**ENGINEERING Management** 2026, 13(2): 265–291
https://doi.org/10.1007/s42524-026-5359-0

RESEARCH ARTICLE

**Hao ZHENG, Ziyue GENG, Xun XU**

# Foundation model-based generative AI for smart manufacturing: A paradigm shift

© The Author(s) 2026

**Abstract** The rise of generative AI (GenAI) has prompted significant attention and discourse across academia and industry, as stakeholders grapple with its capabilities, potential applications, and associated risks. Driven by the aim to address the key question of whether and how GenAI can reshape the manufacturing industry, this paper explores the role, applications and prospects of GenAI for manufacturing. A traditional paradigm of AI implementation in manufacturing is initially outlined, followed by a review of GenAI applications in manufacturing through a proposed five-level framework characterizing the depth of GenAI integration. Building on this review and an analysis of the development trajectory of foundation models, it is argued that GenAI not only enhances each stage of the traditional paradigm but also has the potential to establish a new paradigm in smart manufacturing. In the envisioned paradigm, GenAI functions as a self-contained service provider, capable of directly addressing complex manufacturing needs with innovative solutions, while maintaining a balance between task efficiency, human well-being, environmental sustainability, and societal impacts. Aligned with the core principles of Industry 4.0 and Industry 5.0, this paradigm represents a highly desirable evolution for the manufacturing sector. Following this, a GenAI-driven product design-to-manufacturing framework is introduced to ground the paradigm in practical applications. This research provides a robust framework for understanding GenAI's transformative trajectory in manufacturing and sets forth a research agenda for future exploration. Rather than offering definitive conclusions, this work aims to stimulate ongoing discussions and encourage further exploration in this evolving field.

**Keywords** generative AI, smart manufacturing, large language model, vision language model, Industry 4.0, Industry 5.0

Received Nov. 2, 2025; revised Mar. 20, 2026; accepted Mar. 25, 2026

Hao ZHENG
Department of Mechanical and Mechatronics Engineering, The University of Auckland, 1010, New Zealand; Centre for Transformative Garment Production, Hong Kong SAR, 999077, China

Ziyue GENG, Xun XU (✉)
Department of Mechanical and Mechatronics Engineering, The University of Auckland, 1010, New Zealand
E-mail: x.xu@auckland.ac.nz

This work was supported by the China Scholarship Council (No. 202006420001), the University of Auckland, and the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative.

## 1 Introduction

According to the National Institute of Standards and Technology, smart manufacturing refers to fully-integrated, collaborative manufacturing systems that respond in real-time to meet changing demands and conditions in the factory, in the supply network, and in customer needs (NIST, 2018). While definitions vary in emphasis, from networked data capabilities (Mittal et al., 2019; Zheng et al., 2018; Tao et al., 2018), data analytics (Davis et al., 2012), cyber-physical systems (Kusiak, 2018) to collaborative systems (Lu et al., 2020), the common thread is leveraging connectivity and computational intelligence to optimize production processes, enhance productivity, and improve product quality. Over the past few decades, Artificial Intelligence (AI), a branch of computer science dedicated to creating systems capable of performing tasks that require human intelligence, has played a pivotal role in advancing smart manufacturing (Xu et al., 2026). From early expert systems to contemporary machine learning and deep learning advancements, AI has continuously enhanced manufacturing through real-time data analysis, predictive maintenance, automation, and rapid responsiveness to market changes and customer demands (Arinez et al., 2020; Nti et al., 2022; Plathottam et al., 2023). These advancements have collectively elevated manufacturing efficiency, quality, and flexibility.

Recently, the rise of generative AI (GenAI) has captured significant attention from both industry and

academia due to its successful applications, such as ChatGPT, Copilot, and Midjourney, as well as its transformative potential across various domains. GenAI refers to a set of artificial intelligence techniques and models capable of generating text, images, or other forms of data by learning from existing data sets (Ooi et al., 2025; Sengar et al., 2024). Distinct from discriminative AI models that learn the conditional probability $P(Y|X)$, directly mapping inputs to outputs for tasks such as classification and regression. In contrast, generative models learn the joint probability $P(X, Y)$ or the data distribution $P(X)$ itself, capturing the underlying structure of how data are generated (Cao et al., 2023). Within the broader GenAI landscape, it is important to distinguish between traditional generative models and foundation model (FM)-based GenAI. Traditional generative models, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion models, are typically architectures designed for data generation tasks. These models have been explored in manufacturing over the past decade, with applications including synthetic data generation for fault diagnosis (Ren et al., 2023), design synthesis and topology optimization (Bing et al., 2026), process parameter optimization (Wu et al., 2023), sim-to-real transfer (Li et al., 2024c) and robot motion planning (Wolf et al., 2025). In contrast, FM-based GenAI, exemplified by Large Language Models (LLMs) and Vision-Language Models (VLMs), represents a distinct paradigm characterized by large-scale pre-training on massive, diverse data sets, yielding emergent capabilities in reasoning, natural language interaction, and cross-task generalization (Wei et al., 2022b; Manduchi et al., 2025). While traditional generative models have been established in manufacturing for years and can serve as the underlying mechanisms of FMs, this paper focuses on FM-based GenAI, as its emergent general-purpose capabilities hold the potential to fundamentally reshape how manufacturing problems are understood, approached, and solved.

Despite these promising capabilities, manufacturing engineering problems are largely not to be solved solely by generating text, images, or other forms of data; they usually require the utilization of scientific knowledge, specialized engineering tools, and direct interaction with physical systems. This reveals an intrinsic limitation of current GenAI when applied to the manufacturing industry. In manufacturing context, GenAI should be redefined as *a set of technologies that can understand a problem, chart a plan of attack and generate and execute a targeted solution to a manufacturing problem*. This "version" of GenAI goes beyond aiding the process of arriving at a solution; instead, it directly provides the solution from unstructured semantic prompts. Such solutions should transcend the limitation of digital content creation, focusing instead on the integration and optimization of real-world production resources to fulfil manufacturing needs effectively. The new definition suggests the necessity of developing new forms of GenAI specifically designed for manufacturing.

Scholarly interest in GenAI for manufacturing has surged. Kusiak (2025) clarified the concepts of traditional generative models and FM-based GenAI, and identified several domains where GenAI may impact manufacturing. Shahin et al. (2025) provided a comprehensive review of GenAI's technical architectures and applications across product design, inventory management, quality control, maintenance, marketing, and supply chain management, accompanied by industrial case studies. Other recent works have examined the technical architectures of industrial FMs (Ayyat et al., 2025; Ren et al., 2025a; Zhang et al., 2026; Zhao et al., 2025), industrial agentic AI (Ren et al., 2025b), specific diffusion model applications (Leng et al. 2025), the broader landscape of Artificial Intelligence Generated Content (Leng et al., 2026); applications within Industry 5.0 (Ma et al., 2025; Zhang et al., 2025) and digital twin integration (Chen et al. 2025); and domain-specific uses in product design (Leng et al., 2025; Mustapha, 2025), human-robot collaboration (Fan et al., 2025; Dong et al., 2025) and predictive maintenance (Khan et al., 2025; Mikołajewska et al., 2025). These works establish a foundational understanding of the GenAI's capabilities and constraints in manufacturing. Distinct from these efforts, the present work adopts an exploratory approach with a distinct objective: to ascertain whether, in what form, and how GenAI can reshape manufacturing and usher in a new paradigm in smart manufacturing. To this end, we make the following contributions: (1) outline the traditional paradigm of AI implementation in manufacturing as a baseline; (2) propose a five-level framework characterizing the depth of GenAI integration; (3) analyze the developmental trajectory of the underlying FMs; and (4) introduce a futuristic paradigm in which GenAI functions as a self-contained service provider, accompanied by a design-to-manufacturing framework and a case study. This paradigm-level reconceptualization, grounded in goal-oriented investigation, distinguishes our work from existing reviews and offers a forward-looking research agenda for the manufacturing community.

In the subsequent sections, this topic is systematically explored to expand our forward-looking perspectives at the intersection of smart manufacturing and GenAI. While this paper aims to offer insightful answers, providing definitive conclusions is not its objective. Instead, this paper seeks to stimulate and foster ongoing discussions in this evolving field, contributing to a deeper understanding and further exploration of these topics.

## 2 Traditional AI in manufacturing

AI has significantly influenced the manufacturing industry for decades, from early expert systems to machine learning

and now deep learning, each stage driving intelligent manufacturing forward (Arinez et al., 2020; Nti et al., 2022; Plathottam et al., 2023). To effectively explore how GenAI can reshape the manufacturing industry, it helps to look back on how traditional AI has been applied in this sector. However, after decades of development of AI technologies, AI models have become increasingly complex, with numerous techniques and vast variants tailored to specific tasks across every sector and step of manufacturing. Cataloguing all AI techniques used in manufacturing is beyond this paper's scope. Therefore, this section first takes a problem-oriented view to briefly review AI applications in manufacturing and then outlines the traditional paradigm of AI implementation in manufacturing. This foundation paves the way for a subsequent discussion on how GenAI might transform this paradigm.

### 2.1 Problem-oriented overview of AI in manufacturing

AI encompasses a broad range of technologies, which can be categorized into supervised learning, unsupervised learning, and reinforcement learning based on learning paradigms (Goodfellow et al., 2016). Each paradigm includes a variety of models, e.g., traditional machine learning models, deep learning models, generative models, and reinforcement learning models. These models have been applied to numerous tasks, ranging from fundamental tasks, e.g., classification, regression, and clustering to high-level tasks, e.g., natural language processing, computer vision, and robotic control. Every technology has found applications in manufacturing (Wang et al., 2018).

To elucidate the role of AI in manufacturing, a problem-oriented approach is adopted to review its applications. We have identified six research problems, that can be addressed by AI in manufacturing and highlighted several typical applications for each, as shown in Table 1. Exhaustively listing AI applications in manufacturing is not our intention; rather, our intent is to present a structured classification that offers a clear analytical framework and establishes a foundation for the subsequent analysis of GenAI's role in manufacturing. These six categories were derived based on the distinct functional roles AI serves in manufacturing, representing how AI creates value rather than categorizing by specific techniques or process stages. We acknowledge that this classification is not exhaustive, some applications may span multiple categories, and the boundaries are not mutually exclusive. Nevertheless, this framework captures the major problem types addressed by AI in manufacturing as observed in the current literature. For each category, only a select number of relevant studies are referenced.

### 2.2 Traditional paradigm of AI implementation in manufacturing

The previous section summarized six key problems that AI technologies typically address in the manufacturing industry. Building on above discussion, a traditional paradigm of implementing AI in manufacturing can be outlined, as depicted in Fig. 1. We structured this implementation process into four distinct stages to mirror the human cognitive problem-solving process (Pólya and Conway, 2014), transitioning from need clarification to precise problem abstraction and tailored model development, and finally to practical solution integration:

1) **Need clarification**: *Identifying and defining the manufacturing need*. The need can encompass intentions for improving any facet of manufacturing process,

**Table 1** Six research problems addressed by AI in manufacturing: definitions, typical applications, and some references

| Categories | Definition | Applications | Some references |
|---|---|---|---|
| Pattern recognition and analysis | This category uses AI to recognize patterns, anomalies, or features from complex data. | Fault diagnosis, quality control, defect detection, compliance tracking. | Ge et al. (2025); Jiang et al. (2026), Jiang et al. (2025); Zhang et al.(2020); Zheng et al. (2022); Zheng et al. (2020) |
| Predictive modeling and decision support | This category uses AI to predict future outcomes based on the historical data to support decision-making processes. | Predictive maintenance, demand forecasting, and process prediction. | Liu and Bao (2025); Neu et al. (2022); Serradilla et al. (2022); Lv et al. (2023) |
| Optimization and control | This category uses AI to identify optimal strategies and solutions that enhance the efficiency, precision, and effectiveness of various manufacturing operations. | Production planning, robot path planning, human-robot collaboration, logistics planning. | Usuga Cadavid et al.(2020), Zhou et al. (2022), Rolf et al. (2023); Semeraro et al. (2023); Zheng et al. (2023) |
| Data manipulation and utilization | This category uses AI to retrieve, augment, manage and organize large volumes of data from manufacturing operations. | Document retrieval, data augmentation, inventory management, automated reporting, question answering. | Garouani et al. (2022); Ribeiro et al. (2021); Ruiz et al. (2023) |
| Human-machine Interaction | This category uses AI to facilitate effective interactions between humans and machines. | Collaborative robots, user-friendly machine interfaces. | Bonarini (2020), Guo et al. (2021); Gammulle et al. (2023); Liu et al. (2021); Zheng et al. (2025) |
| Customization and personalization | This category uses AI to tailor the product design and manufacturing process to meet personal requirements. | Custom product design, on-demand manufacturing, human-centric manufacturing. | Jiang et al. (2022); Liu et al. (2022), Yin et al. (2021); Wan et al. (2021) |

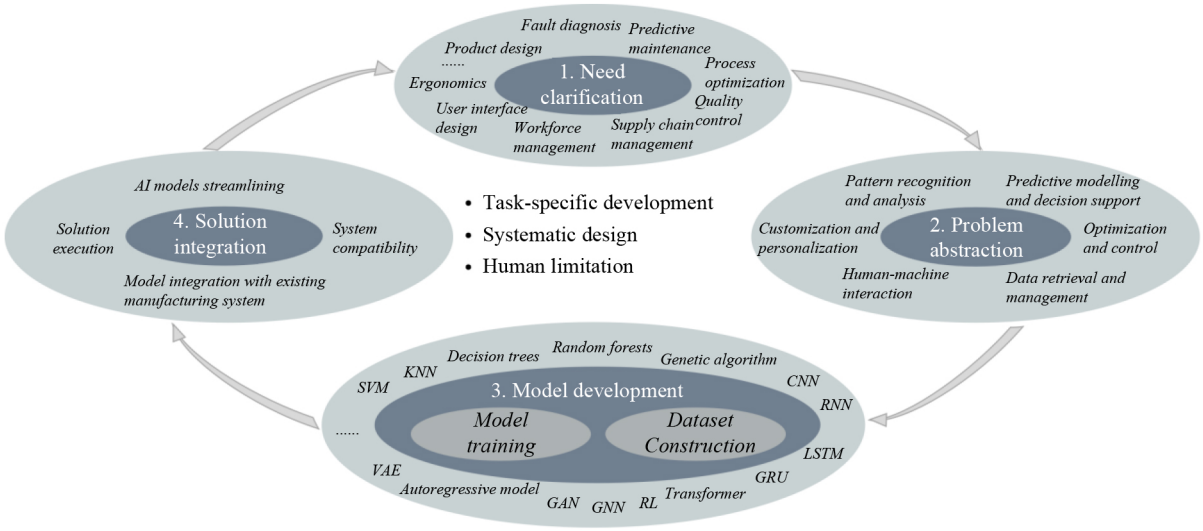


**Fig. 1** The traditional paradigm of AI implementation in manufacturing.

including but not limited to all the applications in Table 1. This stage focuses on application-oriented objectives. The specification remains qualitative and application-centric, distinct from the specific algorithmic formulations utilized in the subsequent stage. Clearly defined needs ensure that subsequent analysis and AI model development are focused and aligned with the overarching goals.

2) **Problem abstraction**: *Abstracting the specific problems that AI can solve based on analyzing the identified need*. This step breaks down the need into specific, mathematically solvable problems, e.g., the problems identified in Table 1, ensuring a targeted and efficient AI solution development. Precise problem abstraction is crucial for developing effective AI solutions.

3) **Model development**: *Developing tailored AI models to address each identified problem*. This step can be further decomposed into data set construction and model design and training. By focusing on specific problems, AI models can be modified and fine-tuned for higher accuracy and effectiveness in solving targeted challenges. Rigorous AI model development is essential for achieving reliable applications.

4) **Solution integration**: *Integrating the developed AI models into a cohesive and process-oriented solution, involving seamlessly connecting the models with existing manufacturing systems and devices*. This integration ensures that the AI-driven approach holistically addresses the initial need, streamlining workflows and enhancing overall efficiency. Effective integration guarantees that AI solutions are practical and beneficial within the manufacturing context.

The traditional paradigm of implementing AI technology in the manufacturing industry exhibits three key features:

1) **Task-specific development**: *Developing distinct AI models for isolated problems*. In this approach, manufacturing needs are decomposed into discrete sub-problems, each requiring a bespoke AI model. Since each model demands data set construction, architecture design, and training, this approach results in substantial resource redundancy and lacks cross-task synergy.

2) **Expert-driven design**: *Reliance on experts to architect the overall solution*. Every step of the traditional paradigm requires careful design by domain experts. The variations in expertise significantly impact how well the final solution meets the manufacturing needs.

3) **Innovation limitation**: *Bounded by pre-defined solution spaces*. In the traditional paradigm, AI functions primarily as an executor or optimizer for expert-crafted solutions. While these models may outperform humans in specific tasks, they lack the agency to question or redefine the underlying solution design. Consequently, the potential for breakthrough innovation is constrained by the initial human conception of the solution.

The analysis in this section reveals that the traditional paradigm of AI implementation in manufacturing mirrors the fundamental process of human analysis and problem-solving, which is universally applicable to general engineering challenges. This indicates that the role of AI is an *auxiliary tool* for humans to replace manual analysis, decision-making, and operations. Consequently, while AI significantly enhances the quality, efficiency, and flexibility of manufacturing, it is inherently constrained by the scope and creativity of human-driven processes. This reliance on expert-crafted solutions limits the potential for AI to achieve transformative innovations.

## 3 GenAI in manufacturing

This section aims to systematically explore the applications of GenAI in manufacturing and the challenges. It begins

with a compact literature review to show the trends of GenAI applications in manufacturing. Then, it categorizes the applications into five levels according to the depth of GenAI integration into manufacturing systems and analyzes how GenAI has been and will be implemented into manufacturing processes at each level. Subsequently, it delves into the technical, practical and social dimensions of the challenges associated with GenAI implementation in manufacturing.

### 3.1 Literature survey: Distribution, trends, and insights

Recognizing that rapid advancements of this emergent field appear in preprint format, literature was sourced from both Scopus and arXiv using the query ("generative AI" OR "LLM" OR "VLM") AND ("manufacturing" OR "industry" OR "engineering" OR "production") across January 2020-December 2025.

To ensure corpus relevance and rigor, the selection process adhered to strict inclusion criteria: (1) the study must explicitly investigate or utilize FM-based GenAI; (2) the content must address specific industrial manufacturing applications; and (3) the work must offer a substantive scholarly contribution (e.g., empirical study, theoretical framework, or systematic review), excluding generic non-technical commentaries.

After manual curation to exclude non-manufacturing studies, 202 articles (2022-2025) were retained. These were classified into 158 empirical contributions, defined as studies presenting original methodologies or systems substantiated by experimental validation, and 44 secondary or conceptual contributions, comprising literature reviews, position papers, and theoretical frameworks without empirical implementation. Eligible empirical contributions were categorized into six application domains (defined in Table 1) based on their primary focus, with their temporal distribution detailed in Table 2. This curated corpus provides a foundation for analyzing GenAI's applications and transformative potential in manufacturing. Details of the selected articles can be found in our Github repository: *GenAI_in_manufacturing*. From the systematic review of the curated literature, several key trends and insights are summarized to characterize the current state.

• **Accelerating yet nascent adoption**: The exponential growth in publications, from only 2 studies in 2022 to 82 in 2025, underscores the increasing recognition of GenAI's potential in manufacturing. Despite this rapid expansion, we found that the volume of GenAI-related research in manufacturing remains relatively limited compared to other sectors such as education, healthcare, and finance. This disparity can be attributed to the physicality gap: unlike digital domains, manufacturing demands that GenAI not only process multi-modal data but also interact with immutable physical constraints and machinery, introducing substantial complexity. Moreover, the scarcity of pretraining data specific to manufacturing further exacerbates this disparity.

• **Concentration in cognitive applications**: The literature reveals a pronounced skew toward cognitive tasks rather than perception tasks. As detailed in Table 2, Customization (42 papers) and Human-Machine Interaction (36 papers) significantly outnumber fundamental tasks like Pattern Recognition (15 papers). This trend indicates that current GenAI is primarily leveraged for its semantic reasoning and code generation capabilities to address tasks that align with the text-centric nature of LLMs (Capitanelli and Mastrogiovanni, 2024; Xia et al., 2024b; Xu et al., 2024a; Timperley et al., 2025). Conversely, applying GenAI to pattern recognition tasks (Chen and Liu, 2024; Lin et al., 2025) requires bridging the gap between semantic tokens and continuous numerical sensor data, a technical barrier that currently limits adoption in these areas.

• **Superficial technical adaption**: A notable observation is that most of the current studies treat GenAI as an off-the-shelf tool that can be simply embedded into existing manufacturing systems. While this enables rapid prototyping of conversational assistants or basic planners, it represents a surface-level adaptation, without domain-specific fine-tuning, architectural modifications, and physical process integrations. Given that current GenAI lacks grounded understanding of physical systems and manufacturing processes, it can manage high-level logic

**Table 2** Literature distribution by application domain and year

| Domain | 2022 | 2023 | 2024 | 2025 | Total |
|---|---|---|---|---|---|
| Pattern Recognition and analysis | – | 1 | 5 | 9 | 15 |
| Predictive modeling and decision support | – | – | 2 | 4 | 6 |
| Optimization and Control | — | 2 | 13 | 15 | 30 |
| Data manipulation and utilization | – | 3 | 11 | 15 | 29 |
| Human-machine interaction | 2 | 2 | 10 | 22 | 36 |
| Customization and personalization | – | 2 | 21 | 19 | 42 |
| **Total** | 2 | 10 | 62 | 84 | 158 |

but remain inadequate for tasks governed by physical laws, material properties, and precise engineering requirements.

This review underscores that while the adoption of GenAI in manufacturing is growing rapidly, it remains at a nascent stage as the majority of studies are exploratory proof-of-concepts rather than validated industrial deployments. Current implementations predominantly represent surface-level adaptation, utilizing GenAI's pre-trained capabilities without domain-specific fine-tuning, architectural modifications, or integrations with physical process feedback loops. Consequently, discussions solely on application domains (defined in Table 1) are insufficient to reveal GenAI's potential in manufacturing. We argue that a more rigorous understanding requires analyzing GenAI through the lens of its depth of integration within manufacturing processes. To this end, the following section proposes a five-level framework to systematically characterize this integration spectrum.

## 3.2 Examining GenAI implementation in manufacturing

While generic LLMs have popularized conversational AI, the industrial potential of GenAI extends deeper, encompassing synthetic data generation, process scheduling, and intelligent agent creation (Ge et al., 2023). To systematically characterize the role of GenAI, we propose a five-level framework (Fig. 2) that maps the technology's trajectory from a supportive tool to an autonomous innovator. As illustrated in the figure, the framework is structured along a continuum of increasing integration depth and autonomy. The diagonal progression depicts the alignment between these GenAI levels and the standard stages of the traditional AI implementation pipeline. Specifically, Conversational Assistance aligns with Need Clarification and Problem Abstraction by providing natural language understanding and reasoning capabilities that support requirements elicitation and problem formulation. Data Enrichment corresponds to Data set Construction by generating synthetic training data. Model Enhancement maps to Model Training by improving model architectures and training strategies. Solution Orchestration aligns with Solution Integration by coordinating multiple AI components into deployable systems. In contrast, the fifth level, *Innovation Creation*, represents a paradigm shift where GenAI transcends the traditional pipeline to autonomously generate innovative solutions. Section 4 elaborates on this paradigm shift and articulating a futuristic vision for GenAI-driven manufacturing. This structure provides a taxonomy for understanding how GenAI is currently optimizing existing processes while paving the way for future smart manufacturing. The distribution of the reviewed literature across these five implementation levels is summarized in Table 3, revealing the current focus and emerging trends within the research community.

### 3.2.1 Conversational assistance

**Objective**: The primary objective at this level is to leverage GenAI as a conversational assistant to enhance human decision-making by knowledge retrieval, data analysis, plan optimization, code generation, and operations support.

**Recent advancements:** Recent literature demonstrates significant strides in adapting generic GenAI for specialized manufacturing contexts through novel frameworks and fine-tuning strategies.

• **Operational support and safety:** Systems like the Popeye chatbot (Colabianchi et al., 2022) have been deployed to assist operators in identifying hazard scenarios in dock areas. Zhou et al. (2024) integrated structured

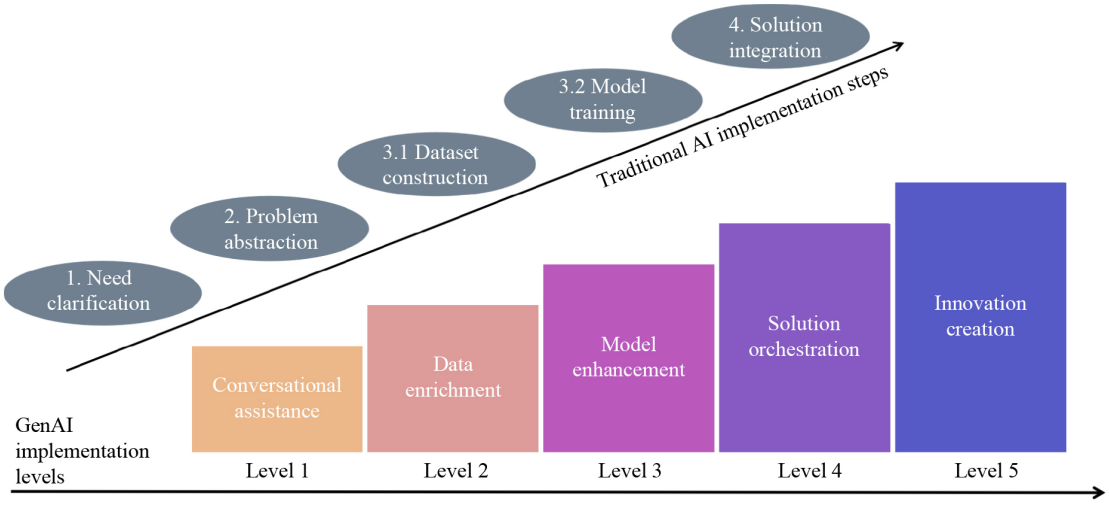


**Fig. 2** The five levels of GenAI implementation in manufacturing, referencing the traditional paradigm of AI implementation in manufacturing.

**Table 3** Literature distribution by implementation level in manufacturing and year

| Domain | 2022 | 2023 | 2024 | 2025 | Total |
|---|---|---|---|---|---|
| Level 1 | 1 | 6 | 28 | 36 | 71 |
| Level 2 | – | 1 | 8 | 4 | 13 |
| Level 3 | – | – | 5 | 10 | 15 |
| Level 4 | – | 1 | 11 | 15 | 27 |
| Level 5 | 1 | 2 | 10 | 19 | 32 |

Below, we analyze each level in terms of its operational objective, recent advancements, current maturity, and persistent technical challenges.

causal knowledge graphs with LLMs to enhance root-cause analysis for quality defects in aerospace manufacturing. Li et al. (2023) leveraged GPT-4 to automate the generation of high-fidelity scene descriptions, accelerating the construction of industrial digital twins. Wang et al. (2025) proposed a human-LLM collaborative design framework, where LLM plays a proactive, multi-faceted role as both a knowledge provider and decision maker that integrates multidisciplinary design and manufacturing knowledge to translate vague customer intentions into manufacturable generative designs.

• **Code generation:** To ensure reliable code output, frameworks such as LLM4PLC (Fakih et al., 2024) utilize iterative pipelines for PLC code generation. Similarly, Xia et al. (2024a) introduced error-assisted fine-tuning approaches to adapt models specifically for manufacturing-related coding queries.

• **Process planning**: Extending beyond code, GenAI is being applied to system-level organization. Holland and Chaudhari (2024) leveraged LLMs to tackle process planning challenges, including cycle time estimation and resource allocation.

• **Knowledge retrieval:** Addressing the need for domain precision, researchers have developed advanced retrieval architectures. For instance, the Golden-Retriever method (An et al., 2024) incorporates reflection-based query augmentation to handle industrial jargon, while the KnowPAT framework (Zhang et al., 2024b) aligns model outputs with expert human preferences through knowledgeable fine-tuning. Additionally, hybrid approaches proposed by Yang et al. (2023) utilize small, domain-specific models to guide larger LLMs, ensuring that answers are grounded in verified technical data. Recently, Shi et al. (2025) enhanced manufacturing RAG systems by leveraging the Asset Administration Shell for structured knowledge representation.

**Maturity status:** As detailed in Table 3, Level 1 dominates the research landscape, accounting for 71 studies (approximately 45% of the reviewed corpus). This prevalence is largely attributed to the low technical barrier to entry at the conversation level without requiring the complex modification and physical integration necessary for higher-level automation.

**Technical challenges:** Despite this maturity, widespread deployment is hindered by the challenge of reliability. Generic GenAI is prone to hallucinations, generating plausible but factually incorrect outputs, which poses severe risks in precision-critical environments (Huang et al., 2023; Tonmoy et al., 2024). To mitigate this, the community has pivoted toward Retrieval-Augmented Generation (RAG), such as Golden-Retriever (An et al., 2024) which handles domain-specific jargon, and Knowledge Alignment, exemplified by frameworks like KnowPAT (Zhang et al., 2024b) that align outputs with expert preferences. Beyond factual correctness, ensuring robust engineering reasoning represents another challenge. Recent advancements in reasoning-oriented architectures and Chain-of-Thought (CoT) methodologies (Wei et al., 2022a; Yao et al., 2023; Guo et al., 2025) have notably enhanced the logical consistency of GenAI, allowing models to solve increasingly complex multi-step problems. However, current models still lack the deterministic reliability required for precision-critical manufacturing (Kambhampati, 2024). Bridging the gap between plausible logic and physically guaranteed execution remains an urgent frontier necessitating further research.

### 3.2.2 Data enrichment

**Objective**: The primary objective at this level is to utilize GenAI to overcome the data scarcity in manufacturing by generating high-fidelity synthetic data and constructing complex virtual environments.

**Recent advancements:** Recent literature demonstrates the versatility of GenAI across these two functional domains.

• **Synthetic data augmentation:** To improve the robustness of downstream models, researchers use GenAI to generate diverse data samples that are expensive to collect in reality. For example, Getz and Tong (2025) utilized an LLM to extract problems and solutions from the machine maintenance records, compiling a data set to train a model for deriving insights from operational data. Leveraging the commonsense reasoning of LLMs, Yang et al. (2024) modeled human thermal preferences, generating synthetic behavioral data to train reinforcement learning algorithms for optimizing HVAC control systems.

• **Automated environment generation:** Beyond discrete samples, GenAI is increasingly used to construct entire interactive environments for robotics (Thumm et al., 2024) and digital twins (Martínez-Gutiérrez et al., 2024). Katara et al. (2024) developed Gen2Sim, a method that automates the generation of 3D assets, task descriptions, and reward functions, scaling up robot skill learning. Similarly, Jackson et al. (2024a) leveraged GenAI to generate valid simulation models for logistics systems

from verbal descriptions, while Huang et al. (2024b) integrated GenAI into Digital Twins to simulate unprecedented events for decision support.

**Maturity status:** As shown in Table 3, Level 2 remains an emerging area, accounting for only 13 studies. This limited adoption is largely attributable to the fidelity gap: general-purpose GenAI models, pre-trained on internet data, often lack the granular domain knowledge required for precision manufacturing. Consequently, researchers continue to favor traditional generative models trained on specific industrial data to ensure high-fidelity synthesis.

**Technical challenges:** Despite the promise of data enrichment, widespread adoption is hindered by two critical bottlenecks. First, the indiscriminate use of synthetic data risks model collapse (Shumailov et al., 2024), a degenerative process where recursive training on generated outputs progressively degrades model variance and fidelity. Second, current GenAI operates primarily on statistical correlations rather than causal mechanics, often failing to adhere to strict physical laws. This results in sim-to-real gaps where models trained on synthetic data fail in the physical world. Bridging this gap requires efforts toward World Models and Spatial Intelligence (Cen et al., 2025; Li, 2025) that explicitly encode spatiotemporal dynamics to ensure synthetic data and environments are both digitally effective and physically valid.

### 3.2.3 Model enhancement

**Objective:** The primary objective at this level is to utilize GenAI to augment the structural design, feature representation, and interpretability of traditional discriminative models.

**Recent advancements:** Recent literature demonstrates distinct approaches to this integration:

• **Semantic feature engineering:** A critical advancement lies in bridging textual semantics with time-series sensor data. Tao et al. (2025) fine-tuned an LLM for fault diagnosis by converting vibration signal patches into token embeddings. Similarly, Lin et al. (2025) introduced FD-LLM, which employs modal alignment to map engineering data descriptions directly to equipment status in the feature space, significantly improving fault diagnosis generalizability.

• **Automated architecture design and optimization:** GenAI is also revolutionizing neural architecture search. Nasir et al. (2024) combined LLM code generation with quality-diversity algorithms to autonomously synthesize robust neural networks, while Chen et al. (2023a) utilized LLMs as adaptive mutation and crossover operators in evolutionary search. On the optimization front, Khanghah et al. (2025) established a framework that integrates RAG-based knowledge extraction with iterative LLM-driven refinement to automatically generate analytical models of manufacturing processes.

• **Explainability and robustness:** Beyond performance, GenAI addresses the black-box nature of AI. It enhances explainable AI (XAI) by generating natural language insights into decision-making processes, thereby fostering stakeholder trust and regulatory compliance (Cambria et al., 2024). Furthermore, although currently underexplored, the generation of adversarial examples offers a promising avenue to improve industrial models against anomalies and external disturbances.

**Maturity status:** As indicated in Table 3, Level 3 remains an experimental area (15 studies). This limited adoption reflects a high technical barrier; unlike surface-level semantic applications, successful implementation at this level usually requires deep architectural modifications to align GenAI with discriminative models in the high-dimensional feature space rather than the accessible semantic space.

**Technical challenges:** The central bottleneck is the alignment between the discrete tokens of GenAI, the continuous time-series of physical sensors, and the high-dimensional latent feature spaces of discriminative models. Effectively bridging these disparate formats requires complex embedding strategies, as seen in (Lin et al., 2025) and (Tao et al., 2025). Aligning these modalities without losing signal precision is mathematically non-trivial. Moreover, integrating heavy GenAI modules into lightweight, latency-sensitive control loops introduces computational overhead, creating a barrier to real-time deployment. Future research can focus on distillation techniques to compress these semantic capabilities into efficient, edge-deployable formats.

### 3.2.4 Solution orchestration

**Objective**: The objective at this level is to elevate GenAI from a passive assistant to an autonomous agent capable of generating and executing manufacturing solutions in dynamic environments. This level focuses on tool use and action: agents analyze high-level user queries, assess available resources, and autonomously control production equipment.

**Recent advancements**: Recent literature focuses on enabling agents to bridge the gap between language and physical action through two primary scales:

• **Robotic task planning and execution:** GenAI is increasingly used to translate natural language into machine-executable directives. Zheng et al. (2024) utilized LLMs to autonomously generate and execute robotic task plans in human-robot collaborative environments. Similarly, Fan et al. (2025a) proposed a framework where LLM agents autonomously generate tool paths and G-code for industrial simulations. Zhou et al. (2026) developed a multi-agent VLM architecture to achieve

zero-shot, multi-step robot tool manipulation in industrial environments. In the domain of logistics, Wang et al. (2024b) developed a multi-modal agent enabling automated guided vehicles to navigate based on verbal commands.

• **Multi-agent shop floor orchestration:** At the system level, research has shifted toward multi-agent systems (MAS) to handle complex production flows. Xia et al. (2023) developed a hierarchical MAS where manager agents interpret commands and operator agents execute production plans. This work was later improved to be more scalable (Xia et al., 2024b). Expanding on this, Zhao et al. (2024b) proposed an agentic system enabling dynamic negotiation and decision-making for high-mix, small-batch production. Furthermore, Gkournelos et al. (2024) integrated these agents with manufacturing execution systems and digital twins, while Wang and Qin (2024) leveraged frameworks like LangChain to optimize production sustainability.

**Maturity status**: As shown in Table 3, this level encompasses 27 studies, reflecting a surge of research interest. This momentum is largely catalyzed by the emergence of specialized agentic frameworks, which have significantly lowered the technical barrier to entry for constructing autonomous workflows. However, current implementations remain confined to isolated pilot cells and tolerance-lenient tasks, highlighting the challenges of widespread industrial adoption and precision-critical manufacturing.

**Technical challenges**: First, the heterogeneity of physical interfaces, the sheer diversity of proprietary application programming interfaces (APIs), communication protocols, and mechanical constraints, remains a formidable barrier to creating universal manufacturing agents. Second, current embodied agents face a severe dexterity gap, largely restricted to elementary primitives (e.g., pick-and-place) while struggling with the high-precision manipulation required for manufacturing. Finally, the stochastic inference and high latency of GenAI conflict with the strict deterministic and real-time constraints of industrial control.

### 3.2.5 Innovation creation

**Objective:** At this level, GenAI transcends its role as an auxiliary tool to become a self-contained service provider to provide innovative solutions that directly meet the manufacturing need. Unlike Level 4, which focuses on the autonomous execution of predefined tasks using existing tools (e.g., generating code to move a robot), Level 5 focuses on generating novel product designs, manufacturing methods, system architectures that did not previously exist and exceed human engineering intuition.

**Recent Advancements:** Recent literature indicates that GenAI is beginning to drive innovation across three critical frontiers:

• **Generative product design:** GenAI is redefining the boundaries of product design (Liu et al., 2024; Picard et al., 2024; Xu et al., 2024a). Nafea (2025) utilized GenAI tools to design actuators with complex, non-Euclidean shapes and infill patterns for 4D printing, achieving functionalities previously unattainable through manual design. To ensure manufacturability, Chong et al. (2024) introduced a method that conditions the generative process on feasible CAD imagery, significantly improving the engineering validity of the output. Liu et al. (2026) developed a multi-agent workflow, iDesignGPT, where specific agents are assigned to optimize specific design processes, including problem definition, information gathering, concept generation and evaluation.

• **End-to-end manufacturing workflows:** Moving beyond isolated components, research is now exploring holistic process generation. Makatura et al. (2024) investigated an end-to-end framework where LLMs transform text prompts directly into manufacturing instructions, effectively compiling natural language into physical products. Similarly, Kyaw et al. (2025) developed a system that leverages 3D GenAI to generate objects from speech, discretize them into voxel components, and autonomously optimize robotic toolpaths for fabrication.

• **Factory-level systems engineering:** At the macro scale, Tinsel et al. (2024) explored using LLMs to analyze user requirements and autonomously plan entire factory layouts. Their system identifies necessary production processes, selects machinery, and generates simulation models to validate the proposed production system in a virtual environment before physical implementation.

**Maturity status:** As shown in Table 3, Level 5 encompasses 32 studies, ranking second in publication volume. However, this activity remains at the proof-of-concept stage. Most studies concentrate on product design application as it only involves a small need-solution loop with minimal reliance on other devices and machinery. Crucially, the generated designs often lack the functional novelty required to be classified as true innovation, often falling back on statistical averages of their training data.

**Technical challenges:** The primary barrier to achieving this level is the creative plateau. Because GenAI models are trained on historical data, they tend to converge on known, safe solutions rather than exploring the long tail of novel, high-risk/high-reward designs. Overcoming this requires more than stochastic exploration; it demands the deep integration of domain knowledge to ground generative exploration in feasible innovation. Additionally, while GenAI can understand intricate geometries (as seen in (Nafea, 2025)), verifying that these designs can be manufactured at scale remains a complex constraint satisfaction problem. Future research should focus on physics-embedded GenAI that treats manufacturing constraints (e.g., material stress, thermal dissipation) not as

afterthoughts, but as fundamental laws governing the generation process.

Collectively, GenAI's implementation is concentrated at Level 1, aligning with earlier observations that most studies embed GenAI into manufacturing systems as an off-the-shelf tool with minimal customization. Conversely, Levels 2 and 3 exhibit comparatively fewer articles, largely because these applications usually require deeper integration and adaptation of GenAI models. Although the number of studies at Levels 4 and 5 is higher, this does not imply technical simplicity; rather, it reflects even straightforward adaptations of GenAI can yield substantial impacts. However, as noted, these high-level applications remain largely exploratory, often lacking the robustness for large-scale adoption.

Manufacturing is a complex system characterized by diverse devices, intricate processes and a strong emphasis on interactions within the physical world. From the preceding discussion, three critical technical features emerge as essential for GenAI application in manufacturing: (1) advanced multi-modal data processing capabilities to perceive and interact precisely with the physical environment; (2) robust reasoning and planning abilities to ensure the reliability and effectiveness of outputs; and (3) seamless control and collaboration across diverse devices to enable cohesive and efficient operations. Realizing the transformative potential of GenAI in manufacturing will require innovative research to address these challenges and drive advancements.

In conclusion, although the current literature demonstrates that GenAI has already brought tangible benefits to the manufacturing industry, its potential for deeper integration and innovative solutions remains underexplored. Further strides in scientific reasoning, domain-specific adaptions, and robust control of physical manufacturing processes are needed to move from proof-of-concept studies to systemic transformations. The next section discusses the broader challenges of GenAI implementation in manufacturing, spanning technical, practical, and societal dimensions, and examines ongoing research that aims to address these hurdles.

### 3.3 Technology roadmap for GenAI-enabled smart manufacturing

While Fig. 2 categorizes the integration depth of GenAI within manufacturing systems, it does not illustrate how advancing integration progressively transforms smart manufacturing. To elucidate this developmental trajectory, Fig. 3 presents a technology roadmap that links the integration levels introduced in Fig. 2 with their underlying technological drivers and emergent system capabilities. The upper layer delineates three evolutionary stages of GenAI-enabled manufacturing. The middle rows specify, for each stage, the corresponding system capabilities, technology drivers, and typical applications. The bottom layer maps the integration levels from Fig. 2 onto these three stages, highlighting the progressive advancement in the integration depth, capability and operational autonomy of GenAI as it evolves from near-term task assistant toward a long-term end-to-end innovator.

In the near term, GenAI is expected to function primarily as an AI-assisted tool augmenting specific manufacturing tasks. At this stage, GenAI provides conversational assistance, knowledge retrieval, and design support to engineers, leveraging its contextual understanding and multi-modal reasoning capabilities. These applications correspond to the lower integration levels in Fig. 2, wherein

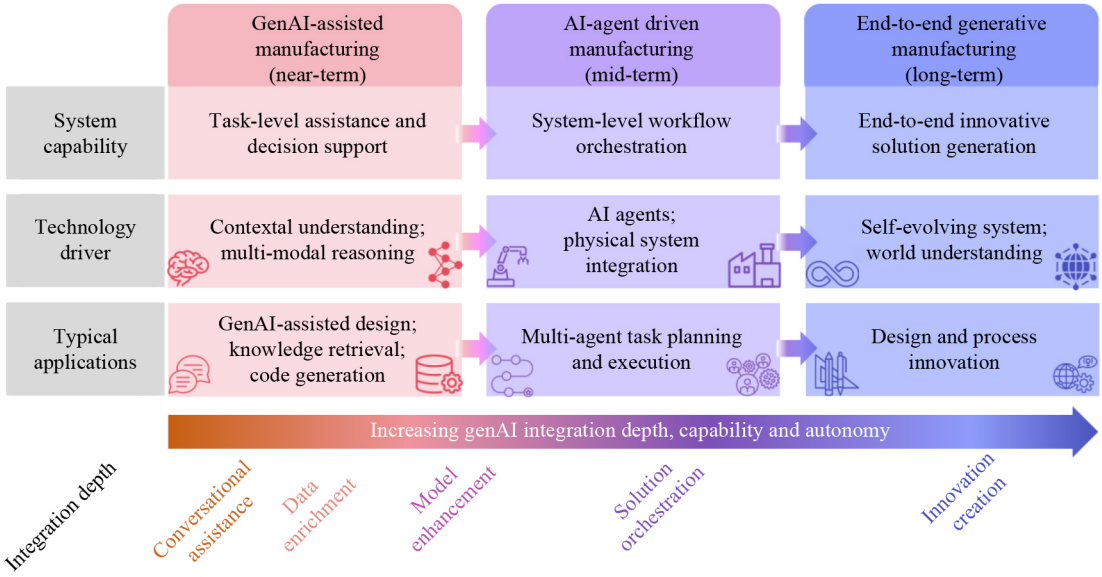


**Fig. 3** Technology roadmap for GenAI-enabled smart manufacturing.

GenAI enhances human decision-making but does not assume direct control over manufacturing workflows.

In the mid-term, the emergence of AI agents capable of interfacing with physical systems enables more advanced forms of system-level coordination. Rather than supporting isolated tasks, GenAI systems at this stage orchestrate multiple manufacturing processes, including design generation, process planning, and production optimization, in an integrated manner. This corresponds to the solution orchestration level in Fig. 2, where GenAI becomes embedded within manufacturing workflows to manage complex production systems directly.

In the long-term, continued advances in FMs may enable end-to-end innovative solution generation. In such systems, GenAI autonomously generates, evaluates, and optimizes manufacturing solutions from general user queries, driven by self-evolving architectures and comprehensive world understanding. These systems could explore novel solution alternatives beyond conventional solution spaces and continuously improve production strategies through closed-loop learning. This final stage corresponds to the highest integration levels in Fig. 2, wherein GenAI transcends task assistance to function as an end-to-end innovator.

Importantly, the roadmap should not be interpreted as a rigid or deterministic timeline. Rather, it highlights how advances in GenAI capabilities and system architectures may progressively enable deeper levels of integration within manufacturing systems, providing a structured perspective to guide future research and development efforts.

### 3.4 Challenges of GenAI implementation in manufacturing

This section provides a structured analysis of the GenAI challenges and the associated research opportunities from technical, practical and social dimensions. Addressing these issues requires not only technical advancements but also a comprehensive approach that includes policy development, stakeholder engagement, and continuous education for industry professionals. The discussion is guided by two principles: comprehensiveness in covering multi-layered challenges, and a specific focus on distinguishing those uniquely posed by GenAI.

#### 3.4.1 Technical challenges

The application of GenAI in manufacturing currently faces several technical challenges, spanning areas such as data processing, model optimization, and system compatibility. Overcoming these challenges will require cutting-edge research and innovation across various technical domains.

**Data scarcity and multi-model alignment**: While data scarcity is a common bottleneck in industrial AI, GenAI introduces unique structural demands regarding the scale and modalities. Unlike traditional AI models that can be trained on smaller, task-specific data sets, GenAI FMs require massive, diverse corpora to develop generalizable reasoning. High-quality industrial data are fragmented, proprietary, and scarce (Chen et al., 2023b; Wang et al., 2023), preventing the creation of a general manufacturing GenAI. Furthermore, while traditional multi-modal learning focuses on fusing compatible numerical features, GenAI faces the specific challenge of semantic-numeric alignment. The model must bridge the discrete, symbolic space of language tokens with the continuous, high-frequency space of sensor signals (Nafea, 2025). This requires fundamentally new representation learning and tokenization strategies (Dai et al., 2023; Mai et al., 2023; Sun et al., 2023; Devillers et al., 2025; Zhao et al., 2024a) to prevent the loss of signal fidelity when translating physical dynamics into semantic descriptions, a hurdle not present in standard sensor fusion models.

**Reliable output:** Unlike traditional deterministic AI, GenAI introduces stochastic uncertainty, manifesting as hallucination (Yin et al., 2023) and probabilistic mimicry in strategic planning. In manufacturing, where decisions must adhere to strict constraints, the risk of generating plausible but physically invalid plans is severe. To mitigate this, developers must apply specific customization strategies based on the nature of the application. RAG (Li et al., 2024b; Romera-Paredes et al., 2024) is selected for knowledge-intensive tasks requiring dynamic factual updates, as it grounds the model in retrieved external data. In contrast, Supervised Fine-Tuning (SFT) is the criterion for structure-intensive tasks, where the model must internalize specialized syntax or jargon through weight adaptation. Furthermore, Reinforcement Learning from Human Feedback (RLHF) is essential for safety-critical scenarios, using reward models to align the system's policy with complex human operational standards (Wang et al., 2024f). Beyond customization, ensuring ultimate reliability demands advanced intrinsic reasoning (Qiao et al., 2023; Sumers et al., 2024). Recent advancements, such as CoT (Wei et al., 2022a), enhance reliability by prompting the model to decompose complex problems into intermediate linear logical steps. For more intricate planning tasks, Tree of Thoughts (ToT) (Yao et al., 2023) further extends this by enabling the exploration of multiple reasoning branches, allowing the system to look ahead and backtrack to find optimal solutions. However, a critical debate persists regarding whether such methods facilitate true logical deduction or merely refine the mimicry of reasoning patterns observed in training data (Yao et al., 2023; Qiao et al., 2023). Consequently, the frontier must shift from probabilistic coherence to establishing principled, verifiable reasoning to accomplish complex, multi-

step manufacturing tasks.

**Scientific explainability:** In manufacturing, the black-box nature of AI poses a severe barrier to accountability (Naqvi et al., 2024). Unlike traditional discriminative models where explainability focuses on feature importance, GenAI introduces the challenge of verifying generative reasoning chains. This encompasses two levels: model explainability, *which involves understanding how the model output specific results*, and result explainability, *which explains why the results meet user needs and to what extent*. Although various methods have been explored to enhance model explainability (Luo and Specia, 2024), including feature attribution analysis (Wang et al., 2024d), transformer block dissection (Modarressi et al., 2022), probing techniques (Li et al., 2024a), and mechanistic interpretability (Zhao et al., 2023), these approaches remain insufficient for current GenAI models and will face greater challenges with larger and more complex future models. In addition, extensive research is needed to enhance result explainability, as it significantly impacts user experience.

**Robustness and generalization:** A critical challenge unique to GenAI is navigating the stability-plasticity dilemma during adaptation. Unlike traditional discriminative models that can be frequently retrained from scratch, FMs risk catastrophic forgetting, the erosion of their core general reasoning capabilities, when aggressively fine-tuned on narrow industrial data sets (Jiang et al., 2024; Parthasarathy et al., 2024). Consequently, achieving robustness requires moving beyond standard adversarial training (Wang et al., 2024e) or data augmentation (Wang et al., 2024a) toward advanced parameter-efficient fine-tuning and continual learning architectures (Zhang et al., 2023). The ultimate objective is to develop systems capable of few-shot adaptation to the long tail of diverse manufacturing scenarios (plasticity) while structurally preserving the broad pre-trained knowledge (stability) essential for handling unseen edge cases.

**Compatibility with existing manufacturing systems:** Integrating GenAI into the vast and complex landscape of both modern and legacy manufacturing systems, which encompass diverse tools, sensors, software, and machines, presents significant challenges (Xia et al., 2023). Xu et al. (2024b) provided a comprehensive review of GenAI-enabled cyber-physical systems and highlighted the key research challenges in this domain. Integrating GenAI into this ecosystem presents a unique semantic-structural mismatch. Unlike traditional software that interacts via rigid, pre-defined APIs, GenAI agents operate on fluid semantic tokens. Bridging this gap requires standardized interface abstraction layers that can deterministically translate agentic intent into the precise, structured commands required by legacy firmware without disrupting critical control loops. Emerging standards like the Model Context Protocol (MCP) offer a promising trajectory by decoupling model logic from system implementation. However, future research must extend such protocols to handle the real-time constraints and stateful dependencies of cyber-physical systems, ensuring that GenAI can serve as a scalable orchestration layer atop the fragmented manufacturing stack.

### 3.4.2 Practical and operational challenges

Beyond technical challenges, the implementation of GenAI in manufacturing raises practical and operational concerns, including data security, high costs, and workforce adaptation.

**Data security and privacy:** Manufacturing data often includes sensitive and proprietary information, and generated content that must be protected against unauthorized access, sometimes even including the GenAI provider. GenAI introduces unique privacy risks, including training data memorization leading to inadvertent disclosure of sensitive information, and susceptibility to prompt injection attacks. Furthermore, clear definitions of content ownership are necessary to alleviate user concerns about utilizing GenAI. Future research involves implementing advanced encryption methods, secure data storage solutions, and blockchain techniques tailored to GenAI in manufacturing (Gupta et al., 2023).

**Cost and investment:** Training large-scale GenAI models demands substantial computational resources, energy, and expertise, often beyond the financial and technical capacity of most manufacturing companies (Berthelot et al., 2024). As a result, current practices typically involve specialized firms developing GenAI models, with manufacturing enterprises fine-tuning them to meet their specific requirements. However, even the process of fine-tuning often remains out of reach for smaller enterprises. To overcome this challenge, research should focus on developing cost-effective, user-friendly tools and third-party services for fine-tuning and deployment, along with advancing GenAI that can address cross-domain manufacturing problems. Recent surveys can be found in Han et al. (2024) and Xu et al. (2025).

**Workforce adaptation:** The integration of GenAI in manufacturing will fundamentally transform job roles and operational workflows, requiring substantial training and upskilling of the existing workforce to effectively utilize GenAI tools. Addressing resistance to change and fostering a culture that embraces technological advancements are also crucial (Prasad Agrawal, 2024). Another risk is skill atrophy, where engineers become over-reliant on GenAI for problem solving, potentially eroding the foundational domain expertise required to verify the GenAI's outputs. Additionally, a recent work revealed GenAI's potential to improve workplace well-being (Yuan et al., 2025), which may, in turn, foster a positive adoption of the technology.

#### 3.4.3 Societal needs

The expected widespread adoption of GenAI in manufacturing will raise significant concerns regarding its societal impact. Ethical, security and environmental issues must be addressed carefully to balance the technological advancements with the well-being of society and the environment.

**Ethical concerns:** Unlike traditional automation which displaces manual labor, GenAI threatens cognitive displacement, automating the decision-making and creative processes that define human expertise. Therefore, there is an unprecedent risk of substantial job losses (Gmyrek et al., 2023). To address this challenge, one direction is to develop human-centric AI technologies that function as copilots rather than autopilots. Another direction is to create workforce transition strategies to help displaced workers find new roles within the evolving industry. Policymakers and industry leaders must engage in dialog to create frameworks that balance technological advancement with social responsibility, ensuring equitable distribution of GenAI benefits.

**Security risks:** With the continuous strengthening of GenAI autonomy in manufacturing, the accompanying security risks escalate proportionally. These include vulnerabilities related to sensitive information leaks, cyber-attacks targeting GenAI systems, and the generation of hazardous outputs (Wach et al., 2023; Mavikumbure et al., 2024). Given the complexity of interconnected devices and systems in manufacturing, securing these networks against breaches is critical. Future research should prioritize strengthening cybersecurity measures tailored to GenAI applications in manufacturing and conducting risk assessments on generated outputs.

**Environmental impact:** A significant contradiction exists between the goal of environment-friendly manufacturing and the reality of the massive carbon footprint associated with training and hosting FMs (Berthelot et al., 2024). As the scale and complexity of FMs continue to grow, addressing this issue becomes increasingly urgent. Mitigating these environmental effects requires research focused on developing smaller-scale models, energy-efficient training methods, and neuromorphic hardware (Narayanan et al., 2021; Schick and Schütze, 2021; Kang et al., 2023). Additionally, leveraging renewable energy sources and establishing regulations and standards to promote environmentally responsible AI development and deployment are also essential.

While these challenges span technical, practical, and social dimensions, their collective impact reveals a fundamental tension: the requirement for general-purpose reasoning versus the rigid, deterministic precision required by industrial systems. Reaching a broader impact in manufacturing will require a transition from off-the-shelf tool usage to physics-grounded GenAI architectures that treat manufacturing constraints as fundamental laws rather than optional prompts.

## 4 Prospective insights into GenAI for manufacturing

Before presenting our prospective insights, it is important to position this work against recent perspective papers that have explored GenAI's future in manufacturing. Ma et al. (2025) investigated how LLMs can be integrated with Industry 5.0 enabling technologies and applied across smart manufacturing stages. Chen et al. (2025) proposed the Interactive-DT framework for LLM-digital twin integration, elaborating on LLM's roles at the edge, digital twin, and service layers to enhance construction, operation, and cloud-edge collaboration. Fan et al. (2025b) systematically reviewed VLM applications in human-robot collaboration, covering robotic task planning, navigation, manipulation, and human-robot skill transfer through multimodal data integration. Shahin et al. (2025) provided a comprehensive survey of GenAI applications across manufacturing domains and discussed GenAI's potential to enhance each domain. Zhang et al. (2025a) presented a new-generation intelligent manufacturing LLM framework to emphasize the LLM's potential to enhance five processes of self-perception, self-learning, self-decision, self-execution and self-adaption in manufacturing. These works offer their insights into how GenAI can reshape manufacturing landscape. However, they predominantly adopt an enhancement perspective, examining how GenAI improves existing processes, integrates with established systems, or augments existing technologies. In contrast, the present work adopts a paradigm perspective, arguing that GenAI's significance extends beyond incremental enhancement to fundamentally reshaping how manufacturing problems are approached and solved. Specifically, we propose that GenAI can transition from functioning as an auxiliary tool that enhances individual applications to serving as a self-contained service provider that directly addresses manufacturing needs through holistic solution generation. This paradigm-level reconceptualization distinguishes our work from existing enhancement-focused perspectives and offers a forward-looking vision for the manufacturing community.

### 4.1 Overview of GenAI and its outlook

The generation capability of GenAI, which is rooted in the deep understanding of the complex multi-modal data, stems from the underlying Foundation Models (FMs) (Zhou et al., 2023) which are large-scale models pre-trained on extremely extensive data sets. Analyzing the development trajectory of FMs is crucial for understanding the future potential of GenAI.

While current terminology surrounding FMs can be

ambiguous, especially with some researchers also referring to models capable of processing visual content as LLMs (Jin et al., 2024), this paper presents a distinct classification of FMs based on the data types they process.

The initial FMs emerged as LLMs (Chang et al., 2024), designed to interact with human language, including textual and audio formats. Recognizing the value of multi-modal perception, Vision Language Models (VLMs) (Zhang et al., 2024a), also known as large multi-modal models or Multi-modal LLMs (Yin et al., 2024), are subsequently developed to integrate diverse data modalities, including text, image, video, audio, and sensory data. By enhancing environmental perception, VLMs play a crucial role in advancing embodied AI (Mu et al., 2023). Integrated with action modules, recent developments in vision-language-action (VLA) models have enabled physical agents, including various robots and autonomous vehicles, to achieve unprecedented generalist capabilities (Brohan et al., 2023; Cui et al., 2024). Building on this trajectory, it is anticipated that future FMs will holistically understand the complex dynamics of the real world, thereby providing more trustworthy and capable AI. Therefore, the concept of Unified World Models (UWMs) is proposed. Unlike the world models that primarily focus on predicting world states in response to different actions in the context of robotics (Ha and Schmidhuber, 2018), UWM integrates the abilities of LLM and VLM, while also emphasizing a scientific understanding of both the physical world and human society. Fig. 4 illustrates the development trajectory of FMs, highlighting the increasing capabilities of FMs without suggesting a sequential development dependency. This classification is depicted as a nested hierarchy, illustrating that advanced models encompass and extend the capabilities of their predecessors rather than merely replacing them. Furthermore, the figure highlights a critical inverse relationship for industrial application: as the supported prompt modality expands (from text-only to multimodal sensory inputs), the required prompt complexity diminishes. This trend suggests that as FMs evolve, manufacturing interactions will shift from rigid, highly specific technical prompting to natural, intuitive human-machine interaction.

The key features of LLMs, VLMs, and UWMs are summarized in three aspects: data modality to be interacted, content to be understood, and knowledge to be mastered.

#### 4.1.1 Key features of LLM

• **Language interaction**: LLMs excel in understanding and generating human language. The essence of language is sequential data, which explains why LLMs can adeptly process not only text but also audio (Huang et al., 2024a), code (Ni et al., 2023), and other forms of sequential data (Yu et al., 2023).

• **Contextual understanding**: LLMs demonstrate a profound ability to comprehend and maintain context over extended sequences, enabling them to generate coherent and contextually appropriate responses. However, because LLMs can only process linguistic prompts, which inherently possess ambiguity, these prompts must be precise and detailed to ensure high-quality outputs. This highlights the importance of prompt engineering as a crucial technique when working with LLMs (White et al., 2023).

• **Knowledge integration**: LLMs integrate vast amounts of information from diverse sources, providing knowledgeable responses on a wide array of topics. They can also be fine-tuned for specific domain knowledge

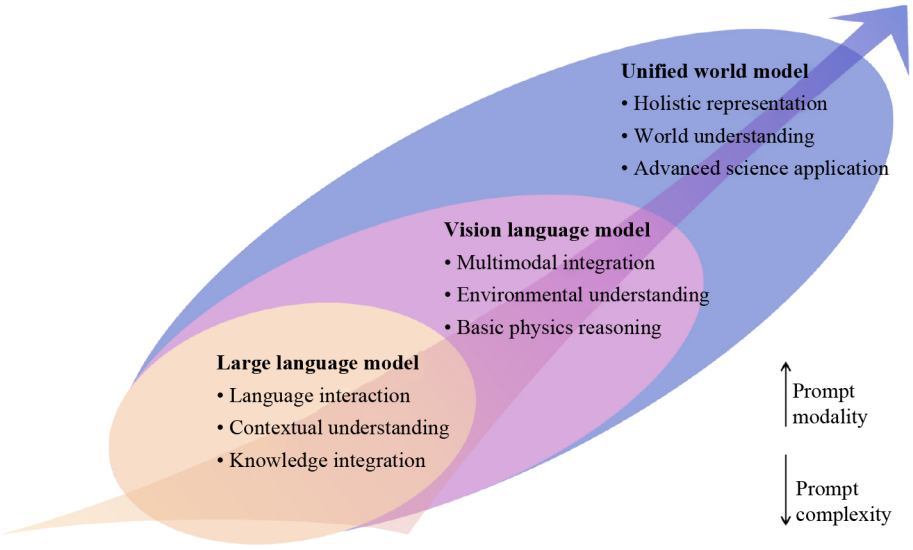


**Fig. 4** The development trajectory of foundation models in GenAI: LLM, VLM, and UWM. As the trajectory progresses from LLMs to UWMs, the supported prompt modalities expand, while the complexity and specificity required in the prompts diminish.

(Shen et al., 2024). However, as probabilistic models, they sometimes produce probabilistically plausible but nonsensical answers, highlighting a limitation in ensuring output accuracy (Ji et al., 2023).

#### 4.1.2 Key features of VLM

• **Multimodal integration**: VLMs can process and interact with both visual and linguistic data. In this context, visual and linguistic data represent broader categories of spatial and sequential data, respectively.

• **Environmental understanding**: Beyond contextual understanding, VLMs can interpret and comprehend the visual environment. This encompasses recognizing the states of various elements, their interrelationships, and their dynamic changes in the environment. Consequently, the required complexity of prompts is reduced, as VLMs can intuitively grasp both environmental and contextual cues.

• **Basic physics reasoning**: Unlike LLMs, which emphasize textual knowledge understanding, VLMs place higher demands on accurate application of physics knowledge to generate high-quality visual content.

#### 4.1.3 Key features of UWM

• **Holistic representation**: UWMs create holistic internal representations that encompass various aspects of the world, combining visual, auditory, linguistic, and other sensory information.

• **World understanding**: UWMs emphasize a comprehensive understanding of the physical world, human interpretation of reality, and the dynamics of human society. They can handle multi-modal prompts with lower complexity, and even enable prompt-free applications, allowing for natural interactions with the world guided by the implanted general purpose. For example, a robot empowered by a UWM would instinctively catch a falling human to prevent injury. In manufacturing, the general purpose can encompass human well-being, environmental sustainability, and societal values.

• **Advanced science application**: Beyond basic physics, UWMs can adeptly integrate and apply advanced knowledge from both natural and social sciences. When addressing real-world challenges, UWMs apply advanced scientific knowledge to develop solutions that benefit tasks, humans, the environment and society as a whole.

It should be noted that the FMs discussed here refer to the general-purpose FMs. Researchers have also been developing FMs in various specific domains such as bioscience (Hao et al., 2024), autonomous driving (Wang et al., 2024c), and robotics (Firoozi et al., 2025). These domain-specific FMs are variants of the general-purpose FMs.

The development trajectory of FMs, as depicted in Fig. 4, underscores the future advancements of GenAI toward multimodal data interaction, comprehensive and scientific world understanding, and advanced knowledge application. These advancements are critical for the manufacturing industry, which inherently deals with multimodal data, requires precise interaction with the physical world, and relies on extensive domain knowledge.

### 4.2 A futuristic paradigm of GenAI implementation in manufacturing

Recent research highlights researchers' efforts to advance GenAI beyond conversational assistance toward deep integration with manufacturing systems, enabling end-to-end automation from design to production while driving innovation (Kyaw et al., 2025; Makatura et al., 2024; Nafea, 2025). Based on these trends, we take a step further to envision a futuristic paradigm (Fig. 5) in which interprets user needs, expressed through any modal prompts, and subsequently generates and executes innovative solutions. This futuristic paradigm is rooted in the principles of Industry 5.0, moving beyond traditional operational metrics to adopt a value-driven framework. It reconceptualizes prompts to integrate functional manufacturing needs with broader human, environmental, and societal values. The values can be embedded as implicit constraints, ensuring GenAI aligns outputs with them even for purely functional requests. Consequently, the generated solutions are required to exhibit four core attributes:

• **Task optimization:** addressing manufacturing needs efficiently and with high quality.

• **Human well-being:** catering to workers' physical and mental wellbeing.

• **Environmental sustainability:** minimizing resource consumption, reducing emissions, and promoting eco-friendly practices.

• **Societal impacts:** upholding ethical standards and contributing to broader societal sustainability.

Synthesizing these four diverse attributes presents a complex multi-objective optimization (MOO) challenge. Traditionally, MOO is typically resolved through mathematical formulations. In contrast, the proposed GenAI paradigm approaches this optimization through semantic reasoning and constraint satisfaction. By internalizing these values as high-dimensional context within the latent space, the GenAI implicitly navigates the trade-offs, aiming to generate solutions that logically co-satisfy the functional prompt and the embedded value constraints. This represents a fundamental shift from numerical trade-offs to holistic value alignment. However, achieving consistent, optimal trade-offs through the implicit semantic approach, remains a significant research challenge, necessitating further investigation.

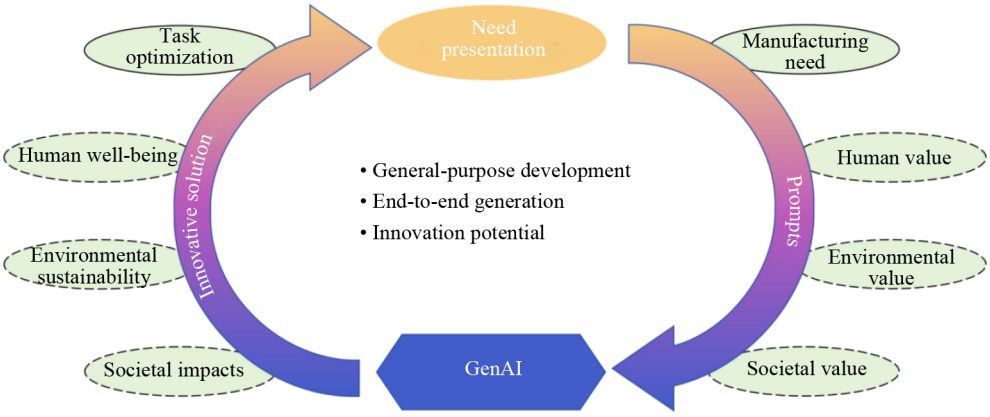


**Fig. 5** A futuristic paradigm of GenAI implementation in manufacturing.

Compared to the traditional paradigm, the futuristic paradigm of GenAI implementation in manufacturing exhibits three transformative features:

1) **General-purpose development**: *One GenAI model addressing diverse applications.* Unlike the task-specific approach, FMs are inherently general-purpose. A single model can seamlessly transition between tasks without the need for task-specific retraining. This capability directly addresses the industrial need for scalable intelligence, eliminating the high cost and complexity of maintaining a fragmented ecosystem of bespoke AI tools. This shift advocates for researchers to transition from developing specialized AI models to collaboratively advancing general GenAI for manufacturing.

2) **Holistic generation**: *Holistically generating a solution to meet the manufacturing need*. In contrast to the expert-driven solution design, GenAI functions as a holistic service provider to generate and execute solutions that fulfil the manufacturing needs directly from the prompt. This addresses the industry need for operational agility by mitigating the complexity of manual system integration. The quality of generated solutions averts the impact of the experts' expertise variations but is highly dependent on the proficiency of the GenAI.

3) **Innovation potential**: *Transcending expert-defined boundaries*. By leveraging broad knowledge and reasoning capabilities, GenAI can explore solution spaces beyond human intuition, conceptualizing novel solutions that redefine the boundaries of manufacturing productivity and product performance. Early studies have shown promising results of GenAI innovations in manufacturing (Chong et al., 2024; Timperley et al., 2025); however, it is highlighted to ensure that generated solutions align with human, environmental, and societal values. To fully harness the innovation potential of GenAI, the development of UWM is essential.

Compared to the traditional paradigm, the new paradigm has revolutionized the manufacturing industry in three aspects: *Development mode* — shifting from a task-specific to a general-purpose development approach, where a single model can address diverse applications; *Solution Formulation* — transitioning from expert-driven solution design to AI-driven holistic solution generation, thereby enhancing efficiency and reducing reliance on human expertise; *Solution Potential* — breaking expert-defined boundaries to foster unprecedented innovation, addressing needs more effectively and comprehensively. Overall, in the traditional paradigm, AI functions as an *auxiliary tool* extending human capabilities. In contrast, within the futuristic paradigm, GenAI takes on the role of a *self-contained service provider* that directly and holistically fulfils manufacturing needs.

### 4.3 Evaluation framework

The proposed paradigm aims to guide future technology development. Consequently, it is essential to evaluate both the targeted technology's adherence to the paradigm and its practical effectiveness in real-world applications. The evaluation can be approached from two perspectives:

(1) Evaluating the alignment of the given technology with the three key features of the proposed paradigm

• **General-purpose development**: Evaluating the model's ability to seamlessly transition between different applications without requiring significant retraining or customization.

o *Metrics*: Zero-shot task success rate (ZSR) and Cross-task performance retention (CPR) when transferring to unseen tasks without fine-tuning.

o Measurement: $ZSR = N_{\text{successful tasks}}/N_{\text{total unseen tasks}}$, where $N_{\text{successful tasks}}$ denotes the number of tasks whose outputs satisfy predefined task constraints. $CPR = Performance_{\text{unseen}}/Performance_{\text{trained}}$, where $Performance_{\text{trained}}$ represents the model's average performance on tasks within the training domain, and $Performance_{\text{unseen}}$ represents its performance on the unseen tasks. Performance can be measured using task-specific metrics such as accuracy, constraint satisfaction

rate, or planning success rate.

• **Holistic generation**: Assessing the system's ability to autonomously generate and execute manufacturing solutions that directly meet user needs without requiring human intervention.

o *Metrics*: Human intervention ratio (HIR), defining as the number of mandatory human corrections divided by the total number of process steps generated; End-to-End Task Completion Rate (ECR).

o Measurement: A human-in-the-loop logging protocol where every manual override of the GenAI's output is recorded. $HIR = N_{\text{manual corrections}}/N_{\text{total generated steps}}$ and $ECR = N_{\text{successful autonomous workflows}}/N_{\text{total tasks}}$.

• **Innovation potential**: Examining whether the generated solutions offer innovative approaches that go beyond existing human-designed solutions.

o *Metrics*: Design Novelty Index (DNI); Performance Improvement Ratio (PIR).

o *Measurement*: Compute novelty using distance from existing design database, $DNI = 1 - \text{similarity}(generated, nearest_{\text{existing}})$. Similarity methods include geometric similarity for design, sequence similarity for process planning. $PIR = (Performance_{\text{generated}} - Performance_{\text{baseline}})/Performance_{\text{baseline}}$. Performance could be cycle time, energy consumption, production throughput.

(2) The effectiveness of the technology can be validated through metrics that focus on four dimensions

• **Task performance**: Evaluating the system's efficiency, accuracy, and quality in executing manufacturing tasks.

o *Metrics*: Overall equipment effectiveness (OEE) (Singh et al., 2013) for production efficiency; First Pass Yield (FPY) for quality generation; Inference Latency (ms) for real-time control applicability; Task success rate (TSR) for reliability.

o *Measurement*: OEE is calculated as the product of availability, performance, and quality: $OEE = Availability \times Performance \times Quality$, where $Availability$ is defined as operating time divided by planned production time, $Performance$ is defined as actual output divided by theoretical maximum output, and $Quality$ is defined as the number of good units divided by the total units produced. FPY measures the proportion of products that meet quality standards without rework, $FPY = N_{\text{good units}}/N_{\text{total units produced}}$, where $N_{\text{good units}}$ represents units passing inspection in the first attempt. Inference Latency is measured as the average time difference between receiving an input query and producing the corresponding output. $TSR = N_{\text{sucessful tasks}}/N_{\text{total tasks}}$, where successful tasks are defined as those completed without system failure or manual override.

• **Human well-being**: Assessing the system's impact on workers' physical and mental health by analyzing metrics such as human fatigue level, task workload reduction, and job satisfaction.

o *Metrics*: NASA Task Load Index (TLX) (Hart and Starveland, 1988) for cognitive workload; RULA/REBA Scores (Kee 2021) for physical ergonomic risk.

o *Measurement*: After completing each task, operators rate six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration) on a scale from 0 to 100. $TLX = (\sum_{i=1}^{6} \omega_i \cdot r_i)/(\sum_{i=1}^{6} \omega_i)$, where $r_i$ represents the operator's rating for dimension $i$, and $\omega_i$ denotes the weighting factor obtained through pairwise comparison of the workload dimensions.

• **Environmental sustainability**: Quantifying the system's environmental impacts by evaluating the metrics related to resource utilization, energy consumption and other sustainability factors.

o *Metrics*: Carbon footprint; Specific energy consumption (SEC) (Lawrence et al., 2019) per unit; Material buy-to-fly (BTF) ratio (material efficiency).

o *Measurement*: Conduct life cycle assessment (ISO14040) on the GenAI-generated process plans to calculate the theoretical carbon footprint (kg $CO_2$) compared to traditional process plans. SEC is defined as the energy used per unit of production, expressed in kg $CO_2$ per unit. BTF is defined as the mass of raw material divided by the mass of the finished part.

• **Societal impact**: Evaluating the system's adherence to societal values such as net employment balance, workforce adaptation readiness, data privacy and decision transparency.

o *Metrics*: Explainability score (SHAP/LIME values); Workforce transition readiness score.

o *Measurement*: Apply interpretability tools (SHAP/LIME (Givisis et al., 2025)) to assess decision transparency. Conduct training adoption surveys to evaluate workforce readiness.

This section establishes a structured evaluation framework to assess GenAI-driven manufacturing systems. However, the proposed framework is intended to provide a structured baseline for evaluating GenAI-driven manufacturing systems from multiple dimensions rather than an exhaustive standard. The proposed metrics and measurement methods serve as a foundational reference that can be further adapted and extended according to specific application requirements. Given the rapid evolution of FMs, these measurements cannot fully capture the complex, emergent behaviors of autonomous agents. Therefore, establishing a universal benchmarking protocol remains an ongoing challenge, requiring continuous metric refinement and the development of large-scale, open-source data sets.

### 4.4 A GenAI-driven product design-to-manufacturing framework and case study

In this section, a holistic GenAI-driven product design-to-manufacturing framework is proposed to demonstrate GenAI's capability to function as a self-contained service

provider. This framework integrates distinct operational phases, ranging from product design and production planning to manufacturing execution and system maintenance, into a unified, intelligent workflow. Subsequently, to ground these theoretical components in the physical world, a case study regarding the autonomous design and fabrication of a customized phone stand is conducted. This empirical experiment validates the framework's feasibility and illustrates its unique ability to embed human and environmental values directly into the manufacturing lifecycle.

#### 4.4.1 A GenAI-driven product design-to-manufacturing framework

A GenAI-driven product design-to-manufacturing framework is presented in Fig. 6. This conceptual framework synthesizes GenAI's demonstrated capabilities in product design (Chong et al., 2024), process planning (Wang and Qin, 2024), automation (Xia et al., 2024b), maintenance (Kiangala and Wang, 2024), and supply chain management (Jackson et al., 2024b), integrating these advancements into a cohesive system. It demonstrates how GenAI holistically designs and manufactures products that meet user requirements while simultaneously maintaining the production system.

In this framework, the user interacts with the system solely by providing an unstructured, semantic prompt describing the desired products. GenAI then operates in three sequential stages:

1) **Product design**: Based on the user's prompt, GenAI generates a highly tailored product design that aligns with the specified requirements. GenAI synthesizes entirely new geometry from abstract descriptions and proactively infers implicit constraints (e.g., user preferences, ergonomics, sustainability) without explicit specification.

2) **Production planning**: The generated design is transformed into an implicit prompt, combined with the initial prompt, enabling GenAI to generate a dynamic production plan that incorporates existing manufacturing resources and constraints. Through cross-domain reasoning, GenAI infers implicit constraints and adapts to resource availability, capabilities beyond rule-based scheduling approaches.

3) **Manufacturing**: The production plan is further processed as an implicit prompt, guiding GenAI to coordinate and manage manufacturing resources to execute the production tasks. GenAI generates machine-specific instructions while handling exceptions through reasoning, rather than relying on pre-programmed error-handling routines.

Simultaneously, GenAI performs real-time analysis of sensor data to monitor and maintain the production system, ensuring operational stability and efficiency. To fulfill the above applications effectively, GenAI needs to engage in real-time communication and interaction with

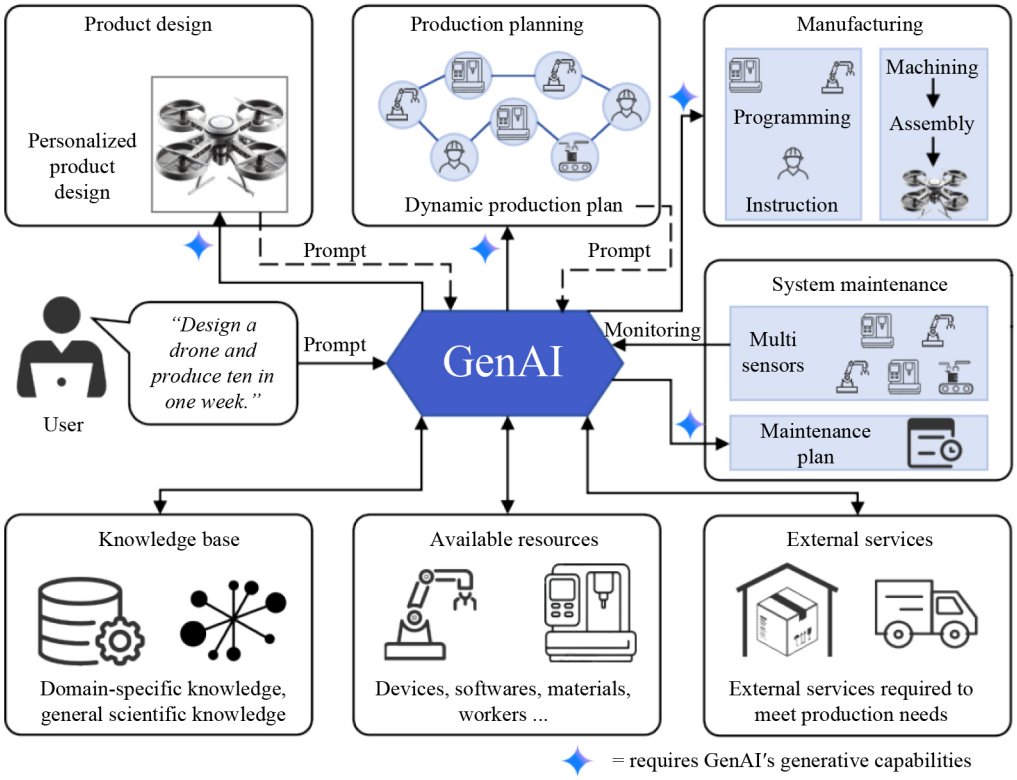


**Fig. 6** A GenAI-driven product design-to-manufacturing framework.

knowledge bases, available production resources and necessary external services.

Figure 6 offers a conceptual visualization of how GenAI integrates seamlessly into the design-to-manufacturing workflow in the proposed paradigm fashion. It is important to emphasize that this framework fundamentally depends on GenAI's unique generative capabilities. Traditional discriminative models map inputs to outputs within predefined parameter spaces, whereas GenAI *generates* novel solutions through semantic reasoning in open-ended spaces. Specifically, traditional AI cannot: (1) interpret ambiguous natural language prompts without structured preprocessing; (2) generate novel product designs from abstract user intent; (3) infer implicit constraints (e.g., user preferences, ergonomics, sustainability) without explicit specification; or (4) maintain coherent cross-domain reasoning across design, planning, and execution stages within a single FM. These generative capabilities are what enable the paradigm shift from AI as an auxiliary tool to GenAI as a self-contained service provider (Section 4.2).

#### 4.4.2 Case study: Value-driven design-to-manufacturing of a phone stand

To validate the Design-to-Manufacturing framework (Fig. 6) and ground its theoretical components in physical reality, we implemented a case study: the autonomous design and fabrication of a customized phone stand. This case study demonstrates the system's ability to translate abstract user intent into physical product without human intervention. This experiment highlights the framework's value-embedded architecture where agents are not merely functional executors but are conditioned to proactively optimize for human-centric ergonomics, societal responsibility, and environmental sustainability. The experimental setup comprises a MAS driven by Claude Opus 4.5 as the reasoning core, integrated with a Creality Ender 3 3D printer managed by an OctoPrint server. The workflow proceeds through four phases, as shown in Fig. 7.

**Phase 1**: Intent parsing & value alignment: The workflow commences with a user prompt: *"I need a desk phone stand for my iPhone 16 that allows for charging while in use."* The Requirement Analysis Agent parses this input via a value-embedded Meta-Prompt, which injects value constraints into the design brief alongside the explicit functional requirements. In this case, it contains two types of value constraints:

- *Human value (Ergonomics):* The agent proactively infers the need for usability, appending the constraint: *"Design a 60-degree angle from horizontal to optimize user posture and reduce neck strain."*
- *Environmental value (Sustainability):* The agent interprets the generic need for a stand through a sustainability lens, appending the structure type constraint: *"Frame with large cutouts to reduce plastic waste."*

**Phase 2**: Generative product design: The Design Agent translates this enriched design brief into executable geometry. Leveraging the coding capabilities of Claude Opus 4.5, the agent generates a parametric OpenSCAD script. This code-based approach enables mathematical control over functional features, including accommodating the dimensions of iPhone 16, integrating a bottom cutout for charging cable, enforcing the ergonomic 60-degree inclination, and designing a frame structure to save materials.

**Phase 3**: Sustainability optimization: To strictly enforce the system's environmental mandate, the design is passed to a specialized Sustainability Agent. The agent applies a Biomimetic Bone Structure optimization algorithm to excavate the internal volume but saving the solid skin to handle mechanical loads. This strategy reduces filament consumption by 30.1% compared to a solid counterpart. This Agent is also designed by Claude Opus 4.5.

**Phase 4**: Automated 3D printing: The 3D Printing Agent utilizes a headless slicing engine (CuraEngine CLI) to convert the optimized geometry into machine-specific G-code. Then, it executes the fabrication using the OctoPrint REST API. The agent utilizes a POST

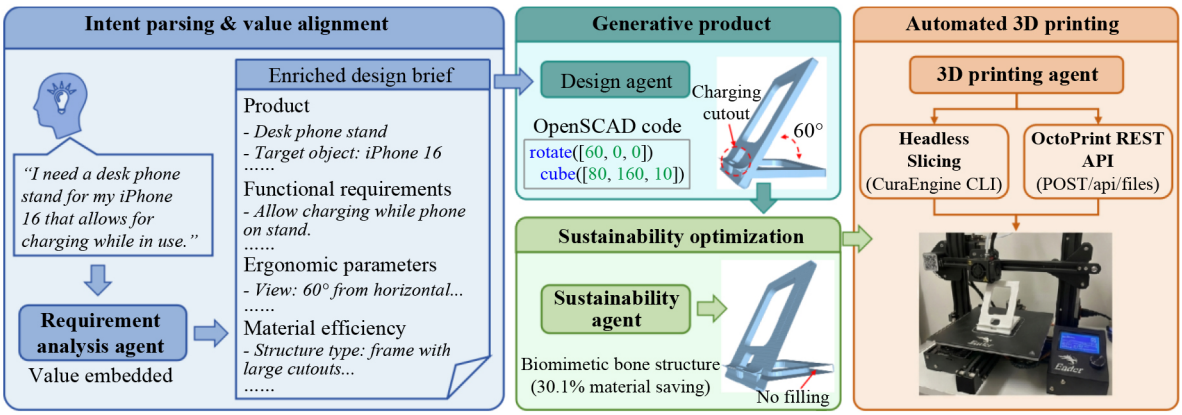


**Fig. 7** Workflow of the value-driven autonomous manufacturing of a phone stand.

request to autonomously upload the G-code and triggers the print job on the connected printer.

The fabricated phone stand demonstrates that the system successfully addressed several core requirements: accommodating the iPhone 16 dimensions, implementing the ergonomic 60-degree viewing angle, and incorporating a bottom cutout for cable access. However, physical testing revealed two design shortcomings attributable to GenAI's limited spatial reasoning: (1) the front lip supporting the phone is insufficiently deep, resulting in unstable retention; and (2) the charging port cutout fails to account for the charger head length, providing inadequate vertical clearance.

Despite these shortcomings, this case study confirms that the proposed framework can successfully orchestrate the entire Design-to-Manufacturing cycle with embedded values, demonstrating a shift from passive automation to responsible, value-aligned autonomy. The Meta-Prompt, generated OpenSCAD code, Sustainability Agent, and product files are available at our Github repository.

Nevertheless, the observed deficiencies highlight the limitations inherent to current GenAI. While the agents excelled at semantic interpretation of functional requirements and sustainability optimization, they struggled with precise dimensional reasoning for physical interfaces, which is a manifestation of the spatial grounding challenges discussed in Section 3.4. Furthermore, the lack of a deterministic physics engine prevented structural verification via Finite Element Analysis (FEA). Consequently, the current framework is best suited for conceptual design and prototyping. Advancing toward production-ready systems will require GenAI with robust spatial reasoning capabilities for precise mating interfaces and kinematic constraints.

### 4.5 Further discussion

This section analyzes how the proposed paradigm supports the core principles of Industry 4.0 and Industry 5.0, and outlines critical directions for future research.

#### 4.5.1 Synergy with Industry 4.0 and Industry 5.0

According to (Hermann et al., 2016), Industry 4.0 is characterized by four key principles: *Interconnectivity*, *Information Transparency*, *Decentralized Decisions*, and *Technical Assistance*. The proposed paradigm supports them as follows:

• **Interconnectivity**: The proposed paradigm integrates GenAI with manufacturing devices, sensors, machines and systems, enabling seamless communication between interconnected objects. This establishes GenAI as a tool for device interconnectivity, eliminating the need for additional efforts to unify communication protocols across various devices. Additionally, GenAI's natural language interaction capabilities also ensure seamless communication between humans and interconnected objects.

• **Information transparency**: GenAI enhances transparency by providing comprehensive analysis, interpretation, and visualization of data from multiple sources, presenting it in a clear format for informed decision-making.

• **Decentralized decisions**: By improving interconnectivity and information transparency, GenAI enables decentralized decision-makers to access local and global data and devices, optimizing decision-making processes across different levels of the organization.

• **Technical assistance**: GenAI offers flexible and intuitive assistance, allowing users without specialized expertise to solve a wide range of problems efficiently.

Industry 5.0 emphasizes three principles: *Human-Centricity*, *Sustainability*, and *Resilience* (Xu et al. 2021). The proposed paradigm supports them as follows:

• **Human-centricity**: GenAI's natural interaction capabilities are inherently aligned with the principles of human-centric design. Furthermore, the proposed paradigm highlights that the generated solutions should prioritize human well-being.

• **Sustainability**: The general-purpose development mode of the proposed paradigm also enhances environmental sustainability by reducing redundant development efforts that consume significant energy (Samsi et al., 2023). Besides, the proposed paradigm requires the generated solutions to be energy-efficient.

• **Resilience**: By leveraging comprehensive data analysis from various sources and integrating diverse manufacturing equipment, GenAI enables effective responses to disruptions through global optimization and proactively maintenance scheduling, thus enhancing the resilience of manufacturing systems.

In conclusion, the proposed paradigm not only aligns with but also enhances the principles of both Industry 4.0 and Industry 5.0, making it highly valuable for the future of manufacturing.

#### 4.5.2 Future prospects and research directions

While the proposed paradigm offers a transformative vision, realizing its full potential requires addressing several frontier challenges. We identify three critical directions for future research:

• **From passive models to agentic AI:** The frontier lies in developing autonomous manufacturing agents capable of long-horizon planning, tool usage, and self-correction. Future research must focus on endowing these agents with embodied intuition, the ability to perceive and manipulate the physical world through active sensorimotor loops.

• **Federated foundation models for privacy:** Manu-

facturing data are highly proprietary. Developing Federated Learning frameworks for Foundation Models is essential. Future work should explore mechanisms to train shared UWMs across organizational boundaries without exposing sensitive local process data, balancing collective intelligence with corporate privacy.

• **Standardization and trustworthiness:** As GenAI takes on control roles, establishing rigorous industrial benchmarks is imperative. Future efforts must move beyond generic NLP metrics to define manufacturing-specific standards for reliability, safety, and physical validity. This includes the development of frameworks that embed immutable industrial safety protocols directly into the model's inference logic.

## 5 Conclusions

At the dawn of GenAI cutting a striking figure, this paper explores the role, applications, and prospects of GenAI in the manufacturing industry, highlighting its transformative potential. Through a proposed five-level framework characterizing the depth of GenAI integration, our analysis reveals that GenAI not only enhances every phase of the traditional AI implementation paradigm but also has the potential to establish a new paradigm that reshapes the future of manufacturing. In this new paradigm, GenAI directly addresses diverse needs with innovative solutions that transcend human design limitations, as well as address task, human, environmental and societal needs holistically. Furthermore, it enhances the core principles of Industry 4.0 and 5.0. The evaluation framework of the proposed paradigm is discussed. Following this, a GenAI-driven product design-to-manufacturing framework is introduced to ground the paradigm in practical applications.

Compared to the traditional paradigm, the new paradigm demonstrates three key advancements: general-purpose development approach, where a single model addresses multiple applications; AI-driven holistic solution generation, reducing reliance on human expertise; and the ability to generate innovative solutions that surpass human design limitations. As a result, this paradigm has the potential to reshape the manufacturing industry. In addition, GenAI's user-friendly nature lowers the entry barrier, enabling rapid adoption even among non-experts, further accelerating the industry's transformation.

This paper makes a bold proposition on the development trajectory of FMs, which underpins the envisioned futuristic paradigm. It is anticipated that FMs will evolve from current LLMs and VLMs to UWMs, enhancing capabilities in multimodal data interaction, comprehensive and scientific world understanding, and advanced knowledge application. However, these propositions are our prospection based on the analysis of literature and industry needs rather than a definitive conclusion. Readers are encouraged to critically evaluate our findings and engage in rigorous discourse. Moreover, we invite further substantive studies to validate, refine, and expand upon our propositions, fostering a deeper understanding and more robust advancement of GenAI applications in manufacturing.

Finally, the technical, operational, and social challenges of integrating GenAI into manufacturing have been thoroughly discussed. Successfully overcoming these challenges will require not only technological innovations but also active collaboration among stakeholders to advance management practices, policy frameworks, and legal regulations. It is necessary to address these challenges holistically to unlock the full potential of GenAI, ensuring it delivers widespread benefits to both industry and society.

**Acknowledgement** The authors are grateful for the valuable support from the Laboratory for Industry 4.0 Smart Manufacturing Systems, Department of Mechanical and Mechatronics Engineering, The University of Auckland.

**Competing Interests** The authors declare that they have no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

An Z, Ding X, Fu Y C, Chu C C, Li Y, Du W (2024). Golden-Retriever: High-Fidelity Agentic Retrieval Augmented Generation for Industrial Knowledge Base. arXiv preprint arXiv:2408.00798

Arinez J F, Chang Q, Gao R X, Xu C, Zhang J (2020). Artificial Intelligence in Advanced Manufacturing: Current Status and Future Outlook. Journal of Manufacturing Science and Engineering, 142(11): 110804

Ayyat M, Osman M, Nadeem T (2025). Opportunities and Challenges of Foundation Models in Industrial Manufacturing. IEEE Access: Practical Innovations, Open Solutions, 13: 138745–138775

Gmyrek P, Berg J, Bescond D, International Labour Organization. Research Department, (2023). Generative AI and jobs: a global analysis of potential effects on job quantity and quality. ILO, Geneva

Berthelot A, Caron E, Jay M, Lefèvre L (2024). Estimating the environmental impact of Generative-AI services using an LCA-based methodology. Procedia CIRP, 122: 707–712

Bing Y, Yu L, Li S, Cho Y, Li C (2026). A novel product shape innovation

design method based on Kansei Engineering and GAN model with limited sample data. Journal of Engineering Design, 37(3): 981–1006

Bonarini A (2020). Communication in Human-Robot Interaction. Current Robotics Reports, 1(4): 279–285

Brohan A, Brown N, Carbajal J, et al. (2023). RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv preprint arXiv:2307.15818

Cambria E, Malandri L, Mercorio F, Nobani N, Seveso A (2024). XAI meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models. arXiv preprint arXiv:2407.15248

Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu P S, Sun L (2023). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv preprint arXiv:2303.04226

Capitanelli A, Mastrogiovanni F (2024). A framework for neurosymbolic robot action planning using large language models. Frontiers in Neurorobotics, 18: 1342786

Cen J, Yu C, Yuan H, Jiang Y, Huang S, Guo J, Li X, Song Y, Luo H, Wang F, Zhao D, Chen H (2025). WorldVLA: Towards Autoregressive Action World Model. arXiv preprint arXiv:2506.21539

Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, Ye W, Zhang Y, Chang Y, Yu P S, Yang Q, Xie X (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3): 1–45

Chen A, Dohan D, So D (2023a). EvoPrompting: Language models for code-level neural architecture search. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 7787–7817

Chen C, Zhao K, Leng J, Liu C, Fan J, Zheng P (2025). Integrating large language model and digital twins in the context of industry 5.0: Framework, challenges and opportunities. Robotics and Computer-integrated Manufacturing, 94: 102982

Chen Y, Liu C (2024). Remaining useful life prediction: A Study on multidimensional industrial signal processing and efficient transfer learning based on Large Language Models. arXiv preprint arXiv:2410.03134

Chen Y T, Hsu C Y, Yu C M, Barhamgi M, Perera C (2023b). On the Private Data Synthesis Through Deep Generative Models for Data Scarcity of Industrial Internet of Things. IEEE Transactions on Industrial Informatics, 19(1): 551–560

Chong L, Rayan J, Dow S, Lykourentzou I, Ahmed F (2024). CAD-prompted generative models: A pathway to feasible and novel engineering designs. arXiv preprint arXiv:2407.08675

Colabianchi S, Bernabei M, Costantino F (2022). Chatbot for training and assisting operators in inspecting containers in seaports. Transportation Research Procedia, 64: 6–13

Cui C, Ma Y, Cao X, Ye W, Zhou Y, Liang K, Chen J, Lu J, Yang Z, Liao K D, Gao T, Li E, Tang K, Cao Z, Zhou T, Liu A, Yan X, Mei S, Cao J, Wang Z, Zheng C (2024). A survey on multimodal Large Language Models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, 958–979

Dai Y, Yan Z, Cheng J, Duan X, Wang G (2023). Analysis of multimodal data fusion from an information theory perspective. Information Sciences, 623: 164–183

Davis J, Edgar T, Porter J, Bernaden J, Sarli M (2012). Smart manufacturing, manufacturing intelligence and demand-dynamic performance. Computers & Chemical Engineering, 47: 145–156

Devillers B, Maytié L, VanRullen R (2025). Semi-supervised multimodal representation learning through a global workspace. IEEE Transactions on Neural Networks and Learning Systems, 36(5): 7843–7857

Dong W, Li S, Zheng P (2025). Toward embodied intelligence-enabled human–robot symbiotic manufacturing: A Large Language Model-based perspective. Journal of Computing and Information Science in Engineering, 25(5): 050801

Fakih M, Dharmaji R, Moghaddas Y, Quiros G, Ogundare O, Al Faruque M A (2024). LLM4PLC: Harnessing Large Language Models for Verifiable Programming of PLCs in Industrial Control Systems. In: Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice, 192–203

Fan H, Liu X, Fuh J Y H, Lu W F, Li B (2025a). Embodied intelligence in manufacturing: Leveraging large language models for autonomous industrial robotics. Journal of Intelligent Manufacturing, 36(2): 1141–1157

Fan J, Yin Y, Wang T, Dong W, Zheng P, Wang L (2025b). Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey. Engineering Management, 12(1): 177–200

Firoozi R, Tucker J, Tian S, Majumdar A, Sun J, Liu W, Zhu Y, Song S, Kapoor A, Hausman K, Ichter B, Driess D, Wu J, Lu C, Schwager M (2025). Foundation Models in Robotics: Applications, Challenges, and the Future. International Journal of Robotics Research, 44(5): 701–739

Gammulle H, Ahmedt-Aristizabal D, Denman S, Tychsen-Smith L, Petersson L, Fookes C (2023). Continuous Human Action Recognition for Human-machine Interaction: A Review. ACM Computing Surveys, 55(13s): 1–38

Garouani M, Ahmad A, Bouneffa M, Hamlich M, Bourguin G, Lewandowski A (2022). Towards big industrial data mining through explainable automated machine learning. International Journal of Advanced Manufacturing Technology, 120(1-2): 1169–1188

Ge C, Yu X, Zheng H, Fan Z, Chen J, Shum P P (2025). A dual reverse distillation scheme for image anomaly detection. Neurocomputing, 624: 129479

Ge Y, Hua W, Mei K, Ji J, Tan J, Xu S, Li Z, Zhang Y (2023). OpenAGI: When LLM meets domain experts. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc. 36: 5539–5568

Getz N, Tong X (2025). Large Language Model accelerated aintenance insights. In: Proceedings of Annual Conference of the PHM Society, 17(1)

Givisis I, Kalatzis D, Christakis C, Kiouvrekis Y (2025). Comparing Explainable AI Models: SHAP, LIME, and Their Role in Electric Field Strength Prediction over Urban Areas. Electronics (Basel), 14(23): 4766

Gkournelos C, Konstantinou C, Makris S (2024). An LLM-based approach for enabling seamless Human-Robot collaboration in assembly. CIRP Annals, 73(1): 9–12

Goodfellow I, Bengio Y, Courville A (2016). Deep Learning. MIT Press

Guo D, Yang D, Zhang H et al. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 645(8081): 633–638

Guo L, Lu Z, Yao L (2021). Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review. IEEE Transactions on Human-Machine Systems, 51(4): 300–309

Gupta M, Akiri C, Aryal K, Parker E, Praharaj L (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. IEEE Access: Practical Innovations, Open Solutions, 11: 80218–80245

Ha D, Schmidhuber J (2018). World Models. doi:10.5281/zenodo.1207631

Han Z, Gao C, Liu J, Zhang J, Zhang S Q (2024). Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv preprint arXiv:2403.14608

Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, Wang T, Ma J, Zhang X, Song L (2024). Large-scale foundation model on single-cell transcriptomics. Nature Methods, 21(8): 1481–1491

Hart S, Staveland L (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in Psychology. Elsevier, 139–183

Hermann M, Pentek T, Otto B (2016). Design Principles for Industrie 4.0 Scenarios. In: Proceedings of 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, Koloa, HI, USA, 3928–3937

Holland M, Chaudhari K (2024). Large language model based agent for process planning of fiber composite structures. Manufacturing Letters, 40: 100–103

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv preprint arXiv:2311.05232

Huang R, Li M, Yang D, Shi J, Chang X, Ye Z, Wu Y, Hong Z, Huang J, Liu J, Ren Y, Zou Y, Zhao Z, Watanabe S (2024a). AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. Proceedings of the AAAI Conference on Artificial Intelligence, 38(21): 23802–23804

Huang Y, Zhang J, Chen X, Lam A H F, Chen B M (2024b). From Simulation to Prediction: Enhancing Digital Twins with Advanced Generative AI Technologies. In: Proceedings of 2024 IEEE 18th International Conference on Control and Automation (ICCA). IEEE, Reykjavík, Iceland, 490–495

Jackson I, Ivanov D, Dolgui A, Namdar J (2024b). Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation. International Journal of Production Research, 62(17): 6120–6145

Jackson I, Jesus Saenz M, Ivanov D (2024a). From natural language to simulations: applying AI to automate simulation modelling of logistics systems. International Journal of Production Research, 62(4): 1434–1457

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y J, Madotto A, Fung P (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12): 1–38

Jiang J, Li J, Zhao W X, Song Y, Zhang T, Wen J R (2024). Mix-CPT: A Domain Adaptation Framework via Decoupling Knowledge Learning and Format Alignment. arXiv preprint arXiv:2407.10804

Jiang R, Yuan Z, Wang H, Fan Z, Zhang Y, Liang N, Yu X (2025). MCFPred: A Novel Multichannel Signal Adaptive Fusion Framework for Fault Diagnosis in Hydraulic Systems. IEEE Transactions on Instrumentation and Measurement, 74: 1–13

Jiang Y, Liu T, Bao J (2026). What are the eigen visual features for penetration state recognition? Expert Systems with Applications, 299: 130169

Jiang Z, Wen H, Han F, Tang Y, Xiong Y (2022). Data-driven generative design for mass customization: A case study. Advanced Engineering Informatics, 54: 101786

Jin P, Takanobu R, Zhang W, Cao X, Yuan L (2024). Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13700–13710

Kambhampati S (2024). Can large language models reason and plan? Annals of the New York Academy of Sciences, 1534: (1)15–18

Kang M, Lee S, Baek J, Kawaguchi K, Hwang S J (2023). Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 48573–48602

Katara P, Xian Z, Fragkiadaki K (2024). Gen2Sim: Scaling up Robot Learning in Simulation with Generative Models. In: Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, Yokohama, Japan, 6672–6679

Kee D (2021). Comparison of OWAS, RULA and REBA for assessing potential work-related musculoskeletal disorders. International Journal of Industrial Ergonomics, 83: 103140

Khan Z U, Nasim B, Rasheed Z (2025). Generative AI-based predictive maintenance in aviation: A systematic literature review. CEAS Aeronautical Journal, 16(2): 537–555

Khanghah K N, Patel A, Malhotra R, Xu H (2025). Large Language Models for Extrapolative Modeling of Manufacturing Processes. arXiv preprint arXiv:2502.12185

Kiangala K S, Wang Z (2024). An experimental hybrid customized AI and generative AI chatbot human machine interface to improve a factory troubleshooting downtime in the context of Industry 5.0. International Journal of Advanced Manufacturing Technology, 132(5-6): 2715–2733

Kusiak A (2018). Smart manufacturing. International Journal of Production Research, 56(1-2): 508–517

Kusiak A (2025). Generative artificial intelligence in smart manufacturing. Journal of Intelligent Manufacturing, 36(1): 1–3

Kyaw A H, Jeon S H, Smith M, Gershenfeld N (2025). Speech to Reality: On-Demand Production using Natural Language, 3D Generative AI, and Discrete Robotic Assembly. In: Proceedings of the ACM Symposium on Computational Fabrication, 16: 1–12.

Lawrence A, Thollander P, Andrei M, Karlsson M (2019). Specific Energy Consumption/Use (SEC) in Energy Management for Improving Energy Efficiency in Industry: Meaning, Usage and Differences. Energies, 12(2): 247

Leng J, Su X, Liu Z, Zhou L, Chen C, Guo X, Wang Y, Wang R, Zhang C, Liu Q, Chen X, Shen W, Wang L (2025). Diffusion model-driven smart design and manufacturing: Prospects and challenges. Journal of Manufacturing Systems, 82: 561–577

Leng J, Zheng K, Li R, Chen C, Wang B, Liu Q, Chen X, Shen W (2026). AIGC-empowered smart manufacturing: Prospects and challenges. Robotics and Computer-integrated Manufacturing, 97: 103076

Li F F (2025). From Words to Worlds: Spatial Intelligence is AI's Next Frontier. Available from the website of Substack

Li K, Patel O, Viégas F, Pfister H, Wattenberg M (2024a). Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. Preprint at arXiv. Available from the website of arxiv.org

Li M, Wang R, Zhou X, Zhu Z, Wen Y, Tan R (2023). ChatTwin: Toward Automated Digital Twin Generation for Data Center via Large Language Models. In: Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. ACM, Istanbul Turkey, 208–211

Li Y, Wu Z, Zhao H, Yang T, Liu Z, Shu P, Sun J, Parasuraman R, Liu T (2024c). ALDM-Grasping: Diffusion-aided Zero-Shot Sim-to-Real Transfer for Robot Grasping. arXiv preprint arXiv:2306.03341

Li Y, Zhang R, Liu J (2024b). An Enhanced Prompt-Based LLM Reasoning Scheme via Knowledge Graph-Integrated Collaboration. arXiv preprint arXiv:2403.11459

Lin L, Zhang S, Fu S, Liu Y (2025). FD-LLM: Large language model for fault diagnosis of complex equipment. Advanced Engineering Informatics, 65: 103208

Liu C, Tian W, Kan C (2022). When AI meets additive manufacturing: Challenges and emerging opportunities for human-centered products development. Journal of Manufacturing Systems, 64: 648–656

Liu M, Ene T D, Kirby R, et al. (2024). ChipNeMo: Domain-Adapted LLMs for Chip Design. arXiv preprint arXiv:2311.00176

Liu Q, Liu Z, Xiong B, Xu W, Liu Y (2021). Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function. Advanced Engineering Informatics, 49: 101360

Liu S, Shen Y, Zhang Y, Hou Z, Wang X, Luo J, Zhang Z (2026). iDesignGPT enhances conceptual design via large language model agentic workflows. Nature Communications, 17(1): 1997

Liu T, Bao J (2025). A Novel Period-Sensitive LSTM for Laser Welding Quality Prediction. IEEE Transactions on Industrial Informatics, 21(1): 830–838

Lu Y, Xu X, Wang L (2020). Smart manufacturing process and system automation – A critical review of the standards and envisioned scenarios. Journal of Manufacturing Systems, 56: 312–325

Luo H, Specia L (2024). From Understanding to Utilization: A Survey on Explainability for Large Language Models. arXiv preprint arXiv:2401.12874

Lv Y, Guo X, Zhou Q, Qian L, Liu J (2023). Predictive maintenance decision-making for variable faults with non-equivalent costs of fault severities. Advanced Engineering Informatics, 56: 102011

Ma Y, Zheng S, Yang Z, Zheng P, Leng J, Hong J (2025). Leveraging large language models in next generation intelligent manufacturing: Retrospect and prospect. Journal of Manufacturing Systems, 82: 809–840

Mai S, Sun Y, Zeng Y, Hu H (2023). Excavating multimodal correlation for representation learning. Information Fusion, 91: 542–555

Makatura L, Foshey M, Wang B, Hähnlein F, Ma P, Deng B, Tjandrasuwita M, Spielberg A, Owens CE, Chen PY, Zhao A, Zhu A, Norton WJ, Gu E, Jacob J, Li Y, Schulz A, Matusik W (2024). Large Language Models for Design and Manufacturing. An MIT Exploration of Generative AI

Manduchi L, Meister C, Pandey K, et al. (2025). On the Challenges and Opportunities in Generative AI. arXiv preprint arXiv:2403.00025

Martínez-Gutiérrez A, Díez-González J, Perez H, Araújo M (2024). Towards industry 5.0 through metaverse. Robotics and Computer-integrated Manufacturing, 89: 102764

Mavikumbure H S, Cobilean V, Wickramasinghe C S, Drake D, Manic M (2024). Generative AI in Cyber Security of Cyber Physical Systems: Benefits and Threats. In: Proceedings of the 2024 16th International Conference on Human System Interaction (HSI). IEEE, Paris, France, 1–8

Mikołajewska E, Mikołajewski D, Mikołajczyk T, Paczkowski T (2025). Generative AI in AI-Based Digital Twins for Fault Diagnosis for Predictive Maintenance in Industry 4.0/5.0. Applied Sciences (Basel, Switzerland), 15(6): 3166

Mittal S, Khan M A, Romero D, Wuest T (2019). Smart manufacturing: Characteristics, technologies and enabling factors. Proceedings of the Institution of Mechanical Engineers. Part B, Journal of Engineering Manufacture, 233(5): 1342–1361

Modarressi A, Fayyaz M, Yaghoobzadeh Y, Pilehvar M T (2022). GlobEnc: Quantifying Global Token Attribution by Incorporating the Whole Encoder Layer in Transformers. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, 258–271

Mu Y, Zhang Q, Hu M, Wang W, Ding M, Jin J, Wang B, Dai J, Qiao Y, Luo P (2023). EmbodiedGPT: Vision-Language Pre-Training via Embodied Chain of Thought. In: Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 25081–25094

Mustapha K B (2025). A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing. Advanced Engineering Informatics, 64: 103066

Nafea M (2025). 4D printing of generative AI-assisted designs. Smart Materials and Structures, 34(2): 025029

Naqvi M R, Elmhadhbi L, Sarkar A, Archimede B, Karray M H (2024). Survey on ontology-based explainable AI in manufacturing. Journal of Intelligent Manufacturing, 35(8): 3605–3627

Narayanan D, Shoeybi M, Casper J, LeGresley P, Patwary M, Korthikanti V, Vainbrand D, Kashinkunti P, Bernauer J, Catanzaro B, Phanishayee A, Zaharia M (2021). Efficient large-scale language model training on GPU clusters using megatron-LM. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, St. Louis Missouri, 1–15

Nasir M U, Earle S, Togelius J, James S, Cleghorn C (2024). LLMatic: Neural Architecture Search Via Large Language Models And Quality Diversity Optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference. ACM, Melbourne VIC Australia, 1110–1118

NIST (2018). National Institute of Standard and Technology Product Definitions for Smart Manufacturing Available from nist.gov

Neu D A, Lahann J, Fettke P (2022). A systematic literature review on state-of-the-art deep learning methods for process prediction. Artifi-

cial Intelligence Review, 55(2): 801–827

Ni A, Iyer S, Radev D, Stoyanov V, Yih W T, Wang S, Lin X V (2023). LEVER: Learning to Verify Language-to-Code Generation with Execution. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research. PMLR, 26106–26128

Nti I K, Adekoya A F, Weyori B A, Nyarko-Boateng O (2022). Applications of artificial intelligence in engineering and manufacturing: a systematic review. Journal of Intelligent Manufacturing, 33(6): 1581–1601

Ooi K B, Tan G W H, Al-Emran M, et al. (2025). The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. Journal of Computer Information Systems, 65(1): 76–107

Parthasarathy V B, Zafar A, Khan A, Shahid A (2024). The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. arXiv preprint arXiv:2408.13296

Picard C, Edwards K M, Doris A C, Man B, Giannone G, Alam M F, Ahmed F (2024). From Concept to Manufacturing: Evaluating Vision-Language Models for Engineering Design. arXiv preprint arXiv:2311.12668

Plathottam S J, Rzonca A, Lakhnori R, Iloeje C O (2023). A review of artificial intelligence applications in manufacturing operations. Journal of Advanced Manufacturing and Processing, 5(3): e10159

Pólya G, Conway J H (2014). How to solve it: a new aspect of mathematical method. Princeton University Press, Princeton, NJ, 1 pp

Prasad Agrawal K (2024). Towards Adoption of Generative AI in Organizational Settings. Journal of Computer Information Systems, 64(5): 636–651

Qiao S, Ou Y, Zhang N, Chen X, Yao Y, Deng S, Tan C, Huang F, Chen H (2023). Reasoning with Language Model Prompting: A Survey. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistic. Association for Computational Linguistics, Toronto, Canada, 5368–5393

Ren L, Wang H, Wang Y, Huang K, Wang L, Li B (2025a). Foundation Models for the Process Industry: Challenges and Opportunities. Engineering (Beijing), 52: 53–59

Ren Y, Liu Y, Ji T, Xu X (2025b). AI Agents and Agentic AI–navigating a plethora of concepts for future manufacturing. Journal of Manufacturing Systems, 83: 126–133

Ren Z, Zhu Y, Liu Z, Feng K (2023). Few-Shot GAN: Improving the Performance of Intelligent Fault Diagnosis in Severe Data Imbalance. IEEE Transactions on Instrumentation and Measurement, 72: 1–14

Ribeiro J, Lima R, Paiva S (2021). Document Classification in Robotic Process Automation Using Artificial Intelligence—A Preliminary Literature Review. In: Sharma H, Gupta MK, Tomar GS, Lipo W (Eds), Communication and Intelligent Systems. Lecture Notes in Networks and Systems. Springer Singapore, Singapore, 211–221

Rolf B, Jackson I, Müller M, Lang S, Reggelin T, Ivanov D (2023). A review on reinforcement learning algorithms and applications in supply chain management. International Journal of Production Research, 61(20): 7151–7179

Romera-Paredes B, Barekatain M, Novikov A, Balog M, Kumar M P, Dupont E, Ruiz F J R, Ellenberg J S, Wang P, Fawzi O, Kohli P, Fawzi A (2024). Mathematical discoveries from program search with large language models. Nature, 625(7995): 468–475

Ruiz E, Torres M I, Del Pozo A (2023). Question answering models for human–machine interaction in the manufacturing industry. Computers in Industry, 151: 103988

Samsi S, Zhao D, McDonald J, Li B, Michaleas A, Jones M, Bergeron W, Kepner J, Tiwari D, Gadepally V (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. In: 2023 IEEE High Performance Extreme Computing Conference (HPEC). IEEE, Boston, MA, USA, 1–9

Schick T, Schütze H (2021). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. arXiv preprint arXiv:2009.07118

Semeraro F, Griffiths A, Cangelosi A (2023). Human–robot collaboration and machine learning: A systematic review of recent research. Robotics and Computer-integrated Manufacturing, 79: 102432

Sengar S S, Hasan A B, Kumar S, Carroll F (2024). Generative artificial intelligence: a systematic review and applications. Multimedia Tools and Applications, 84(21): 23661–23700

Serradilla O, Zugasti E, Rodriguez J, Zurutuza U (2022). Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. Applied Intelligence, 52(10): 10934–10964

Shahin M, Hosseinzadeh A, Chen F F (2025). Generative artificial intelligence in manufacturing: applications, case studies, and future directions for next-generation intelligent production systems. International Journal of Advanced Manufacturing Technology, 141(3-4): 1159–1265

Shen Y, Lv O, Zhu H, Wang Y G (2024). ProteinEngine: Empower LLM with Domain Knowledge for Protein Engineering. In: Finkelstein J, Moskovitch R, Parimbelli E (Eds), Artificial Intelligence in Medicine. Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, 373–383

Shi D, Li J, Meyer O, Bauernhansl T (2025). Enhancing retrieval-augmented generation for interoperable industrial knowledge representation and inference toward cognitive digital twins. Computers in Industry, 171: 104330

Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022): 755–759

Singh R, Shah D B, Gohil A M, Shah M H (2013). Overall Equipment Effectiveness (OEE) Calculation - Automation through Hardware & Software Development. Procedia Engineering, 51: 579–584

Sumers T R, Yao S, Narasimhan K, Griffiths T L (2024). Cognitive Architectures for Language Agents. arXiv preprint arXiv:2309.02427

Sun Z, Shen S, Cao S, Liu H, Li C, Shen Y, Gan C, Gui L Y, Wang Y X, Yang Y, Keutzer K, Darrell T (2023). Aligning Large Multimodal Models with Factually Augmented RLHF. arXiv preprint arXiv:2309.14525

Tao F, Qi Q, Liu A, Kusiak A (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48: 157–169

Tao L, Liu H, Ning G, Cao W, Huang B, Lu C (2025). LLM-based framework for bearing fault diagnosis. Mechanical Systems and Signal Processing, 224: 112127

Thumm J, Trost F, Althoff M (2024). Human-Robot Gym: Bench-

marking Reinforcement Learning in Human-Robot Collaboration. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, Yokohama, Japan, 7405–7411

Timperley L R, Berthoud L, Snider C, Tryfonas T (2025). Assessment of large language models for use in generative design of model based spacecraft system architectures. Journal of Engineering Design, 36(4): 550–570

Tinsel E F, Lechler A, Riedel O, Verl A (2024). Concept of an Initial Requirements-Driven Factory Layout Planning and Synthetic Expert Verification for Industrial Simulation Based on LLM. In: 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN). IEEE, Beijing, China, 1–6

Tonmoy S M T I, Zaman S M M, Jain V, Rani A, Rawte V, Chadha A, Das A (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. arXiv preprint arXiv:2401.01313

Usuga Cadavid J P, Lamouri S, Grabot B, Pellerin R, Fortin A (2020). Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. Journal of Intelligent Manufacturing, 31(6): 1531–1558

Wach K, Duong C D, Ejdys J, Kazlauskaitė R, Korzynski P, Mazurek G, Paliszkiewicz J, Ziemba E (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. Entrepreneurial Business and Economics Review, 11(2): 7–30

Wan J, Li X, Dai H N, Kusiak A, Martinez-Garcia M, Li D (2021). Artificial-Intelligence-Driven Customized Manufacturing Factory: Key Technologies, Applications, and Challenges. Proceedings of the IEEE, 109(4): 377–398

Wang H, Liu M, Shen W (2023). Industrial‐generative pre‐trained transformer for intelligent manufacturing systems. IET Collaborative Intelligent Manufacturing, 5(2): e12078

Wang J, Ma Y, Zhang L, Gao R X, Wu D (2018). Deep learning for smart manufacturing: Methods and applications. Journal of Manufacturing Systems, 48: 144–156

Wang K, Zhu J, Ren M, Liu Z, Li S, Zhang Z, Zhang C, Wu X, Zhan Q, Liu Q, Wang Y (2024a). A Survey on Data Synthesis and Augmentation for Large Language Models. arXiv preprint arXiv:2410.12896

Wang T, Fan J, Zheng P (2024b). An LLM-based vision and language cobot navigation approach for Human-centric Smart Manufacturing. Journal of Manufacturing Systems, 75: 299–305

Wang T H, Maalouf A, Xiao W, Ban Y, Amini A, Rosman G, Karaman S, Rus D (2024c). Drive Anywhere: Generalizable End-to-end Autonomous Driving with Multi-modal Foundation Models. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, Yokohama, Japan, 6687–6694

Wang X, Jiang Z, Xiong Y, Liu A (2025). Human-LLM collaboration in generative design for customization. Journal of Manufacturing Systems, 80: 425–435

Wang Y, Zhang T, Guo X, Shen Z (2024d). Gradient based Feature Attribution in Explainable AI: A Technical Review. arXiv preprint arXiv:2403.10415

Wang Z, Ramnath K, Bi B, et al. (2024f). Reinforcement Learning for LLM Post-Training: A Survey. arXiv preprint arXiv: 2407.16216

Wang Z, Qin H (2024). Intelligent industrial production process automatic regulation system based on LLM agents. In: 2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA). IEEE, Shenzhen, China, 133–137

Wang Z, Wang W, Chen Q, Wang Q, Nguyen A (2024e). Generating Valid and Natural Adversarial Examples with Large Language Models. In: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, Tianjin, China, 1716–1721

Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi E H, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W (2022b). Emergent Abilities of Large Language Models. arXiv preprint arXiv:2206.07682

Wei J, Wang X, Schuurmans D, Bosma M ichter brian, Xia F, Chi E, Le QV, Zhou D (2022a). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 24824–24837

White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt D C (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv preprint arXiv:2302.11382

Wolf R, Shi Y, Liu S, Rayyes R (2025). Diffusion models for robotic manipulation: a survey. Frontiers in Robotics and AI, 12: 1606247

Wu Y, Feng Y, Peng S, Mao Z, Chen B (2023). Generative machine learning-based multi-objective process parameter optimization towards energy and quality of injection molding. Environmental Science and Pollution Research International, 30(18): 51518–51530

Xia L, Li C, Zhang C, Liu S, Zheng P (2024a). Leveraging error-assisted fine-tuning large language models for manufacturing excellence. Robotics and Computer-integrated Manufacturing, 88: 102728

Xia Y, Jazdi N, Zhang J, Shah C, Weyrich M (2024b). Control Industrial Automation System with Large Language Models. arXiv preprint arXiv:2409.18009

Xia Y, Shenoy M, Jazdi N, Weyrich M (2023). Towards autonomous system: flexible modular production system enhanced with large language model agents. In: 2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, Sinaia, Romania, 1–8

Xu M, Cai D, Yin W, Wang S, Jin X, Liu X (2025). Resource-efficient Algorithms and Systems of Foundation Models: A Survey. ACM Computing Surveys, 57(5): 1–39

Xu S, Wei Y, Zheng P, Zhang J, Yu C (2024a). LLM enabled generative collaborative design in a mixed reality environment. Journal of Manufacturing Systems, 74: 703–715

Xu W, Liu M, Sokolsky O, Lee I, Kong F (2024b). LLM-Enabled Cyber-Physical Systems: Survey, Research Opportunities, and Challenges. In: 2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys). IEEE, Hong Kong, Hong Kong, 50–55

Xu X, Ji T, Zheng P, Wang L (2026). Human-centric manufacturing: Re-thinking, Re-justifying, and Re-envisioning. Journal of Manufacturing Systems, 84: 259–268

Xu X, Lu Y, Vogel-Heuser B, Wang L (2021). Industry 4.0 and Industry 5.0—Inception, conception and perception. Journal of Manufacturing Systems, 61: 530–535

Yang F, Zhao P, Wang Z, Wang L, Zhang J, Garg M, Lin Q, Rajmohan S, Zhang D (2023). Empower Large Language Model to Perform Better on Industrial Domain-Specific Question Answering. arXiv preprint arXiv:2305.11541

Yang H, Siew M, Joe-Wong C (2024). An LLM-Based Digital Twin for Optimizing Human-in-the Loop Systems. In: 2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems and Internet of Things (FMSys). IEEE, Hong Kong, Hong Kong, 26–31

Yao S, Yu D, Zhao J, Shafran I, Griffiths T L, Cao Y, Narasimhan K (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv preprint arXiv:2305.10601

Yin R, Wang D, Zhao S, Lou Z, Shen G (2021). Wearable Sensors‐Enabled Human–Machine Interaction Systems: From Design to Application. Advanced Functional Materials, 31(11): 2008936

Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E (2024). A Survey on Multimodal Large Language Models. arXiv preprint arXiv:2306.13549

Yin Z, Sun Q, Guo Q, Wu J, Qiu X, Huang X (2023). Do large language models know what they don't know? arXiv preprint arXiv:2305.18153

Yu X, Chen Z, Ling Y, Dong S, Liu Z, Lu Y (2023). Temporal data meets LLM – Explainable financial time series forecasting. arXiv preprint arXiv:2306.11025

Yuan A, Garcia Colato E, Pescosolido B, Song H, Samtani S (2025). Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots. ACM Transactions on Management Information Systems, 16(1): 1–26

Zhang A, Fei H, Yao Y, Ji W, Li L, Liu Z, Chua T S (2023). VPGTrans: Transfer Visual Prompt Generator across LLMs. In: Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 20299–20319

Zhang C, Xu Q, Yu Y, Zhou G, Zeng K, Chang F, Ding K (2025). A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing. Robotics and Computer-integrated Manufacturing, 92: 102883

Zhang H, Semujju S D, Wang Z, Lv X, Xu K, Wu L, Jia Y, Wu J, Liang W, Zhuang R, Long Z, Ma R, Ma X (2026). Large scale foundation models for intelligent manufacturing applications: a survey. Journal of Intelligent Manufacturing, 37(1): 119–170

Zhang J, Huang J, Jin S, Lu S (2024a). Vision-Language Models for Vision Tasks: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8): 5625–5644

Zhang S, Zhang S, Wang B, Habetler T G (2020). Deep Learning Algorithms for Bearing Fault Diagnostics—A Comprehensive Review. IEEE Access: Practical Innovations, Open Solutions, 8: 29857–29881

Zhang Y, Chen Z, Fang Y, Lu Y, Li F, Zhang W, Chen H (2024b). Knowledgeable Preference Alignment for LLMs in Domain-specific Question Answering. arXiv preprint arXiv:2311.06503

Zhao F, Zhang C, Geng B (2024a). Deep Multimodal Data Fusion. ACM Computing Surveys, 56(9): 1–36

Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M (2023). Explainability for Large Language Models: A Survey. arXiv preprint arXiv:2309.01029

Zhao S, Liu S, Jiang Y, Zhao B, Lv Y, Zhang J, Wang L, Zhong R Y (2025). Industrial Foundation Models (IFMs) for intelligent manufacturing: A systematic review. Journal of Manufacturing Systems, 82: 420–448

Zhao Z, Tang D, Zhu H, Zhang Z, Chen K, Liu C, Ji Y (2024b). A Large Language Model-based multi-agent manufacturing system for intelligent shopfloor. arXiv preprint arXiv:2405.16887

Zheng H, Chand S, Keshvarparast A, Battini D, Lu Y (2023). Video-Based Fatigue Estimation for Human-Robot Task Allocation Optimisation. In: Proceedings of the 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE). IEEE, Auckland, New Zealand, 1–6

Zheng H, Cheng G, Li Y, Liu C (2020). A fault diagnosis method for planetary gear under multi-operating conditions based on adaptive extended bag-of-words model. Measurement, 156: 107593

Zheng H, Cheng G, Lu Y, Liu C, Li Y (2022). A general fault diagnosis framework for rotating machinery and its flexible application example. Measurement, 199: 111497

Zheng H, Xia W, Xu X (2025). A human-robot collaborative assembly framework with quality checking based on real-time dual-hand action segmentation. Robotics and Computer-integrated Manufacturing, 94: 102976

Zheng P, Li C, Fan J, Wang L (2024). A vision-language-guided and deep reinforcement learning-enabled approach for unstructured human-robot collaborative manufacturing task fulfilment. CIRP Annals, 73(1): 341–344

Zheng P, Wang H, Sang Z, Zhong R Y, Liu Y, Liu C, Mubarok K, Yu S, Xu X (2018). Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives. Frontiers of Mechanical Engineering, 13(2): 137–150

Zhou B, Li X, Liu T, Xu K, Liu W, Bao J (2024). CausalKGPT: Industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. Advanced Engineering Informatics, 59: 102333

Zhou C, Huang B, Fränti P (2022). A review of motion planning algorithms for intelligent robots. Journal of Intelligent Manufacturing, 33(2): 387–424

Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, Zhang K, Ji C, Yan Q, He L, Peng H, Li J, Wu J, Liu Z, Xie P, Xiong C, Pei J, Yu P S, Sun L (2023). A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv preprint arXiv:2302.09419

Zhou Q, Gu Y, Li J, Feng B, Li B, Bi Y (2026). Towards zero-shot robot tool manipulation in industrial context: A modular VLM framework enhanced by multimodal affordance representation. Robotics and Computer-integrated Manufacturing, 98: 103161