ASME Journal of Computing and Information Science in Engineering
Online journal at:
https://asmedigitalcollection.asme.org/computingengineering

Wenhang Dong
Research Institute for Advanced Manufacturing,
Department of Industrial and Systems Engineering,
The Hong Kong Polytechnic University,
Hong Kong SAR, China
e-mail: wenhang.dong@connect.polyu.hk

Shufei Li[1]
Department of Industrial and Systems Engineering,
The Hong Kong Polytechnic University,
Hong Kong SAR, China;
State Key Laboratory of Digital Manufacturing Equipment and Technology,
Huazhong University of Science and Technology,
Wuhan, China
e-mails: shufei.li@connect.polyu.hk;
shufei.li@outlook.com

Pai Zheng[1]
Research Institute for Advanced Manufacturing,
Department of Industrial and Systems Engineering,
The Hong Kong Polytechnic University,
Hong Kong SAR, China;
State Key Laboratory of Digital Manufacturing Equipment and Technology,
Huazhong University of Science and Technology,
Wuhan, China
e-mail: pai.zheng@polyu.edu.hk

Toward Embodied Intelligence-Enabled Human–Robot Symbiotic Manufacturing: A Large Language Model-Based Perspective

Human–robot collaborative manufacturing is crucial in modern production landscapes for flexible automation. However, existing human–robot systems face several challenges, including a lack of human-centric autonomy for adjustments, limited generalization for production variants, and ineffective synchronous teamwork with feedback. Emerging large language models (LLMs) offer a viable solution. Despite the growing interest in LLMs, their deployment within the manufacturing domain remains unexplored. This article delves into the potential of LLMs for embodied intelligence-enabled human–robot symbiotic manufacturing (HRSM). HRSM is a paradigm characterized by human centricity, generalizability, and seamless integration. LLMs can facilitate human-centric interactions, generalizable collaboration, and ensure seamless execution, paving the way for realizing HRSM. Finally, the main challenges and potential opportunities are further discussed to enable the readiness toward HRSM. [DOI: 10.1115/1.4068235]

Keywords: large language models, human robot collaboraton, human robot symbiotic manufacturing

1 Introduction

In contemporary smart manufacturing scenarios, existing automation systems fail to meet the demands of flexible and dynamic manufacturing tasks, such as assembling complex products with frequent changes [1]. To satisfy the diverse and personalized needs of these working conditions, human–robot collaborative manufacturing has emerged as a promising approach. By combining the strengths of robots, such as repeatability and accuracy, with the cognitive abilities, flexibility, and adaptability of humans, a symbiotic working environment is created, resulting in enhanced overall productivity [2].

The development of human–robot collaborative manufacturing has gone through several stages. Humans and robots were first separated by fences in manufacturing tasks. Then, researchers attempted to establish connectivity between human and robot through various means such as gestures [3], speech [4], and even brain–machine interfaces [5]. However, these interactions were often one-way and conveyed limited information [6]. Subsequently, the human and robot team evolved into a collaborative relationship [7], instead of a one-way command–obedience mode. In this stage, robots could infer human intentions to a certain extent and human–robotic agents collaborate toward the same task goal. Despite this progress, the learning and cognitive capabilities of robots remained at a low level, impeding a flexible, adaptable production environment [8].

Furthermore, symbiotic collaboration was proposed, involving robots with advanced reasoning and contextual awareness [2]. This mode enabled collaboration between humans and robots to transcend fixed scenarios, catering to complex and dynamic applications in instructed spaces [2]. Following this evolutionary trend, several key goals are pursued: the sense of human belonging, the adaptability of human–robot collaboration (HRC) systems, and the seamless integration between human operations and robot manipulations. Ultimately, this convergence leads to the realization of human–robot symbiotic manufacturing (HRSM), characterized by a high level of human centricity, generalization, and seamlessness.

In this context, the HRSM concentrates on three pivotal elements. First, the interactive performance between humans and robots must be enhanced, fostering human initiative and participation, ultimately creating a sense of belonging [9]. Subsequently, the system's proactive planning and learning abilities should be bolstered, enabling robots to quickly understand and adjust to complex environments [10]. Finally, the attainment of seamless integration between offline programming and online interaction is essential to improve the efficiency of the human–robot system [11].

[1]Corresponding authors.

Manuscript received July 18, 2024; final manuscript received March 8, 2025; published online April 2, 2025. Assoc. Editor: Chih-Hsing Chu.

To achieve these goals, a promising pathway has emerged with the increasing research on large language models (LLMs). LLMs possess remarkable natural language interaction capability, including natural language understanding and generation. On the one hand, this capability enhances the interaction experience between humans and robots, strengthening the sense of human belonging. On the other hand, natural language carries rich semantic information, enabling finer granularity in collaboration between humans and robots and empowering humans to exert greater agency. Simultaneously, LLMs have the potential to enhance the cognitive collaboration capability of robots, enabling them to attain embodied intelligence and function as intelligent agents with the ability to adapt to complex and dynamic work scenarios and diverse tasks, akin to their human counterparts [12]. Furthermore, LLMs can empower human–robot systems with online development capability, facilitating synchronous interaction and programming, offering a powerful tool for seamless collaboration between humans and robots.

At present, LLMs have been discussed in artificial intelligence (AI) domains [13–15], which have put major efforts into providing a detailed understanding of their network structures and algorithms. Beyond this, research has also been conducted on applications in areas like robotics [16], medicine [17], design [18,19], material [20,21], and human–robot interaction [22]. However, there is still a gap in the literature when it comes to reviews in the manufacturing field, especially regarding human–robot systems. While motivated by this, this article aims to conduct a systematic investigation and perspective on the LLM-enabled HRSM. The rest of this article is organized as follows: Sec. 2 introduces the concept of embodied intelligence-enabled HRSM. Section 3 analyzes how LLMs improve the key capabilities of HRSM, including interaction, collaboration, and execution. Section 4 illustrates the challenges of the current LLM-enabled HRSM. Section 5 provides valuable directions for further exploration in the future. Finally, in Sec. 6, we conclude by synthesizing all of the aforementioned discussions.

2 Toward Embodied Intelligence-Enabled HRSM

2.1 The Characteristic of HRSM. In order to enhance the human–robot system, we defined the characteristics of HRSM from the following three perspectives: human centricity, generalization, and seamlessness. First, human centricity emphasizes the dominant position and needs of people in the system [23], ensuring that the system design and operation always revolve around people's comfort and efficiency [24]. This not only helps to improve production efficiency [25] but also improves the satisfaction and safety of operators [26]. Second, generalization requires robots to be able to adapt to different tasks and environments, thereby improving the flexibility and adaptability of the system. This feature enables the system to maintain efficient operation in a changing production environment and reduces dependence on specific tasks [27]. Finally, seamlessness focuses on efficient and smooth information interaction between humans and robots [28], ensuring that data and instructions can be transmitted in real time, reducing misunderstandings and delays [29]. This efficient information flow is the key to achieving HRSM and can significantly improve the response speed and coordination of the overall system. These three characteristics of the HRSM system will be further explained below.

2.1.1 Human Centricity. Industry 5.0 shifts the center of manufacturing from system-oriented manufacturing systems to human-centric manufacturing systems [30]. To achieve this goal, it is necessary to identify the needs of humans in manufacturing. Human needs are considered to largely follow Abraham Maslow's "Hierarchy of Needs" [31]. According to Maslow's theory, human needs can be divided into five levels from low to high, including physiological, safety, love/belonging, esteem, and self-actualization [32]. However, human needs in the manufacturing industry have not yet been clearly classified. In the manufacturing field, a full understanding of human needs plays a guiding role in clarifying the future development direction of future human–robot systems [33]. Therefore, after referring to Maslow's human needs pyramid [32] and the industrial human needs pyramid [9] and discussing with manufacturing practitioners, we proposed a human needs pyramid for manufacturing. Human needs for manufacturing systems can be divided into five levels: safety, health, assistance, belonging, and self-actualization, as illustrated in Fig. 1.

- *Safety*: It represents the fundamental human needs within the working environment. Potential hazards in the manufacturing environment (i.e., high-speed rotating gears) may pose risks to worker safety. Thus, ensuring worker safety is the precondition for the coexistence of humans and robots.
- *Health*: It pertains to the impact of the working environment on workers. During manufacturing activities, workers may encounter mental and physical issues (i.e., muscle strain and anxiety) during manufacturing activities. Thus, ensuring the health of operators during extended manufacturing tasks is vital to eliminating occupational disease.
- *Assistance*: It denotes the companionship and support between workers and robots. The robot should provide human-desired assistance, and it can alleviate human workloads and enhance human skills and capabilities, and ultimately creates a more convenient and comfortable working environments for manufacturing workers.
- *Belonging*: It represents the human satisfaction experience in manufacturing processes. In human-centric smart manufacturing, humans should be an active part of a human–robot system with a distinctive role in team success [34]. Enhancing human mental belonging can improve one's proactive responsibility and problem-solving ability in new tasks.
- *Self-actualization*. It encompasses self-fulfillment, personal growth, and peak experiences. This level is essential for the coevolution of humans and robots. This represents the ultimate requirement for workers in the realm of human-centric manufacturing.

In manufacturing, various studies have focused on human needs, including obstacle avoidance [35], ergonomics [36], and assisted work [37]. However, research on the aspect of belonging remains relatively limited. One of the distinctive features of HRSM is the increased focus on belonging, which aims to maximize bidirectional empathic teamwork in the manufacturing system [6].

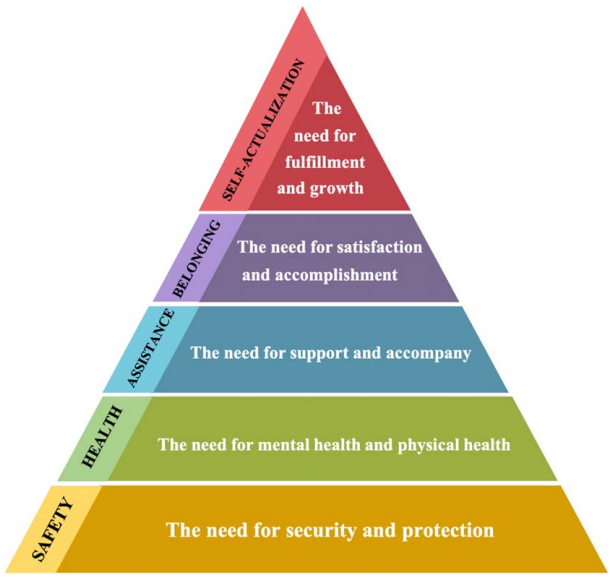

Fig. 1 Human needs pyramid architecture model

2.1.2 Generalization. Generalizable HRSM means the ability to achieve adaptive production processes across diverse tasks and new scenarios. An emerging trend in flexible and re-configurable manufacturing is the introduction of collaborative robots that work alongside humans with complementary skills [38–40]. AI technology is essential for this vision [41] by providing robots with the ability of autonomous perception, optimized decision-making, and precise execution, bringing new vitality to manufacturing [42].

In improving the system's environmental adaptability, existing studies leverage dynamic environmental information captured by sensors. For example, vision-based digital twins [43] or additional torque information [44] are used to optimize the robot's motion trajectory. In deconstructing tasks, some studies use meta-reinforcement learning [45] to enhance the adaptability of collaborative robots to new tasks through task modularization. Other studies use the sequential decision-making framework [46] to break down complex tasks into a series of decisions, which are then optimized to complete the tasks effectively.

Although these studies have achieved a certain level of generalization in human–robot systems, they often rely on prior knowledge. For example, complex environmental information need selection from human for robot path optimization or devising policies as the basis for task decomposition. This reliance on manual input limits the high-level "intelligence" of HRSM. In HRSM, robots should be able to infer the information necessary for interaction with the environment, thereby accomplishing comprehensive and dynamic planning for the task.

2.1.3 Seamlessness. Seamless HRSM means that the development and interaction can be carried out online at the same time, achieving continuity and high integration of the production process. HRSM typically occurs through two main approaches [47]: (1) Offline programming of robot tasks by demonstration [48]). (2) Online interaction between humans and robots enabled by external sensor systems. However, an asynchronous issue arises between these two approaches as the online interaction in the human–robot system is constrained by the offline programming of the robot, which does not allow for instantaneous feedback of interaction and execution. This issue restricts industrial companies from achieving resilient manufacturing (i.e., transformable production without high changeover times when new products are introduced by manufacturers [49]). The key to addressing this issue is improving the ability of seamless execution in the new manufacturing processes without significant delays or disruptions. One approach to bridge this gap is to utilize triggered action programming to achieve online programming [50,51], which is a programming method that uses events to trigger behaviors. However, this approach essentially aims to shorten the offline programming time, rather than truly enabling the robot to adapt directly to new environments. Achieving seamless HRSM requires developing more natural and cost-effective techniques for robotic programming to adjust and adapt in real time based on human input, and integrate advanced sensor systems to provide accurate and timely feedback to the robot.

2.2 Embodied Intelligence-Enabled HRSM. Embodied intelligence can endow robots with advanced perception, cognition, and action capabilities, promote more natural interactions with humans and their environment, achieve high levels of collaborative efficiency, and ultimately promote the development of manufacturing toward intelligence and high productivity. Therefore, embodied intelligence is expected to empower HRSM. This section first explains the basic definition of embodied intelligence, then conducts a comparative analysis of human–robot systems based on embodied intelligence and collaborative intelligence, and finally elaborates on the implementation of HRSM enabled by embodied intelligence.

2.2.1 The Concept of Embodied Intelligence. *Embodied intelligence refers to the purposeful exchange of physical or informational elements between an intelligent agent and the physical environment* [12]. A key aspect of embodied intelligence is *the integration of physical and cognitive capabilities* [52]. Intelligent agents, such as robots, will not only have the ability to understand scenes and generate responses to surrounding elements but also be able to physically manipulate objects and perform complex tasks [53].

In manufacturing, embodied intelligence implies the ability of autonomous learning to form cognition and action in industrial environments, realizing human–robot system coexistence, collaboration, and coevolution for HRSM. In this context, collaborative manufacturing elements, such as industrial robots, can work alongside humans as empathetic companions, complementing human skills and augmenting human capabilities [54].

2.2.2 Collaborative Intelligence Versus Embodied Intelligence. With HRSM implying human–robotic agents, its roadmap includes two pivotal aspects: embodied intelligence and collaborative intelligence. They overlap in enhancing the physical and cognitive capabilities of human–robot systems, with each encompassing an even broader scope, as shown in Fig. 2.

Collaborative intelligence emphasizes a complementary relationship between humans and robots [55]. In HRC scenarios, the collaborative intelligence-based approach is mainly in two paradigms, human-assisting robot and robot-assisting human [1]. The former refers that humans guide robots to adjust their operational strategies in response to novel situations, explain what decisions are made, and maintain updated knowledge of tasks. For example, human operators can use augmented reality technology to guide robots in assembly tasks [56]. The latter means that robots are designed to extend human work capabilities and provide flexible decision support, ensuring that human-centric needs are met [35]. For instance, robots can help humans perform delicate operations and reduce human burdens [57].

Embodied intelligence focuses on enhancing autonomous thinking and decision-making in robots through improved reasoning and generative capacities. This intelligence enables robots to adjust their behaviors in real time to accommodate changes in their human partners and task requirements. For example, in the manufacturing environment, embodied intelligence enables robots to autonomously identify and classify different parts of the production and adjust assembly order and strategy according to real-time production needs [58], thus achieving a highly flexible production process. When an abnormal situation occurs on the production line, robots with embodied intelligence can autonomously analyze the problem and take corresponding measures, such as replanning [59] or adjusting the operating parameters [60], to ensure the continuity and stability of production.

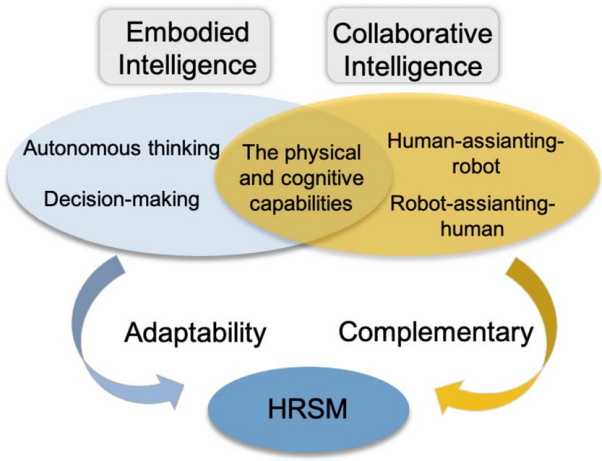

Fig. 2 Collaborative intelligence versus embodied intelligence

Hence, both collaborative intelligence and embodied intelligence can elevate the overall performance of human–robot systems, paving two pathways to achieve HRSM. Collaborative intelligence emphasizes the complementarity between humans and robots, while embodied intelligence focuses on the autonomy and adaptability of robots within human–robotic agents. In this article, by aligning with the principles of collaborative intelligence, embodied intelligence-enabled HRSM not only promotes bidirectional assistance for human–robot swarms but also optimizes their shared autonomy and hybrid intelligence augmentation in collaborative manufacturing environments.

2.3 Embodiments of Embodied Intelligence-Enabled HRSM. For human–robotic agents in manufacturing, embodied intelligence promotes human-like cognitive intelligence in three embodiments: perception, cognition, and action [61]. This allows human–robot swarms to adapt to all elements in manufacturing systems, including external environments, internal resources, tasks, and processes.

As shown in Fig. 3, perception involves sensing the unstructured environment and interpreting sensor data.

Robots utilize multiple sensors to acquire multimodal environmental information, including worker instructions, and production scenario. Leveraging the language understanding capability of LLMs, semantic information including worker intentions [62], equipment status [8], and material locations [63] can be extracted from the environmental information, thereby enhancing the perception capability of robots. Cognition refers to the process of reasoning, recommending, and providing feedback. The logical reasoning capability of LLMs can also enhance the task parsing, procedure understanding, and resource optimization ability of robots, enabling them to adapt better to complex production environments. Action involves physical interaction with others and task execution in the real environments. LLMs enable robots with planning capabilities, allowing them to decompose tasks into multiple subtasks and further generate a series of robot control instructions, and facilitating the seamless transition from instructions to executed actions. Hence, as an important technology for realizing embodied intelligence-enabled HRSM, the specific applications of LLMs will be discussed next.

3 Large Language Models and Connotation in HRSM

The most crucial part of HRSM is the interaction, collaboration, and execution between humans and robots, enabling them to accomplish a wide range of manufacturing tasks such as complex product assembly, flexible drilling with online learning, and human–robot skill transfer. The emergence of LLMs has reshaped these three modes of operation with human centricity, generalization, and seamlessness. In the following content of this section, the recent advances of mainstream LLMs are investigated. Then, we will analyze the connotation of LLMs in HRSM in detail according to the framework shown in Fig. 4.

3.1 The Introduction of Large Language Models

3.1.1 Large Language Models. At present, there is no uniformly recognized accurate definition of LLMs. Carlini et al. [64] point out that LLMs are composed of neural networks with a large number of parameters and are trained on a large amount of unlabeled data using self-supervised or semi-supervised learning. They have certain generalization capabilities and can perform a wide range of natural language processing. Zhao et al. [13] believe that LLMs refer to language models trained on massive text corpus and containing at least billions of parameters. These parameters allow them to learn very complex language patterns and perform tasks that would be difficult or impossible for smaller models. They can understand the statistical relationships between words and phrases, allowing them to generate grammatically correct and semantically meaningful text [65]. In general, LLMs include the following key points: (1) pretraining on a large amount of data, (2) the parameters are above one billion levels, and (3) they have a certain generalization ability. Therefore, in this article, LLM is a generalized LLM, including LLMs pretrained with large text-based data, vision language models (VLMs) pretrained with large image-based data, and multimodal large language models (mLLMs) trained based on large multimodal data.

The training process of LLMs is shown in Fig. 5. The first stage of LLMs training is pretraining process, which is unsupervised training using large amounts of data [13]. It includes not only public unlabeled data like Wikipedia but also professional text materials, such as technical standards and specifications [66]. After pretraining, the model is only a general language generation model and cannot perform nuanced tasks. Therefore, the base model needs to undergo fine-tuning progress. Using the base model's previous knowledge as a starting point, fine-tuning makes small adjustments to model parameters by training it on a narrow dataset, such as fault diagnosis records during assembly [67]. The performance of fine-tuning stage can be enhanced by incorporating human feedback through reward training [17] and reinforcement learning [17,68]. In order to obtain the responses of LLM for specific tasks, prompting techniques are needed. Prompting does not change the parameters and structure of the model. It focuses on adjusting the original input and adding some templates or prompts so that LLMs can output the result required by the target task [19]. For example, in the downstream navigation task, a code template based on the target location of the instruction is given, and the required target location only needs to be extracted from the instruction for completion [63].

3.1.2 Mainstream Large Language Models. LLMs are built upon the transformer architecture [69], which uses a multilayer neural network to understand context and learn meaning in sentences and long paragraphs using a mechanism called "attention" [13]. The transformer consists of encoders and decoders [70]. The encoder is responsible for comprehending and extracting the relevant information from the input text, and understanding the contextual relationships in the text through the self-attention mechanism.

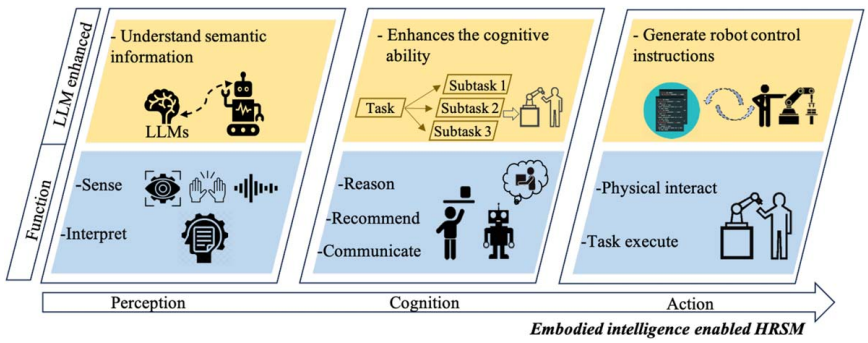

Fig. 3 Embodiments of embodied intelligence-enabled HRSM

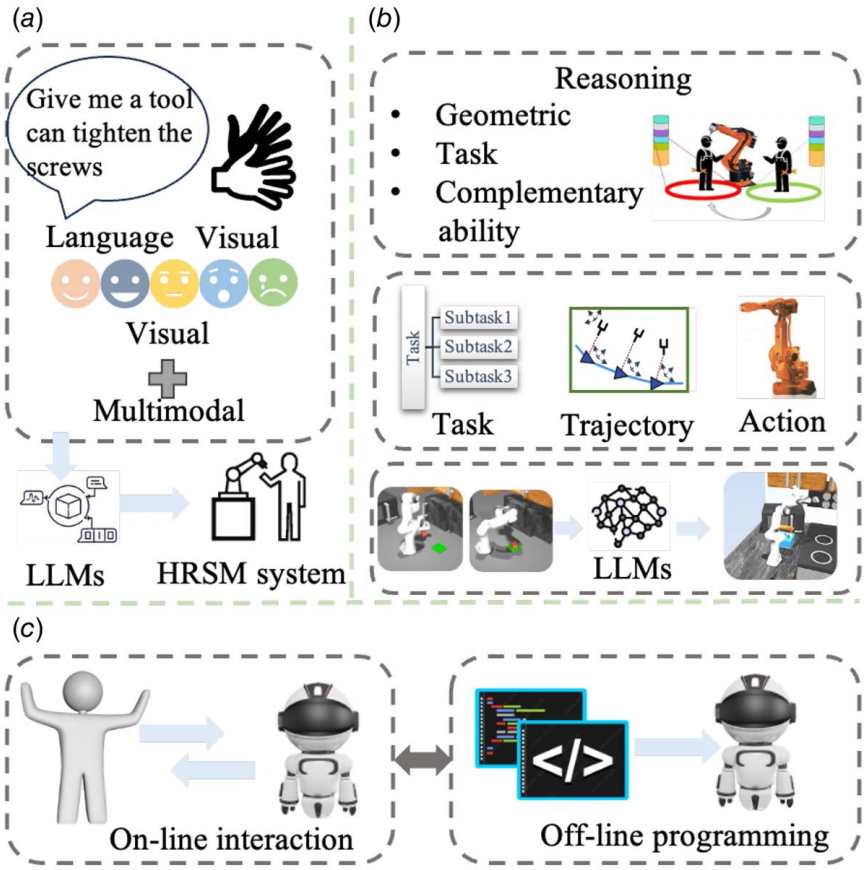

Fig. 4 LLM-based interaction, collaboration, and learning in HRSM: (*a*) interaction, (*b*) collaboration, and (*c*) execution.

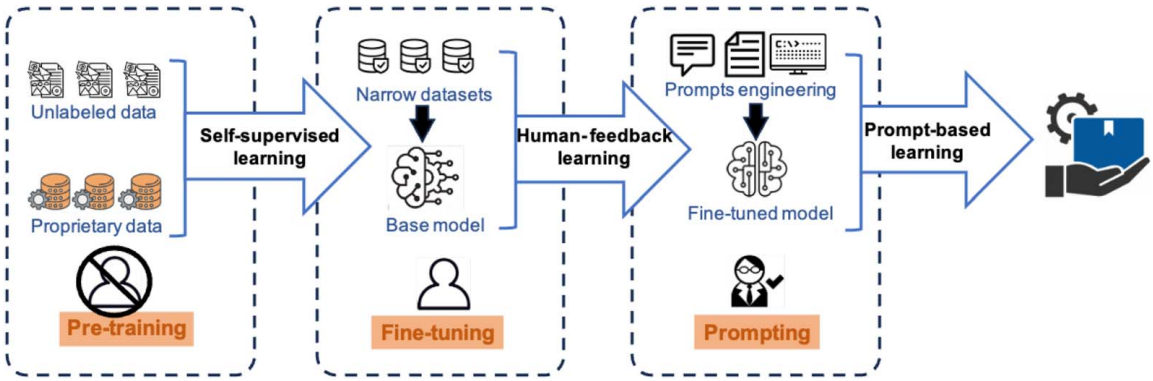

Fig. 5 The training progress of LLMs

The decoder's role is to generate translated text (in the target language) based on the embedding received from the encoder, also using a self-attention mechanism [71].

Thus, structurally, LLMs can be categorized into three categories: encoder only, decoder only, and encoder–decoder. Table 1 provides a summary of LLMs based on these three distinct structures.

- *Encoder Only*: Models solely consist of encoders in their architecture. The feature is encoded from input by the neural network and passed to postprocess. The advantage of this type is a deep understanding of input text. However, it cannot directly generate textual output. Hence, this type of LLMs concentrates on doing task-specific outputs, such as text categorization, as shown in Fig. 6(*a*).
- *Encoder–Decoder*: Models, as the name suggests, incorporate both encoder and decoder components. It encodes the input to feature information and passes it to the decoder. Then the decoder can generate according to the sequence as output. This structure has a good ability to deal with the connection between input sequence and output sequence [72]. It is suitable for translation and text summarizing, as shown in Fig. 6(*b*).
- *Decoder Only*: Models, solely feature decoder components. This type of LLMs uses the encoder to generate a corresponding sequence from the input encoding, focusing on generating or predicting output from a series of inputs [73]. It specializes in generation tasks, such as knowledge question-answering [18], and is the dominant architecture today, as shown in Fig. 6(*c*).

Table 1 Mainstream LLMs

Type	Name	Time	Company	Open/close
Encoder only	BERT	2018	Google	Open
	RoBERTa	2019	Meta	Open
	ALBERT	2019	Meta	Open
	DistillBERT	2019	Hugging Face	Open
	ERINE	2019	Baidu	Open
	ELECTRA	2020	Hugging Face	Open
	DeBERTa	2020	Microsoft	Open
Encoder–decoder	BART	2019	Meta	Open
	T5	2019	Google	Open
	mT5	2020	Google	Open
	T0	2020	BigScience	Open
	ST-MoE	2022	Google	Close
	FlanT5	2022	Google	Open
	TK	2022	Hugging Face	Open
	ChatGLM	2023	Tsinghua University	Open
	FlanUL2	2023	Google	Open
Decoder only	GPT-1	2018	OpenAI	Open
	GPT-2	2019	OpenAI	Open
	GPT-3	2020	OpenAI	Close
	CodeX	2021	OpenAI	Close
	Gopher	2021	DeepMind	Close
	ERNIE3.0	2021	Baidu	Close
	LaMDA	2021	Google	Close
	ChatGPT	2022	OpenAI	Close
	PaLM	2022	Google	Close
	LLaMA	2023	Meta	Open
	Bard	2023	Google	Open
	GPT-4	2023	OpenAI	Close
	Claude	2023	Anthropic	Close

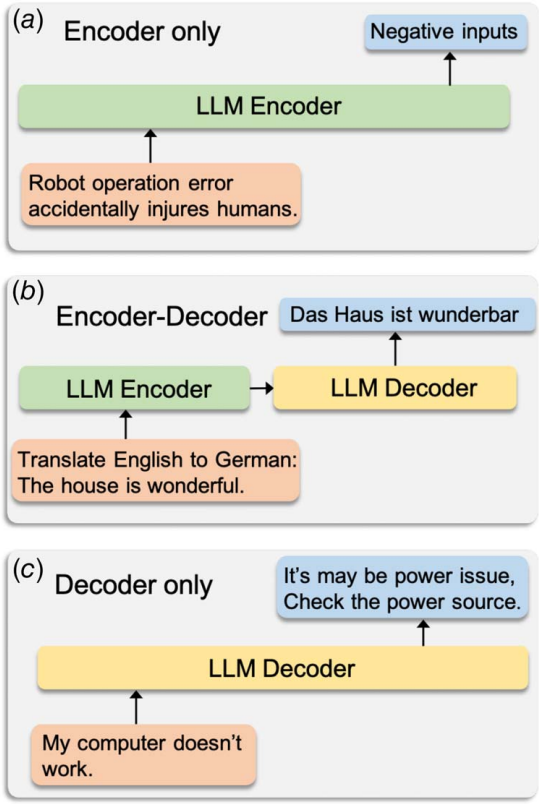

Fig. 6 The different structure of LLMs

3.2 Human-Centric Interaction. Human–robot interaction technology has developed rapidly in the past few decades, enabling more efficient interaction between robots and humans. However, a key challenge in this field is improving the human interaction experience, making humans feel like they are interacting with an intelligent agent rather than a machine, thus fostering a sense of belonging. Recently, human–robot interaction based on LLMs has garnered increasing attention. LLM-based interaction aims to use the powerful reasoning ability of LLM to understand the various forms of information present in the HRSM system. Humans can provide interaction cues in various forms, including language, vision, and multimodal types. The robot then performs the corresponding tasks based on these clues. In this section, we elaborate on the latest progress in latest progress in human-centric interaction based on LLMs, as shown in Table 2.

3.2.1 Language Interaction. Language interaction proves to be one of the most direct approach to information communication. In the future manufacturing industry, LLMs can convert vague cues in human natural language expressions into semantic knowledge that robots can understand. Many studies were devoted to realizing this prospect. For example, in order to simplify autonomous underwater vehicle (AUV) piloting processes by abstracting technical complexities with natural language, Yang et al. [74] employed speech as interaction clues and propose OceanChat, an AUV control system based on GPT-4. GPT-4 translated abstract human instructions into high-level goals, which were further transformed into task sequences with logical constraints by the task planner. Besides, interaction through language not only includes direct instructions but also requires understanding the hidden meanings in the language. Kakodkar and coworkers [76] used GPT-3 to infer the implicit

Table 2 Human-centric interaction

Interaction type	Field	Method	Function	Ref.
Language interaction	Visual language navigation	GPT-4	Translate human instructions into high-level goals	[74]
	Visual language navigation	RL, BLIP, GPT-4	Translate human instructions into RL format; interpret the contextual information of the environment	[75]
	Natural language navigation	GPT-3	Infer the implicit instructions	[76]
Visual interaction	Robot manipulation	ChatGPT	Interpret gestures and language instructions	[77]
	Communication	GPT-3.5	Processing related facial expressions of the conversation	[78]
	Communication	ChatGPT	Generate appropriate responses to marks and gestures	[79]
Multimodal interaction	Robot navigation	GPT-2, CLIP	Parsing visual and language	[80]
	Robot manipulation	GPT-3.5, SAM	Summarize target instructions	[81]
	Robot navigation	CLIP, GPT-2	Parsing visual and language for remote interaction	[82]
	Robot navigation	CLIP, GPT-4	Parsing visual and language information	[83]
	Robot manipulation	GPT-4	Understand the overall scene	[84]
	Communication	GPT-4	Interpret both textual and visual information	[85]

instructions in the user's intuitive instructions and associated the reasoning information with physical points, which were retrieved from the environment. In addition, GPT-4 can also convert language instructions into robot language format, thus helping robots understand human language and perform robot motion tasks and avoid human instructions from excessively dominating the robot's decision-making process [75].

3.2.2 Visual Interaction. In manufacturing, due to noise and pollution, visual interaction is more effective than verbal interaction. Visual modalities data, such as gestures and facial expressions, can be used to transfer information for human–robot agents through visual recognition algorithms [77–79]. Gestures are important for interaction as they contain human intention information. A typical framework, which was based on the gesture recognition module and ChatGPT, was proposed to flexibly interpret gestures and language instructions. LLMs promote it no longer limited to gesture libraries and predefined gesture meanings [77]. In addition, a GPT-based method called Mediapipe can not only understand human gestures but also make robot reply through real-time training, enabling bidirectional gesture communication between robots and humans [79]. Human facial expression is also an essential manner for interaction. Li and Kim [78] proposed a social robot framework based on facial expressions, which used GPT-3.5 to process the context of the conversation, personalized descriptions, and relevant facial expressions to achieve a more natural and intuitive conversation experience.

3.2.3 Multimodal Interaction. Unimodal interaction predefined instructions often falls short in accuracy for limited information. Therefore, it is necessary to achieve natural and effective multimodal interaction between human instructions and robot control systems [34].

Two LLM-related methods pave the way to multimodal interaction. The first one is to use LLM for parsing language inputs and VLM for understanding visual inputs. In this way, agents often need the following types of information: (1) Environment representation: the agent's visual perception of the environment, which may include images, point clouds, or other forms of representation of the environment; (2) natural language instructions: a textual description of the target that the agent should achieve. The task of the agent is to parse the meaning of instructions based on natural language instructions and match the environmental perception content and provide feedback information to complete human–robot interaction. Qiao et al. [82] used CLIP as a scene parser and GPT-2 as an in-context learning model to build an MiC (March in Chat) model for remote interaction. Similarly, Nwankwo and Rueckert [80] introduce ROS2-based CLIP and GPT-2 for avoiding erroneous instructions during LLMs interaction. Wang et al. [81] designed a WALL-E model based on GPT-3.5 and SAM, which could complete the interactive task by multiround visual language interactive dialogue. Schumann et al. [83] proposed VELMA, a model based on CLIP and GPT-4, which uses verbal descriptions of trajectories and visual environment observations as contextual prompts for the next action.

The second one is to use mLLMs, which can directly implement multimodal input and output. Social robot research emphasizes conversational qualities. Abbo and Belpaeme [85] used the functions of GPT-4 to enhance the conversation system. The conversation system can simultaneously obtain environmental image information and natural language information, thereby improving its contextual understanding ability and realizing social companionship function. Hu et al. [84] used GPT-4V for visual observations of the environment and high-level language instructions, without combining LLMs with independent visual modality information processing models, and achieved a deep understanding of common sense knowledge based on the visual world. This is a newly proposed application that has not yet been used in the industrial field, but it has great potential in the direction of industrial end-to-end embodied robotics.

3.3 Generalizable Collaboration. LLMs can analyze task requirements and resource allocation, optimize task distribution and coordination in human–robot collaboration, and promote knowledge sharing between humans and robots, enhancing teamwork.

In the work loop, LLM-based generalizable collaboration aims to align the needs of humans and the responses of industrial robots. With reasoning, planning, and learning capabilities, LLMs can dissect the task by parsing the hidden intention information to provide the robot with clear and understandable planning and cognitive strategies.

3.3.1 Human–Robot Relationship Reasoning. Similar to the reasoning function of the human brain, the HRSM system based on LLMs can reason various knowledge in production processes. In detail, LLMs are introduced to the tasks of reasoning geometric relationships, task relationships, and complementary capabilities relationships in collaborative scenarios. Reasoning types are presented as follows, as illustrated in Table 3:

- *Reasoning About Geometric Relationship*: Geometric relationship reasoning provides maps of the shared workspace of HRSM settings for navigation when handling complex tasks. It contains reasoning about collaborative scenarios and parsing rich scene semantic information. However, the variability and complexity of the geometric information make it difficult to identify and reason about the overall knowledge of the environment. One potential solution is to use LLMs to perform geometric relationship reasoning. Liu et al. [86] proposed an approach to achieve comprehensive environment recognition through the integration of yolo and GPT-2 for motion correction, which can be applied to table-top manipulation tasks. Another key geometric reasoning task is localization, which requires spatial reasoning abilities. Graule and Isler [87] integrated multimodal input sources, including video streams from cameras and data streams from connected devices, to infer human next actions and language, connecting them to a semantic map of the

Table 3 Human–robot relationship reasoning

Reasoning type	Field	Method	Function	Ref.
Geometric relationship	Robot manipulation	Yolo, GPT-2	Realize overall environment recognition	[86]
	Visual language navigation	LLaMA2	Predict actions and map positioning	[87]
	Visual language navigation	GPT-3, BLIP	Observe and understanding of the overall scenario	[88]
Tasks relationship	Robot manipulation	T5	Infer execution steps from instructions and environment	[89]
	Autonomous driving	GPT-3.5	Divide into multiple subproblems	[90]
	Safety constraints	GPT-4	Supervise whether subtasks comply with constraints	[91]
Complementary capabilities relationship	Robot manipulation	GPT-4	Compile human instructions into RAP; interact with humans to eliminate ambiguity	[92]
	Robot manipulation	GPT-3.5-turbo	Analyze human intention	[93]
	Robot manipulation	GPT-4	Respond humans language correction feedback	[94]
	Robot manipulation	LLama	Infer the state; guiding the robot	[95]

Table 4 Human–robot collaboration planning

Planning type	Field	Method	Function	Ref.
Task planning	Motion control	GPT-4	Understand logs and diagnostic information	[96]
	Manipulation	GPT-3+Codex	Generate geometrically feasible plans	[97]
	General task	GPT-3.5	Generate feasible plans	[98]
	Navigation	CLIP	Remote embodied referring expression	[99]
	Manipulation	RL+LLM+CLIP	Guide for high-level strategies	[68]
	Navigation	GPT-3	Generating plans in environment	[100]
	General task	Codex	Generating plans in PDDL block world	[101]
	Manipulation	CLIPort+ViLD+GPT-3	Generating plans in constrained environment	[102]
	General task	ChatGPT-4	Generate task sequence	[103]
	Manipulation	GPT-3.5 GPT-4	Generate initial plan	[104]
	Manipulation	GPT-4V	Divide into subtask and replan	[59]
Trajectory planning	Obstacle avoidance	GPT-3.5-turbo	Few-shot near-optimal path planning	[105]
	Manipulation	BERT+CLIP	Guide the correction of the robot trajectory	[106]
	Navigation	GPT-3.5	Extract semantic map information	[107]
	Manipulation	GPT-4+OWL-ViT+SAM	Inferring affordances and constraints	[108]
	Communication	GPT-4	Generate trajectory from environment feedback	[109]
Action planning	Motion control	GPT-4	Convert natural language instructions into desired contact patterns	[110]
	Motion control	ChatGPT	Generate robot behaviors	[111]
	Articulated object manipulation	GPT-3.5-Turbo/GPT-4	Generate 3D operation path points	[112]
	Robot manipulation	GPT-3 text-davinci003	Judging targets from multimodal perception	[113]
	Manipulation	GPT-3.5	Generate adaptive actions	[114]

environment for geolocation of predicted actions. Besides, LLM could be combined with the image subtitle model BLIP to achieve cross-language navigation by aligning input instructions with action decision sequences [88].

- *Reasoning About the Task Relationship*: An important step in achieving generalization is decomposing complex tasks into subtasks, which requires utilizing LLMs for reasoning about task relationships. For this goal, Zhao et al. [89] introduced an embodied learning architecture named ERRA (Embodied Representation and Reasoning Architecture), using T5 to reason about subtask sequences based on natural language instructions and environment states. Meanwhile, Sha et al. [90] used LLM to decompose the analysis and decision-making process of driving scenarios into multiple subproblems, including identifying the vehicles requiring attention and reasoning the situation and offering action guidance. For safety, Yang et al. [91] employed Lang2LTL to establish safety constraints, supervising the compliance of generated subtasks with the constraints through GPT-4 inference and ensuring adherence to safety requirements.
- *Reasoning About the Complementary Capabilities Relationship*: In HRSM, humans and robots possess distinct advantages. Humans excel in high-level task planning and have the ability to adapt their decisions promptly in response to changing environments or unknown obstacles. On the other hand, robots excel in precise and efficient task execution. By leveraging the complementary strengths of both, the collaboration system as a whole can become more efficient and adaptable. Current researches in LLMs focused on complementary capabilities relationships are primarily in two main areas. First, LLMs can leverage their exceptional natural language processing and reasoning capabilities to bridge the communication gap between humans and robots. For example, Hori et al. [92] utilized GPT-4 to transform high-level cognitive instructions from humans into robot low-level action instructions (RAP) and employ uncertainty analysis to eliminate ambiguities in the instructions. Zhang et al. [93] utilized GPT-3.5 to analyze language states and historical information, enabling an understanding of humans' true intentions and providing robots with accurate decision-making. Zha et al. [94] also utilized GPT-4 to facilitate the conversion between high-level task planning and low-level skill primitives. Second, LLMs can empower robots with human-like thinking patterns, enabling them to actively acquire information from the environment or analyze task failures through logical inference, thereby optimizing task execution. For instance, the DROC (Distillation and Retrieval of Online Corrections) system [94] leveraged GPT-4 to learn from corrections and improve future task execution success rates. Sun et al. [95] used LLama to enable robots to interact with sensors and gather missing information from the environment to complete tasks.

3.3.2 Collaboration Planning. With the assistance of LLMs, human–robot systems can realize collaborative planning of complex tasks for generalizable HRSM. As shown in Table 4, robots undertake three types of planning in collaborative tasks: task planning, trajectory planning, and action planning.

- *Task Planning*: Task planning is a key intelligent function that should be implemented in HRSM systems to form solutions to complex tasks. Unlike reasoning, which involves inferring solutions based on existing clues, task planning is the process of developing a sequence of subtasks to achieve a specific goal. Zhang et al. [98] used a knowledge graph (KG) containing fire prevention, fire warning, and early fire intervention as the contextual input of LLM to enhance the planning ability of LLM based on dynamic fire scenarios. Tagliabue et al. [96] leveraged the prior knowledge possessed by GPT-4 to enhance adaptability in real-world scenarios, allowing for effective task planning with a minimal user input. Lin et al. [97] further explored GPT-3 planning by interweaving it with policy sequence optimization, which not only generates robot plans but also ensures the geometric feasibility of the planned skills. In terms of planning efficiency, Prakash et al. [68] combined LLMs with RL (reinforcement learning) to build a hierarchical agent that uses LLMs to solve complex tasks while using RL to learn from the environment, thereby achieving efficient planning of long-horizon tasks.

A key challenge for embodied intelligence-enabled HRSM is to grounding these plans in the robot operation environments. Song et al. [100] explored the combination of object

detectors and GPT-3 to generate robot plans with environmental information. Silver et al. [101] tried to use LLM for planning domain definition language (PDDL) representation to achieve robot-embodied planning. Brohan et al. [102] provided GPT-3 with real-world constraints through pretrained value functions, enabling planning in a given environment. Mohammadi et al. [99] introduced the ACK framework, which utilizes commonsense information structured as a spatiotemporal KG to enhance agent navigation. Within this framework, the CLIP model was employed to gather and prioritize the most relevant knowledge concerning the scene and identified objects. Besides, the verification of plans generated by LLMs is of significance, especially in manufacturing scenarios. For this purpose, a framework named ISR-LLM (Iterative Self-Refined Large Language Model) was introduced, which employed GPT-3.5 to generate an initial plan, and utilized a validator to evaluate and enhance the plan [104]. Qi et al. [103] combined control instructions with LLM to generate the necessary task sequences and then combined them with task KG for verification before executing the tasks. Meanwhile, Skreta et al. [59] employed GPT-4V to obtain feedback from the environment and replan failed tasks.

- *Trajectory Planning*: Trajectory planning involves determining the optimal path for the robot to follow while performing tasks. LLMs can assist in locating the target geographical location and planning the path during the manufacturing process, without the need for explicit command guidance. Xiao et al. [105] combined GPT-3.5 Turbo and utility-optimal A* for few-shot near-optimal path planning by incorporating basic environment and task information with a small amount of mid-term manual feedback. To get rid of dependence on people, Bucker et al. [106] utilized CLIP and BERT as encoders, with CLIP capturing the semantics of environmental objects and BERT encoding language commands to calculate the similarity between the two embeddings. This information was then used to guide the correction of the robot's trajectory by matching it with the planned path and the position information of environmental objects. Meanwhile, Wu et al. [107] developed a method to infer the most appropriate navigation path using semantic map information, paths, and long-range descriptions extracted by GPT-3.5. Furthermore, Huang et al. [108] proposed voxposer, which is a new framework that possesses embodied intelligence. VoxPoser's planning precision reaches the waypoint level of robotic arm movement trajectories, making manipulation tasks highly flexible. For multirobot planning task, Mandi et al. [109] utilized GPT-4 for communication in multirobot planning, incorporating environment feedback to improve subtask planning and trajectory generation, thus enabling the generation of optimal trajectory planning.
- *Action Planning*: Action planning involves generating sequences of poses and actions for the robot's end effector during task execution. LLMs were able to predict the pose sequence of the robot's end effector by analyzing objects in images and generating task-related semantic reasoning steps based on given language cues to complete specific operating tasks [110]. Meanwhile, Tang et al. [111] used ChatGPT to convert kinematic knowledge and natural language instructions as prompts into the required contact patterns and used RL to generate robot actions based on the output of ChatGPT. Xia et al. [112] designed a unified kinematic knowledge parser to provide articulated object kinematic structure for GPT-4, thereby being able to generate accurate 3D operating points. To improve the adaptation of action planning and environment, Zhao et al. [113] utilized GPT-3 for multimodal perception (vision, sound, touch, proprioception) and interaction with the environment, generating action commands and responding accordingly based on environmental feedback. Hu et al. [114] utilized GPT-3.5 to generate execution actions based on previous observations and environmental information, with the observed results being fed back to GPT-3.5 to generate the subsequent action.

3.3.3 Cognitive Learning. Traditionally, rule-based approaches have been widely used in manufacturing systems, which used detailed rules and codes to control robot activities. However, these methods have limited learning capabilities, especially in unstructured and dynamic environments where preprogrammed actions are insufficient. A potential solution is to allow robots to acquire new skills from human demonstrations. This human-centric skill transfer paradigm offers more benefits than traditional robot programming methods, enabling robots to have the rich skills and intuition that humans possess. Among all the learning-from-demonstration approaches, LLM-based skill transfer has garnered increasing attention. For example, Hori et al. [115] proposed a multimodal imitation learning method that uses demonstration datasets including video, audio, and text, using the audio–visual transformer with style transfer learning to achieve efficient robot learning. Similarly, Shao et al. [116] proposed an imitation learning framework that combines natural language instructions and visual input to enhance the ability of robot transfer learning, providing better generalization, efficient learning, and handling of complex operations. Sontakke et al. [117] presented RoboCLIP, an innovative imitation learning method that uses pretrained VLMs to generate reward functions from a single demonstration (video or text). It reduces the reliance on expert demonstrations and complex reward engineering, leading to superior zero-shot performance and efficient fine-tuning. Murray et al. [118] also introduced a highly modular neural-symbolic framework for synthesizing robotic skills from visual demonstrations and natural language instructions.

Another probable solution is to guide robot learning by leveraging the cognitive capabilities of LLMs. LLMs possess the ability to comprehend commonalities and generalities across multiple tasks, showing remarkable contextual understanding. As a result, robots can realize cognitive learning with LLMs, eliminating the need for predefined intricate rules. Aristeidou et al. [119] used the zero-shot detection capability of LLMs to dynamically update detection objects and complete the predicted localization coordinates. When the environment changes, Yow et al. [120] employed BERT to comprehend semantic instructions and utilized CLIP to generate and describe scene and object features in the environment, thereby achieving the matching of instructions with features and the acquisition of new planning skills of trajectory. In addition, Quartey et al. [121] employed InstructGPT to generate context-aware object embeddings for constructing and learning auxiliary tasks. Meanwhile, a framework, SAGE, was proposed to leverage GPT-4 to guide objects and provide feedback updates based on environmental information and task progression [122]. When complex tasks are detected as impossible to complete, Parakh et al. [123] utilized GPT-4 to proactively initiate requests for the incorporation of new skills. This allowed GPT-4 to dynamically expand its skill library based on task requirements and subsequently reuse these newly acquired skills in subsequent tasks. Ming et al. [124] empowered GPT-3.5 to detect errors and automatically generate error corrections by integrating environmental perception information with hierarchical error states.

3.4 Seamless Execution. The objective of LLM-based seamless execution is to integrate online interaction and offline programming for robots. Although LLM-based HRSM seamless execution is still in its infancy stages, it has shown tremendous potential in manufacturing activities because it can quickly implement innovative product design without affecting the original production plan. Several works proposed to realize this prospect. Ye et al. [125] fine-tuned ChatGPT, and the output results could be accurately integrated with robotic control modules. Liu et al. [60] attempted to use GPT-4 to adjust robot error behavior online, and the robot directly generates code through these language feedbacks to correct errors. Qiu et al. [126] learned to synchronously generate

executable code from visual observations through LLM. Besides, Google conducted a series of works to explore embodied robot models for seamless robotic execution. They integrate visual perception into LLM by training on mixed robotic data and web-scale vision-language data [127] and retraining the LLM to directly output robot action parameters in a language format [128]. This requires a large amount of computing resources because of the full training of novel LLM structures. It is worth noting that in the process of seamless interaction, the quality of the prompt is a key influencing factor. Formatted prompt content is a key to efficiently convert verbal instructions into python code [60,126,129]. In addition, setting guiding questions is also a key method to improve prompt [130]. This means that humans must have a clear understanding and accurate description of the desired function. Furthermore, providing a comprehensive description of the context is an effective approach to enhancing prompts. Qiu et al. [126] utilized LLMs to generate environment explanations from visual observations and textual descriptions of scenes, using them as prompts for inputting LLMs until programmable executable actions.

3.5 Summary. We elaborate on the application of embodied intelligence driven by LLMs in HRSM. In terms of interaction, LLMs enable humans to communicate with robots more naturally and intuitively through linguistic and visual modalities, and assist robots in understanding human emotional states via emotion recognition, thereby facilitating more appropriate responses and enhancing the human interaction experience. In terms of collaboration, LLMs can analyze task requirements and resource allocation, optimize task distribution and coordination in human–robot collaboration, promote knowledge sharing between humans and machines, and enhance teamwork. In terms of execution, LLMs leverage their reasoning and generative capabilities to assist robots in automatically generating task execution codes, thereby reducing the burden on human developers. Through these mechanisms, LLMs significantly enhance the capabilities of human–robot systems in interaction, collaboration, and execution, propelling human–robot symbiotic manufacturing toward greater efficiency and intelligence.

However, current research still exhibits certain limitations. In terms of interaction, existing LLM research predominantly focuses on visual and linguistic modalities, lacking in-depth exploration and application of force feedback modalities, which constrains the performance of robots in complex manufacturing tasks. For instance, in precision assembly tasks, force feedback can aid robots in better sensing and adjusting the applied force to prevent damage to precision components. In terms of generalization, scenario-based generalization is largely predicated on premodeling of scenarios [131]. Although this approach performs well in specific contexts, it often proves inadequate in dynamic and unknown environments. Consequently, further research is required on LLMs integrated with technologies such as real-time updatable navigation and three-dimensional assembly scenes to enhance their generalization capabilities in evolving environments. In terms of seamlessness, a unified framework for online programming in cross-platform languages for different robots has yet to be established. This results in poor interoperability between different robotic systems, limiting their collaborative efficacy in complex tasks. A unified cross-platform programming framework could streamline communication and collaboration among diverse robots, thereby improving the overall efficiency and flexibility of the system.

4 Challenges

As an emerging technology with great potential, there are still many challenges that need to be addressed. This section mainly discusses the challenges that researchers and workers may encounter in the application of LLMs.

4.1 Security and Privacy. Manufacturing typically involves sensitive and proprietary data, including production flows, process parameters, and fault diagnosis. Ensuring data privacy during training and deploying LLMs is critical for manufacturing applications. Furthermore, data security is also a concern, as human operators may require cloud interactions during task execution in industrial production processes, such as online monitoring [132] and remote diagnostics [133], which could potentially lead to data leakage. Therefore, it is imperative to address security and privacy concerns through further investigation for LLM-based HRSM.

4.2 Artificial Intelligence Hallucinations. LLMs use a large amount of multimodal data during pretraining and overfitting [134]. However, all the data are based on existing data or prior knowledge without guarantee of its accuracy. As a result, LLMs may produce inaccurate or misleading outputs. This phenomenon is known as AI hallucinations. Under the influence of AI hallucinations, the generated outputs may lead to serious consequences if used without proper verification. For example, during drilling processes, equipment may receive erroneous outputs from LLMs and harm the operators.

4.3 Evaluation. Although research on robotic applications related to LLMs is flourishing, there is currently no universally accepted evaluation metric due to the diverse range of tasks. At present, different evaluation indicators may be utilized for assessing the performance of LLMs in HRSM depending on the specific task. For instance, in manipulation tasks, the planning success rate and execution success rate are employed to gauge the system's task execution effectiveness [135]. In navigation tasks, the success rate, navigation error, and CLS (Coverage Weighted by Length Score) are utilized [136,137]. Planning tasks encompass the utilization of task success rate, task duration, task diversity, task difficulty, and task dependencies [95,138]. Establishing different evaluation criteria for various tasks proves to be time consuming and challenging when it comes to demonstrating the effectiveness of the research.

4.4 Heterogeneity of Scenarios. In practical manufacturing settings, the manufacturing domain exhibits distinctive characteristics, including diverse data sources, a wide array of equipment, and intricate processes [139]. Within heterogeneous manufacturing environments, the diversity in data, formats, and modalities hinders the achievement of human-centric multimodal interactions. The existence of information silos and communication barriers among different devices and processes restricts the system's capacity for collaborative and comprehensive engagement. Furthermore, the intricate mechanisms underlying the code pose difficulties in seamlessly integrating interaction and programming. In such circumstances, the heterogeneity leads to substantial losses in terms of time and finances, thereby presenting formidable obstacles in the development of HRSM.

4.5 Limited Data. LLMs undergo fine-tuning through the utilization of specialized domain expert datasets, demonstrating the considerable potential for logical reasoning and strategic planning across a multitude of problem domains, such as materials [140]. However, the deployment of these models in manufacturing encounters the challenge of insufficient training data availability. Unlike domains like natural language processing or computer vision, which often benefit from amply annotated datasets, the manufacturing industry typically grapples with the challenge of data privacy. On the one hand, manufacturing procedures are often intricate, involving intricate steps and parameters that are laborious to capture and annotate exhaustively. On the other hand, the confidential and sensitive nature of manufacturing information hampers data sharing and accessibility. The dearth of training data lacks of generalization and prediction capabilities of the models. Consequently, the collection of diverse and substantial datasets for fine-tuning LLMs materializes as a formidable obstacle.

4.6 Ethical Considerations. Although LLMs harbor the potential to bring technological transformation to the manufacturing industry, the ethical considerations of their implementation warrant equal consideration. Given that the data used for training are unfiltered, biases and stereotypes related to race, ethnicity, and gender present may lead to discriminatory outcomes [141]. Although existing research has focused on benchmarks used to assess the ability of existing LLMs to address specific biases [142], eliminating biases and discrimination deserves more attention because they could potentially impact the mental health of workers and even lead to legal violations.

4.7 Contextual Restrictions. A significant challenge in applying LLMs to manufacturing is the constraint of context length [143]. LLMs are restricted by the maximum number of tokens that can be processed in a single input. This limitation presents a substantial obstacle in manufacturing settings, where intricate, multistep processes and extensive documentation are prevalent. Detailed manufacturing protocols, comprehensive maintenance logs, and complex design specifications frequently exceed the token limits of current LLMs. Consequently, the inability to process lengthy sequences of text in one instance results in incomplete understanding and responses, thereby undermining the reliability and effectiveness of LLMs in delivering accurate and contextually relevant insights. Furthermore, manufacturing often necessitates the integration of diverse data sources, including real-time sensor data, historical performance records, and predictive maintenance plans. Context length constraints hinder LLMs' ability to synthesize and correlate information across these varied inputs, potentially leading to suboptimal decisions. One possible solution to this problem is to leverage retrieval-augmented generative models to improve the ability to process long text sequences [143]. Overcoming this challenge is crucial to unlocking the full potential of LLMs in manufacturing, enabling them to more effectively manage the complexity and the scale of industrial applications.

5 Future Opportunity

HRSM represents a future paradigm of smart manufacturing. This transformation aims to optimize the efficacy, productivity, and adaptability of the manufacturing process. In this vision, humans and robots work together to combine their unique strengths to create an efficient and flexible manufacturing ecosystem. As shown in Fig. 7, in the future, enabled by optimizing LLMs, LLMs industrial agents (including industrial robots, unmanned aerial vehicles, and unmanned ground vehicles) will adaptively complete manufacturing tasks in the production workshop in a low-code development. Human operators can use industrial wearable devices to achieve immersive interaction in the industrial virtual world, freeing themselves from tedious and physically demanding production tasks. The various production departments on the overall production line are efficiently integrated, and production resource scheduling is completed automatically. In order to achieve this goal, this section points out four specific research directions.

5.1 Large Language Model-Based Low Code Collaboration. Rapid product iteration is of great significance to future manufacturing. As product iteration requires a lot of time in code programming, the speed of product updates is reduced. Using low-code programming is an effective solution. Low-code programming primarily utilizes the human-intuitive approach to control robots, such as visual, language-based, or natural language descriptions, which are closer to human thinking, rather than traditional textual programming [144]. LLMs possess the capability to generate code based on textual prompts [145], and current research has already focused on leveraging this ability to invoke underlying robot functions [146]. On the one hand, the utilization of low-code empowers both expert and nonexpert users to actively engage in application development, resulting in more efficient, user-oriented manufacturing systems within industrial environments. On the other hand, this approach allows developers to allocate greater attention to the high-level logic and functionalities of the application by streamlining the development process. Currently, there is still a lack of a universal paradigm for different robot systems and tasks, making it time consuming and less user-friendly for developers. Moreover, it is important to ensure the safety and robustness of the generated code, as errors or misunderstandings would result in unintended consequences. This entails the implementation of rigorous testing, verification, and validation procedures.

5.2 Manufacturing Metaverse. It should be noted that the metaverse has huge potential in manufacturing [147]. It cannot only offer users a sense of immersion and presence by integrating the physical and virtual realms [148] but also enable the complete execution of product life-cycles within virtual worlds for the unique spatiotemporal characteristics [149]. As a result, every stakeholder in the manufacturing industry could interact and

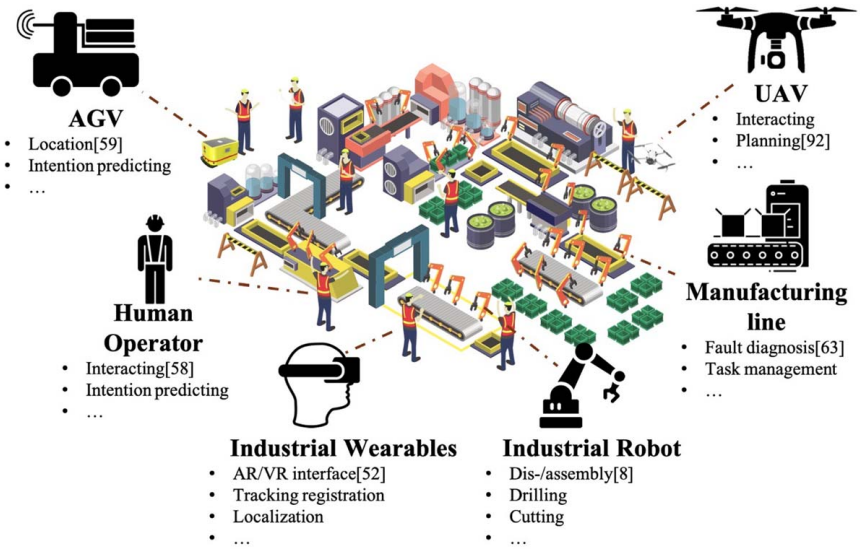

Fig. 7 A grand vision for futuristic HRSM

collaborate in the metaverse. However, establishing a highly robust and personalized metaverse proves to be a complex challenge, necessitating the incorporation of content generation platforms and powerful rendering engines for providing an immersive, interactive virtual environment [150]. For the former, utilizing the generative capabilities of LLMs to understand user needs and generate design solutions has become an important way for metaverse content generation. As for the latter, notable progress in metaverse research has been observed among several renowned companies. For instance, NVIDIA's Omniverse virtual platform [151] encompasses a highly realistic physics simulation engine and high-performance rendering capabilities. Additionally, Apple's Vision Pro products [152] show diverse natural interactions based on XR. These advancements serve as a testament to the promising future research direction of the manufacturing metaverse.

5.3 Enhancing Vertical Integration Through Technology Reorganization. Smart manufacturing aims at achieving the intelligence of the shop-floor, the enterprise, and even the whole supply chain [41]. Current research has already employed LLMs as potent tools in diverse subdomains of manufacturing, encompassing detection, analysis, prediction, and decision-making for their capability to schedule and reason completed tasks. However, in the broader vertical domain, tasks at different stages possess distinct boundaries, resulting in technological isolation, let alone the overall interconnection system based on transformable technologies. Hence, stakeholders at different stages of the manufacturing system should adopt overall system management to achieve the reintegration of vertical fields under the technological changes led by LLMs.

5.4 Optimizing Large Language Models for Manufacturing Deployment. Although LLMs have great application potential in the manufacturing field, their high requirements for computing power and storage capacity make it difficult to directly deploy LLMs on IoT devices and embedded systems. To address this issue, researchers have primarily optimized software and training methodologies to enable real-time and low-power applications of LLMs in manufacturing environments. Regarding software, researchers employ techniques such as quantization and pruning to compress the model [153]. For training, they utilize distributed training frameworks to enhance training efficiency [154]. Currently, research on MIC-LLMs (Machine Learning Compiler) [155] focuses on compiler acceleration and runtime optimization for cross-platform local deployment, successfully enabling the deployment of LLMs on mobile devices such as smartphones.

6 Discussion and Conclusion

HRSM is envisioned as a potential future manufacturing paradigm emphasizing human-centricity, generalization, and seamlessness. However, current human–robot system development is observed to have limitations in terms of human centricity, generalizability, and seamless integration. The emerging technology of LLMs has garnered attention for its advanced reasoning capabilities and extensive knowledge base. Some initial research suggests that LLMs show promising effectiveness in interaction, collaboration, and execution, indicating a potential pathway towards realizing HRSM.

In this article, we offer a comprehensive overview of the primary objectives for achieving HRSM and explore how LLMs could potentially support these objectives. Initially, we outline the key characteristics of the objectives for HRSM and provide a definition of embodied intelligence. Subsequently, we introduce the concept of LLMs and examine their potential contributions to HRSM in terms of interaction, collaboration, and execution. Finally, we discuss the challenges and potential future directions for LLMs within the domain of HRSM. It is important to acknowledge that this article's discussion is limited to the engineering field and does not cover areas such as human factors engineering and psychology. This limitation may restrict our understanding and exploration of LLMs in broader applications.

It is hoped that this article can shed light on the future research and development of embodied intelligence-enabled HRSM with more in-depth discussions and discoveries.

Acknowledgment

The authors would like to express their sincere thanks to the financial support from the Research Institute for Advanced Manufacturing (RIAM) of The Hong Kong Polytechnic University (project code: 1-CD9V). This work was mainly supported by the funding support from the RIAM (project code: 1-CDJT), and the Intra-Faculty Interdisciplinary Project (1-WZ4N) of The Hong Kong Polytechnic University. This work was approved by the Human Subjects Ethics Sub-committee (HSESC) of The Hong Kong Polytechnic University on Nov 20th, 2022 for 5 years (Reference No. HSEARS20210927012). This work was partially supported by the General Research Fund (GRF) from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. PolyU 15210222 and PolyU15206723), Collaborative Research Fund (CRF) from the Research Grants Council (Hong Kong) (Grant No. C6044-23GF), State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology (No.: IMETKF2024010), COMAC International Collaborative Research Project (COMAC-SFGS-2023-3148), and the PolyU-Rhein Köster Joint Laboratory on Smart Manufacturing (H-ZG6L).

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

No data, models, or code were generated or used for this paper.

Nomenclature

AI = artificial intelligence
AGV = automated guided vehicle
AUV = autonomous underwater vehicle
CLS = coverage weighted by length score
GPT = generative pretrained transformer
HRC = human–robot collaboration
HRSM = human–robot symbiotic manufacturing
IoT = Internet of things
KG = knowledge graph
LLM = large language model
mLLM = multimodal large language model
MLC-LLMs = machine learning compiler for large language models
PDDL = planning domain definition language
Q&A = question and answer
RAP = robot action planning
RL = reinforcement learning
ROS = robot operating system
VLM = vision language model
VR = virtual reality
XR = extended reality

References

[1] Zheng, P., Li, S., Fan, J., Li, C., and Wang, L., 2023, "A Collaborative Intelligence-Based Approach for Handling Human-Robot Collaboration Uncertainties," CIRP Ann., **72**(1), pp. 1–4.

[2] Wang, L., Gao, R., Váncza, J., Krüger, J., Wang, X. V., Makris, S., and Chryssolouris, G., 2019, "Symbiotic Human-Robot Collaborative Assembly," CIRP Ann., **68**(2), pp. 701–726.
[3] Nuzzi, C., Pasinetti, S., Pagani, R., Ghidini, S., Beschi, M., Coffetti, G., and Sansoni, G., 2021, "Meguru: A Gesture-Based Robot Program Builder for Meta-Collaborative Workstations," Rob. Comput. Integr. Manuf., **68**, p. 102085.
[4] Bingol, M. C., and Aydogmus, O., 2020, "Performing Predefined Tasks Using the Human–Robot Interaction on Speech Recognition for an Industrial Robot," Eng. Appl. Artif. Intell., **95**, p. 103903.
[5] Buerkle, A., Bamber, T., Lohse, N., and Ferreira, P., 2021, "Feasibility of Detecting Potential Emergencies in Symbiotic Human-Robot Collaboration With a Mobile EEG," Rob. Comput. Integr. Manuf., **72**, p. 102179.
[6] Li, S., Wang, R., Zheng, P., and Wang, L., 2021, "Towards Proactive Human–Robot Collaboration: A Foreseeable Cognitive Manufacturing Paradigm," J. Manuf. Syst., **60**, pp. 547–552.
[7] Wang, L., Liu, S., Liu, H., and Wang, X. V., 2020, "Overview of Human-Robot Collaboration in Manufacturing," 5th International Conference on the Industry 4.0 Model for Advanced Manufacturing, Belgrade, Serbia, June 1–4, Springer, pp. 15–58.
[8] Zheng, P., Li, C., Fan, J., and Wang, L., 2024, "A Vision-Language-Guided and Deep Reinforcement Learning-Enabled Approach for Unstructured Human-Robot Collaborative Manufacturing Task Fulfilment," CIRP Ann, **73**(1), pp. 341–344.
[9] Lu, Y., Zheng, H., Chand, S., Xia, W., Liu, Z., Xu, X., Wang, L., Qin, Z., and Bao, J., 2022, "Outlook on Human-Centric Manufacturing Towards Industry 5.0," J. Manuf. Syst., **62**, pp. 612–627.
[10] Wang, L., 2019, "From Intelligence Science to Intelligent Manufacturing," Engineering, **5**(4), pp. 615–618.
[11] Angleraud, A., Ekrekli, A., Samarawickrama, K., Sharma, G., and Pieters, R., 2024, "Sensor-Based Human–Robot Collaboration for Industrial Tasks," Rob. Comput. Integr. Manuf., **86**, p. 102663.
[12] Gupta, A., Savarese, S., Ganguli, S., and Fei-Fei, L., 2021, "Embodied Intelligence via Learning and Evolution," Nat. Commun., **12**(1), p. 5721.
[13] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., et al., 2023, "A Survey of Large Language Models," arXiv Preprint arXiv:2303.18223.
[14] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J., 2024, "Large Language Models: A Survey," arXiv Preprint arXiv:2402.06196.
[15] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., and Mian, A., 2023, "A Comprehensive Overview of Large Language Models," arXiv Preprint arXiv:2307.06435.
[16] Zeng, F., Gan, W., Wang, Y., Liu, N., and Yu, P. S., 2023, "Large Language Models for Robotics: A Survey," arXiv Preprint arXiv:2311.07226.
[17] Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J., and Daneshjou, R., 2023, "Large Language Models in Medicine: The Potentials and Pitfalls," arXiv Preprint arXiv:2309.00087.
[18] Jiang, Z., Tan, X., Sun, L., and Zhang, L., 2025, "A Conceptual Design Method Based on Concept–Knowledge Theory and Large Language Models," J. Comput. Inf. Sci. Eng., **25**(2), p. 021001.
[19] Agarwal, V., Jablokow, K., and McComb, C., 2025, "Putting the Ghost in the Machine: Emulating Cognitive Style in Large Language Models," J. Comput. Inf. Sci. Eng., **25**(2), p. 021002.
[20] Naghavi Khanghah, K., Wang, Z., and Xu, H., 2025, "Reconstruction and Generation of Porous Metamaterial Units Via Variational Graph Autoencoder and Large Language Model," ASME J. Comput. Inf. Sci. Eng., **25**(2), p. 021003.
[21] Grandi, D., Jain, Y. P., Groom, A., Cramer, B., and McComb, C., 2025, "Evaluating Large Language Models for Material Selection," ASME J. Comput. Inf. Sci. Eng., **25**(2), p. 021004.
[22] Zhang, C., Chen, J., Li, J., Peng, Y., and Mao, Z., 2023, "Large Language Models for Human-Robot Interaction: A Review," Biomimetic Intell. Rob., **3**(4), p. 100131.
[23] Ivanov, D., 2023, "The Industry 5.0 Framework: Viability-Based Integration of the Resilience, Sustainability, and Human-Centricity Perspectives," Int. J. Prod. Res., **61**(5), pp. 1683–1695.
[24] Pacaux-Lemoine, M.-P., Trentesaux, D., Rey, G. Z., and Millot, P., 2017, "Designing Intelligent Manufacturing Systems Through Human-Machine Cooperation Principles: A Human-Centered Approach," Comput. Ind. Eng., **111**, pp. 581–595.
[25] Malik, A. A., and Brem, A., 2021, "Digital Twins for Collaborative Robots: A Case Study in Human-Robot Interaction," Rob. Comput. Integr. Manuf., **68**, p. 102092.
[26] Bilberg, A., and Malik, A. A., 2019, "Digital Twin Driven Human–Robot Collaborative Assembly," CIRP Ann., **68**(1), pp. 499–502.
[27] Katuk, N., Vergallo, R., and Sugiharto, T., 2024, *The Future of Human-Computer Integration: Industry 5.0 Technology, Tools, and Algorithms*, CRC Press, Boca Raton, FL.
[28] Gkournelos, C., Konstantinou, C., Angelakis, P., Michalos, G., and Makris, S., 2023, "Enabling Seamless Human-Robot Collaboration in Manufacturing Using LLMS," European Symposium on Artificial Intelligence in Manufacturing, Springer, pp. 81–89.
[29] Ghobakhloo, M., Iranmanesh, M., Foroughi, B., Tirkolaee, E. B., Asadi, S., and Amran, A., 2023, "Industry 5.0 Implications for Inclusive Sustainable Manufacturing: An Evidence-Knowledge-Based Strategic Roadmap," J. Cleaner Prod., **417**, p. 138023.
[30] Pizoń, J., and Gola, A., 2023, "Human-Machine Relationship-Perspective and Future Roadmap for Industry 5.0 Solutions," Machines, **11**(2), p. 203.
[31] McLeod, S., 2007, "Maslow's Hierarchy of Needs," Simply Psychol., **1**(1–18).
[32] Maslow, A. H., 1958, *A Dynamic Theory of Human Motivation*, Howard Allen Publishers.
[33] Turja, T., Särkikoski, T., Koistinen, P., and Melin, H., 2022, "Basic Human Needs and Robotization: How to Make Deployment of Robots Worthwhile for Everyone?Technol. Soc., **68**, p. 101917.
[34] Wang, T., Zheng, P., Li, S., and Wang, L., 2024, "Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing: A Survey," Adv. Intell. Syst., **6**(3), p. 2300359.
[35] Li, C., Zheng, P., Yin, Y., Pang, Y. M., and Huo, S., 2023, "An AR-Assisted Deep Reinforcement Learning-Based Approach Towards Mutual-Cognitive Safe Human-Robot Interaction," Rob. Comput. Integr. Manuf., **80**, p. 102471.
[36] Zhang, X., Fan, J., Peng, T., Zheng, P., Lee, C. K., and Tang, R., 2022, "A Privacy-Preserving and Unobtrusive Sitting Posture Recognition System via Pressure Array Sensor and Infrared Array Sensor for Office Workers," Adv. Eng. Inf., **53**, p. 101690.
[37] Yin, Y., Zheng, P., Li, C., Cong, J., and Pang, Y. M., 2023, "An Empirical Study of an MR-Enhanced Kinematic Prototyping Approach for Articulated Products," Adv. Eng. Inf., **58**, p. 102203.
[38] Lu, Y., Adrados, J. S., Chand, S. S., and Wang, L., 2021, "Humans Are Not Machines—Anthropocentric Human–Machine Symbiosis for Ultra-Flexible Smart Manufacturing," Engineering, **7**(6), pp. 734–737.
[39] Bauer, M., Lecrubier, Y., and Suppes, T., 2008, "Awareness of Metabolic Concerns in Patients With Bipolar Disorder: A Survey of European Psychiatrists," Eur. Psychiatry, **23**(3), pp. 169–177.
[40] Christo, C., and Cardeira, C., 2007, "Trends in Intelligent Manufacturing Systems," 2007 IEEE International Symposium on Industrial Electronics, Vigo, Spain, June 4–7, IEEE, pp. 3209–3214.
[41] Zhang, M., Tao, F., Zuo, Y., Xiang, F., Wang, L., and Nee, A., 2023, "Top Ten Intelligent Algorithms Towards Smart Manufacturing," J. Manuf. Syst., **71**, pp. 158–171.
[42] Vatankhah Barenji, A., Liu, X., Guo, H., and Li, Z., 2021, "A Digital Twin-Driven Approach Towards Smart Manufacturing: Reduced Energy Consumption for a Robotic Cell," Int. J. Comput. Integr. Manuf., **34**(7–8), pp. 844–859.
[43] Fan, J., Zheng, P., and Lee, C. K., 2023, "A Vision-Based Human Digital Twin Modeling Approach for Adaptive Human–Robot Collaboration," ASME J. Manuf. Sci. Eng., **145**(12), p. 121002.
[44] Gams, A., Nemec, B., Ijspeert, A. J., and Ude, A., 2014, "Coupling Movement Primitives: Interaction With the Environment and Bimanual Tasks," IEEE Trans. Rob., **30**(4), pp. 816–830.
[45] Chen, Q., Heydari, B., and Moghaddam, M., 2021, "Leveraging Task Modularity in Reinforcement Learning for Adaptable Industry 4.0 Automation," ASME J. Mech. Des., **143**(7), p. 071701.
[46] Ammar, H. B., Eaton, E., Ruvolo, P., and Taylor, M., 2014, "Online Multi-Task Learning for Policy Gradient Methods," International Conference on Machine Learning, PMLR, pp. 1206–1214.
[47] Yang, C., Zhu, Y., and Chen, Y., 2021, "A Review of Human–Machine Cooperation in the Robotics Domain," IEEE Trans. Hum.-Mach. Syst., **52**(1), pp. 12–25.
[48] Ravichandar, H., Polydoros, A. S., Chernova, S., and Billard, A., 2020, "Recent Advances in Robot Learning From Demonstration," Annu. Rev. Control Rob. Auton. Syst., **3**(1), pp. 297–330.
[49] Pedersen, M. R., Nalpantidis, L., Andersen, R. S., Schou, C., Bøgh, S., Krüger, V., and Madsen, O., 2016, "Robot Skills for Manufacturing: From Concept to Industrial Deployment," Rob. Comput. Integr. Manuf., **37**, pp. 282–291.
[50] Senft, E., Hagenow, M., Radwin, R., Zinn, M., Gleicher, M., and Mutlu, B., 2021, "Situated Live Programming for Human-Robot Collaboration," The 34th Annual ACM Symposium on User Interface Software and Technology, pp. 613–625.
[51] Ikeda, B., and Szafir, D., 2024, "Programar: Augmented Reality End-User Robot Programming," ACM Trans. Hum.-Rob. Interact, **13**(1), pp. 1–20.
[52] Howard, D., Eiben, A. E., Kennedy, D. F., Mouret, J.-B., Valencia, P., and Winkler, D., 2019, "Evolving Embodied Intelligence From Materials to Machines," Nat. Mach. Intell., **1**(1), pp. 12–19.
[53] Fan, H., Liu, X., Fuh, J. Y. H., Lu, W. F., and Li, B., 2024, "Embodied Intelligence in Manufacturing: Leveraging Large Language Models for Autonomous Industrial Robotics," J. Intell. Manuf., **36**, pp. 1–17.
[54] Koditschek, D. E., 2021, "What Is Robotics? Why Do We Need It and How Can We Get It?, Annu. Rev. Control Rob. Auton. Syst., **4**(1), pp. 1–33.
[55] Wang, B., Zheng, P., Yin, Y., Shih, A., and Wang, L., 2022, "Toward Human-Centric Smart Manufacturing: A Human-Cyber-Physical Systems (HCPS) Perspective," J. Manuf. Syst., **63**, pp. 471–490.
[56] Yin, Y., Zheng, P., Li, C., and Wan, K., 2024, "Enhancing Human-Guided Robotic Assembly: AR-Assisted DT for Skill-Based and Low-Code Programming," J. Manuf. Syst., **74**, pp. 676–689.
[57] Wang, F., Zhou, X., Wang, J., Zhang, X., He, Z., and Song, B., 2020, "Joining Force of Human Muscular Task Planning With Robot Robust and Delicate Manipulation for Programming by Demonstration," IEEE/ASME Trans. Mechatron., **25**(5), pp. 2574–2584.
[58] You, Y., Shen, B., Deng, C., Geng, H., Wang, H., and Guibas, L., 2023, "Make a Donut: Language-Guided Hierarchical EMD-Space Planning for Zero-Shot Deformable Object Manipulation," arXiv Preprint arXiv:2311.02787.
[59] Skreta, M., Zhou, Z., Yuan, J. L., Darvish, K., Aspuru-Guzik, A., and Garg, A., 2024, "Replan: Robotic Replanning With Perception and Language Models," arXiv Preprint arXiv:2401.04157.
[60] Liu, H., Chen, A., Zhu, Y., Swaminathan, A., Kolobov, A., and Cheng, C.-A., 2023, "Interactive Robot Learning From Verbal Correction," 2nd Workshop on Language and Robot Learning, Atlanta, GA, Sept. 6.

[61] Ogmen, H., Shibata, K., and Yazdanbakhsh, A., 2020, "Perception, Cognition, and Action in Hyperspaces: Implications on Brain Plasticity, Learning, and Cognition," Front. Psychol., **10**, p. 485654.
[62] Fan, J., and Zheng, P., 2024, "A Vision-Language-Guided Robotic Action Planning Approach for Ambiguity Mitigation in Human–Robot Collaborative Manufacturing," J. Manuf. Syst., **74**, pp. 1009–1018.
[63] Wang, T., Fan, J., and Zheng, P., 2024, "An LLM-Based Vision and Language Cobot Navigation Approach for Human-Centric Smart Manufacturing," J. Manuf. Syst, **75**, pp. 299–305.
[64] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., and Roberts, A., 2021, "Extracting Training Data From Large Language Models," 30th USENIX Security Symposium, Online, Aug. 11–13, pp. 2633–2650.
[65] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., et al., 2021, "On the Opportunities and Risks of Foundation Models," arXiv Preprint arXiv:2108.07258.
[66] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., et al., 2023, "LLaMA: Open and Efficient Foundation Language Models," arXiv Preprint arXiv:2302.13971.
[67] Peifeng, L., Qian, L., Zhao, X., and Tao, B., 2024, "Joint Knowledge Graph and Large Language Model for Fault Diagnosis and Its Application in Aviation Assembly," IEEE Trans. Ind. Inf, **20**(6), pp. 8160–8169.
[68] Prakash, B., Oates, T., and Mohsenin, T., 2023, "LLM Augmented Hierarchical Agents," NeurIPS 2023 Foundation Models for Decision Making Workshop, New Orleans, LA, Dec. 15.
[69] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., et al., 2023, "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage," Authorea Preprints.
[70] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., 2017, "Attention Is All You Need," 31st Conference on Neural Information Processing Systems, Long Beach, CA, Dec. 4.
[71] Fu, Z., Lam, W., Yu, Q., So, A. M.-C., Hu, S., Liu, Z., and Collier, N., 2023, "Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder," arXiv Preprint arXiv:2304.04052.
[72] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M., 2019, "fairseq: A Fast, Extensible Toolkit for Sequence Modeling," NAACL-HLT 2019, Minneapolis, MN, June 2.
[73] Wang, T., Roberts, A., Hesslow, D., Le Scao, T., Chung, H. W., Beltagy, I., Launay, J., and Raffel, C., 2022, "What Language Model Architecture and Pretraining Objective Works Best for Zero-Shot Generalization?" International Conference on Machine Learning, Baltimore, MD, July 17, PMLR, pp. 22964–22984.
[74] Yang, R., Hou, M., Wang, J., and Zhang, F., 2023, "OceanChat: Piloting Autonomous Underwater Vehicles in Natural Language" arXiv preprint arXiv:2309.16052.
[75] Shek, C. L., Wu, X., Manocha, D., Tokekar, P., and Bedi, A. S., 2023, "Lancar: Leveraging Language for Context-Aware Robot Locomotion in Unstructured Environments," arXiv preprint arXiv:2310.00481.
[76] Rivkin, D., Kakodkar, N., Hogan, F., Baghi, B. H., and Dudek, G., 2024, "Cartier: Cartographic Language Reasoning Targeted at Instruction Execution for Robots," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, May 13.
[77] Lin, L.-H., Cui, Y., Hao, Y., Xia, F., and Sadigh, D., 2023, "Gesture-Informed Robot Assistance Via Foundation Models," 7th Annual Conference on Robot Learning, Atlanta, GA, Nov. 6, IEEE, pp. 5615–5621.
[78] Li, C., and Kim, T., "Social Robot for the Depressed and Lonely".
[79] Lim, J., Sa, I., MacDonald, B., and Ahn, H. S., 2023, "A Sign Language Recognition System With Pepper, Lightweight-Transformer, and LLM", arXiv Preprint arXiv:2309.16898.
[80] Nwankwo, L., and Rueckert, E., 2024, "The Conversation Is the Command: Interacting With Real-World Autonomous Robots Through Natural Language," ACM/IEEE International Conference on Human-Robot Interaction, Edinburgh, Scotland UK, Mar. 11, pp. 808–812.
[81] Wang, T., Li, Y., Lin, H., Xue, X., and Fu, Y., 2023, "Wall-e: Embodied Robotic Waiter Load Lifting With Large Language Model," arXiv preprint arXiv:2308.15962.
[82] Qiao, Y., Qi, Y., Yu, Z., Liu, J., and Wu, Q., 2023, "March in Chat: Interactive Prompting for Remote Embodied Referring Expression," The IEEE/CVF International Conference on Computer Vision, Paris, France, Oct. 2, pp. 15758–15767.
[83] Schumann, R., Zhu, W., Feng, W., Fu, T.-J., Riezler, S., and Wang, W. Y., 2024, "Velma: Verbalization Embodiment of LLM Agents for Vision and Language Navigation in Street View," The 38th Annual AAAI Conference on Artificial Intelligence., Vancouver, Canada, Feb. 20, Vol. 38, pp. 18924–18933.
[84] Hu, Y., Lin, F., Zhang, T., Yi, L., and Gao, Y., 2024, "Look Before You Leap: Unveiling the Power of Gpt-4v in Robotic Vision-Language Planning," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, May 13.
[85] Abbo, G. A., and Belpaeme, T., 2023, "I Was Blind But Now I See: Implementing Vision-Enabled Dialogue in Social Robots," arXiv Preprint arXiv:2311.08957.
[86] Liu, H., Zhu, Y., Kato, K., Kondo, I., Aoyama, T., and Hasegawa, Y., 2023, "LLM-Based Human-Robot Collaboration Framework for Manipulation Tasks," arXiv Preprint arXiv:2308.14972.
[87] Graule, M. A., and Isler, V., 2024, "GG-LLM: Geometrically Grounding Large Language Models for Zero-Shot Human Activity Forecasting in Human-Aware Task Planning," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, May 13, IEEE, pp. 568–574.
[88] Zhou, K., 2023, "Accessible Instruction-Following Agent," arXiv Preprint arXiv:2305.06358.
[89] Zhao, C., Yuan, S., Jiang, C., Cai, J., Yu, H., Wang, M. Y., and Chen, Q., 2023, "Erra: An Embodied Representation and Reasoning Architecture for Long-Horizon Language-Conditioned Manipulation Tasks," IEEE Rob. Autom. Lett, **8**(6).
[90] Sha, H., Mu, Y., Jiang, Y., Chen, L., Xu, C., Luo, P., Li, S. E., Tomizuka, M., Zhan, W., and Ding, M., 2023, "LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving," arXiv Preprint arXiv:2310.03026.
[91] Yang, Z., Raman, S. S., Shah, A., and Tellex, S., 2024, "Plug in the Safety Chip: Enforcing Constraints for LLM-Driven Robot Agents," 2024 IEEE International Conference on Robotics and Automation (ICRA) , Yokohama, Japan, May 13, pp. 14435–14442.
[92] Hori, K., Suzuki, K., and Ogata, T., 2024, "Interactively Robot Action Planning With Uncertainty Analysis and Active Questioning by Large Language Model," The 2024 16th IEEE/SICE International Symposium on System Integration, Ha Long, Vietnam, Jan. 8, IEEE, pp. 85–91.
[93] Zhang, C., Yang, K., Hu, S., Wang, Z., Li, G., Sun, Y., Zhang, C., et al., 2023, "Proagent: Building Proactive Cooperative AI With Large Language Models," CoRR.
[94] Zha, L., Cui, Y., Lin, L.-H., Kwon, M., Arenas, M. G., Zeng, A., Xia, F., and Sadigh, D., 2024, "Distilling and Retrieving Generalizable Knowledge for Robot Manipulation Via Language Corrections," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, IEEE, pp. 15172–15179.
[95] Sun, L., Jha, D. K., Hori, C., Jain, S., Corcodel, R., Zhu, X., Tomizuka, M., and Romeres, D., 2024, "Interactive Planning Using Large Language Models for Partially Observable Robotic Tasks," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, IEEE, pp. 14054–14061.
[96] Tagliabue, A., Kondo, K., Zhao, T., Peterson, M., Tewari, C. T., and HOW, J. P., 2023, "Real: Resilience and Adaptation Using Large Language Models on Autonomous Aerial Robots," 2nd Workshop on Language and Robot Learning: Language as Grounding.
[97] Lin, K., Agia, C., Migimatsu, T., Pavone, M., and Bohg, J., 2023, "Text2motion: From Natural Language Instructions to Feasible Plans," Auton. Rob., **47**(8), pp. 1345–1365.
[98] Zhang, J., Cai, S., Jiang, Z., Xiao, J., and Ming, Z., 2024, "Firerobbrain: Planning for a Firefighting Robot Using Knowledge Graph and Large Language Model," 2024 10th IEEE International Conference on Intelligent Data and Security (IDS), IEEE, pp. 37–41.
[99] Mohammadi, B., Hong, Y., Qi, Y., Wu, Q., Pan, S., and Shi, J. Q., 2024, "Augmented Commonsense Knowledge for Remote Object Grounding," Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, pp. 4269–4277.
[100] Song, C. H., Wu, J., Washington, C., Sadler, B. M., Chao, W. -L., and Su, Y., 2023, "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents With Large Language Models," International Conference on Computer Vision, Paris, France, pp. 2998–3009.
[101] Silver, T., Hariprasad, V., Shuttleworth, R. S., Kumar, N., Lozano-Pérez, T., and Kaelbling, L. P., 2022, "PDDL Planning With Pretrained Large Language Models," NeurIPS 2022 Foundation Models for Decision Making Workshop.
[102] Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., and Ibarz, J., 2023, "Do as I Can, Not as I Say: Grounding Language in Robotic Affordances," Conference on Robot Learning, Atlanta, GA, PMLR, pp. 287–318.
[103] Qi, Y., Kyebambo, G., Xie, S., Shen, W., Wang, S., Xie, B., He, B., Wang, Z., and Jiang, S., 2024, "Safety Control of Service Robots With LLMS and Embodied Knowledge Graphs," arXiv Preprint arXiv:2405.17846.
[104] Zhou, Z., Song, J., Yao, K., Shu, Z., and Ma, L., 2024, "ISR-LLM: Iterative Self-Refined Large Language Model for Long-Horizon Sequential Task Planning," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, May 13, IEEE, pp. 2081–2088.
[105] Xiao, H., and Wang, P., 2023, "LLM A*: Human in the Loop Large Language Models enabled A* Search for Robotics," arXiv Preprint arXiv:2312.01797.
[106] Bucker, A., Figueredo, L., Haddadinl, S., Kapoor, A., Ma, S., and Bonatti, R., 2022, "Reshaping Robot Trajectories Using Natural Language Commands: A Study of Multi-Modal Data Alignment Using Transformers," 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 978–984.
[107] Wu, P., Mu, Y., Wu, B., Hou, Y., Ma, J., Zhang, S., and Liu, C., 2024, "Voronav: Voronoi-based Zero-Shot Object Navigation With Large Language Model," Forty-First International Conference on Machine Learning.
[108] Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., and Fei-Fei, L., 2023, "Voxposer: Composable 3D Value Maps for Robotic Manipulation With Language Models," Proc. Mach. Learn. Res., **229**.
[109] Mandi, Z., Jain, S., and Song, S., 2024, "Roco: Dialectic Multi-Robot Collaboration With Large Language Models," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, May 13, IEEE, pp. 286–299.
[110] Kwon, T., Di Palo, N., and Johns, E., 2023, "Language Models as Zero-Shot Trajectory Generators," 2nd Workshop on Language and Robot Learning: Language as Grounding.
[111] Tang, Y., Yu, W., Tan, J., Zen, H., Faust, A., and Harada, T., 2023, "Saytap: Language to Quadrupedal Locomotion," Conference on Robot Learning, PMLR, pp. 3556–3570.
[112] Xia, W., Wang, D., Pang, X., Wang, Z., Zhao, B., Hu, D., and Li, X., 2024, "Kinematic-Aware Prompting for Generalizable Articulated Object

Manipulation With LLMS," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, IEEE, pp. 2073–2080.
[113] Zhao, X., Li, M., Weber, C., Hafez, M. B., and Wermter, S., 2023, "Chat With the Environment: Interactive Multimodal Perception Using Large Language Models," IEEE/RSJ International Conference on Intelligent Robots and Systems, Detroit, MI, IEEE, pp. 3590–3596.
[114] Hu, M., Mu, Y., Yu, X. C., Ding, M., Wu, S., Shao, W., Chen, Q., Wang, B., Qiao, Y., and Luo, P., 2023, "Tree-Planner: Efficient Close-Loop Task Planning With Large Language Models," The Twelfth International Conference on Learning Representations.
[115] Hori, C., Peng, P., Harwath, D., Liu, X., Ota, K., Jain, S., Corcodel, R., Jha, D., Romeres, D., and Roux, J. L., 2023, "Style-Transfer Based Speech and Audio-Visual Scene Understanding for Robot Action Sequence Acquisition From Videos," arXiv Preprint arXiv:2306.15644.
[116] Shao, L., Migimatsu, T., Zhang, Q., Yang, K., and Bohg, J., 2021, "Concept2robot: Learning Manipulation Concepts From Instructions and Human Demonstrations," Int. J. Rob. Res., **40**(12–14), pp. 1419–1434.
[117] Sontakke, S., Zhang, J., Arnold, S., Pertsch, K., Bıyık, E., Sadigh, D., Finn, C., and Itti, L., 2024, "Roboclip: One Demonstration Is Enough to Learn Robot Policies," Adv. Neural Inf. Process. Syst., **36**.
[118] Murray, M., Gupta, A., and Cakmak, M., 2024, "Teaching Robots With Show and Tell: Using Foundation Models to Synthesize Robot Policies From Language and Visual Demonstration," 8th Annual Conference on Robot Learning.
[119] Aristeidou, C., Dimitropoulos, N., and Michalos, G., 2024, "Generative AI and Neural Networks Towards Advanced Robot Cognition," CIRP Ann, **73**(1).
[120] Yow, J., Garg, N. P., Ramanathan, M., Ang, W. T., et al., 2024, "Extract–Explainable Trajectory Corrections From Language Inputs Using Textual Description of Features," arXiv Preprint arXiv:2401.03701.
[121] Quartey, B., Shah, A., and Konidaris, G., 2023, "Exploiting Contextual Structure to Generate Useful Auxiliary Tasks," arXiv Preprint arXiv:2303.05038.
[122] Geng, H., Wei, S., Deng, C., Shen, B., Wang, H., and Guibas, L., 2023, "Sage: Bridging Semantic and Actionable Parts for Generalizable Articulated-Object Manipulation Under Language Instructions," arXiv Preprint arXiv:2312.01307.
[123] Parakh, M., Fong, A., Simeonov, A., Chen, T., Gupta, A., and Agrawal, P., 2023, "Lifelong Robot Learning With Human Assisted Language Planners," Conference on Robot Learning, Atlanta, GA.
[124] Ming, C., Lin, J., Fong, P., Wang, H., Duan, X., and He, J., 2023, "HiCRISP: A Hierarchical Closed-Loop Robotic Intelligent Self-Correction Planner," arXiv Preprint arXiv:2309.12089.
[125] Ye, Y., You, H., and Du, J., 2023, "Improved Trust in Human-Robot Collaboration With Chatgpt," IEEE Access, **11**, pp. 55748–55754.
[126] Qiu, J., Xu, M., Han, W., Moon, S., and Zhao, D., 2023, "Embodied Executable Policy Learning With Language-Based Scene Summarization," arXiv Preprint arXiv:2306.05696.
[127] Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., and Wahid, A., 2023, "Palm-e: An Embodied Multimodal Language Model," International Conference on Machine Learning, Honolulu, HI, July 23, PMLR, pp. 8469–8488.
[128] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., and Wu, J., 2023, "Rt-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," Conference on Robot Learning, Atlanta, GA, PMLR, pp. 2165–2183.
[129] Bärmann, L., Kartmann, R., Peller-Konrad, F., Niehues, J., Waibel, A., and Asfour, T., 2024, "Incremental Learning of Humanoid Robot Behavior From Natural Interaction and Large Language Models," Front. Rob. AI, **11**, p. 1455375.
[130] Kumar, K. N., Essa, I., and Ha, S., 2023, "Words Into Action: Learning Diverse Humanoid Robot Behaviors Using Language Guided Iterative Motion Refinement," arXiv Preprint arXiv:2310.06226.
[131] Tideman, M., 2008, "Scenario Based Product Design."
[132] Mangaonkar, M., and Penikalapati, V. K., 2024, "Enhancing Production Data Pipeline Monitoring and Reliability Through Large Language Models (LLMS)," Eduzone: Int. Peer Rev./Ref. Multidisc. J., **13**(1), pp. 51–56.
[133] Lee, S.-H., Lee, D.-W., Song, H.-S., Jeong, S., Ji, Y., Song, J.-S., Kim, J., and Yi, B.-J., 2023, "Robotic Manipulation System Design and Control for Non-Contact Remote Diagnosis in Otolaryngology: Digital Twin Approach," IEEE Access, **11**, pp. 28735–28750.
[134] Park, J.-S., Xiao, X., Warnell, G., Yedidsion, H., and Stone, P., 2023, "Learning Perceptual Hallucination for Multi-Robot Navigation in Narrow Hallways," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, IEEE, pp. 10033–10039.
[135] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., et al., 2022, "Rt-1: Robotics Transformer for Real-World Control at Scale," arXiv Preprint arXiv:2212.06817.
[136] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., and Fox, D., 2020, "Alfred: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks," Conference on Computer Vision and Pattern Recognition, Online, June 14, pp. 10740–10749.
[137] Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W. Y., Shen, C., and Hengel, A. v. d., 2020, "Reverie: Remote Embodied Visual Referring Expression in Real Indoor Environments," Conference on Computer Vision and Pattern Recognition, online, June 14, pp. 9982–9991.
[138] Mathur, P., 2023, "Proactive Human-Robot Interaction Using Visuo-Lingual Transformers," arXiv Preprint arXiv:2310.02506.
[139] ElMaraghy, H. A., 2005, "Flexible and Reconfigurable Manufacturing Systems Paradigms," Int. J. Flexible Manuf. Syst., **17**(4), pp. 261–276.
[140] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., and Jain, A., 2024, "Structured Information Extraction From Scientific Text With Large Language Models," Nat. Commun., **15**(1), p. 1418.
[141] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R., 2023, "Challenges and Applications of Large Language Models," arXiv Preprint arXiv:2307.10169.
[142] Hall, S. M., Gonçalves Abrantes, F., Zhu, H., Sodunke, G., Shtedritski, A., and Kirk, H. R., 2024, "Visogender: A Dataset for Benchmarking Gender Bias in Image-Text Pronoun Resolution," Adv. Neural Inf. Process. Syst., **36**.
[143] Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., and Niforatos, E., 2024, "Knowledge Sharing in Manufacturing Using LLM-Powered Tools: User Study and Model Benchmarking," Front. Artif. Intell., **7**, p. 1293084.
[144] Hirzel, M., 2023, "Low-Code Programming Models," Commun. ACM, **66**(10), pp. 76–85.
[145] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., et al., 2021, "Evaluating Large Language Models Trained on Code," arXiv Preprint arXiv:2107.03374.
[146] Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A., 2023, "Code as Policies: Language Model Programs for Embodied Control," 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp. 9493–9500.
[147] Ren, M., and Zheng, P., 2024, "Towards Smart Product-Service Systems 2.0: A Retrospect and Prospect," Adv. Eng. Inf., **61**, p. 102466.
[148] Aloqaily, M., Bouachir, O., Karray, F., Al Ridhawi, I., and El Saddik, A., 2022, "Integrating Digital Twin and Advanced Intelligent Technologies to Realize the Metaverse," IEEE Consum. Electron. Mag, **12**, pp. 47–55.
[149] Mourtzis, D., 2023, "The Metaverse in Industry 5.0: A Human-Centric Approach Towards Personalized Value Creation," Encyclopedia, **3**(3), pp. 1105–1120.
[150] Yang, C., Wang, Y., Jiang, Y., Lan, S., and Wang, L., 2023, "Metaverse: Architecture, Technologies, and Industrial Applications," IEEE 19th International Conference on Automation Science and Engineering (CASE), Auckland, New Zealand, IEEE, pp. 1–6.
[151] Nvidia, "Nvidia Omniverse," Nvidia, https://www.nvidia.com/en-us/omniverse/, Accessed July 29, 2024.
[152] Apple, "Apple Vision Pro," Apple, https://www.apple.com/apple-vision-pro/, Accessed September 7, 2024.
[153] Chavan, A., Magazine, R., Kushwaha, S., Debbah, M., and Gupta, D., 2024, "Faster and Lighter LLMS: A Survey on Current Challenges and Way Forward," arXiv Preprint arXiv:2402.01799.
[154] Xu, M., Xu, Y. L., and Mandic, D. P., 2023, "Tensorgpt: Efficient Compression of the Embedding Layer in LLMS Based on the Tensor-Train Decomposition," arXiv Preprint arXiv:2307.00526.
[155] "MLC-LLM", 2023, Mlc-llm, https://github.com/mlc-ai/mlc-llm, Accessed November 13.