless, W. (2026) "Breaking the Knowledge Acquisition Bottleneck for System Health Digital Twins: Agentic LLMs, Neuro-Symbolic Reasoning, and Enterprise-Scale Deployment", proceedings of the 2026 International Conference on Applied Computing: Bridging Theory, Innovation, and Real-World Impact, Las Vegas, NV.

Rosen, K. M. and K. R. Pattipati. (2023) "Operating Digital Twins within an Enterprise Process", in: The Digital Twins, A. Drobot, R. Minerva and N. Crespi (Eds.), Springer.