

Large Language Model Agents in Manufacturing: A Survey of Applications, Architectures, and Open Challenges

Tinku Talreja

*Department of Computer
Science & Engineering
IET, SAGE University
Indore, India
Talrejatinku123@gmail.com*

Prof Sapna Patel

*Department of Computer
Science & Engineering
IET, SAGE University
Indore, India
sapna.patel@sageuniversity.in*

Abstract— Large language models (LLMs) have moved from research curiosities to candidate components of production systems in a remarkably short time, and manufacturing has become one of the most active testing grounds for their agentic use. This survey examines how LLMs and LLM-based agents are being applied across manufacturing functions, organizing the literature into a four-part taxonomy: knowledge access, decision support, autonomous orchestration, and human-machine interaction. Rather than cataloging systems, we analyze them critically: we compare architectural choices such as retrieval-augmented grounding, role-specialized multi-agent designs, and digital twin coupling; we examine how published systems are evaluated; and we identify where reported capabilities rest on simulation rather than plant deployment. The analysis reveals recurring structural weaknesses—hallucination in jargon-dense industrial contexts, non-deterministic behavior colliding with real-time constraints, evaluation practices that do not measure safety, and a near-total absence of shared benchmarks. We consolidate these observations into a research agenda, arguing that the field's central open problem is no longer whether language agents can reason about production tasks, but how their proposals can be verified before they act on physical equipment. This survey is intended both as a map of a fast-moving literature and as a critical foundation for research on trustworthy agent deployment on the shop floor.

Keywords—Large Language Models, Multi-Agent Systems, Manufacturing, Retrieval-Augmented Generation, Digital Twin, Industry 5.0, Agents

Introduction

Manufacturing has absorbed successive waves of computational intelligence: rule-based automation in the 1980s, distributed multi-agent control in the 1990s and 2000s, and data-driven machine learning through the Industry 4.0 era. Each wave attacked a different bottleneck. Rule-based systems encoded expert procedures but could not generalize; classical multi-agent systems decentralized decision-making but required painstaking hand-engineering of agent behaviors and communication protocols; machine learning models learned from sensor data but remained narrow, single-task instruments requiring labeled data for every new problem [5], [6].

Large language models represent a fourth wave with a qualitatively different profile. Built on the transformer architecture [1] and trained on broad corpora, LLMs bring three capabilities no earlier technology combined: general-purpose reasoning that can be elicited through techniques such as chain-of-thought prompting [2]; natural-language interfaces that make specialist knowledge conversationally accessible; and, when embedded in agent frameworks that interleave reasoning with tool calls [3], the ability to act—querying databases, running simulations, and commanding equipment. Grounding methods such as retrieval-augmented generation (RAG) [4] connect these general capabilities to the specific manuals, logs, and standards of an individual factory.

The manufacturing research community has responded quickly. Since 2023, publications have proposed LLM-based systems for maintenance question answering, process planning, shop-floor scheduling, robot instruction, and fully agentic production orchestration. The literature is, however, fragmented across venues and framings, and enthusiasm frequently outruns evidence. This survey brings the work together, imposes a taxonomy on it, and—following the critical-review tradition rather than the annotated-bibliography one—asks harder questions of it: What do these systems actually do? How are they evaluated? What breaks when they meet real plants? And what must the field build next?

The remainder of the survey is organized as follows. We first review the trajectory from classical multi-agent manufacturing control to language-model agents, establishing what is genuinely new. We then present a four-category taxonomy of applications and analyze representative systems in each category. Subsequent sections assess evaluation practices, consolidate open challenges, and derive research directions. A short conclusion closes the paper.

From Classical Agents to Language Model Agents

Classical Multi-Agent Manufacturing Control

The idea of organizing production around autonomous, cooperating software agents long predates LLMs. Holonic and agent-based control architectures assigned agents to machines, orders, and products, and let global behavior emerge from local negotiation—typically through contract-net protocols in which task announcements, bids, and awards coordinate work allocation [5]. This literature established durable insights: decentralization buys robustness to disruption at the cost of predictability; negotiation protocols need careful design to avoid pathological outcomes; and industrial adoption stalls when agent behavior cannot be certified. Notably, classical agents were competent but inarticulate—their decision logic was hand-coded, opaque to operators, and expensive to extend to new products or processes.

What Language Models Change

LLMs invert the classical trade-off. Where classical agents had rigid competence and no generality, LLM agents have broad generality whose reliability must be engineered after the fact. Three mechanisms define the modern agent stack. First, prompting strategies such as chain-of-thought elicit multi-step reasoning without task-specific training [2]. Second, the ReAct pattern interleaves reasoning traces with tool invocations, letting a model consult a scheduler, query a machine state, or run a simulation mid-deliberation [3]. Third, retrieval-augmented generation grounds responses in retrieved enterprise documents rather than parametric memory, reducing—though not eliminating—fabrication [4]. Surveys of LLM use across mechanics, design, and manufacturing document rapid uptake of this stack [8], and analyses of foundation-model agents in industrial automation map its purposes and capabilities while cataloging open concerns around reliability and safety [9], [10].

The result is a new design space: agents that can read a maintenance manual, explain their reasoning to an operator in natural language, and participate in the same negotiation protocols the classical literature defined—now with the protocol roles filled by models that can improvise. Whether improvisation is a feature or a hazard on a factory floor is, in effect, the organizing question of this survey.

A Taxonomy of LLM Applications in Manufacturing

We organize the literature into four categories distinguished by the degree of autonomy granted to the model, from answering questions to taking actions. Fig. 1 presents the taxonomy together with the cross-cutting concerns every category inherits; Table 1 compares representative systems.

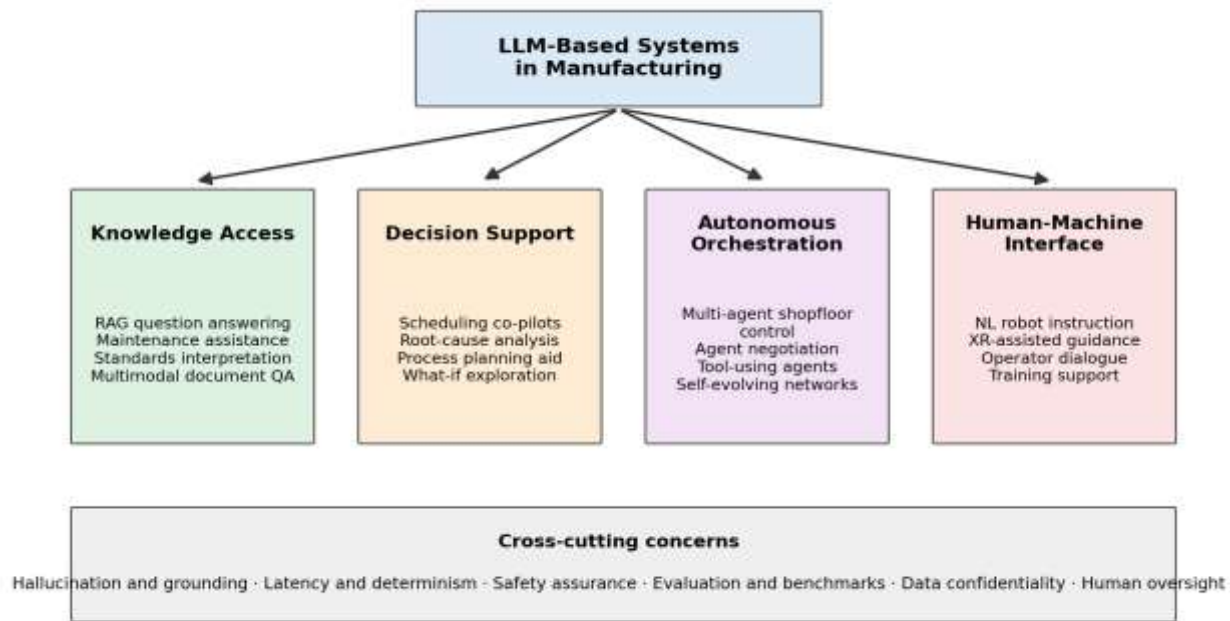


Fig. 1. A taxonomy of LLM-based systems in manufacturing, ordered by increasing autonomy, with cross-cutting concerns shared by all categories.

Knowledge Access

The most deployed category uses LLMs as conversational front-ends to industrial knowledge. RAG pipelines answer technician questions from equipment manuals, maintenance logs, and internal standards [16], and have been extended into extended-reality assistance that delivers retrieved instructions in the technician's field of view [19]. The category's defining engineering problem is domain language: industrial corpora are dense with jargon, abbreviations, and inconsistent formats that defeat generic retrieval. Golden-Retriever addresses this with a reflection step that identifies and expands jargon before retrieval [15]; multimodal approaches extend retrieval to drawings and tables that dominate manufacturing documentation [18]. Evaluations on realistic factory corpora confirm that performance degrades sharply on unprepared data [20]—a finding with an encouraging corollary, since data preparation is a tractable engineering investment.

Decision Support

The second category keeps humans in command but uses LLMs to widen the decision funnel: scheduling co-pilots that translate a planner's natural-language intent into solver configurations, root-cause analysis assistants that correlate anomaly reports with historical incidents [17], and process planning aids that draft routings for engineer review [8], [23]. A recurring design lesson is that the LLM performs best as an interface to specialized machinery rather than as the machinery itself—formulating the optimization problem a solver then solves, or generating candidate hypotheses a statistical model then ranks—because delegation keeps the numerically hard work deterministic while preserving the conversational surface.

Systems in this category nevertheless inherit the knowledge-access problems and add a subtler one: recommendation quality is hard to measure, because the counterfactual—what the human would have decided unaided—is rarely recorded. Published evaluations lean on expert ratings of plausibility, which do not distinguish a genuinely better plan from a fluent one. Automation-impact analyses suggest the largest near-term value lies in exactly these semi-structured cognitive tasks [23], which makes the measurement gap economically consequential, not merely methodological.

Autonomous Orchestration

The most ambitious category grants agents decision authority. The landmark design pattern assigns specialized roles—machine agents that report status, bid-inviting and bidding agents that run allocation auctions, and LLM-empowered thinking and decision agents that reason about the production context rather than executing fixed rules [11]. Extensions incorporate sustainability objectives, using LLM-driven negotiation to trade off production efficiency against carbon emissions [12], and self-evolving agent networks that restructure coordination topologies in response to predictive-maintenance signals [13]. A complementary line couples agents to digital twins so that proposed actions can be simulated before execution [14]—a coupling this survey regards as the category's most consequential idea, for reasons developed below. Across the category, however, validation remains almost entirely simulation-bound, and scheduling benchmarks designed for LLM agents reveal an observability paradox: richer real-time information can degrade agent decisions rather than improve them [21].

Human-Machine Interface

The fourth category uses language as the interface between people and machinery: natural-language robot task specification, dialogue-based operator support, and training assistance for new workers [8], [10]. Its promise is inclusion—lowering the expertise barrier to sophisticated equipment, consistent with Industry 5.0's human-centric mandate [7], [24]. Its risk is miscommunication at the boundary between ambiguous human language and precise machine action, which pushes the category toward the same verification requirements as autonomous orchestration whenever the interpreted command moves metal.

TABLE 1. REPRESENTATIVE LLM-BASED SYSTEMS ACROSS THE TAXONOMY

Category	Representative approach	LLM role	Evaluation setting	Key reported limitation
Knowledge access	Jargon-aware agent; RAG over industrial knowledge base [15], [16], [18]	Query understanding; grounded answer synthesis	Document QA benchmarks; use studies	Jargon and formatting degradation; retrieval [20]
Decision support	Scheduling and root cause co-pilots [17] [23]	Intent translation; hypothesis generation	Expert plausibility ratings	No counterfactual baseline; fluency bias
Autonomous orchestration	Role-specialized multi-agent shopfloor control [11]–[13]	Reasoning and negotiation inside agent roles	Simulated shopfloors	Simulation-only validation; observability paradox [21]
Human-machine interface	NL robot instruction; XR maintenance guidance [19]	Language-to-action interpretation	Task-completion trials	Ambiguity at the language-action boundary

Table 1 compares representative systems by the autonomy granted to the model, how claims were evaluated, and the limitation most often reported.

Critical Analysis of Evaluation Practices

A survey that stopped at capabilities would flatter the field. The sharper picture emerges from how systems are evaluated. Three patterns recur. First, simulation dominance: orchestration papers evaluate on synthetic or simulated shopfloors, and almost none report brownfield deployment; the gap between simulated negotiation

and a plant with legacy PLCs, network partitions, and union work rules is unexamined. Second, metric mismatch: papers report task-success rates and makespan improvements, but rarely measure the quantities that determine deployability—frequency of unsafe or infeasible proposals, latency distributions under load, and behavior when retrieval returns nothing relevant. Studies focused on hallucination in industrial settings show error rates that would be intolerable in closed-loop control [22], yet most orchestration papers do not report an equivalent statistic. Third, benchmark absence: unlike vision-based inspection, which matured around shared datasets, LLM-in-manufacturing research has no widely adopted testbed; early efforts such as calibrated scheduling benchmarks [21] and safe factory evaluation environments [20] are recent and not yet standard. Until they are, cross-paper comparison is essentially impossible, and reported gains cannot accumulate into scientific knowledge.

A fourth observation concerns architectural choices themselves. The literature divides between single-agent designs, in which one model with tools handles a task end to end, and multi-agent designs that decompose work across specialized roles [10], [11]. Multi-agent decomposition is usually justified by modularity and by mapping naturally onto organizational functions, but it multiplies inference cost, introduces coordination failure modes that single agents cannot exhibit, and complicates attribution when something goes wrong. Few papers compare the two designs on equal tasks, so the field currently cannot say when multi-agent structure pays for its overhead. Similarly, centralized orchestration (one coordinating agent delegating to workers) and decentralized negotiation (peers bidding under a protocol [5], [11]) embody different robustness-predictability trade-offs that echo debates settled empirically in the classical agent literature—debates the LLM generation appears to be re-running without citing the earlier results.

Industrial Readiness and Adoption Pathways

For practitioners, the taxonomy implies a staged adoption path that matches autonomy to demonstrated reliability. The natural entry point is knowledge access: a RAG assistant over maintenance documentation requires no actuation authority, produces measurable value in reduced search time, and—critically—builds the curated, machine-readable document corpus that every later stage depends on [15], [16]. Decision support follows once the corpus is trustworthy, with recommendations logged against subsequent human decisions to accumulate exactly the counterfactual data whose absence this survey criticized. Orchestration authority should arrive last, if at all, and only behind verification mechanisms of the kind discussed below.

Two structural observations condition adoption. First, small and medium-sized manufacturers—the majority of the sector—cannot staff prompt-engineering or MLOps functions, which favors packaged, connector-based deployments and lightweight implementation methods over bespoke systems; early work on agile RAG adoption in industrial SMEs points in this direction. Second, integration effort concentrates not in the model but at its boundaries: connecting to MES and historian systems, normalizing legacy document formats, and fitting outputs into existing approval workflows. Published system papers rarely report this integration cost, which likely understates the real barrier to the deployments they envision. Workforce response is equally under-reported: the Industry 5.0 framing predicts that assistants which visibly defer to human judgment will be adopted faster than those positioned as replacements [7], [24], but field evidence on either side of that prediction is thin.

Open Challenges

Table 2 consolidates the challenges surfaced across categories and the mitigation approaches the literature currently offers. Two deserve narrative emphasis. Grounding under industrial language remains the load-bearing problem for every category: no amount of agent architecture compensates for a model that misreads a torque specification. And verification before actuation is the unresolved problem specific to the autonomous end of the taxonomy: current systems largely trust the model's proposal, with digital twin coupling [14] the most promising—but least standardized—countermeasure. The energy cost of large-scale inference [26] and

the confidentiality of process data add operational constraints that favor smaller, locally deployed models over frontier-scale APIs.

TABLE 2. OPEN CHALLENGES AND CURRENT MITIGATION APPROACHES

Challenge	Manifestation	Current mitigation and maturity
Hallucination and grounding	Fabricated parameters and procedures in jargon-dense contexts [20], [22]	RAG with jargon handling [15], [18] procedure-level consistency engineering [22]. Partially mature.
Real-time constraints	Non-deterministic latency conflicts with cycle-time and control deadlines	Hierarchical designs reserving LLMs for slow loops [9]. Early.
Verification before actuation	Agent proposals executed without independent feasibility or safety checks	Digital twin simulation gating [14] Conceptual; not standardized.
Evaluation and benchmarks	Simulation-only validation; no shared testbeds; safety unmeasured [21]	Emerging benchmarks and sandboxes [20], [21]. Immature.
Confidentiality and cost	Process data leaves the plant via API calls; inference energy and cost [26]	On-premise open-weight models retrieval-side redaction. Practical but under-studied.

Table 2 links each systemic challenge to the mitigation approaches reported in the surveyed literature and our assessment of their maturity.

Research Gaps and Future Directions

Verification-gated autonomy. The clearest gap sits between the orchestration literature, which grants agents authority, and the safety literature, which shows why that authority is premature. Architectures in which every agent proposal must pass an independent verification layer—digital twin simulation, rule-based safety envelopes, and calibrated risk scoring—before touching equipment are sketched in early work [14] but lack reference designs, failure-mode analyses, and empirical study. We regard this as the field's most valuable near-term research target, and a natural subject for a dedicated design paper.

Benchmarks that measure what matters. Future testbeds should score agents not only on task success and makespan but on unsafe-proposal rates, escalation quality, latency percentiles, and graceful degradation—extending early calibrated benchmarks [20], [21] into a community standard.

Small and specialized models. The economics and confidentiality of plants favor compact models fine-tuned on industrial corpora over frontier APIs; systematic studies of the capability-cost frontier for manufacturing tasks are scarce.

Human-agent teaming. As agents join negotiations that were previously human, questions of authority allocation, explanation quality, and accountability move from philosophy to engineering [7], [24]. Field studies of operators working alongside agent systems are almost entirely absent from the literature.

Sustainability-aware agents. Early evidence that LLM-driven negotiation can internalize carbon objectives [12] suggests a broader program: making environmental cost a first-class term in agent deliberation rather than a post-hoc report.

Conclusion

Language models have given manufacturing agents something they never had: the ability to read, explain, and improvise. This survey organized the resulting literature into four categories of increasing autonomy, compared the architectural ideas that define each, and scrutinized the evidence behind reported capabilities. The pattern is consistent: knowledge access is approaching practical maturity; decision support is useful but poorly measured; autonomous orchestration is architecturally inventive and empirically thin; and interfaces are promising wherever interpretation errors cannot reach actuators. The field's frontier is therefore not greater model capability but engineered trustworthiness—grounding, verification, and evaluation infrastructure that lets an improvising agent be safely wrong. The companion challenge for the research community is to build and study reference architectures with exactly that property.

References

- [1] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [2] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [3] S. Yao et al., “ReAct: Synergizing reasoning and acting in language models,” in *Proc. International Conference on Learning Representations (ICLR)*, 2023.
- [4] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [5] P. Leitão, “Agent-based distributed manufacturing control: A state-of-the-art survey,” *Engineering Applications of Artificial Intelligence*, vol. 22, no. 7, pp. 979–991, 2009.
- [6] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, “Machine learning in manufacturing: advantages, challenges, and applications,” *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23–45, 2016.
- [7] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, “Industry 4.0 and Industry 5.0—Inception, conception and perception,” *Journal of Manufacturing Systems*, vol. 61, pp. 530–535, 2021.
- [8] “A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing,” *Advanced Engineering Informatics*, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S1474034624007171> *
- [9] “Foundation-Model-Based Agents in Industrial Automation: Purposes, Capabilities, and Open Challenges,” *arXiv preprint*, 2026. [Online]. Available: <https://arxiv.org/pdf/2605.02592> *
- [10] “LLM-Powered AI Agent Systems and Their Applications in Industry,” *arXiv preprint*, 2025. [Online]. Available: <https://arxiv.org/pdf/2505.16120> *
- [11] “A Large Language Model-based multi-agent manufacturing system for intelligent shopfloor,” *arXiv preprint arXiv:2405.16887*, 2024. [Online]. Available: <https://arxiv.org/abs/2405.16887> *
- [12] “Large language model-driven multi-agent systems for improving production efficiency and reducing carbon emissions in manufacturing,” *Computers & Industrial Engineering*, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0360835225004450> *
- [13] “Self-Evolving Multi-Agent Network for Industrial IoT Predictive Maintenance,” *arXiv preprint*, 2026. [Online]. Available: <https://arxiv.org/pdf/2602.16738> *
- [14] “Integrating Large Language Model Agents with Digital Twins for Industrial Autonomous Systems,” *arXiv preprint*, 2026. [Online]. Available: <https://arxiv.org/pdf/2606.20761> *

- [15] “Golden-Retriever: High-Fidelity Agentic Retrieval Augmented Generation for Industrial Knowledge Base,” arXiv preprint, 2024. [Online]. Available: <https://arxiv.org/pdf/2408.00798> *
- [16] “Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model,” Computer Standards & Interfaces, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0920548925000248> *
- [17] “Retrieval-augmented generation enhanced LLM for industrial anomaly detection,” Discover Applied Sciences, 2026. [Online]. Available: <https://link.springer.com/article/10.1007/s42452-026-08750-7> *
- [18] “ManuRAG: Multi-modal Retrieval Augmented Generation for Manufacturing Question Answering,” arXiv preprint, 2026. [Online]. Available: <https://arxiv.org/pdf/2601.15434> *
- [19] “Cross-Format Retrieval-Augmented Generation in XR with LLMs for Context-Aware Maintenance Assistance,” arXiv preprint, 2025. [Online]. Available: <https://arxiv.org/pdf/2502.15604> *
- [20] “FactoryLLM: A Safe and Open-Source AI Playground for Evaluating LLMs in Smart Factories,” arXiv preprint, 2026. [Online]. Available: <https://arxiv.org/pdf/2606.14119> *
- [21] “DynaSchedBench: Calibrated Dynamic Scheduling Benchmarks and Observability Paradox in LLM-based Scheduling Agents,” arXiv preprint, 2026. [Online]. Available: <https://arxiv.org/pdf/2605.27566> *
- [22] “Toward Epistemic Stability: Engineering Consistent Procedures for Industrial LLM Hallucination Reduction,” arXiv preprint, 2026. [Online]. Available: <https://arxiv.org/pdf/2603.10047> *
- [23] “The Impact of Large Language Models on Task Automation in Manufacturing Services,” arXiv preprint, 2025. [Online]. Available: <https://arxiv.org/pdf/2505.10581> *
- [24] “Generative artificial intelligence in manufacturing: opportunities for actualizing Industry 5.0 sustainability goals,” Journal of Manufacturing Technology Management, vol. 35, no. 9, 2024. [Online]. Available: <https://www.emerald.com/jmtm/article/35/9/94/1228980> *
- [25] “Multi-agent reinforcement learning for flexible shop scheduling problem: a survey,” Frontiers in Industrial Engineering, 2025. [Online]. Available: <https://www.frontiersin.org/journals/industrial-engineering/articles/10.3389/fieng.2025.1611512/full> *
- [26] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in Proc. 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645–3650.

** Entries marked with an asterisk were verified by title, venue, and year against the linked source, but author names and complete pagination were not confirmed and must be completed from the linked source before submission. No bibliographic details have been invented.*