Journal of Manufacturing Systems 82 (2025) 809–840

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys

Review

Leveraging large language models in next generation intelligent manufacturing: Retrospect and prospect

Yunfei Ma [a], Shuai Zheng [a,*], Zheng Yang [a], Pai Zheng [b], Jiewu Leng [c], Jun Hong [d]

[a] School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, Shannxi, China
[b] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, 999077, Hong Kong Special Administrative Region of China
[c] Guangdong Provincial Key Laboratory of Computer Integrated Manufacturing, Guangdong University of Technology, Guangzhou, 510006, Guangdong, China
[d] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, 710049, Shannxi, China

ARTICLE INFO

Keywords:
Industry 5.0
Large language model
Human–robot collaboration
Human centered intelligent manufacturing

ABSTRACT

Industry 5.0, as the guiding ideology of the new generation intelligent manufacturing, points the way for global industrial transformation. It emphasizes the collaborative cooperation between humans, machines and intelligent systems, and places humans at the core of the industrial production process, aiming to create a more flexible, personalized and sustainable production paradigm. Large language model, as an advanced natural language processing technology, has received attention from researchers related to Industry 5.0 due to its ease of use and powerful language processing capability. LLM is considered to be one of the key enabling technologies to drive the development of Industry 5.0 and has great application potential. After a rigorous review of existing approaches, we find there is few existing survey papers that focuses on how LLM will drive the development of Industry 5.0 applications. Therefore, this paper provides a comprehensive review of the application of LLM in the field of Industry 5.0. Firstly, we conduct a literature review to explore the current state of research related to Industry 5.0. Subsequently, we analyze LLM-based technologies, synergizing LLMs with Industry 5.0 enablers and the applications of LLM in various domains of intelligent manufacturing. Finally, we explore the challenges of LLM in real-world scenarios and future research directions in the context of Industry 5.0. It is hoped that this study will contribute to the further development of LLM-based solutions in the context of Industry 5.0 and unite various efforts to achieve the vision of Industry 5.0.

1. Introduction

Technological progress has been a central driver of the industrial revolution. As shown in Fig. 1 , the development of the steam engine marked the beginning of the First Industrial Revolution, fundamentally changing human production methods and social structures. The widespread use of electricity triggered the Second Industrial Revolution, leading to a clear division of labor in society. It also marked the beginning of mass production and modern industry. Industry 3.0 achieved mass customization at the information technology level [1]. With the development of automation, the Internet of Things, and intelligent systems, Germany proposed Industry 4.0 in 2011. Industry 4.0 refers to the intelligent networking of industrial machines based on CPS (Cyber–Physical Systems) that enables intelligent control through embedded network systems [2]. CPS-based systems can make intelligent decisions by enabling real-time communication and collaboration between manufacturing things to achieve high-quality, high-efficiency, personalized, and flexible production. However, the implementation of Industry 4.0 exhibits a tendency to advance technological aspects at the expense of the social or human aspect, as demonstrated, among others, by [3,4]. Thus, Industry 5.0 emerged as a response.

Industry 5.0 is significantly different from the above industrial revolution. Industry 5.0 is the long-term vision for future industries, aiming to establish a human-centric, sustainable, and resilient manufacturing system. It highlights the collaboration between human operators and intelligent systems, positioning humans at the center of industrial activities. In Industry 5.0, machines are not just tools for completing tasks. They also intelligently adapt to human operating habits and work rhythms to enhance the interaction and cooperation between humans and machines. The overarching goal of Industry 5.0 is to return industry to humanization by integrating artificial intelligence and intelligent systems into human operations in order to enhance human capabilities and creativity. The collaboration between humans and machines not only boosts production efficiency but also provides more possibilities

* Corresponding author.
E-mail address: shuaizheng@xjtu.edu.cn (S. Zheng).

https://doi.org/10.1016/j.jmsy.2025.07.019
Received 22 February 2025; Received in revised form 14 July 2025; Accepted 23 July 2025
Available online 5 August 2025

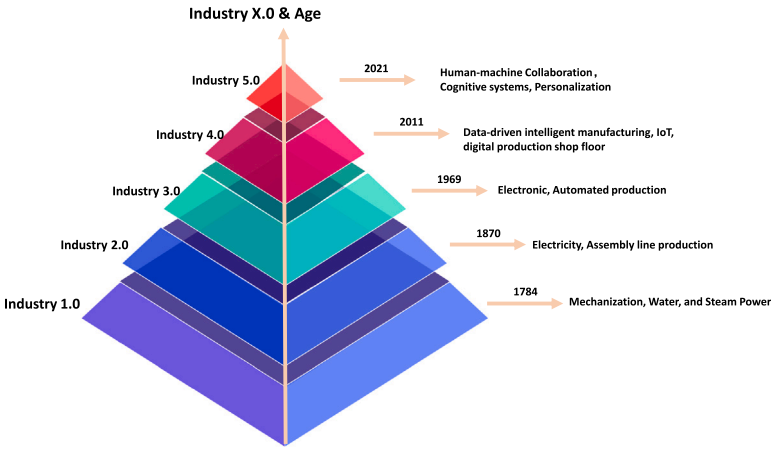

Fig. 1. Illustration of industrial evolution.

for personalized customization and flexible manufacturing, enabling products to meet people's individual needs. As economist Leo Cherne said: *The computer is incredibly fast, accurate, and stupid. Man is unbelievably slow, inaccurate, and brilliant. The marriage of the two is a force beyond calculation.* Through such integration, Industry 5.0 aims not only at technological progress but also at expanding and improving human capabilities. It presents a more human-centered future, where intelligent systems are powerful supporters of human work, rather than as replacements for humans.

In addition, technology has not stopped advancing. Recently, pre-trained language models (PLMs) based on the Transformer architecture, such as BERT [5] and T5 [6], have shown superior performance in NLP tasks. Research shows that the performance of the models improves with increasing parameter size. The models not only demonstrate improved performance, but also possess certain characteristics similar to humans, such as in-context learning [7] and emerging abilities [8]. Researchers in the relevant fields have proposed the term "large language model" to differentiate these PLMs with large parameter sizes. Based on the type of generated content, they can be categorized into text-to-text models (ChatGPT [9]), text-to-image models (DALL-E [10]), text-to-code models (GitHub Copilot [11]), text-to-speech models (Tacotron [12]), and text-to-video models (Sora [13]). These models are referred to as Generative Artificial Intelligence (GAI). GAI shows great potential and contributes to human productivity in many complex real-world tasks. For example, a study by Noy and Zhang shows that ChatGPT significantly increased the productivity of workers in mid-level professional collaborative tasks [14]. Another study by Brynjolfsson showed that call center operators using GAI increased their productivity by 14%, and less experienced employees increased their productivity by more than 30% [15]. The reports suggest that GAI could increase global GDP by 7%. The widespread popularity of LLMs is largely due to their ease of use and interactivity. Users can accomplish various tasks using these models through simple conversational interactions, which greatly lowers the technical barrier. Non-professionals can also easily access and utilize these advanced AI technologies. This aligns perfectly with the human-centric concept of Industry 5.0. With the help of LLMs, workers can expand the boundaries of their capabilities, improving production efficiency and product quality, while also promoting more sustainable and personalized production methods.

The combination of Industry 5.0 and large language models offers new opportunities for future industrial development. Industry 5.0 advocates for a human-centered industrial ecosystem, committed to achieving a deep integration of technology and humanity. The interactivity, intelligence, and generalizability of LLMs perfectly match this vision. Hence, we review LLM-based solutions in the context of Industry 5.0 and present the challenges currently faced, along with potential solutions. We searched major databases including Spring, Scopus, IEEE, and arXiv for nearly five years (2020 2025) using keywords such as 'Industry 5.0', 'large language model', and 'smart manufacturing'. We use 'OR' as a retrieval method between the above keywords to reveal publications, which are relevant to Industry 5.0 and/or LLM. Based on the retrieval results, we focused on a total of 197 excellent and representative papers.

As shown in Fig. 2, the rest of this paper is organized as follows. Section 2 reviews the related review works. Section 3 introduces the advanced technologies of LLMs. Section 4 introduces the characteristics of LLMs. Section 5 introduces the integration of LLMs with Industry 5.0 enabling technologies. Section 6 presents the applications of LLMs across different stages of smart manufacturing. Section 7 presents the current challenges of applying LLMs in smart manufacturing and their future research directions. Section 8 concludes this paper.

2. Related review works

With the transition from Industry 4.0 to Industry 5.0, many researchers focus on exploring how emerging technologies can contribute to human-centered manufacturing environments. Existing review articles primarily focus on exploring the realization paths of human-centered production systems, flexible production, and sustainable production from multiple dimensions, including the conceptual framework of Industry 5.0, enabling technologies, smart manufacturing, and human–machine collaboration. However, despite these studies providing valuable perspectives on understanding the development of Industry 5.0, there are still notable gaps in the existing review literature. Therefore, this section firstly discusses the differences and connections between Industry 5.0 and Industry 4.0 (Section 2.1 The concept of Industry 5.0), then reviews existing review articles in related fields (Section 2.2 Previous reviews), and finally identifies the gaps and shortcomings in current articles (Section 2.3 Review gaps).

2.1. The concept of industry 5.0

Industry 4.0 is a German initiative launched in 2013 and described as an industrial revolution [16]. The core aim of the initiative is to improve the competitiveness of German industry and give it a head start in the new industrial revolution. Industry 4.0 is seen as an extension of past trends in automation, which has led to the development of technologies such as cyber–physical systems and the Internet of Things. The goal of Industry 4.0 is to employ the above technologies to significantly increase the sophistication of automation and interconnectivity, thereby improving the efficiency of manufacturing. In addition, many countries have launched similar strategic initiatives and spent significant amounts of money to develop and implement some of the Industry 4.0 technologies [17]. Examples include China's "Made in China 2025",

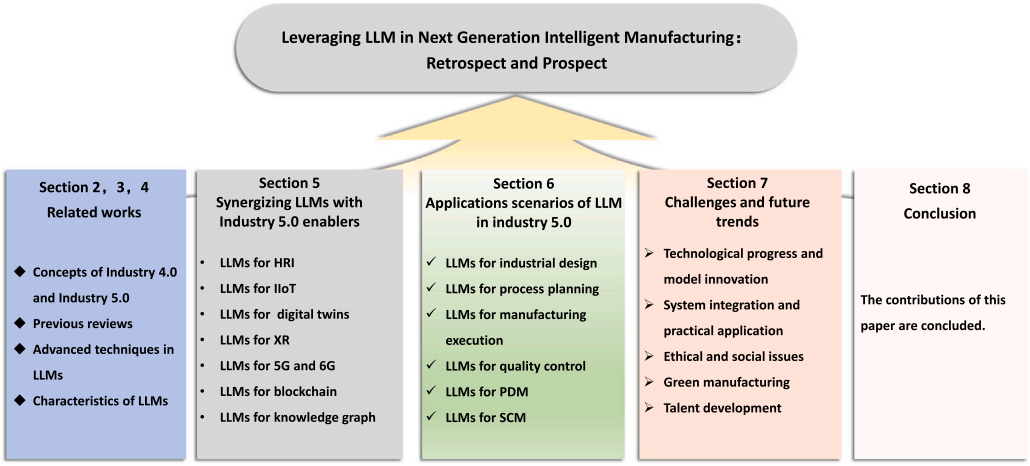

Fig. 2. The overall structure of the paper.

the United States' "National Strategic Plan for Advanced Manufacturing", and many others. Over the past decade, with the development of machine learning, augmented reality, collaborative robotics, and other technologies, there has been a general trend in many countries and regions to digitize more aspects of life and society as a whole. This trend not only reflects the widespread adoption of Industry 4.0-related technologies but also represents a deepening of the application of Industry 4.0 and its successor technologies across a wide range of industries, including manufacturing. But many of the jobs affected by Industry 4.0 have not become interesting over the last decade. Instead, the vast amounts of big data and the interconnectedness of machines, processes, and systems used for model training mean that there is a diminished human touch. This leads to challenges in operations, performance, and efficiency. Many scholars are also realizing that Industry 4.0 is less focused on social equity and sustainability and more on digitization and AI-driven technologies, as well as a lack of attention to the human factor in the development of Industry 4.0 [18]. This has also led to workers becoming uncomfortable with technological developments and the new skills they require, and a decline in acceptance of Industry 4.0 technologies such as robotics. It has even led to panic about some new technologies, such as the difficulty of skill adaptation for a large number of front-line workers in the face of new technologies due to a lack of appropriate training and support, which has led to occupational anxiety and insecurity. Additionally, due to limited knowledge of automation and intelligent systems, workers have shown low acceptance of Industry 4.0 technologies, including collaborative robots (cobots) and intelligent manufacturing systems, and even fear of AI replacing human jobs in some contexts [2].

In 2021, the European Commission officially called for a Fifth Industrial Revolution and officially released the report entitled "Industry 5.0: Towards a Sustainable, Human-centric, and Resilient European Industry" on January 4, 2021 [19]. This is similar to the introduction of Industry 4.0 in 2011 by the German government, which devised a top-down initiative in response to the changing societal and geopolitical landscape. Industry 5.0 is defined in the EU White Paper as follows: Industry 5.0 recognizes the power of industry to achieve societal goals beyond jobs and growth to become a resilient provider of prosperity, by making production respect the boundaries of our planet and placing the well-being of the industry worker at the center of the production process. In other words, Industry 5.0 should not be understood as a replacement for the existing Industry 4.0 paradigm, but rather as an evolutionary and logical continuation of the existing Industry 4.0 paradigm. The concept of Industry 5.0 can also be said to reintroduce the "human-centered/value-centered Industry 4.0" or even be called Industry 4.1. The concept of Industry 5.0 is therefore not based on technology, but is centered on values, not a narrow focus on profit that fails to properly consider environmental and social costs and benefits. Furthermore, the core technologies of Industry 5.0 are broadly aligned with those of Industry 4.0, which focuses more on human-centered technologies. The focus of the technology used should not be on replacing shop-floor workers, but rather on supporting their capabilities and bringing about a safer and more satisfying working environment, to the mutual benefit of industry and workers. In summary, Industry 5.0 overcomes the weaknesses of Industry 4.0 by focusing explicitly on the human element of the system and placing greater emphasis on social responsibility and sustainability in industrial development. Through the application of more humane technologies, Industry 5.0 aims to create a sustainable, human-centric, and resilient industry for the society of the future.

2.2. Previous reviews

The existing review articles on Industry 5.0 are generally categorized into two types: applications of Industry 5.0 in different industry sectors and the use of specific technologies to promote the realization of Industry 5.0. As shown in Table 1, the following will review the relevant articles based on these categories.

2.2.1. Sector-specific applications and extensions of industry 5.0

The first category of reviews focused on the application of Industry 5.0 ideas in specific industries. It analyzes how related fields are incorporating Industry 5.0 ideas to achieve more efficient, personalized, and sustainable operations. Vrutti et al. provided an in-depth discussion on the role of smart wearable devices in Healthcare Industry 5.0, focusing on how they use machine learning and digital technologies to deliver personalized patient care solutions. This technological advancement has not only improved the quality of healthcare delivery but also facilitated more efficient interactions between patients and healthcare providers [26]. Eleni et al. explored the application of digital twin technology in Agriculture 5.0, particularly how this technology can optimize livestock production systems. The research reviewed the latest advancements in digital twins and discussed their potential and challenges in Agriculture 5.0 [23]. Andreas et al. provided an overview of the evolution of Industry 5.0 and its transition to Forestry 5.0. They proposed a framework to clarify the current state of these developments, identify beneficial technologies, particularly artificial intelligence, and reveal the challenges hindering the effective adoption of these technologies in Forestry 5.0 [24]. Abdo et al. reviewed the recent research progress on Food Industry 4.0 and identified enabling technologies for future Food Industry 5.0 [25]. Guilherme added valuable insights to researchers and practitioners

Table 1
Classification of the relevant and recent review.

No.	Author	Year	Title	Domain-specific or enabler	Review category
1	Guilherme et al. [20]	2021	From Supply Chain 4.0 to Supply Chain 5.0: findings from a systematic Literature Review and Research Directions	Supply Chain 5.0	Applications of Industry 5.0
2	Foivos et al. [21]	2023	Envisioning maintenance 5.0: Insights from a systematic literature review of Industry 4.0 and a proposed framework	Envisioning maintenance 5.0	Applications of Industry 5.0
3	Marina et al. [22]	2023	From Industry 4.0 to Construction 5.0: Exploring the path towards human–robot Collaboration in Construction	Construction 5.0	Applications of Industry 5.0
4	Eleni et al. [23]	2024	Recent Advances in Digital Twins for Agriculture 5.0: Applications and Open Issues in Livestock Production Systems	Agriculture 5.0	Applications of Industry 5.0
5	Andreas et al. [24]	2024	From Industry 5.0 to Forestry 5.0: Bridging the gap with Human-Centered Artificial Intelligence	Forestry 5.0	Applications of Industry 5.0
6	Abdo et al. [25]	2024	From Food Industry 4.0 to Food Industry 5.0——Identifying technological enablers and potential future application in the food sector	Food Industry 5.0	Applications of Industry 5.0
7	Vrutti et al. [26]	2024	Intelligent wearable-assisted digital healthcare Industry 5.0	Healthcare 5.0	Applications of Industry 5.0
8	Ravdeep et al. [27]	2024	Cybersecurity for Industry 5.0: trends and gaps	Cybersecurity 5.0	Applications of Industry 5.0
9	Bartlomiej et al. [28]	2023	Current development on the Operator 4.0 and transition towards the Operator 5.0: A systematic literature review in light of Industry 5.0	Operator 5.0	Applications of Industry 5.0
10	Leng et al. [29]	2024	Unlocking the power of industrial artificial intelligence towards Industry 5.0: insights, pathways, and challenges	IndAI	Enablers for Industry 5.0
11	Guo et al. [30]	2024	Industrial metaverse towards Industry 5.0: Connotation, architecture, enablers, and challenges	Metaverse	Enablers for Industry 5.0
12	Muhammed et al. [31]	2024	Exploring the synergies between collaborative robotics, digital twins, augmentation, and Industry 5.0 for smart manufacturing	Collaborative robotics, DT, and augmentation	Enablers for Industry 5.0
13	Wang et al. [32]	2024	Human Digital Twin in the context of Industry 5.0	Human Digital Twin	Enablers for Industry 5.0
14	Fang et al. [33]	2023	Head-mounted display augmented reality in manufacturing: A systematic review	Augmented reality	Enablers for Industry 5.0
15	Lv et al. [34]	2023	Digital Twins in Industry 5.0	Digital Twins	Enablers for Industry 5.0
16	Alejandro et al. [35]	2022	A Review of Deep Reinforcement Learning Approaches for Smart Manufacturing in Industry 4.0 and 5.0 Framework	Deep Reinforcement Learning	Enablers for Industry 5.0

by approaching the newest and revolutionary concept of the Industry 5.0 phenomenon in the supply chain context [20]. Foivos et al. proposed the Maintenance 5.0 framework, which emphasized the integration of human-centered and human-driven strategies to achieve efficient and sustainable maintenance in zero-defect manufacturing systems [21]. Marina et al. reviewed the prospects for the use of human–computer collaboration in construction and presented the challenges facing human–computer collaboration. They also provided a realistic assessment of the potential for the Industry 5.0 paradigm to evolve into Construction 5.0 [22].

2.2.2. Technological enablers for advancing industry 5.0

This section focuses on the role of certain specific technologies in Industry 5.0 and analyzes how these technologies drive this new industrial revolution. Leng et al. provided an overview of the key characteristics and enabling technologies of Industrial Artificial Intelligence (IndAI), explaining how IndAI is applied at various stages, from industrial product design to product maintenance, to foster a more efficient and human-centered manufacturing process. Lastly, the review also addressed how to overcome the technical and societal challenges of IndAI in the implementation of Industry 5.0 [29]. Guo et al. examined the development of the industrial metaverse and its role in supporting Industry 5.0, highlighting the significant contribution of technologies such as virtual reality and augmented reality in advancing Industry 5.0 [30]. Muhammed et al. reviewed the synergy between collaborative robots, digital twins, and augmented technologies in enabling smart manufacturing. The research showed that the combination of these technologies not only enhanced flexibility and efficiency in production but also promoted the realization of the "human-centered" concept in Industry 5.0 [31]. Wang et al. conducted a comprehensive survey on the Human Digital Twin (HDT) in the context of Industry 5.0, summarizing its correct evolution and proposing the proper definition of HDT [32]. Fang et al. reviewed the application of Head-Mounted Display (HMD) augmented reality (AR) in manufacturing, pointing out that HMD AR can free workers' hands, help them operate more intuitively, and reduce the psychological burden of decision-making, further promoting a "human-centered" manufacturing environment [33].

2.3. Review gaps

Although the existing review literature provides a rich context for understanding Industry 5.0 and provides in-depth theoretical discussions and practical examples of the development of the field, there is a lack of discussion on how LLM can play a role in the context of Industry 5.0. The gap stems from two reasons: Firstly, the application of LLMs in the field of industrial smart manufacturing is still in its early stages. Although LLMs show great potential in areas such as natural language processing, knowledge management, and automated decision-making, specific application cases in the context of Industry 5.0 are still relatively rare. Secondly, although LLM technology has

broad application prospects in areas such as automation, intelligent decision support, and human–machine collaboration, most of the existing research focuses on the application of traditional Industry 4.0 technologies. Compared to existing work, the innovation of this paper lies in its attempt to fill this gap by exploring how LLMs can be integrated with other enabling technologies of Industry 5.0 (Section 5 Synergizing LLMs with Industry 5.0 enablers) and how they can play a role in the stages of smart manufacturing (Section 6 Application scenarios of LLMs in Industry 5.0). Lastly, we summarize the challenges faced by LLMs in real-world applications and outline future research directions (Section 7 Challenges and future trends). This paper aims to provide new perspectives for both academia and industry, drive the development of future LLM-based industrial solutions, and establish a theoretical foundation for the implementation of Industry 5.0.

3. Characteristics of LLMs

With the rapid development of Industry 4.0, smart manufacturing, automated production, and data-driven decision-making systems have become mainstream. In this process, classical machine learning methods such as Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks, and other methods have been widely applied, effectively supporting the progress of production optimization, quality control, fault diagnosis, and other fields. However, with the further increase in system complexity and the introduction of the Industry 5.0 paradigm, industrial systems are faced with new types of challenges such as mass customization, human-centric solutions, and resilient supply chain management. At this point, classical machine learning gradually shows its limitations in dealing with these complex and variable tasks. In this context, the large language model shows unique advantages over classical rule-based reasoning engines and classical machine learning aspects. This section will provide a brief overview of the main capability features of LLM, especially for the performance advantages that traditional models do not have.

Rich world knowledge. Compared with classical machine learning methods, large language models can learn rich world knowledge after pre-training with ultra-large-scale text data, and have stronger knowledge reasoning and context understanding capabilities [36]. This advantage is particularly important in the context of Industry 5.0 for smart manufacturing, as it can effectively support complex applications such as mass personalization and human–machine interaction. Early expert systems aimed to solve domain-specific application tasks by designing reasoning engines based on knowledge bases and knowledge representations. However, the approach relies heavily on logic, rules, and classical machine learning algorithms, which makes the system relatively limited in its ability to adequately model and leverage extensive world knowledge. In addition, early models such as BERT, RNN, and GPT-1 were relatively small in scale and data size, preventing them from adequately learning the vast amount of world knowledge.

General task solving capability. The second representative capability feature of LLMs is a strong generalized task-solving capability. Large language models are learned primarily through the pretraining task of predicting the next tokens [37]. Although not specifically optimized for a particular downstream task, their general task-solving ability far exceeds that of traditional models. It is able to respond effectively to a wide range of natural language processing tasks (e.g., question answering, translation, information extraction, text categorization, etc.) without specialized fine-tuning. Due to the general task-solving capability of LLMs, it simplifies various natural language processing tasks into text generation, realizing the unification of various task forms. In Industry 5.0, the advantages of this capability are outstanding. This means that LLMs can be used as an 'intelligent interface' to seamlessly connect to a variety of data sources, system platforms, and human users, enabling them to be embedded in a variety of industrial scenarios.

Complex task reasoning ability. LLM shows excellent reasoning ability when dealing with complex tasks. Research shows that LLMs can effectively answer reasoning questions involving multi-hop knowledge relationships and solve mathematical problems that require complex mathematical derivations [38]. Traditional natural language processing and machine learning algorithms tend to have limited performance when faced with such problems, and usually need to adapt their model architectures individually for each task or rely on specially designed datasets for training. The reasoning capability of LLMs is not only reflected in the improvement of model performance, but is also a key technical support to promote human-centered solutions and industrial cognitive intelligence in Industry 5.0.

Follow instructions with human feedback. LLMs establish a unified task-solving paradigm in the form of natural language: both task inputs and execution results are expressed through natural language. Through the pre-training and fine-tuning phases, LLMs have a strong ability to follow human instructions. Users can give task instructions directly through natural language (so known as "prompt engineering") [39]. In early dialogue systems, instruction following was a widely studied direction. However, traditional models lack generalized task understanding and execution capabilities and still need to rely on artificial rules or prior information to assist the design and training of instruction understanding modules. It leads to the fact that dialogue systems based on classical machine learning are not widely used. However, LLMs provide a natural and generic solution for human–computer interaction, which is important for building many human-centered application services (e.g., smart speakers, fault diagnosis decision, etc.).

Human alignment ability. The safety of AI has always been an important research direction in the field of AI. However, in the context where classical AI models have limited intelligence and weaker generalizability, researchers often focus more on improving model performance, while research on security, reliability, and ethics tends to lag behind. With the emergence of large language models, this situation has undergone a fundamental change. Due to its powerful text generation and task-solving capabilities, the lack of effective behavioral constraint mechanisms may lead to serious consequences such as misleading information dissemination, harmful content generation, and privacy leakage. To address this, LLM alignment methods have adopted reinforcement learning from human feedback. By incorporating human evaluation signals in the sample generation process, the model can continuously adjust its behavioral tendencies during training to reinforce outputs that align with human values and task objectives, while suppressing errors, inappropriate, or harmful behaviors [40]. The approach significantly enhances the model's safety and behavioral predictability. Currently, many deployed LLMs (such as the GPT series, Claude, Gemini, etc.) are capable of automatically blocking responses when faced with sensitive, dangerous or inappropriate instructions, effectively mitigating common abuse risks in real-world applications. In the context of Industry 5.0, LLM can facilitate deeper human–machine collaboration through stable, consistent, and human-value-aligned response behaviors that enhance user trust in AI systems.

Expandable tool-use capabilities. Classical machine learning models are often limited by fixed architectural designs, inductive assumptions, and the range of training data, making it difficult to adapt to new tasks that change in real time or go beyond the training corpus. Similarly, LLMs have certain limitations. For example, they still struggle to effectively answer questions that fall outside the time range of their pretraining data and face challenges when performing precise numerical calculations. To break through these limitations, thanks to LLM's natural language-driven task-solving framework, LLM can learn to utilize various external tools through in-context learning or fine-tuning, such as search engines, calculators, code executors, and more [41]. In Industry 5.0, this ability to use tools demonstrates significant practical value. LLMs can leverage search tools to obtain the latest standards, technical specifications, or regulations, assisting engineers in making compliance judgments and design decisions. In addition, LLM can also be used as a unified interface for industrial automation platforms to coordinate multiple systems, such as MES, ERP, etc., for linked operations.

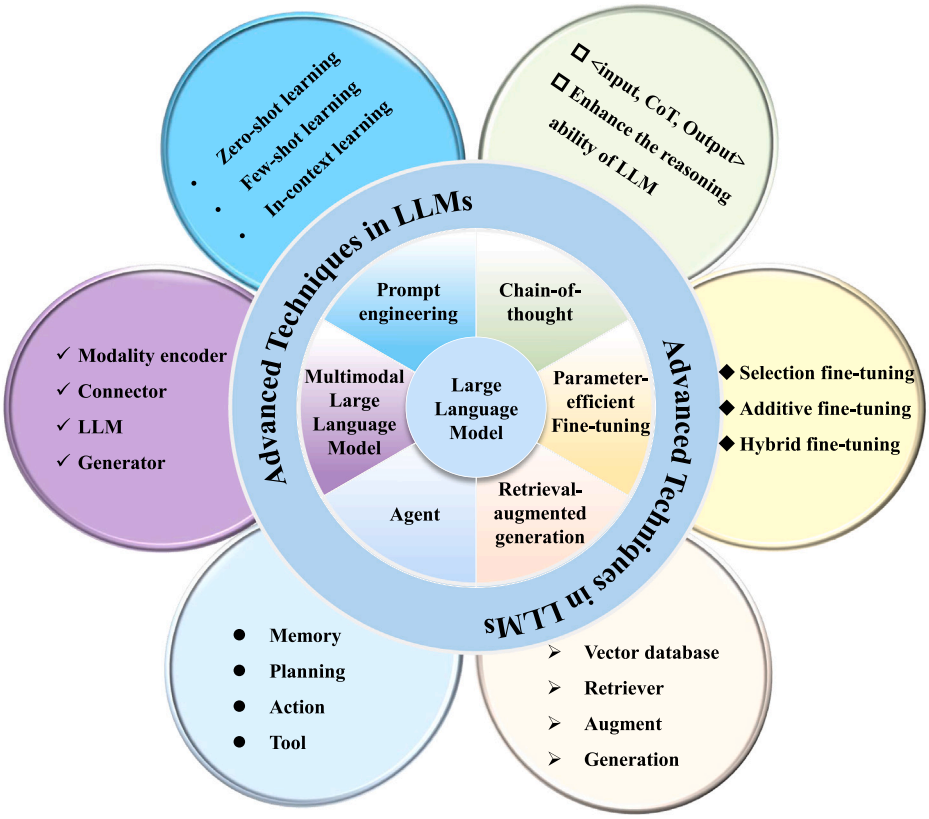

Fig. 3. The advanced techniques in LLMs.

4. Advanced techniques in LLMs

Large language models have been a significant breakthrough in the field of natural language processing in recent years. LLMs are Transformer-based natural language processing models that can understand and generate natural language text. LLMs demonstrate excellent performance in applications such as dialog generation, content creation, and language translation. However, to further enhance the performance and utility of LLMs, researchers have developed a series of advanced techniques to optimize the performance of these models in specific tasks, as shown in Fig. 3.

Prompt engineering. After being pretrained on large-scale text data, LLMs have the capability to serve as general task solvers. However, these capabilities may not explicitly manifest when performing certain specific tasks [42]. OpenAI proposed Prompt engineering for the first time in the GPT-3 report [43]. The approach suggests that designing appropriate linguistic instructions in the LLM input can help stimulate LLMs' abilities. This method does not require changing the model parameters but instead reorganizes downstream tasks by providing additional context to the data through "prompts". There are three forms of prompt engineering: (1) Few-shot learning refers to adding a small number of examples after a natural language prompt as input to the LLM. The LLM can use these limited examples to understand the task's requirements and patterns. (2) Zero-shot learning refers to not using any example data and relying only on a well-designed prompt to activate the knowledge and competencies in the LLM that are relevant to the target task. (3) In-context learning can be viewed as a special form of few-shot learning that implicitly includes the target task and the format information in the problem. Overall, both few-shot learning and zero-shot learning are specific forms of in-context learning, and their key differences are whether or not sample examples are provided and the number of examples.

Chain-of-thought. Chain-of-Thought (CoT) is a technique that enhances the reasoning capabilities of large language models (LLMs) [44]. It is particularly suited to tasks that require complex reasoning and multi-step analysis, such as logical reasoning and arithmetic computation. CoT is also a prompt technique that encourages LLMs to generate intermediate reasoning chains for problem solving, as opposed to the traditional "one-step" approach of generating the final answer directly from LLMs. In chain-of-thought prompting, examples of intermediate natural language reasoning steps replace the [input, output] pairs from few-shot and in-context learning, creating a [input, chain of thought, output] triplet structure. This approach helps the model to show its thinking process, showing each step of reasoning as if it were 'talking to itself', thus simulating the chain of human thought. This not only improves the model's ability to understand and reason about complex problems but also enhances the interpretability and reliability of the generated content. Practical evidence indicates that CoT can significantly improve the performance of LLM in tasks such as arithmetic problems, logical judgments, and complex reasoning. In the manufacturing industry, CoT effectively improves the efficiency and diagnostic accuracy of equipment maintenance by gradually guiding maintenance engineers to systematically analyze and troubleshoot equipment [45].

Parameter-efficient fine-tuning. Early fine-tuning methods, represented by BERT, involved adding a task-adaptation layer on top of the base language model, followed by full fine-tuning. However, as the parameters of pre-trained language models grow larger, the memory required for full-scale training also increases. When performing full-scale fine-tuning of LLM, the parameters mainly come from three parts: the model's parameters, the gradients, and the optimizer. Assuming a language model has 1.5B parameters, the memory required for full fine-tuning would be 24 GB. The memory required for full fine-tuning is approximately 16 times the size of the language model's parameters (with slight variations depending on the optimizer used). The parameter size of current open-source large language models ranges from 6B or 7B to as large as 314B. Full fine-tuning of the above models is essentially impossible for individual researchers or organizations. To

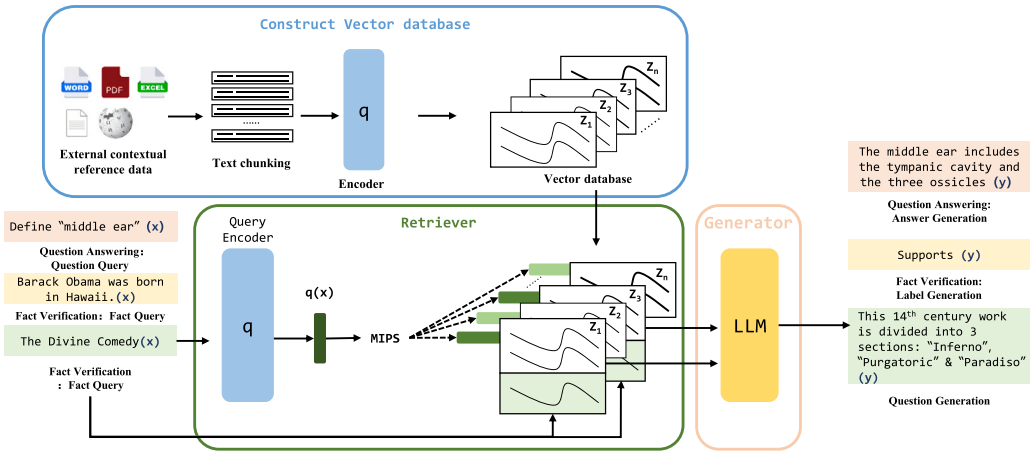

Fig. 4. Retrieval-augmented generation framework.

address these issues, Parametric Efficient Fine-Tuning (PEFT) proposes to optimize the LLMs' fine-tuning process by reducing the number of parameters to be tuned. For example, BitFit adapted to new tasks by fine-tuning a small number of key parameters of the model, rather than updating all parameters of the entire model [46]. LoRA performed parameter updates by introducing low-rank matrices, which were much smaller than the original parameters of the model, to drastically reduce the computational and storage overhead [47]. With the continuous advancement of technology, PEFT is likely to be more widely applied across multiple sectors in both industry and academia.

Retrieval-augmented generation. LLMs have been shown to learn a substantial amount of in-depth knowledge from data [48]. LLM can be used as a parameterized implicit knowledge base. But storing knowledge in the form of parameters leads to a limited ability of LLM to access and manipulate knowledge. To this end, researchers have proposed the Retrieval-augmented Generation (RAG) framework [48,49]. RAG introduces a vector retriever and external databases, addressing the above issues by combining parametric and non-parametric memory (retrieval-based memory). The knowledge in external databases can be easily modified and expanded. And, it can also provide citations for LLM responses, increasing the interpretability of LLM outputs and enhancing trust in the results produced by LLMs. As shown in Fig. 4, RAG consists of four components. (1) Vector database: The original data files are usually in TXT, PDF, and other formats. Since the length of the original document is too long and cannot be put into the context window of LLMs, the data must be divided into smaller chunks, embedding model transforms each chunk into vectors. Finally, the vectors are stored in the vector database. (2) Retrieval: Embedding the user query into the same vector space as the chunks in the vector database using an embedding model. Retrieve the relevant chunks in the Vector database using the similarity measure algorithm and return the top-k chunks with the highest similarity to the user query from the Vector database. (3) Augment: Put the user's query and the retrieved chunks (additional context) into a prompt. (4) Generation: Input the augmented prompt into the LLM for generation.

RAG is the most lightweight of the methods for constructing domain-specific LLMs. In actual industrial production scenarios, RAG can realize better control of data and guarantee data security and privacy by setting role permissions and security access in the database [50].

Agent. Agents are increasingly seen as catalysts for the development of Artificial General Intelligence (AGI) [41]. Agent refers to an autonomous system powered by an LLM, with the LLM serving as the central controller, granting the agent human-like decision-making capabilities. It can interact with the environment, carry out tasks, and learn from them, and can also operate within the human-centric manufacturing framework of Industry 5.0. The main idea is to enable LLMs to acquire key human abilities, such as memory, tool use, planning, and autonomous action, allowing them to behave like humans and efficiently complete various tasks in environments. Specifically, the overall structure of the agent is illustrated in Fig. 5.

The agent consists of four modules: memory, planning, action, and tool [51]. The memory module is responsible for recording past experiences and interactions, providing the agent with historical context to help make more informed decisions. The planning module allows the agent to develop future action plans based on its memory and the present environment. The action module is responsible for converting the agent's planning into actions, ensuring that the agent can carry out tasks in the real world, such as adjusting machinery limits and coordinating production lines. The tool module allows the agent to use external tools during task execution, such as search engines, calculators, and even IoT devices. This module enhances the agent's problem-solving capabilities. Among these modules, the memory module influences the planning module, and these three modules collectively affect the action module. Their interactions enable the LLM to mimic human thinking and behavior, allowing it to perform tasks efficiently in various environments [52]. Fig. 5 illustrates the structure and interaction mode of the LLM-based agent. In this system, the LLM serves as the core component, interacting with different types of external entities.

Human users: Users, as supervisors, partners, and decision makers of the agent, interact with the agent, providing guidance or supervision to ensure that it aligns with human goals and values, which aligns with the core principles of Industry 5.0. Environment: Agent can interact with external environments (e.g., production lines, inventory systems) to collect the necessary information and data, enabling better decision making and supporting intelligent manufacturing systems. Other agents: Agents can work together and form cooperative relationships so that they can work together to accomplish more complex tasks in the production process [53]. Developer: The developer of the agent designs and implements the various components of the agent, providing infrastructure support for the agent system.

In manufacturing automation system (MAS), an LLM-based agent can inject new vitality into traditional manufacturing processes. Firstly, LLM-driven agent can enhance the flexibility and intelligence of production lines. By interacting with the environment and continuously learning in real-time, agents can adjust production schedules and workflows to adapt to varying production needs and unforeseen circumstances. This not only boosts production efficiency but also improves the flexibility of MAS. Furthermore, LLM-based agents can work in close coordination with a human operator, ensuring automation in production processes while also accounting for human factors. For instance, agents can engage with operators during quality control, offering real-time feedback and recommendations. In summary, agent-based

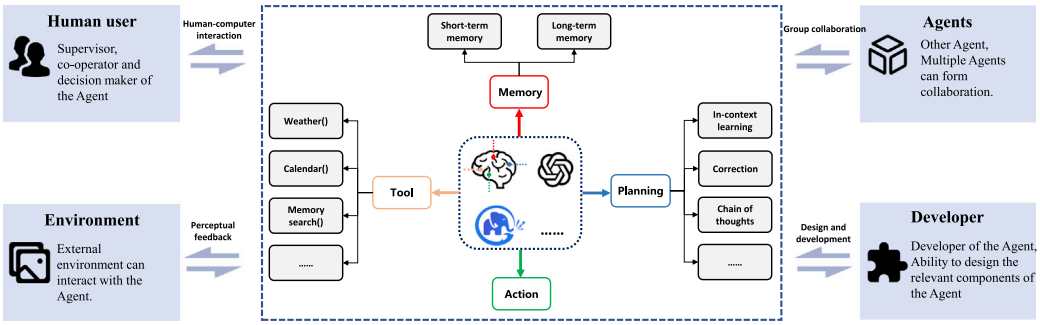

Fig. 5. The overall structure of the agent.

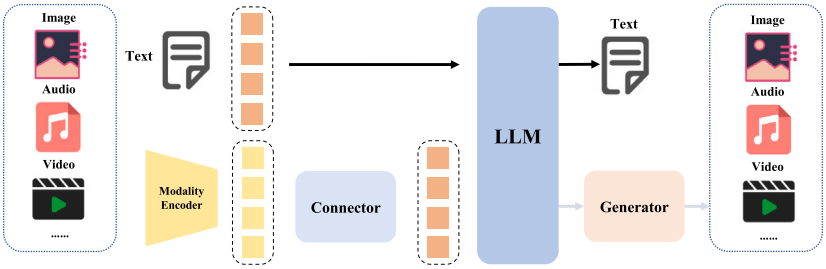

Fig. 6. The structure of MLLM.

MAS can not only raise the level of production automation but also strengthen the adaptability and flexibility of manufacturing systems. By closely collaborating with human operators, agent-based MAS can maintain production efficiency while fully accounting for human factors, fostering the intelligent, human-centered, and personalized nature of production.

Multimodal Large Language Models. Although LLMs have shown strong performance on natural language processing tasks, they can only understand discrete text. In visual tasks, traditional LLMs still have inherent limitations, which have led to the emergence of the new field of multimodal language models (MLLM) [54]. Formally, it refers to LLM-based models that possess the ability to receive, infer, and output multimodal information. As shown in Fig. 6, a typical MLLM consists of four modules that include a pre-trained Modality Encoder, a pre-trained LLM, a connector, and a generator. Unlike traditional LLMs, multimodal LLMs can simultaneously utilize multiple modalities, such as verbal and visual, to generate multimodal outputs, as well as combine text and image prompts to provide a more nuanced response. This takes the interaction beyond the text. The application of MLLM in Industry 5.0 has great potential, especially in human–machine collaboration, intelligent decision support, and process planning. Specifically, in industrial production lines, operators can use MLLM to simultaneously process vision data (e.g., images of the production line status captured by a camera in real time) and natural language-based suggestions (e.g., LLM-generated operation steps, troubleshooting, equipment optimization suggestions, etc.) to make quick decisions and adjustments, increase productivity, and improve human interaction. Therefore, MLLM, as an innovative technology, not only extends the application scope of LLM in Industry 5.0 but also provides new possibilities for the development of future intelligent manufacturing [55].

5. Synergizing LLMs with industry 5.0 enablers

The aim of Industry 5.0 is to tightly integrate humans and technology, achieving an industrial system focused on human-centricity, sustainability, and resiliency through essential enabling technologies.

As shown in Fig. 7, these enablers comprise human–robot interaction (HRI), industrial internet of things (IIoT), digital twins, extended and augmented reality (XR/AR), next-generation mobile communications (5G/6G), blockchain, and knowledge graphs. These enabling technologies play a unique role within the framework of Industry 5.0. However, in practice, they face fragmentation and complexity. For example, language barriers between robots and humans may affect collaborative work efficiency. The massive amount of data generated by IIoT devices is difficult to process and utilize in real time. And, digital twin systems require high-quality semantic models for prediction. This section will explore how LLMs collaborate with these Industry 5.0 enablers to provide innovative solutions for industrial scenarios.

5.1. LLMs for human–robot interaction

The integration of large language models with Industry 5.0 is redefining human–robot interaction. LLMs, with their powerful natural language processing capabilities [56], multimodal integration abilities, and contextual reasoning skills, provide robots with new tools, enabling breakthroughs in human–robot interaction.

As shown in Fig. 8, the LLM serves as the core component that connects and coordinates the relationship between humans, robots, and the environment. Using the code generation capabilities of LLM, humans can influence the behavior of robots through natural language or interventions. Robots realize active cognition and feedback information by sensing the environment. In this framework, LLM not only undertakes language comprehension and generation tasks but also plays a key role in cognition and decision support, thus enhancing the level of intelligence and interaction efficiency of the entire HRI system.

For example, Liu et al. proposed a system called ‘REFLECT’, which converts multisensory observations into a hierarchical summary of the robot’s past experiences and queries LLM progressively for failure explanation. The generated explanation can then guide a language planner to correct the failure and complete the task. This ability allowed robots to learn from past tasks, thereby improving the efficiency and reliability of task completion in industrial scenarios [57]. Wang et al. present an LLM-based vision and language cobot navigation framework for HSM to further enable the robot to fetch tools for humans. This method promotes improvements in ergonomics and safety for human operators in industrial environments [58]. Park et al. proposed a framework to allow human workers to interact with construction

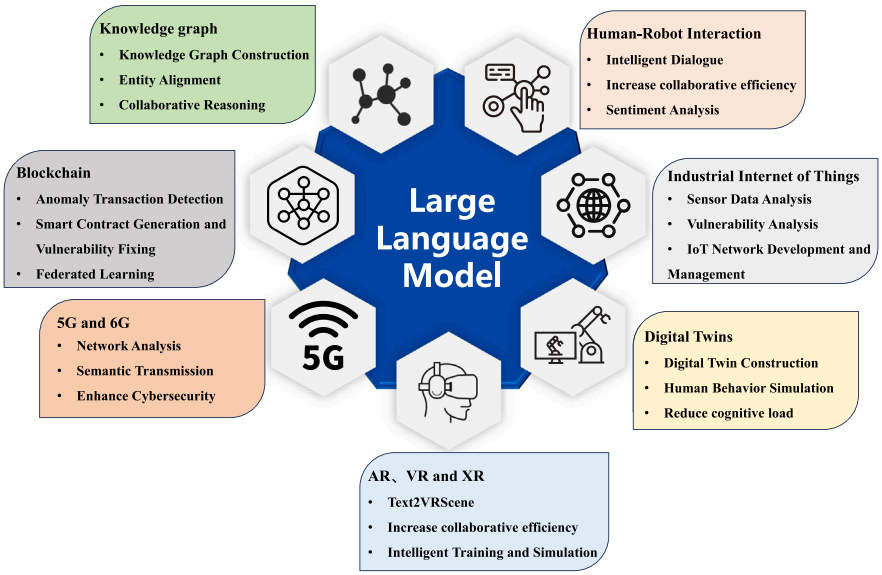

Fig. 7. Synergizing LLMs with Industry 5.0 enablers.

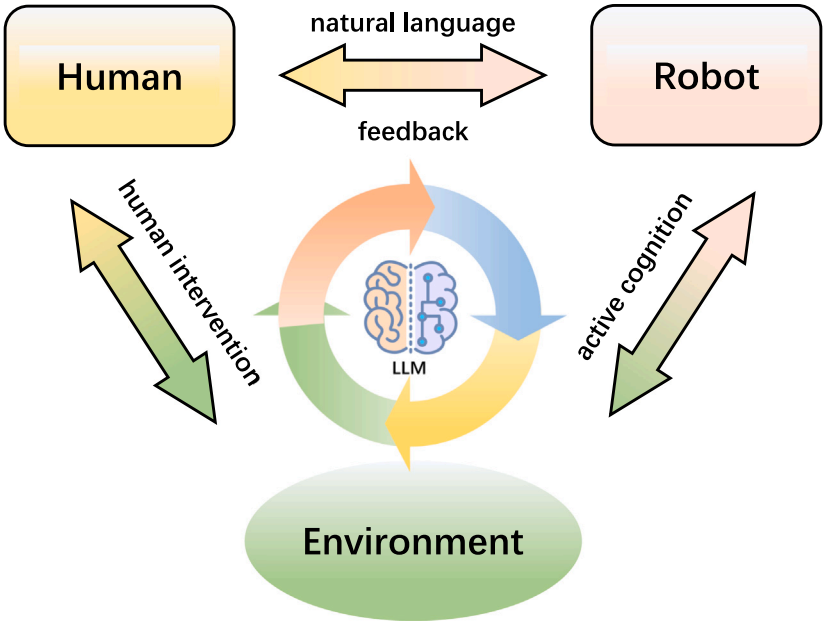

Fig. 8. Human–robot interaction augmented by LLM.

robots based on natural language instructions for pick-and-place construction operations. This collaborative paradigm, combining human flexibility with robotic execution capabilities, was the core vision of Industry 5.0 [59]. Frameworks such as RoboGPT [60] and DAIM-HRI [61] fully utilized the language parsing capabilities of LLMs to achieve efficient control of robots. LLMs enable non-technical users to operate complex industrial robots through simple dialogue, significantly bridging the technological gap and enhancing collaboration efficiency between workshop workers and robots. Additionally, when LLMs are integrated into manufacturing systems, they can optimize the manufacturing process by dynamically adapting to task changes and handling faults. For example, the LLM-based human–robot collaborative assembly framework enhances the ability of industrial robots to handle small-batch customization and dynamic task allocation by integrating sensors and task control mechanisms. It can meet Industry 5.0's requirements for personalized and flexible production [62]. In the field of multimodal interaction, LaMI demonstrated LLMs' ability to integrate multimodal inputs such as speech, vision, facial expressions, and graphical interfaces, enabling robots to engage in more dynamic and emotional interactions with humans. This capability is particularly crucial for the demands of complex human–robot collaboration environments in Industry 5.0 [63]. In task execution, the SayCan robot used LLMs to parse natural language commands, converted them into specific operational instructions, and generated execution plans based on environmental perception. It significantly enhanced the robot's flexibility in dynamic environments [64]. Furthermore, LLM-BRAIN used an LLM-based AI-driven behavior tree generation technology, enabling it to quickly translate human commands into the execution logic for complex tasks [65]. This technology not only supports the operation of mobile robots and drones in Industry 4.0 scenarios but also drives the shift of robotic systems in Industry 5.0 from executing predefined tasks to real-time decision-making. CognitiveOS used generative artificial intelligence to enable robots to autonomously handle complex tasks and dynamically manage manufacturing processes [66]. Industry 5.0

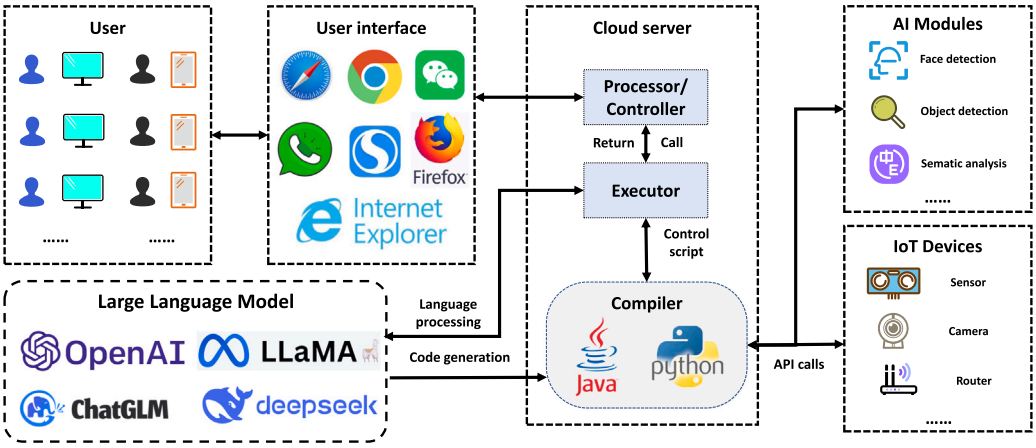

Fig. 9. LLMs enhanced IoT.

aims to achieve highly intelligent and personalized human-centered collaboration, and LLMs bridge the gap between humans and robots in language, perception, and operation by enabling robots with capabilities like natural language understanding, contextual awareness, and task reasoning. The introduction of LLMs marks the shift of robots from passive execution tools to active collaborative partners, providing a solid technological foundation for the full realization of Industry 5.0.

5.2. LLMs for industrial internet of things

In the industrial internet of things, the massive data generated by devices, sensors, and systems often requires efficient processing and analysis. LLMs can play a significant role in this process. On one hand, with their powerful natural language processing capabilities, LLMs can analyze textual data from devices, maintenance logs, etc. On the other hand, by integrating real-time sensor data, LLMs can provide intelligent diagnostics, fault prediction, and optimization recommendations. As shown in Fig. 9, it shows an Industrial Internet of Things (IIoT) system architecture enhanced by LLM. Users can interact with the system through a variety of user interfaces (e.g., browser, WeChat, etc.). The user input is transmitted to the cloud server for processing. The cloud server contains processors/controllers and executors that are responsible for scheduling requests and returning results. The system integrates several mainstream large language models (e.g., OpenAI, LLaMA, ChatGLM, DeepSeek) for performing natural language processing and code generation tasks. The generated code is compiled by a Java or Python compiler and used in the form of control scripts to invoke AI modules (e.g., face recognition, target detection, semantic analysis) and various IoT devices (e.g., sensors, cameras, routers, etc.). The framework enables intelligent and flexible interactions between users and complex IoT devices.

An et al. introduced IoT-LLM, a framework that integrates IoT sensor data with LLMs. This approach enhances the LLM's perception and reasoning capabilities in the physical world [67]. Mo et al. also proposed IoT-LM, a new large multisensory language model with multisensory perception and natural language interaction capabilities over a spectrum of IoT modalities and applications [68]. Zhong et al. proposed a collective intelligent agent system for the IoT (CASIT). This system utilized the collaboration of multiple LLMs to perform sensor data analysis tasks and generates reports through summarization and memory mechanisms. Additionally, operators can interact with IoT data through natural language, making the data understanding and operation process in IoT systems more efficient and intelligent, significantly improving human–machine interaction efficiency [69]. Cui et al. incorporated LLM into a task planning and scheduling framework called LLMind to enable control of IoT devices. LLMind used LLM as a planner, utilizing LLM's code generation capabilities and semantic understanding to enable humans to interact with IoT devices and to enable IoT devices to communicate and collaborate to perform complex tasks. This approach aimed to massively simplify IoT network development and management, thereby increasing productivity [70]. Li et al. proposed an innovative forecasting and health management framework (PHM) that combines blockchain technology with LLM. The immutable and transparent nature of blockchain was utilized to ensure data integrity and security across the IIoT ecosystem. Using LLM to analyze IoT data for maintenance, the framework effectively improved the accuracy of fault prediction and IIoT overall resilience against cyber–physical threats [71]. LLM shows great potential in optimizing IIoT tasks, especially in automated data analytics and decision support etc., but as the number of IIoT devices proliferates, security is gradually becoming a challenge that cannot be ignored. IoT devices are exposed to various cyberattacks and security vulnerabilities due to their openness, distribution, and high degree of integration with the network. Traditional vulnerability detection often relies on manual analysis or fixed rules, facing issues of low efficiency and limited coverage. Recently, LLM-based methods for IoT vulnerability detection have been widely explored. Ferrag et al. presented SecurityBERT, a novel architecture that leverages the BERT model for cyber threat detection in IoT networks [72]. Hibiki et al. proposed a method for generating initial seeds for IoT device fuzzing by effectively utilizing LLMs. This method generated initial seeds for efficient fuzzing of the target of an IoT device by inputting only the specifications of the IoT device and the name of the vulnerability to be inspected into LLMs [73]. Wang et al. proposed an IoT anomaly detection solution based on the content in the data traffic from specific type of IoT devices. The method utilized the embedding layer of LLM to obtain an embedding of network packets and used the embedding to train a deep learning model for IoT traffic anomaly detection, which achieves the unification of large and small models and improves the accuracy of network anomaly activities in IoT environment [74]. Overall, the introduction of LLM not only provides great help in IoT data analysis, but also brings new breakthroughs in security, IoT end device control, and semantic communication. With the continued increase in the number of IIoT devices and the development of LLM, the research and application of LLM in IoT is expected to further deepen in the future, especially in the areas of intelligent control, automated decision-making, and security.

5.3. LLMs for digital twins

With the continuous development of AI and the industrial sector, the integration of LLMs and digital twins is opening up a series of innovative applications. Digital twins are a technology that simulates and analyzes physical systems by creating virtual replicas [75]. When combined with LLMs, they demonstrate immense potential. On one

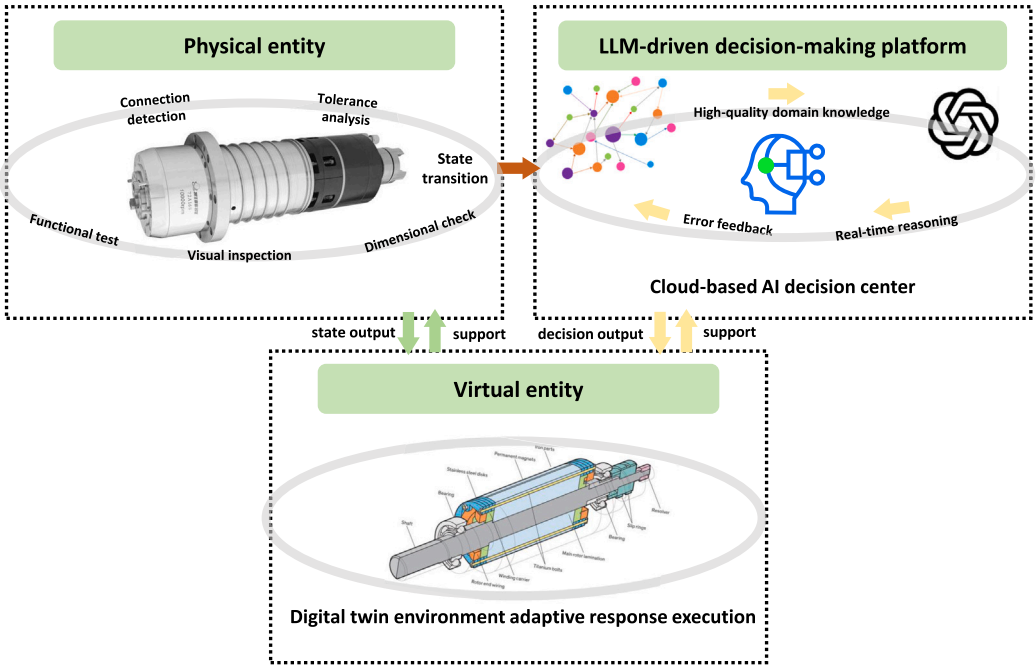

Fig. 10. LLM-enabled digital twins.

hand, LLMs can mimic human thinking processes to understand and solve complex problems. People can track and comprehend their work, intervening at critical moments. On the other hand, LLMs are able to transform from solution implementers to proposers, implementers, and maintainers of their solutions [76]. Under reasonable supervision, LLM will actively change the connection and interaction between the physical and virtual worlds, becoming a key component of the digital twin. Through its powerful data parsing and semantic parsing capabilities, LLM is able to provide deeper decision support for digital twins and help drive the intelligent operation of the system.

Fig. 10 illustrates a digital twin system framework driven by LLM. The system consists of three components: physical entity, virtual entity, and LLM-driven decision-making platform. Physical entity refers to a real-world industrial component or device, whose operational status is comprehensively perceived through various detection methods, including functional tests, connectivity detections, and more. Changes in the state of the physical entity trigger state transition processes and are transmitted in real time to the virtual and decision-making platform, providing data support for subsequent analysis and response. Virtual entity is a digital mapping of a physical entity. It has an adaptive response capability and is able to perform real-time simulation and feedback based on the state changes of the physical entity and the instructions from the decision center. The LLM-driven decision-making platform is the core component of the system. The platform is capable of combining input state information for semantic understanding and decision support, providing intelligent regulation recommendations for both virtual and physical entities, and realizing more efficient system response and troubleshooting.

Researchers are also actively exploring ways to utilize LLMs' code generation and language processing capabilities to simplify the creation of digital twins and enhance their interactivity and accessibility. Li et al. presented ChatTwin, which utilizes the power of LLM to facilitate the generation of comprehensive scene description documents specifically designed for the digital twin [77]. In terms of assisted decision making, Li et al. utilized digital twin technology to incorporate real scene data into LLM and proposed a cognitive digital twin prototype system of Human–robot collaboration manipulation, known as HRC-CogiDT [78]. HRC-CogiDT utilized digital twin technology to integrate LLM into the decision-making loop of the HRC system. HRC-CogiDT can quickly perceive scene changes and make high-level decisions based on different task requirements, such as task planning, anomaly detection, and schedule reasoning. Luo et al. presented ChatTwin, which utilizes LLM to create a more accessible and intuitive user interface for digital twins. By providing a more accessible, intuitive, and user-friendly interface, LLM-driven interactions reduce the learning curve required to interact with intricate infrastructure digital twin systems, empowering decision-makers at different levels, from engineers to policymakers, with timely and contextually relevant insights [79]. Jam et al. presented a conceptual architecture design aimed at enhancing interactions with cognitive digital twins of countries through an LLM agent [80]. LLM-based agents enable users to access and interact with digital twins naturally. In addition, through advanced natural language processing and prompt engineering, agents can understand and process complex queries and translate them into actionable insights for the digital twin model, increasing transparency in decision making and reducing the cognitive load on the user. Xia et al. proposed a multi-agent framework based on LLM that applies LLMs to automate the parametrization of simulation models in digital twins. The framework used specialized LLM agents responsible for observation, reasoning, decision-making, and summarization, enabling them to automatically interact with the digital twin simulation dynamics [81]. Sun et al. proposed an LLM-driven DT multi-agent architecture, integrating LLMs into the digital twin system. The system utilized LLM-driven agents to integrate heterogeneous data from multiple sources coming from the physical system and achieved global awareness of the temporal attributes of the physical system through interactions between multiple agents [76]. In addition, LLM can also assist in building human-in-the-loop (HITL) digital twin systems by simulating human behavior and thermal preferences. Yang et al. developed an LLM-powered digital twin to simulate user behavior in a building and apply it to temperature control [82].

5.4. LLMs for XR

Industry 5.0 emphasizes a human-centered approach, and immersive technologies such as AR are key tools to achieve this goal. The combination of immersive technologies such as AR and VR with LLM

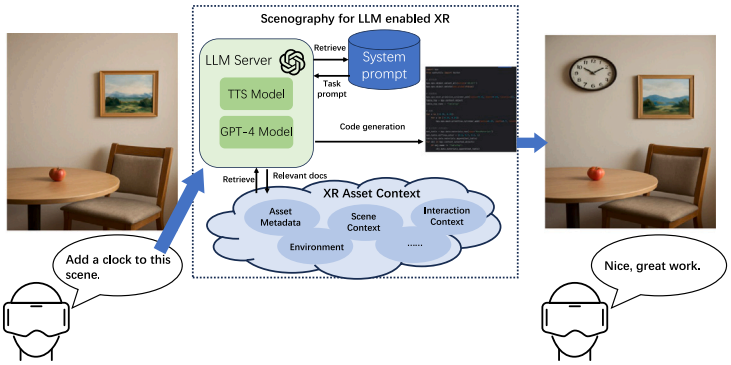

Fig. 11. LLM-enabled XR.

is driving the industrial sector towards greater intelligence and personalization. The Fig. 11 shows a framework of an LLM-enhanced XR system. The framework can process voice/audio input and convert voice context into text. It can perform natural language interaction through the GPT-4 model. In addition, based on understanding the user input and the current XR scene information, the system retrieves the relevant XR Asset Context and system prompts, and then generates precise code to drive the semantic enhancement and content update of the XR scene. The generated code can be directly used in the XR runtime environment to dynamically add, modify, or reconstruct virtual objects, thereby achieving intelligent enhancement and adaptive updating of the scene.

Giunchi et al. proposed DreamCodeVR, which is an AI-code-based behavior generator triggered by the user's speech. It can change the appearance, behavior, and state of a live running Unity VR application [83]. Yin et al. developed Text2VRScene, a VR scene generation system based on well-designed prompts. Text2VRScene generates VR scenes without clear boundaries from simple user-prompted text, and shows creativity through the dynamics of different 3D objects [84]. Fernanda et al. presented LLM for Mixed Reality (LLMR), a framework for the real-time creation and modification of interactive Mixed Reality experiences using LLMs. After evaluation by Unity developers, LLMR was intuitive and easy to use, capable of generating the required outputs for iterative user needs without much manual scripting, greatly simplifying the workload of Unity developers [85]. In XR applications, LLMs can also process and analyze complex data, providing precise operational recommendations and supporting intelligent production processes, as well as personalized customization needs. This human–machine collaboration model promotes the deep integration of the digital and physical worlds, driving Industry 5.0's focus on more efficient, flexible, and human-centered production approaches, resulting in a new work experience and improved productivity. Xu et al. proposed a new method to simplify and optimize AR-based industrial maintenance tasks using LLM. By integrating LLM into an AR-based operating system, the system can provide real-time guidance based on user input, device status, and even potential user behaviors, rather than relying on explicit pre-programmed instructions [86]. To allow operators to monitor the working state of the robot after issuing task instructions in high-level language and demonstrate the robot's motion plan safely, Zhao et al. designed an AR-assisted human–robot interaction system called SeeIt [87]. This system displayed the working state of a robotic arm and predicted its motion trajectory regarding LLM output. The AR interaction module allowed operators to choose among potential running trajectories and adjust erroneous motion plans, leveraging human decision-making to enhance the system's intelligence. Chen et al. [88] proposed an interactive AR/VLM-based query system, named Visual Construction Safety Query (VCSQ) for providing on-demand safety information to workers by processing on-site images. Fan et al. introduced an innovative training system for the metal printer, leveraging the synergy of advanced VLM and AR within the digital twins framework [89]. The AR module significantly enriched the user's learning experience by providing a realistic interactive environment that closely mirrors actual manufacturing conditions. The VLM module dealt with multimodal data and provided nuanced and contextually relevant guidance to learners. The system met the manufacturing industry's urgent need for efficient and easy-to-use training methods, paving the way for wider adoption of these technologies and more sophisticated applications.

5.5. LLMs for 5G and 6G

In the framework of Industry 5.0, 5G and 6G as key communication technologies can provide powerful support for smart manufacturing, personalized production, and human–machine collaboration, while LLM can lead to the further development and optimization of these network technologies through natural language processing techniques. The Fig. 12 illustrates four typical application scenarios of LLMs in 5G/6G communication systems, including intrusion detection, network configuration, network optimization, and semantic communication. By leveraging the powerful semantic understanding and contextual reasoning capabilities of LLMs, the system can automatically identify abnormal behaviors, interpret natural language instructions, dynamically adjust network parameters, and efficiently carry out task-level semantic communication [90].

For example, LLM can automate network optimization, analyze complex data flows, and propose optimization strategies. Kan et al. build Mobile-LLaMA by instruction fine-tuning LLaMA2-13B with network analysis data collected from publicly available, real-world 5G network datasets [91]. Mobile-LLaMA had three main functions: packet analysis, IP routing analysis, and performance analysis, enabling it to provide network analysis and contribute to the automation and artificial intelligence required for 5G network management and data analysis. Xiao et al. proposed a novel paradigm for training and applying LLM agents for task-oriented 6G physical layer automation. By employing a specifically designed two-stage continual training approach, the trained domain-adapted LLM effectively aided in understanding the upper-level requirements of users and accordingly recommended the workflow and the optimal system configurations [92]. Wang et al. proposed the novel concept, NetLLM, which integrates various technologies using ChatGPT technology to support the network management and control architecture for 6G on-demand service. The framework can also provide natural language results in edge computing and smart manufacturing, simplifying the communication between operators and machines and improving the autonomy and intelligence of the system [93]. Mekrache et al. proposed a novel LLM-centric intent life-cycle (LC) management architecture designed to configure and manage network services using natural language. The framework significantly simplified network management, eliminating the need for expertise in low-level configurations and allowing network administrators to focus on high-level network objectives [94]. Shen et al. proposed a multifunctional framework for

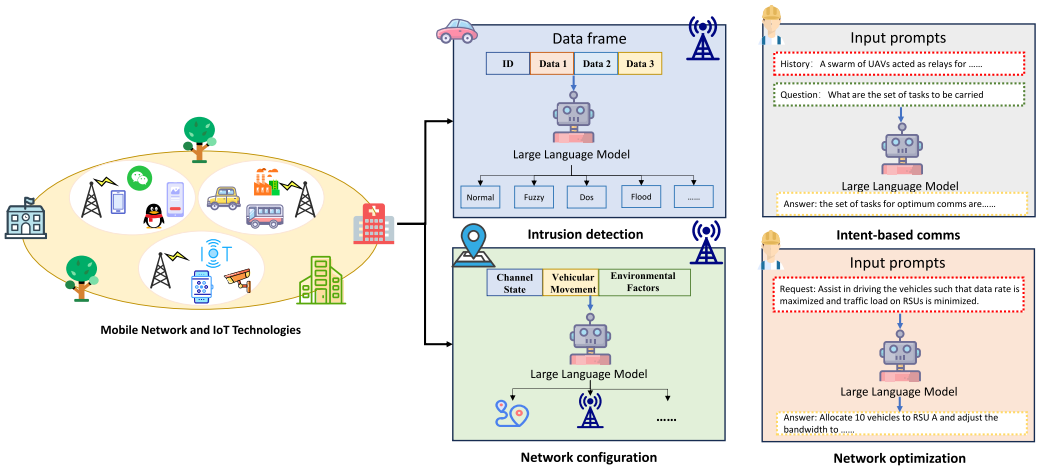

Fig. 12. LLM-enabled 5G, 6G.

the combination of IoT and LLM, where GPT is used to understand the user's requests in the form of natural language and route these requests to appropriate edge AI models [95]. Ali et al. introduced a novel framework, HuntGPT, aimed at integrating actionable, interpretable, and explainable AI in cybersecurity operations. HuntGPT combines the advanced capabilities of GPT-3.5-turbo with a user-friendly dashboard and adeptly elucidates the latent details of detected anomalies [96]. Kaheh et al. presented a GPT-4-based conversational agent called Cyber Sentinel, which can be used for streamlining cybersecurity. Cyber Sentinel helped security analysts perform a range of cybersecurity tasks from querying a cyber threat intelligence feed to managing an SIEM's configuration [97].

5.6. LLMs for blockchain

With the continuous development of blockchain technology and LLM, their integration provides strong technical support for advancing the realization of Industry 5.0 [98]. As is shown in Fig. 13, LLM simplifies the development of blockchain platforms and enhances user experience through its code generation and language processing capabilities. On the other hand, blockchain provides secure data protection and management for LLM through its decentralized and transparent data management capabilities.

In Industry 5.0, how to efficiently and securely manage and process massive amounts of data to guarantee transparency and reliability in the production process is a critical challenge. The combination of blockchain and LLM is an innovative solution to these challenges. Smart contracts are a key application of blockchain technology [99]. LLM is used to automate smart contract generation and vulnerability repair. However, developing smart contracts for blockchain platforms is a time-consuming and labor-intensive task due to the specificity of blockchain platforms. Nenad et al. proposed an approach for automated model-driven smart contract generation leveraging the LLM-based ChatGPT to reduce the time and effort required for their development [100]. Ortu et al. proposed an automated program repair approach based on LLMs for solidity smart contracts, which trains on vulnerable code snippets and manual patches. The results showed that the LLM effectively fixes specific vulnerabilities [101]. Ma et al. proposed iAudit, which combines fine-tuning and LLM-based agents for intuitive smart contract auditing with justifications. This method improved the performance and stability of smart contract vulnerability detection [102]. In addition, LLM has significant applications in the security of blockchain networks. LLM can effectively detect abnormal transactions and automatically trigger the pause mechanism of smart contracts to prevent malicious attacks and improve the intrusion detection capability of blockchain networks in industrial

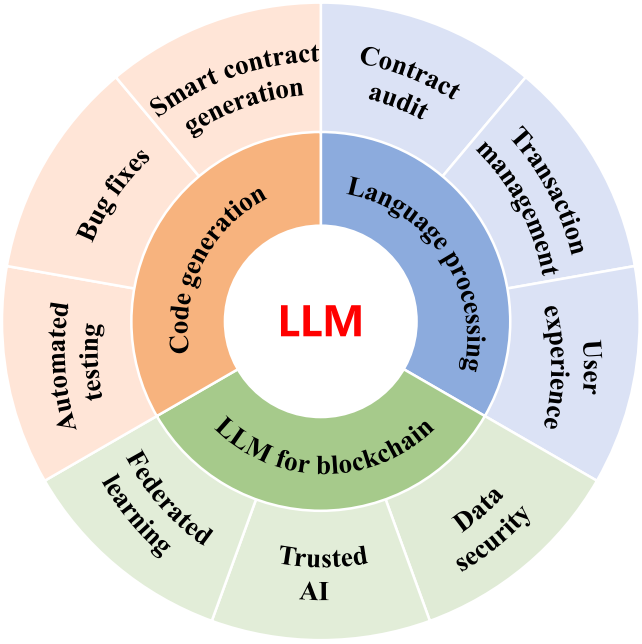

Fig. 13. The application of LLM on the task of blockchain.

applications [103]. Additionally, LLMs are used to enhance the user experience of blockchain-based applications. Mboma et al. proposed an approach to integrate LLM into blockchain platforms that aims to improve and optimize the interaction between users and blockchains by better understanding their intentions, thus paving the way for large-scale use and adoption of decentralized systems [104]. Benzinho et al. integrated a conversational agent driven by RAG-based LLM technology into the blockchain system. The system can retrieve relevant information from the blockchain through semantic search by interpreting user queries. The conversational interface allows users to interact with the blockchain system using natural language, effectively smoothing the learning curve and greatly reducing the learning threshold for users [105]. Additionally, blockchain can also provide security for LLM training and data protection. For instance, BC4LLM can realize the whole process security of LLM through the security attributes of blockchain, including a reliable learning corpus, secure training process, and identifiable generated content [106]. Zuo et al. presented a new blockchain-based federated learning framework for LLM, which achieves highly effective unlearning, enabling the selective forgetting

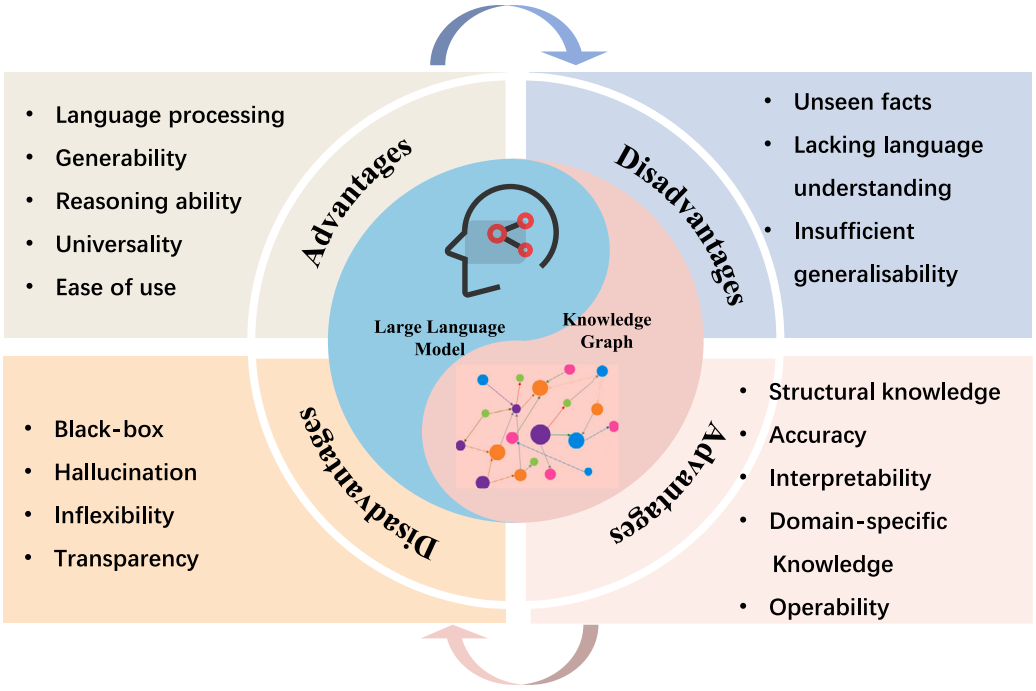

Fig. 14. LLM and KG reinforce each other.

of specific data points while preserving the model's performance on the remaining data [107]. The security and verifiability of LLM training data were ensured. These studies show that the combination of LLM and blockchain can not only promote the innovative application of blockchain technology in Industry 5.0 but also provide higher security and reliability for large-scale distributed artificial intelligence.

5.7. LLMs for knowledge graph

Knowledge graphs serve as an important tool for capturing and representing complex relationships and knowledge in industry [108]. Knowledge graphs are widely used in Industry 5.0. Industrial processes involve a large amount of heterogeneous data from multiple sources. But, knowledge graphs can effectively integrate this data to provide transparent and interpretable knowledge representation and reasoning support [109]. However, knowledge graphs encounter challenges in practical application, such as the complexity of knowledge extraction, semantic understanding, and natural language processing. These questions hinder the widespread application of knowledge graphs in the industrial field. By combining LLMs with knowledge graph, leveraging the powerful semantic understanding and generative capabilities of LLMs can significantly enhance the construction, updating, and reasoning capabilities of knowledge graphs. In the face of heterogeneous and unstructured data from multiple sources, information extraction(IE) is an important step in knowledge graph construction. Wei et al. proposed ChatIE, a multi-turn QA framework for zero-shot information extraction based on ChatGPT. Through this interactive mode, ChatIE can decompose complex IE tasks into several parts and compose the results of each turn into information extraction results [110]. Gui et al. proposed IEPile, an approach to fine-tune open-source LLMs using a schema-based instruction generation strategy, which improves the zero-shot generalization of LLMs in instruction-based information extraction [111]. Zhang et al. proposed the QA4RE framework, which aligns RE (relation extraction) with question answering. By converting input sentences to questions and possible relation types to multiple-choice options, LLMs were able to perform RE by selecting the option representing the correct relation type [112]. Zhang et al. presented AutoAlign, which is the first fully automatic method for KG alignment enabled by LLM. AutoAlign utilized LLMs to implement predicate embeddings and entity embeddings, and significantly improved the performance of the knowledge graph entity alignment task [113]. This is especially important for Industry 5.0 applications, as data from different departments or systems needs to be integrated into a unified knowledge base in order to support global decision-making from supply chain management to production optimization. Finally, knowledge graphs provide LLMs with high-quality, structured background knowledge, thereby making the generated results more accurate and suitable for real-world application scenarios. Zhou et al. proposed an industrial structure causal knowledge graph-enhanced LLM named CausalKGPT for the cause analysis of quality defects in aerospace product manufacturing [114]. Liu et al. proposed a knowledge-enhanced joint model that incorporates aviation assembly KG embedding into LLMs. This framework utilized graph-structured big data within KGs to conduct prefix-tuning of the LLMs [115]. This approach reduced the computational burden of the model and improved the accuracy of the LLM for aviation fault diagnosis. Qi et al. proposed FoodGPT, which utilizes a knowledge graph in the field of food detection to improve the accuracy and rationality of LLM-generated content in this domain [116].

The integration of LLMs and knowledge graphs was not just a simple sum, but resulted in a synergistic enhancement. As shown in Fig. 14, this combination greatly enhanced the performance of LLMs within the food domain and informed the creation of domain-specific LLMs. Knowledge graphs can provide LLMs with structured and high-quality domain knowledge, thereby improving the accuracy and reliability of LLM output. And, LLMs can bring stronger natural language processing capabilities to knowledge graphs, making knowledge acquisition and reasoning more convenient. This integration has a particularly strong potential in industrial scenarios, such as equipment maintenance, production optimization, and supply chain management, where various applications can benefit from this synergy. In the future, with the continuous development of LLMs and knowledge graph technologies, their synergy will be further deepened to drive the evolution of industrial systems towards more intelligent, efficient, and human-centered [117].

6. Applications scenarios of LLM in industry 5.0

With the concept of Industry 5.0, the manufacturing industry is not only pursuing productivity and automation, but also emphasizing

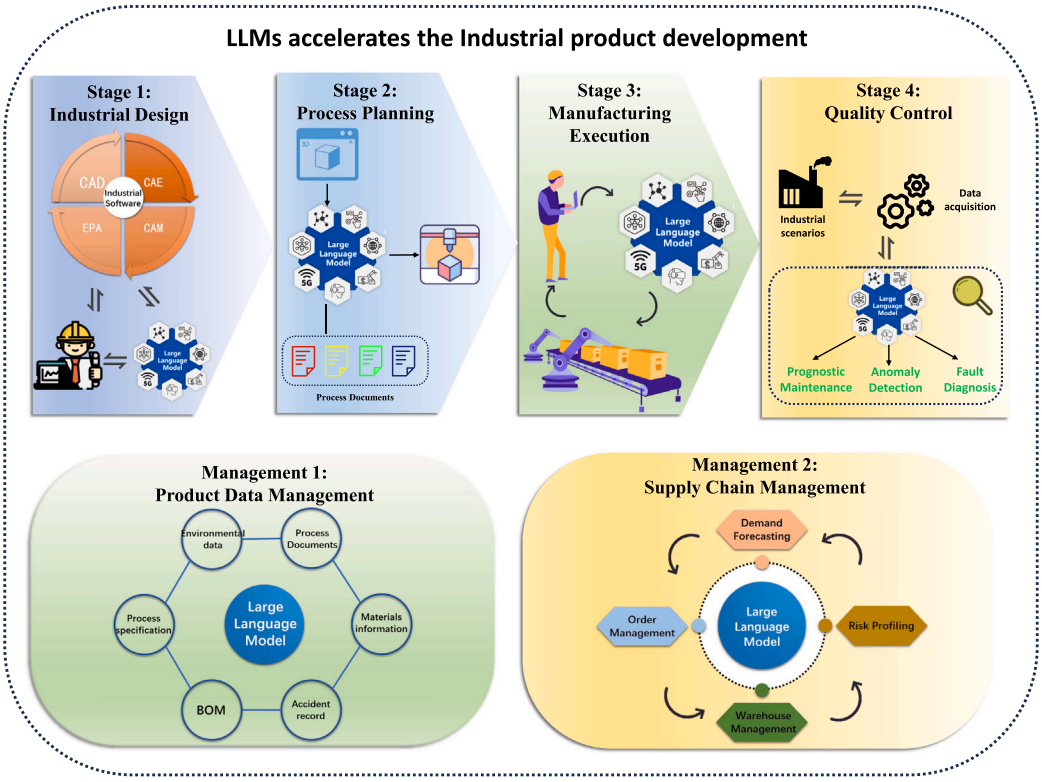

Fig. 15. Applications scenarios of LLM in Industry 5.0.

human–machine collaboration, as well as personalized and customized production [118]. In this context, LLMs, with their powerful natural language processing and generation capabilities, have become a crucial technology supporting Industry 5.0. LLMs can not only improve the efficiency of the production process by optimizing information flow and decision support, but can also promote deep collaboration between people and intelligent systems, enabling the perfect integration of intelligent manufacturing and humane design. As shown in Fig. 15, this section will explore the application of LLMs in smart manufacturing, with a focus on six key areas: industrial design, process planning, manufacturing execution, quality control, product data management, and supply chain management. In the industrial design phase, LLMs can take advantage of their natural language processing capabilities to help designers generate customized design solutions. LLMs can also facilitate real-time feedback and interaction between humans and machines. In the process planning phase, LLMs are able to intelligently handle complex production process data and automatically generate and optimize process flows, making production planning more flexible and adaptable to meet the needs of customized production. In the manufacturing execution and production quality control phases, LLMs can analyze production data to provide real-time decision support, helping operators and managers make more accurate judgments, ensuring product quality, and precise execution of production. In product data management and supply chain management, the application of LLMs can optimize information flow, improve data transparency, and ensure the flexibility and responsiveness of the entire supply chain to cope with rapid market demand changes.

Through the analysis of these application scenarios, this section aims to demonstrate how LLMs, within the Industry 5.0 framework, drive intelligent upgrade and collaborative optimization of various processes, injecting new vitality into traditional manufacturing and facilitating industrial transformation in Industry 5.0. Next, we will explore each application scenario in detail and analyze how LLMs play specific roles in areas such as industrial design, process planning, and manufacturing execution.

6.1. LLMs for industrial design

Industrial design, as the upstream of the product development process, plays the role of constructing the initial form and function of the product. In the context of Industry 5.0, LLM is emerging as one of the key technologies that will change the way traditional design software works. Traditional CAD, CAE, and EDA software, while playing an important role in the engineering design, simulation, and optimization process, rely on specialized knowledge and sophisticated operational knowledge for their use. For non-professional and small enterprises, this often creates a technological barrier, limiting their potential for efficient design and innovation. The introduction of LLMs not only accelerates the design process but also enhances user interaction with the software through natural language interfaces. In the field of computer-aided design (CAD), LLMs can enhance user experience by simplifying the design process. For example, GenCAD [119] and Img2CAD [120] can generate CAD models from images. Query2CAD [121] used natural language to generate executable CAD macros using LLM without supervised data and additional training. Xu et al. proposed CAD-MLLM, which can generate parametric CAD models based on textual descriptions, images, point clouds, or any combination of these inputs, thus facilitating the use by non-expert users [122]. Deng et al. proposed a framework for chaining design requirement analysis, engineering computation, and model creation with LLM for CAD workflow automation. The framework can alleviate some of the human burden in the design process [123]. Timo et al. proposed CADGPT, an innovative plugin integrating NLP with Rhino3D, enhancing 3D modeling in computer-aided design environments [124]. CADGPT simplified the CAD interface with LLM, enabling users to perform complex 3D modeling tasks through natural language.

In the field of Electronic Design Automation (EDA), LLM primarily optimizes design workflows, auto-generates scripts, and simplifies user interaction with complex tools through natural language interfaces. In assisting engineers in code debugging, Wu et al. proposed an LLM-powered Autonomous agent for EDA, which enables a conversational

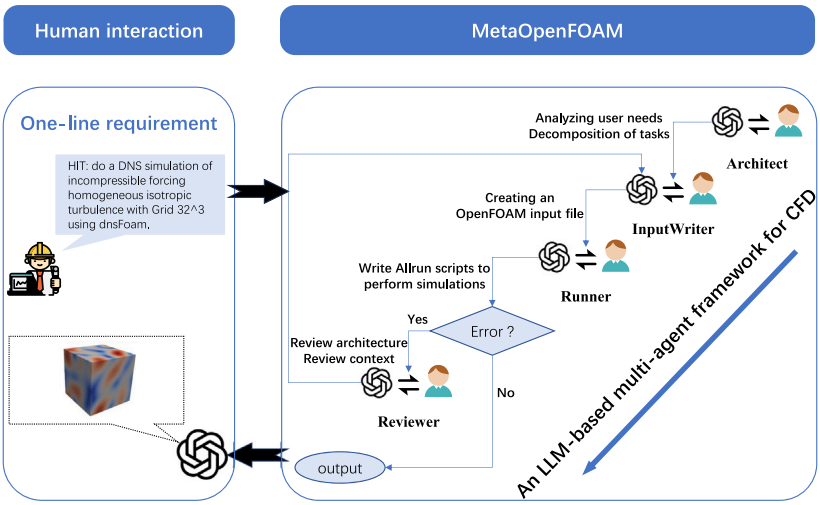

Fig. 16. An example of LLM enhanced industrial design [131].

interface for designers to interact with the design flow. Technically, ChatEDA integrated a fine-tuned AutoMage, which orchestrates the design flow through task decomposition, script generation, and task execution [125]. In order to improve the efficiency of EDA design engineers, Shi et al. proposed Ask-EDA, which utilizes LLM, hybrid RAG, and abbreviation de-hallucination techniques to provide real-time design guidance to engineers [126]. Verilog is a hardware description language (HDL) widely used in electronic design automation software. It is used to describe and model the behavior and structure of digital circuits. Xu et al. introduced a systematic automated debugging framework named MEIC. The framework demonstrated that it is feasible to employ the LLMs for the purpose of debugging Verilog code, encompassing both syntax and function errors [127]. Thakur et al. presented AutoChip, the first feedback-driven fully-automated approach for utilizing LLMs to generate HDL. It combined LLMs with the output from Verilog compilers and simulations to iteratively generate Verilog modules [128].

In the field of Computer-Aided Engineering (CAE), although current research is still in its early stages, LLMs have already demonstrated their application potential. Sun et al. proposed an intelligent heat treatment process design and simulation assistant system based on LLM, named Chat-IMSHT. Chat-IMSHT can not only impart knowledge and recommend processes, but also optimize the interaction between humans and Computer-Aided Engineering software [129]. Kumar et al. proposed MyCrunchGPT, which integrates the entire phase of scientific machine learning under the umbrella of LLM. LLMs played the role of a conductor orchestrating the entire workflow of SciML based on simple prompts by the user [130]. The application of LLM in the field of industrial design is driving the change in traditional engineering design software.

Example I: As an important branch of Computer-Aided Engineering (CAE), Computational fluid dynamics (CFD) plays a crucial role in industrial design. By accurately simulating fluid flow, heat transfer, and multi-physics coupling, CFD-related software such as openFOAM, Fluent, and others enables engineers to optimize product geometry, enhance aerodynamic performance, and improve thermal management. For example, in aircraft and automotive design, CFD is used to analyze gas flow and optimize the shape of the body or wing to reduce drag and noise. Integrating LLM technology into these software programs not only lowers the barrier to use but also makes the simulation setup easier. As shown in Fig. 16, Chen et al. present a MetaOpenFOAM, which is a novel multi-agent collaboration framework. It aims to complete CFD simulation tasks with only natural language as input. These simulation tasks include mesh preprocessing, simulation, and post-processing [131]. The proposed MetaOpenFOAM utilizes MetaGPT's assembly line paradigm, which assigns different roles to different LLMs to effectively decompose complex CFD tasks into manageable subtasks. MetaOpenFOAM is divided into four primary roles. The architect's role is responsible for parsing requirements in natural language form and translating them into actionable tasks. The inputwriter role focuses on generating and refining the necessary input files for CFD simulation. The Runner executes the CFD simulation using OpenFOAM. The Reviewer's role is to analyze any errors that occur during the simulation, identify the relevant file that caused the error, and report these findings back to the InputWriter.

6.2. LLMs for process planning

As a key link connecting product design and manufacturing, process planning is important for cost reduction and quality improvement and has direct impacts on all industrial manufacturing activities. As shown in Fig. 17, LLM, which stands for AI, is accelerating change in this field with its data analysis and generation capabilities, bringing unprecedented innovation to process planning.

In traditional process planning, engineers typically need to manually compile and review complex process documents, which is not only time-consuming and labor-intensive but is also prone to errors [132]. In the context of Industry 5.0, the introduction of LLMs has brought a huge boost to this process. LLM is able to automate many tedious and repetitive tasks, such as document organization, process requirement extraction, and information integration. It can greatly reduce manual intervention and improve work efficiency by reducing error rates. Lee et al. presented a unified ILKM framework to address the complex needs of industrial applications by integrating advanced AI, ML, and LLM technologies with specialized industrial knowledge. ILKM can assimilate knowledge from multimodalities (such as text, 2D image, and 3D shapes) gathered from historical products and provide possible optimization directions for designers and engineers to enhance the performance of new products [133]. Ni et al. introduced LLM Adaptive Process Management (LLMAPM), a strategy that employed LLMs to transform user descriptions into structured manufacturing task flows to achieve accurate manufacturing process planning [134]. Leonie et al. emphasized the great potential of generative AI, such as LLMs, in the context of Industry 5.0, where users can utilize LLMs to automatically generate data for guiding manufacturing based on a range of data related to process knowledge, such as process documentation and historical experience [135]. Mandvikar et al. proposed process automation 2.0 based on LLM, which achieves seamless integration of multiple process data sources and produces well-organized and contextually appropriate content. It demonstrated the great potential

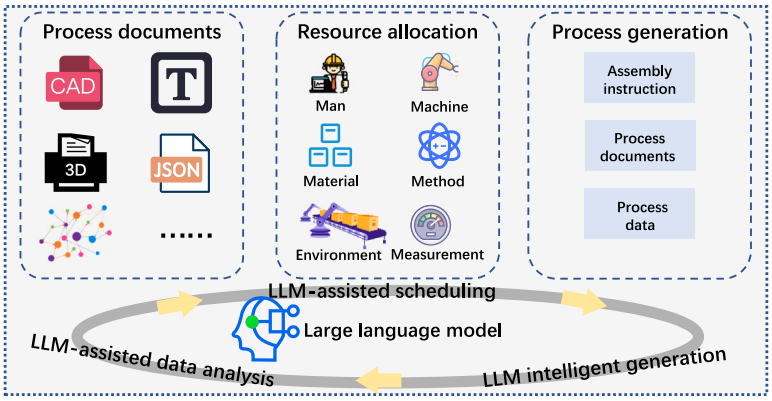

Fig. 17. LLM-enhanced process planning.

of generative AI, such as LLM for the automatic generation of process documentation [136].

The introduction of LLM not only helps users to analyze information in data such as historical process files, but also improves the intelligence of process planning. For example, Meyer et al. proposed an LLM for automatic generation of assembly instructions, a model that automatically generates detailed assembly instructions based on upstream CAD data [137]. Ahmed et al. proposed CAPP-GPT, taking CAD data instead as input and generating a process plan at different levels [138]. Tim et al. demonstrated that LLM can help process designers quickly and easily retrieve the specific products and resource modules required for components, enhancing the efficiency of the process design [139]. Holland et al. demonstrated the feasibility of LLM-based process planning agents, providing adaptive planning capabilities to nonexperts. The framework successfully addressed the tasks of job selection, process chain setup, cycle time estimation, and resource allocation in the area of process planning [140]. Xu et al. proposed an intelligent process planning framework that integrates ProcessGPT and digital twins. ProcessGPT enabled users to input concise natural language text to generate high-quality knowledge in the field of process planning or machining process solutions. Digital twin can serve as a crucial technology for validating the process knowledge or plans generated by ProcessGPT, enhancing their reliability [132].

Overall, the application of LLM in process planning not only improves the intelligence of process design but also enhances the seamless connection between industrial design and manufacturing execution, which promotes the manufacturing industry to transform from the traditional labor-intensive production mode to a more efficient, flexible, and personalized manufacturing intelligence.

Example II: Conventional intelligent process planning systems integrate traditional artificial intelligence techniques, such as knowledge-based systems (KBS) based CAPP systems, continue to exhibit significant shortcomings. These problems include unfriendly user interfaces, limited flexibility, and more. As shown in Fig. 18, Xu et al. proposed an intelligent process planning framework integrating large language models and digital twins, named GIPP. The framework consists of ProcessGPT and a digital twin-based process verification module. Process engineers specify the technical requirements based on product specifications, including raw material and geometric parameters [141]. Process engineers specify the technical requirements based on product specifications, including raw material and geometric parameters. Leveraging the powerful language understanding and generation capabilities of ProcessGPT, the framework generates process knowledge and planning schemes that meet practical requirements. The final process planning solution is transferred to the digital twin-based Computer Numerical Control (CNC) system, enabling digital reproduction and verification of the process workflow.

6.3. LLMs for manufacturing execution

Manufacturing execution, as a downstream part of the product development process, plays a crucial role in translating design into the final product. In manufacturing execution, LLM can fully utilize the human-centered concept. It not only reduces the cognitive load of workers, but also allows workers to make full use of their knowledge and creativity to act as decision makers. For example, Christors et al. proposed an LLM-based manufacturing execution system that provides a natural language operator interface for operators, integrates with Digital Twins for real-time data acquisition, and employs behavior-based control for industrial robots. This integration demonstrated its great potential in real assembly scenarios [62]. Wang et al. proposed a vision and language cobot navigation approach based on LLM, which can help operators retrieve tools more efficiently and thus improve productivity [143]. Li et al. proposed a natural language-enabled virtual assistant named Max. Max can support active interactions in a variety of manufacturing tasks and provide task-relevant advice, and easily control their Industrial robots [144]. Omkar et al. proposed a robotics foundation model based on the integration of LLM and a vision model. This system used text-based task specifications and image-based observations to generate control actions for robot assembly tasks, helping workers perform complex and high-precision assembly tasks [145].

In addition, LLM can also facilitate the automation and efficiency of production in the smart workshop. Mathew presented COSMADS, which utilizes LLM to simplify access to shop floor data. Workshop workers can use natural language to submit queries, which are then collected from different data sources and materialized into tables, thus improving the efficiency of shop floor management [146]. Zeydan et al. integrated LLMs into an intent-based industrial automation control system that utilizes a domain-fine-tuned LLM for processing and executing business intents, as well as a RAG framework for generating informed decisions. This system translated high-level business goals into precise, actionable tasks on the shop floor [147]. Ahn et al. proposed a novel federated digital twin scheduling that combines LLM and deep reinforcement learning algorithms. The framework used large language model-based literacy module to analyze requirements in natural language and assign weights to digital twin attributes to achieve highly relevant KPIs, which are used to guide shop floor scheduling decisions [148]. Xia et al. integrated digital twins and LLM to plan and control manufacturing automated control systems. Given task instructions as inputs, LLM accomplished manufacturing tasks by orchestrating a series of atomic functions and skills. The approach demonstrated the potential of LLM agents in making informed decisions to make the future shop floor smarter [149]. Zhao et al. proposed an LLM-based multi-agent manufacturing system for intelligent shop floors. By defining agents for manufacturing resources on the physical shop floor. Not only data collection and training processes in conventional

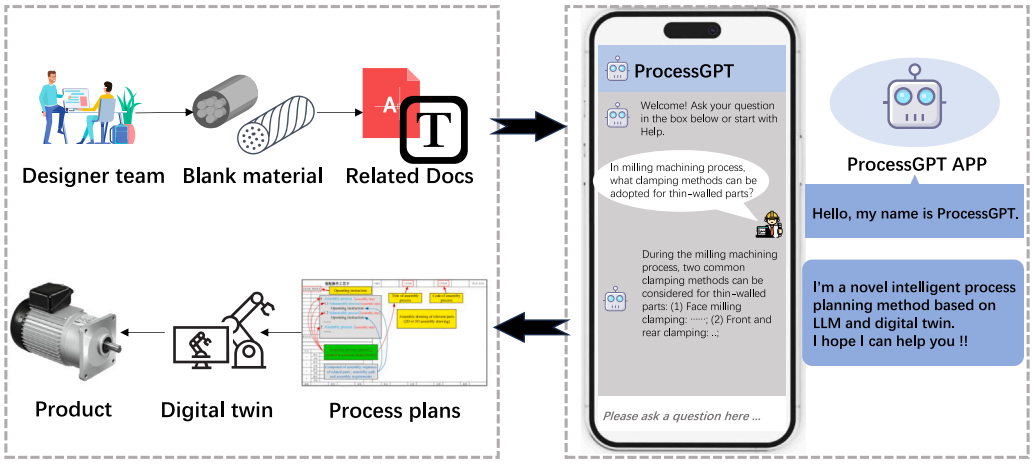

Fig. 18. An example of LLM enhanced process planning [141,142].

AI methods were avoided, but also the complexity of scheduling was notably reduced [150].

In the manufacturing execution, LLMs connect humans and machines to enable collaborative manufacturing and flexible production. They bridge human language through cognitive links at the higher level and connect with machines via instructions at the lower level, creating a continuous and systematic production process. This approach not only fully harnesses human cognitive decision-making capabilities but also maximizes the task execution potential of machines, reflecting the core principle of modern manufacturing: integration.

Example III: Manufacturing execution systems under Industry 5.0 emphasize collaboration between AI and human workers. However, the data architecture of human-centered manufacturing execution systems is often highly complex, requiring advanced frameworks to handle large volumes of data. As shown in Fig. 19, Christos et al. proposed an LLM-based Human-Robot Collaboration (HRC) manufacturing execution system, which enables production engineers to effortlessly orchestrate process plans and assign tasks to resources using textual commands [151]. The system comprises two dedicated agents: the "Interaction Agent" and the "Manufacturing Agent". The Interaction Agent is primarily responsible for handling dialogue with the user. It employs a "plan-execute" strategy to break down user-specified tasks into a series of subtasks. This agent is based on the GPT-3.5 model and has been fine-tuned using Chain-of-Thought (CoT) prompting, enabling it to effectively determine whether the input should be redirected to the Manufacturing Agent, invoke sensor modules, or perform other appropriate actions. The Manufacturing Agent addresses questions related to the HRC assembly deployed in the system. It is responsible for generating comprehensive production plans or modifying ongoing assembly plans. At the core of this agent is HRC-GPT. It is based on the GPT-4 0613 model. HRC-GPT can interface with respective modules that manage resource control through its code generation capabilities.

6.4. LLMs for production quality control

In the context of Industry 5.0 and smart manufacturing, production quality control plays a crucial role as a key link in ensuring that products and equipment are always of a high standard and quality throughout their life cycle. Traditional quality control methods typically rely on manual inspections and empirical judgment, and the process often lacks real-time data support, resulting in an inability to identify potential problems in a timely manner, increasing production costs and equipment downtime events. With the continuous development of LLM, sensors, and other technologies, the way of quality control has changed. Prognostic and health management (PHM), fault diagnosis, and anomaly detection are the three core areas in a modernized quality control system, as the shown in Fig. 20. The three parts of the technology are interdependent and synergistic, helping companies to monitor equipment health in real time and predict the risk of potential failures, thus improving equipment reliability and productivity. The introduction of LLMs can further promote the application of these three technologies in quality control, enhancing the intelligence and automation of the entire process.

PHM is a preventive maintenance method that assesses the health of industrial assets to predict their remaining useful life and develop preventive maintenance plans. The introduction of LLM can enhance the performance and efficiency of PHM systems in many aspects. For example, Tao et al. proposed a preventive maintenance plan generation method for wind power equipment based on maintenance knowledge fusion large model [152]. Chen et al. introduced an innovative regression framework utilizing LLM for remaining useful life prediction, which can effectively capture complex temporal dependencies and improve prediction accuracy [153]. Tao et al. proposed the concept of generative multimodal PHM-LM, which provides an algorithmic basis for realizing fundamental changes in PHM [154]. Lukens et al. presented a practical Prognostics and Health Management workflow and self-evaluation framework that utilizes the LLM agent to perform tasks such as data processing, failure mode discovery, etc., thus enabling the monitoring of physical asset health [155]. Abhay et al. evaluated the application of TimeGPT and Time-LLM for predictive maintenance of electric submersible pumps (ESPs) and showed that both models can significantly improve equipment operating efficiency and reduce unnecessary downtime, thus providing a more efficient maintenance plan for oil and gas production [156].

Unlike PHM, fault diagnosis is the rapid and accurate identification of the cause and location of faults through real-time monitoring and analysis of equipment and system operating status. The goal is to find the root cause as soon as possible after a failure occurs and to take remedial action, thereby interfering with the reduction of downtime and maintenance costs. Traditional fault diagnosis methods usually rely on manual experience and have problems such as inaccurate judgment and slow response. But, LLMs can quickly locate the cause of the fault and provide relevant solutions by analyzing data such as equipment logs, alarm information, and maintenance records. For example, Wang et al. combined LLM and a domain-specific knowledge base to develop a real-time updating system that provides specialized and specific advice during the operation and maintenance of industrial equipment, resulting in more accurate equipment maintenance [157]. Similarly, XCoalChat optimized LLMs in the field of coal mining equipment maintenance by knowledge graph and the triple LoRA fine-tuning mechanism for effective fault diagnosis and maintenance decision analysis [158]. Wang et al. presented an innovative approach to transform a generic LLM into a domain-specific intelligent aircraft maintenance

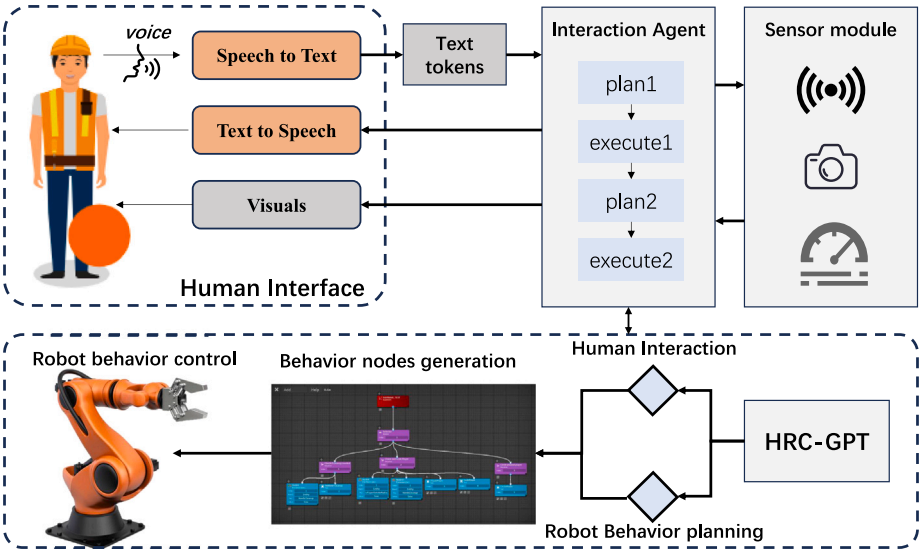

Fig. 19. An example of LLM enhanced manufacturing execution [151].

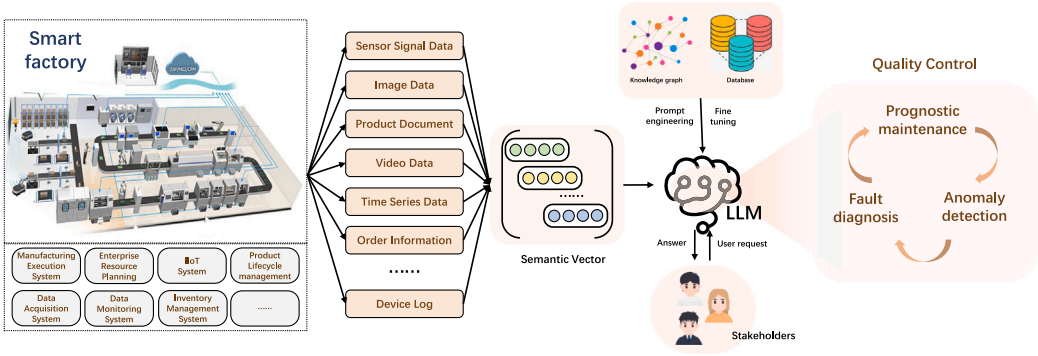

Fig. 20. The schematic diagram of the LLM-based QC platform, take smart factory as an example.

tool, which incorporates the structural ontology of the aircraft into the GPT-3.5 fine-tuning process, improving the model's ability to identify aircraft-related components and diagnose faults [159].

Anomaly detection is also one of the key techniques in quality control and is used to detect the presence of anomalies in equipment, production processes, or operations that deviate from normal operating conditions. These anomalies may indicate potential equipment failures or quality problems in the production process. By detecting anomalies, it is possible to provide early warning of problems before they occur, thus reducing the probability of failure and improving the quality of products and services. Traditional anomaly detection methods tend to rely on threshold settings and rule engines, which are less generalizable to new anomaly types or different production environments. LLMs are able to learn from a large amount of historical data and identify not only known anomalies, but also potential new types of anomalies. Gu et al. proposed AnomalyGPT, which not only detects and localizes industrial anomalies without manually adjusting thresholds, but also demonstrates strong generalization ability on new datasets using few-learning [160]. Wang et al. proposed an intelligent IVM and maintenance framework (IVMMF) empowered by large-scale visual language models. First, the method used LoRA to fine-tune industrial know-how and expertise into the vision model, enhancing the vision model's ability to analyze industrial images and identify defects more accurately. Then, LLM was optimized with a proprietary industry knowledge base that enables the language model to generate actionable maintenance recommendations and decision support when receiving data from the vision model [161]. Fu et al. integrated vision–language model (VLM) and large language model technologies into the steel manufacturing domain to create a novel steelmaking management system. VLM provided textual descriptions for slab defect detection, while LLM supported production data analysis and intelligent question answering. This intelligent system realized real-time anomaly detection in the steelmaking process [162]. Recently, Mercedes-Benz integrated LLM into vehicle production to improve exception identification, quality management, and process optimization [163].

Quality control in the era of Industry 5.0 not only relies on traditional manual inspection and empirical judgment, but also promotes intelligent and precise production and equipment management by integrating advanced technologies such as predictive maintenance, fault diagnosis, and anomaly detection. With the further development of technology, quality control will become more flexible, intelligent, and efficient, driving the manufacturing industry into a new era of greater refinement and customization.

Example IV: Combining the semantic understanding capabilities of large language models (LLMs) with the structured data of knowledge graphs can enhance the intelligence and user experience of fault diagnosis systems, enabling more efficient and accurate reasoning and decision support. As shown in Fig. 21, Ma et al. present Fault Diagnostic Reasoning Knowledge Graph LLM (FDRKG-LLM) [164]. The framework leverages an LLM to perform named entity recognition and intent recognition tasks, eliminating the need for users to create datasets or train models. By utilizing the graph structure of the fault diagnosis knowledge graph, the LLM performs knowledge graph-based reasoning, thereby mitigating limitations related to hallucination and interpretability. When a fault occurs, worker can enter a description of fault

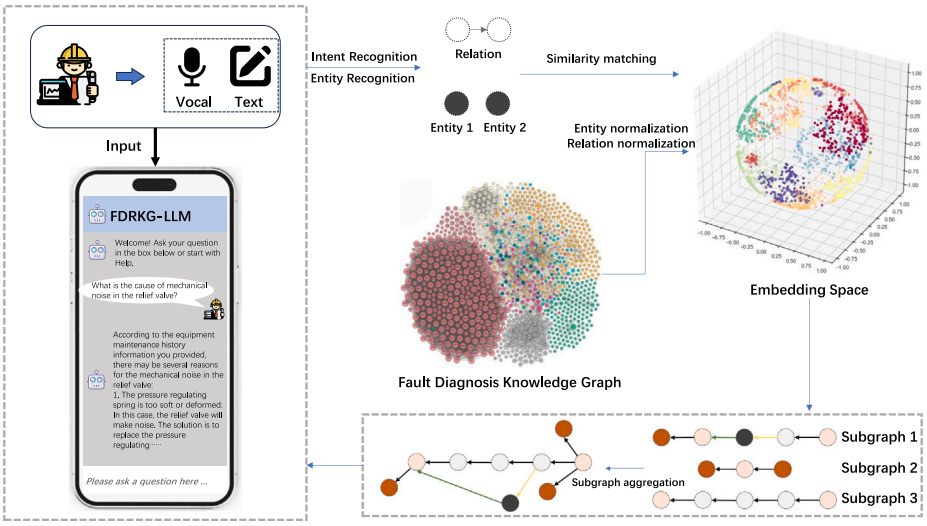

Fig. 21. An example of LLM enhanced fault diagnostic reasoning [164].

phenomenon into the FDRKG-LLM by voice or text. FDRKG-LLM accomplishes named entity recognition, intent recognition, and knowledge graph-based reasoning. Finally, FDRKG-LLM explains the specific cause of the fault based on the fault diagnosis knowledge subgraph. It is worth mentioning that FDRKG-LLM is a human-centered, knowledge-based, and user-friendly fault diagnostic system that extends the boundaries of an engineer's capabilities.

6.5. LLMs for product data management

Product data management (PDM) refers to the process and systems used to manage and control all the data and information related to the product lifecycle. The main goal is to ensure that accurate and consistent data is available across all stages of a product, from design to production, sales, and even decommissioning, while enabling efficient collaboration and management [165]. In summary, PDM is essential for managing product design and development data efficiently, particularly in industries with complex product creation processes, helping ensure smooth operations and better collaboration across teams. In the context of Industry 5.0, LLM will further promote the intelligent development of the product data management system. LLMs can facilitate more seamless and natural interactions between human operators and data management systems, thus enhancing collaboration and promoting knowledge transfer. This, in turn, will lead to more efficient and effective data utilization, resulting in better decision-making and business outcomes. Xue et al. proposed DBGPT, which can understand natural language queries and generate complex SQL queries with high accuracy, allowing users to express queries in natural language, thus achieving more natural and intuitive database interactions [166]. Li et al. introduced Table-GPT, which established a novel "Table-tuning" paradigm to fine-tune LLMs. This approach allowed LLMs to more effectively understand tables and perform table tasks [167]. Xiong et al. proposed Interactive-KBQA, a novel framework that harnesses the reasoning capabilities of LLMs for semantic parsing, enabling multiturn interactions with KBs [168]. Jiang et al. introduced StructGPT, which designs dedicated data access interfaces for three different types of structured data: knowledge graphs, tables, and databases. This allowed LLMs to gather relevant data from structured sources and perform reasoning based on this structured data [169].

In addition, users can help manufacturers provide customized product design solutions by interacting with data from the LLM-based PDM system and providing real-time feedback during the design phase. Du et al. introduced LLM-MANUF, an intelligent decision-making framework that integrates fine-tuned LLMs specialized in the manufacturing domain. It aims to enhance the comprehension and parsing of manufacturing decision-making requirements by leveraging the advanced contextual semantic reasoning capabilities and pre-trained knowledge based on LLMs [170]. Ren et al. presented CockpitGemini. The framework integrated LLM and human digital twins to provide customized designs and services based on user preferences and the user's real-time state, paving the way for future innovations in personalized design [171]. Sun et al. proposed Persona-DB, which is designed to enhance the accuracy and contextual efficiency of retrieval-based LLM personalization by encapsulating extensive user contexts and histories [172]. Shang et al. proposed a new personalized recommender system that integrates semantic understanding with user preferences using the natural language processing capabilities of LLM to improve the performance and accuracy of the recommender system [173].

In Industry 5.0, there is a need for efficient collaboration between various departments such as product design, manufacturing, quality, supply chain, etc. LLM can also help cross-functional teams share knowledge in the PDM system through natural language processing technology. For example, the design team can get feedback from the manufacturing team by simply asking questions, or LLM can automatically generate communication records between different departments to ensure timely sharing and collaboration of information at every step of the process. Xu et al. proposed an AI-augmented multimodal collaborative design (AI-MCD) framework, which provides an intelligent, real-time, and specific collaborative design platform for stakeholders with different research backgrounds. The platform can help stakeholders form a unified understanding and optimize traditional collaborative design processes [174]. Luo et al. proposed a digital intelligent platform architecture for nuclear power LLM, which integrates LLM into the data insights process in industrial sectors allowing for mutual enhancement on both technical and managerial levels to provide robust support for the operations management and decision-making process [175]. Gao et al. proposed CollabCoder [176], a CQA (Collaborative Qualitative Analysis) workflow that integrates the LLM model for developing code solutions. It not only facilitates real-time data synchronization and centralized management, avoiding complex data exchanges between staff members, but also provides individual and shared functionality, facilitating a seamless transition between individual and collaborative settings.

The introduction of LLM not only enhances the intelligence of product data management but also promotes personalization, cross-functional collaboration, and knowledge sharing. By combining with databases, knowledge graphs, design tools, and personalized recommendation systems, LLM provides more efficient and intelligent support for all aspects of product lifecycle management.

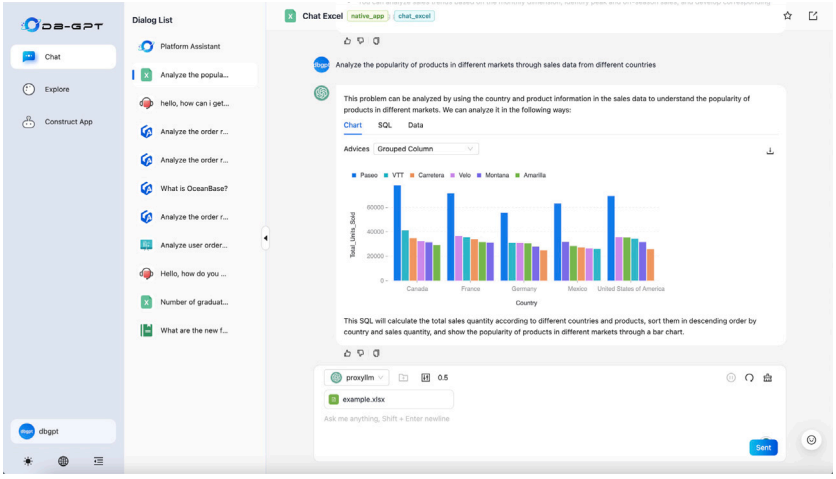

Fig. 22. An example of LLM enhanced data management [177,178].

Example V: The recent breakthroughs in LLM will drive transformation in many areas of software. The development of database technology as a foundational platform for data management and data analysis has a direct impact on the ability of businesses and organizations to gain insight and make decisions. And LLM breathes new life into database technology. Xue et al. proposed DB-GPT, which integrates large language models (LLMs) into traditional data interaction tasks, enabling the system to understand data interaction tasks described in natural language and provide responses also expressed in natural language [177,178]. For example, Fig. 22 illustrates DB-GPT's data analysis capabilities. The process is initiated by the user starting a new chat session and entering a command. DB-GPT utilizes its multi-agent framework to execute the task. First, it invokes a planner to generate a strategy for addressing the query. The data retrieval agent is responsible for performing the text-to-SQL task, extracting sales data for the relevant countries. The chart agent is responsible for invoking interfaces such as ECharts to generate visualizations. The final agent is responsible for aggregating the charts, collecting and organizing the results, and presenting them on the front-end interface. DB-GPT supports user interaction with the displayed charts. If users need to perform further data interaction tasks, they can continue to engage with their data through natural language input.

6.6. LLMs for supply chain management

Supply chain management plays a crucial role in the industrial product development process [179]. It not only affects product quality, cost, and delivery time, but also determines the efficiency and synergy of all aspects of the product from design to production. Guided by the principles of Industry 5.0, LLMs are rapidly transforming various aspects of supply chain management, particularly offering great potential in supply chain knowledge management, risk and security management, supply chain optimization, and customer service. In knowledge management, integrating LLMs into supply chain knowledge management systems can help decision-makers improve decision-making and provide valuable insights. This is because LLMs can store and capture relevant supply chain data, enhancing the accessibility of supply chain knowledge. For example, LLM can develop delivery plans that consider factors such as cost, service levels, and disruptions by analyzing large amounts of historical and real-time data. It can also collaborate as a logistics manager to formulate contingency plans for supply chain disruptions [180]. Furthermore, the IBM Watson system also includes NLP and AI solutions, capable of analyzing historical contract data, supplier performance records, as well as news data and social media, to offer supply chain risk analysis.

Supply chain operations are exposed to various risks and security threats, which can lead to disruptions. By integrating LLMs into supply chain management systems, users can predict and monitor potential risks, enabling manufacturers to take preemptive action. Damian et al. used ChatGPT to forecast the demand for products ordered by manufacturing companies. In comparison to the ARIMA algorithm, ChatGPT provided more accurate demand predictions, offering smart support for inventory management and supply chain scheduling [181]. Microsoft's Dynamics 365 Copilot can be integrated into the supply chain management platform. By utilizing LLMs to analyze potential disruption risks in the supply chain, it generates notifications that are automatically sent to relevant suppliers, helping companies improve their ability to manage supply chain risks. Additionally, as LLMs integrate more supply chain management data, they contribute to enhancing the resilience of the supply chain and assist organizations in building more robust and flexible supply chain networks.

In supply chain optimization, LLMs can be utilized for inventory control and identifying customer query patterns, seamlessly integrating with supply chain software and warehouse management systems to further improve management efficiency [182]. Li et al. proposed OptiGuide—a framework that employs LLMs to interpret supply chain optimization solutions. The LLM translated human queries into "optimization code", which is used by the optimization solver to generate the necessary output. The final output was then converted into human-readable language by the LLM, enabling businesses to automatically generate optimization solutions in supply chain management. This framework offered visual explanations to decision-makers, improving decision transparency and assisting in cost control and process optimization in the supply chain [183].

In customer service, LLMs can significantly improve both the efficiency and quality of services. Walmart, for example, uses LLMs to enhance customer service efficiency by automatically responding to common inquiries such as order status and returns, thereby reducing the workload of customer service agents. The BERT-based NLP model proposed by Mingyan et al. [184] was capable of converting natural language customer queries into order formats that machines can easily understand, enabling contactless order processing.

Guided by the principles of Industry 5.0, LLMs can be deeply integrated into all parts of the supply chain. This will not only enhance supply chain efficiency but also strengthen companies' capacity to deal with unforeseen events. With continuous advancements in LLM technology, the future of supply chain management will be smarter, more adaptable, and more efficient.

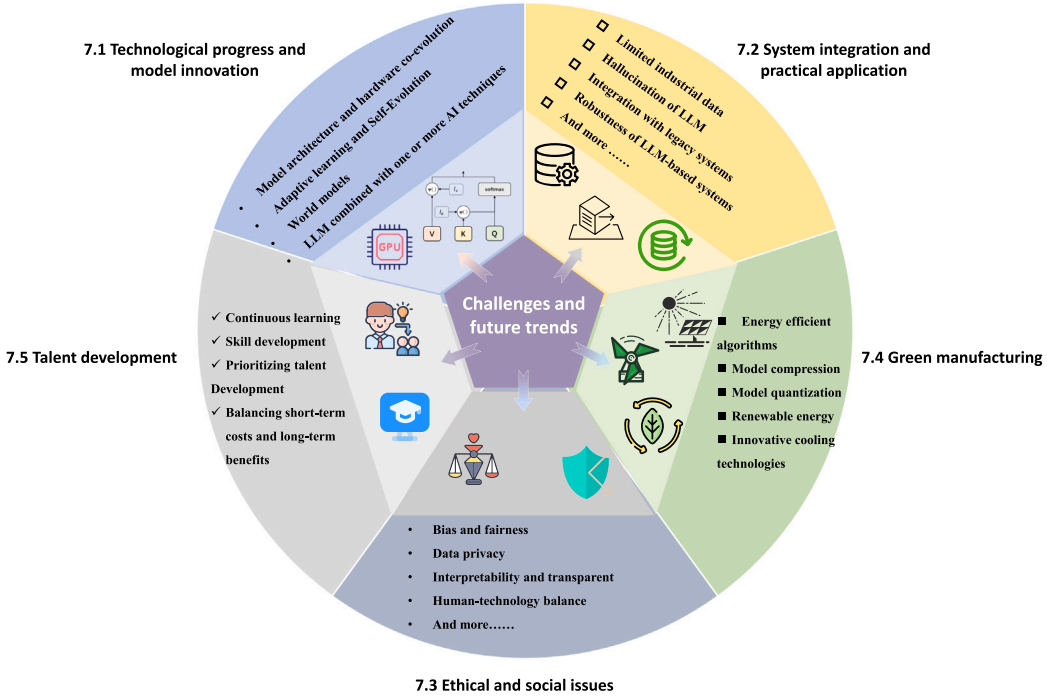

Fig. 23. Challenges and future trends.

7. Challenges and future trends

As Industry 4.0 transitions to Industry 5.0, the focus of manufacturing changes. From the technology-driven Industry 4.0 to human welfare-oriented Industry 5.0, manufacturing not only emphasizes improvements in efficiency and automation but also focuses more on human-centricity, sustainability, and resiliency. In this context, LLM has not only achieved technological optimization in Industry 4.0 but also demonstrated significant potential in transitioning towards Industry 5.0. On one hand, LLM is the latest research achievement in the field of artificial intelligence. Its powerful language comprehension and generation abilities can be applied across various industries. From automated customer service to data analysis and decision support, LLM has shown great potential in enhancing the resilience of MAS systems. On the other hand, the ease of use of LLM gives it a natural advantage in being human-centered. Due to its ability to understand and generate natural language, users can interact efficiently with an intelligent system, without needing complex programming skills and a deep technical background. The integration of generative AI, such as LLM, into smart manufacturing has become inevitable to maintain industrial leadership in an increasingly dynamic and challenging market environment. With its advantages in natural language processing and semantic understanding, LLM is able to show great potential in the areas of industrial design, process planning, and manufacturing execution.

However, as shown in Fig. 23, there are many challenges to fully realizing the value of LLMs in complex industrial environments. From technological innovation and architectural optimization to integration with existing systems, from digital privacy protection to issues of fairness and explainability, and from promoting green manufacturing to cultivating interdisciplinary talent, the successful application of LLMs requires addressing a range of challenges. At the same time, with the continuous growth of industrial data, how to effectively handle heterogeneous data and ensure the robustness and security of the system are still technical obstacles that need to be urgently overcome. Therefore, future research should not only focus on the technological advancements of LLMs but also their comprehensive impact in areas such as ethics, society, and the environment. This section explores these challenges in detail, offering insights into potential solutions and highlighting areas for future research. It is hoped to further stimulate research on LLM-based solutions in Industry 5.0.

7.1. Technological progress and model innovation

With the continuous progress of artificial intelligence, the LLM model architecture and the combination of LLM and other AI technologies have gradually become the key to improve industrial intelligence. To address the complexity and variability of industrial environments, the capabilities of LLMs must be continuously optimized. Therefore, exploring the evolution and developmental path of LLM technology itself will be an important research direction in the future.

7.1.1. Model architecture and hardware co-evolution

In the context of Industry 5.0, LLMs face some obvious limitations in industrial applications. Industry 5.0 emphasizes human–machine collaboration and intelligent manufacturing, which requires LLMs to simultaneously process data from different sensors and industrial textual information, and even understand human actions, thereby providing more accurate decision support. To achieve this goal, existing LLMs use the Transformer architecture, which may need to be redesigned, particularly in developing multimodal Transformer architectures or other network architectures capable of handling multiple input forms. The self-attention mechanism relied upon by the current Transformer has a computational complexity of $O(n^2)$ [185]. This means that as the input sequence length increases, the computational load grows quadratically. This will lead to a significant decrease in training and inference speed, thus requiring a more lightweight model architecture. The progress of LLMs has benefited from the rapid development of hardware technology, particularly the support of dedicated AI chips (such as NVIDIA's A100, H100, and Google's TPU). However, in the era of Industry 5.0, future research should not only focus on hardware improvements but also emphasize the co-evolution of algorithms and hardware. By thoroughly optimizing hardware architecture and integrating more lightweight neural network structures, the processing speed of LLMs can be improved, thereby promoting the widespread application of LLM-based industrial solutions.

7.1.2. Adaptive learning and self-Evolution

LLMs show powerful human–machine interaction capabilities and general-purpose abilities in smart manufacturing. But over time and with the continuous changes in production environments, LLMs' performance faces the risk of degradation. In the manufacturing industry, LLMs need to have the ability to continuously optimize and adapt to new tasks due to the dynamic changes in the production environment, rapid technological updates, and the diversity of task requirements. Without effective updating, LLM may not be able to respond to these new changes promptly, which may lead to inaccurate outputs and degradation of decision quality. To cope with this problem, RAG and incremental learning [186] are key tools to address the performance degradation of LLMs in smart manufacturing applications. For example, the RAG framework is able to incorporate external knowledge sources into the LLMs to enhance its adaptability. Incremental learning enables continuous optimization of the model and avoids the problem of model staleness. Incremental learning is significant in enhancing the adaptability of LLMs. However, real-time training of the model remains a complex and costly task. Real-time training not only requires the model to be able to continuously monitor and receive large amounts of data, but also requires powerful computational resources to support incremental learning. Each model update involves the processing and analysis of large-scale data, as well as the need to reassess the model's performance and make the necessary validations and adjustments. Therefore, how to balance the high cost of real-time training with its benefits is a major challenge in future smart manufacturing applications.

7.1.3. World models

In the field of intelligent simulation and computer-aided engineering, traditional industrial simulation techniques have relied on limited experimental data and specific mechanism models. These approaches often fail to accurately model the complexities and ever-changing real world. Additionally, industrial simulations require a large amount of physical experimental data and detailed models to predict and optimize system behavior. So, they often struggle when dealing with dynamic, nonlinear, and highly complex systems. Particularly in extreme conditions or with complex multi-factor interactions, traditional simulation models often face limitations in accuracy and effectiveness, failing to meet the high-precision and high-flexibility demands of modern manufacturing environments. However, in March 2024, OpenAI's Sora demonstrated great potential in the field of industrial simulation. Sora demonstrated strong world modeling capabilities, enabling comprehensive modeling of the real world through LLM and generating synthetic data that is closer to reality [187]. These data not only have a high degree of realism but also encompass performances under extreme conditions, and even simulate complex scenarios arising from multi-factor interactions. Compared to traditional simulation methods, world models can overcome the limitations of physical experimental data and fixed mechanism models, providing more accurate and comprehensive simulation results, especially in scenarios where data is difficult to obtain. By generating synthetic data, models like Sora are becoming important tools in the field of intelligent simulation. As these technologies continue to develop, world models like Sora will greatly drive data-driven decision-making, system optimization, and predictive capabilities in Industry 5.0 applications.

7.1.4. LLM combined with one or more AI techniques

Another key research direction is integrating LLMs with other AI technologies, such as reinforcement learning (RL) [188] and generative adversarial networks (GANs) [189]. Although LLMs can process a wide range of text and language tasks, they often perform worse than smaller models specifically optimized for particular tasks in certain domains. For example, in tasks such as image recognition, object detection, and face recognition, traditional convolutional neural networks (CNNs) typically provide higher accuracy and efficiency. This is due to the fact that these small models are able to efficiently capture fine-grained features by optimizing them for specific tasks. Therefore, future research should not only focus on enhancing the capabilities of LLMs but also explore how to integrate LLMs with specialized technologies like reinforcement learning and generative adversarial networks to leverage the strengths of both, thereby further improving efficiency and intelligence in Industry 5.0 applications. Specifically, reinforcement learning can help LLM optimize the decision-making process in dynamic environments such as production scheduling and real-time monitoring. Through continuous interaction with the environment, LLMs are able to constantly adjust and refine their strategies within the reinforcement learning framework, thereby improving production efficiency and flexibility. GANs can enhance the data generation and simulation capabilities of LLMs, especially in the fields of design innovation and quality control. For example, GANs can generate diverse design solutions and virtual environments, assisting LLMs in creative design and decision support in the context of Industry 5.0 scenarios. Additionally, because LLMs are usually trained on general-purpose datasets, they may fail to fully grasp the complex operations, equipment conditions, process parameters, and production environment details involved in the manufacturing process. LLM can be combined with digital twins, greatly enhancing the application potential in the manufacturing industry. Digital twin can simulate actual production conditions such as production lines, equipment, and material flow, thus providing LLM with real-time and comprehensive domain data. By combining LLM with digital twins, LLM provides real-time access to data from production systems, enabling more accurate reasoning and decision-making. At the same time, LLM-based agents can interact with the actual production environment through digital twins, adjusting decisions in real time to ensure the effective implementation of optimization plans and provide feedback to the production line. In addition, digital twins can simulate production processes under different operational scenarios, helping LLMs adapt to and respond to changing production conditions and task requirements. Through this cross-technology collaboration, LLM will play an increasingly important role in Industry 5.0 applications, which will not only drive the development of smart manufacturing, but also help to realize a more flexible, efficient and sustainable production model.

7.2. System integration and practical application

In practical industrial applications, the successful application of LLMs relies on seamless integration with existing systems. Traditional production and management systems often have their unique data formats and architectures, and LLMs need the ability to handle heterogeneous data to enhance their intelligence. Furthermore, the deployment of LLMs is not just a simple application of the model, but also involves optimizing computational resources, real-time requirements, and a range of other technical details. Therefore, how to effectively carry out system integration and data processing becomes a fundamental and core challenge in the industrial application of LLMs [190].

7.2.1. Limited industrial data

In the industrial sector, the application of LLM faces the challenge of limited data resources. Unlike the massive public text data required for general LLM training, industrial data is often highly professional, sensitive, and dispersed, resulting in a shortage of high-quality annotated data that can be used for model training and fine-tuning. In addition, industrial data covers multi-modal, multi-source heterogeneous information, such as sensor data, equipment logs, images, and text, which increases the complexity of data preprocessing and fusion. This data scarcity not only limits the performance of LLMs in industrial intelligent manufacturing, but also restricts their widespread application. To cope with this problem, current research should focus on techniques such as transfer learning and small-sample learning applied to LLMs to enhance the generalization ability and adaptability of LLMs in limited data environments. At the same time, the construction of knowledge bases and

corpora specialized for industrial scenarios, as well as the promotion of cross-enterprise and cross-industry data sharing, are expected to further promote the in-depth application and innovative development of LLMs in smart manufacturing.

7.2.2. Hallucination of LLM

The hallucination of LLM is that LLM-generated content may be inconsistent with real-world facts or with user input. LLMs are essentially statistical models, and while LLMs are capable of generating contextually relevant content through large-scale training on textual data, they do not have the ability to deeply understand or factually validate knowledge. In short, LLM doesn't know whether the content it generates is correct or not. LLM simply generates the most likely sequence of tokens based on the input. This poses a serious obstacle to the application of LLMs in the vertical domains. The problem of hallucination is particularly acute in industrial scenarios, especially during the transition from Industry 4.0 to Industry 5.0. Industry 5.0 emphasizes human-centered smart manufacturing, focusing on collaboration between humans and machines rather than relying solely on automated systems. LLM can provide strong support in this transformation process, helping to optimize production processes and improve intelligent decision-making. However, LLM itself cannot fully guarantee the correctness of the generated content. In fields such as complex production decision-making, equipment management, and quality control, the hallucination issue may lead to serious production errors and decision-making failures, thereby affecting overall production efficiency and safety. Despite the challenges posed by hallucinations, there are several strategies to mitigate their impact and harness the power of LLM effectively. First, while generating the output, the LLM is combined with industry-specific databases and literature, and multiple rounds of validation are performed after generating the content to ensure its correctness. Secondly, enhance human intervention and feedback mechanisms. For critical decisions, a human feedback loop can be established, where experts review the content generated by the LLM to ensure that the final output aligns with facts and industry standards.

7.2.3. Heterogeneity of industrial data

A key characteristic of industrial environments is the high heterogeneity of data [191]. Industrial data comes from multiple different sources and exists in various formats and structures. For example, internet of things sensor data, product data management systems, and workers' manual records are all common data sources in industrial systems. It is a complex task for LLM to process and record all this data. These data are subject to problems such as missing data and noisy data due to the industrial environment. They will significantly add to the complexity of model training. Furthermore, because of the data's diversity and inconsistency, LLM is susceptible to the influence of data quality problems, which can affect its resulting decisions. To tackle these challenges, a powerful data preprocessing pipeline is required, which includes data cleaning and data normalization techniques. These methods help improve data quality, ensure consistency throughout the training process, and enhance the model's ability to generalize to new data. In addition, the design of the data preprocessing pipeline needs to consider how multimodal data will be processed. Sensor data, production management system data, and manually recorded data each represent a different data model, and how to integrate and synergize these data in LLM is an important research direction. In conclusion, although LLM faces numerous challenges when handling heterogeneous data in industrial environments, constructing a robust data preprocessing pipeline and combining data cleaning and normalization techniques can partially address these challenges. Future research should focus more on optimizing the data preprocessing process and utilizing advanced algorithms and architectures to improve LLMs' performance in handling complex industrial data, in order to better support the development of smart manufacturing and Industry 5.0.

7.2.4. Combination of multimodal sensing techniques with LLM

As a powerful tool for natural language processing, LLMs can play an important role in industrial documents, scheduling production, and optimizing workflow. However, it still faces many limitations when dealing with complex industrial scenarios. In industrial environments, smart manufacturing relies not only on a large amount of textual information, but also involves a wide range of sensory data such as images, sounds, and smells. These different modalities are critical for accurate decision making, but LLM itself does not have the ability to fuse data across modalities, which limits its effectiveness in integrated analysis and multidimensional decision making. Therefore, combining perceptual techniques such as machine hearing [192] and machine smell [193] is the key to break through the limitations of LLMs. Machine hearing is able to analyze the sound signals of the equipment to monitor the operating status of the machinery, identify potential malfunctions, such as machine vibration, sound fluctuations. While the machine smell simulates the human olfactory system and monitors the gas composition in the air through sensors. It is able to detect the leakage of harmful gases or other environmental pollution in time to protect the safety and health of the production environment. Integrating these sensing technologies with LLM in the framework of Industry 5.0 can provide multi-dimensional intelligent decision support for industrial production. Future research directions should focus on developing multimodal learning and cross-modal reasoning algorithms that explore how to effectively fuse different types of data (images, odors, sounds, etc.).

7.2.5. Model lightweighting and deployment

Many scenarios in industrial applications require real-time or near-real-time data analysis, followed by quick decision-making based on the results. However, LLMs, especially those with billions or even hundreds of billions of parameters, often face extremely high computational demands during real-time deployment, which may result in significant delays. This challenge is mainly due to the large computational resources required by LLMs, particularly when dealing with large datasets and complex tasks. Therefore, in real-time or near-real-time industrial applications, such as production line monitoring, predictive maintenance, and quality control, how to effectively reduce computational load and latency is a key issue. To address this problem, a number of model optimization techniques can be employed. First, model distillation [194] is an effective strategy, which trains a smaller and more efficient model to mimic the behavior of LLM, thereby significantly reducing computational and storage costs while maintaining high prediction accuracy [190]. Model distillation enables small models to perform more efficiently in the inference process without sacrificing too much performance, adapting to the needs of real-time data processing in industrial environments. In addition, there are two popular LLM deployment methods, which are cloud-based and edge-based methods [195]. Cloud-based deployment involves hosting LLM on servers managed by cloud service providers such as Amazon Web Services, Google Cloud Platform, Alibaba Cloud, and Baidu AI Cloud. This approach leverages the nearly unlimited computational power and storage capacity of cloud servers. It benefits model management and seamless updates. However, in manufacturing environments that require low latency and high reliability, this approach may lead to higher latency. The edge-based deployment method involves hosting the LLM on local devices or edge servers that are physically closer to the manufacturing process. This method significantly reduces latency by performing data processing and decision-making on edge servers. Additionally, this method enhances the privacy and security of the data. Furthermore, edge-based deployment is particularly well-suited for environments with limited or unreliable network connectivity, ensuring continuous operation even without internet access. However, edge devices are often limited by restrictions on resources, memory, and storage capacity. The hybrid deployment [196] strategy combines the advantages of cloud and edge-based approaches, emerging as a

promising solution that allows manufacturers to balance performance, scalability, and security in their LLM deployment. In summary, to tackle the computational and latency issues of LLM in real-time industrial applications, the integration of cloud–edge hybrid architecture offers an effective solution. Additionally, emerging technologies such as quantum computing [197] could radically transform the training and deployment of LLMs. Quantum computing has the unprecedented ability to perform parallel computations at scale, which can significantly reduce both the time required to train LLMs and the time for LLM inference. This will help in developing more powerful and efficient LLMs, which can handle larger datasets and more complex tasks. While quantum computing is still in the early stages of development, research and collaboration between experts in the field of quantum computing and practitioners in the manufacturing industry may pave the way for future practical applications of quantum-enhanced LLM.

7.2.6. Integration with legacy systems

The scalability of LLM-based solutions is another important challenge that needs to be carefully addressed. First, many enterprises' production management systems (such as PDM, MES, ERP, etc.) still rely on outdated technical architectures. These systems are deeply embedded in the daily operations of the business, and any changes could introduce risks. Especially when attempting to integrate LLM technology with these traditional systems, it is not only necessary to consider technical compatibility but also to assess the impact on existing processes. Upgrading existing systems may involve complex data migration, functionality reconstruction, and even adjustments to business processes, which could pose potential threats to the operational efficiency and financial stability of the enterprise. Therefore, it has become a major challenge for the industry to achieve seamless integration of LLMs with legacy systems and avoid disruption to existing business processes. Additionally, with the rapid increase in the volume, variety, and velocity of data in industrial environments, traditional technical architectures are struggling to meet these demands. When dealing with large-scale, high-concurrency data processing, enterprises require more flexible and scalable solutions to efficiently deploy LLM models in response to evolving business demands and data environments. To address this, containerization technology and microservices architecture have become key solutions. Through containerization, LLM models and their associated services can be packaged into independent, deployable containers, enabling these models to run efficiently in different computational environments. Meanwhile, using container orchestration tools such as Kubernetes enables the automated management, elastic scaling, and load balancing of multiple microservices. It can ensure that models can collaborate efficiently in complex production environments and dynamically adjust resources based on demand. Additionally, the use of API gateways and data bridging tools can effectively bridge the gap between traditional systems and modern AI technologies. With these tools, enterprises can seamlessly connect the data flow of existing systems to the new technology platform and achieve a smooth transition to intelligent upgrades. This approach enables organizations to avoid replacing all of their legacy systems at once, thereby reducing the risks and costs of the migration process. It is a more viable technical solution, especially for small and medium-sized enterprises (SMEs). Additionally, along with these integration efforts, enterprises need to adopt a series of strategies to ensure success. Integration testing, change management, and rollback strategies are essential components. Integrated testing helps organizations verify the compatibility of old and new systems and the correct flow of data. Change management helps to monitor changes throughout the integration process and avoid potential risks due to technical changes. A rollback strategy ensures that if a problem occurs during the integration process, the enterprise is able to quickly return to a stable state and avoid any impact on the production environment. Most importantly, the LLM must be integrated in a way that complements and enhances the current workflows, rather than introducing unnecessary complexity or redundancy. This requires business practitioners to have a comprehensive understanding of the specific use cases and pain points in the manufacturing process and to work closely with domain experts, engineers, and IT teams to ensure consistency and maximize value.

7.2.7. Robustness of LLM-based systems

Another common limitation of LLMs in industrial applications is that performance often becomes unstable when the system is required to switch between different base LLMs. Since different LLM models may vary in training data, architectural design, and optimization objectives, directly switching models can lead to inconsistencies in output. In some cases, this can trigger performance fluctuations and errors in the system. This instability is especially evident in industrial applications, particularly in dynamic manufacturing environments where the system needs to cope with varying data inputs, changes in scenarios, and task modifications. To address this issue, introducing humans as supervisors has become a key strategy for ensuring the stability of LLMs. Introducing human supervision in LLM-based systems not only helps maintain the stability of model outputs but also provides timely feedback and adjustments when the system faces changes. The role of human supervision can be multifaceted: for example, validating and reviewing system outputs, adjusting strategies during model training, and intervening when biases arise. Additionally, as multiple users and teams collaborate and continuously improve the LLM system, coordinating the update requirements of all parties has become an important challenge. Different users may have different needs and goals, and their expectations and preferences for models may vary over time. The involvement of human supervisors not only helps to harmonize the needs of all parties but also ensures the stability and consistency of the system during the update process. In summary, human supervision plays a critical role in LLM-based systems. By encouraging human involvement in the system's continuous improvement, it not only enhances the stability and adaptability of LLMs but also ensures their ongoing optimization and evolution, ultimately leading to more precise and reliable industrial applications.

7.3. Ethical and social issues

With the widespread application of LLMs in industrial fields, issues like bias, interpretability, and transparency are becoming more critical [198,199]. Industrial decisions are typically high-risk and high-value. Thus, ensuring that the model's decision-making process is fair, transparent, and interpretable is not just a technical challenge, but also a question of social trust and ethical responsibility. Moreover, maintaining an appropriate balance in human–machine collaboration and avoiding over-reliance on technology at the expense of human intervention is also a significant challenge.

7.3.1. Bias and fairness

Bias and fairness are key concerns in the current application of LLMs. As LLMs are usually trained on large-scale datasets from the internet and other public sources, these datasets may contain historical biases and inequities. Therefore, LLMs can easily inherit these biases during the training process. These biases may manifest as discriminatory tendencies related to gender, race, culture, and other factors, especially when dealing with tasks related to human behavior. In the context of Industry 5.0, with the widespread adoption of smart manufacturing and automation systems, this issue becomes particularly important. Intelligent systems, while handling large volumes of data and making decisions, may unintentionally magnify social biases, thus impacting fairness in production and management processes. For example, in data analytics-based decisions on workforce deployment or production line optimization, if the LLM system inherits gender or racial bias during training, it may result in certain groups being overlooked or disadvantaged in resource allocation. Similarly, in supply chain management, if the model is potentially biased against certain

regions or groups of suppliers, it may affect the fairness of decision-making and thus pose a business risk. To tackle these problems and reduce the impact of bias on LLM performance, it is crucial to carefully manage the training data. Particular attention must be paid to the diversity and representativeness of sources when collecting data, ensuring that over-reliance on data from one particular group is avoided. Additionally, during the data cleaning and preprocessing stages, potential biases need to be identified and corrected to ensure that the training set better reflects the diversity and fairness of society. Finally, mitigating bias is not limited to the data layer. The design and evaluation of the model are equally important. It is essential to introduce continuous monitoring and regular evaluation mechanisms to track the performance of LLMs in different tasks and applications, especially in scenarios involving sensitive decisions such as personnel management, recruitment, and compensation distribution. By introducing fairness metrics and bias mitigation techniques, the model's outputs can be regularly reviewed to ensure they do not lead to unfair outcomes. At the same time, businesses and research institutions should advocate for more open and transparent AI development standards, encouraging society at large to participate in practices aimed at reducing AI bias and enhancing fairness.

7.3.2. Data privacy

Data privacy is another critical issue that cannot be overlooked, especially in the context of increasingly strict data protection regulations and environments involving sensitive manufacturing data [200]. Product designs, industrial formulas, customer information, and various types of knowledge from the production process are core assets of a company. These pieces of information are not only crucial to the company's competitiveness but are also strictly protected by law. With the development of AI, especially the wide application of LLM, how to ensure the privacy and security of data and avoid data leakage and data misuse has become one of the key challenges to realize the technology. The sensitivity of industrial data requires companies to adopt a range of advanced security measures when using LLMs to improve production processes, ensuring compliance with laws and regulations, and protecting the company's intellectual property. When handling sensitive data, techniques such as data encryption, access control, and anonymization are fundamental to protecting data privacy. For LLMs, data encryption ensures that data is not accessed or stolen by unauthorized individuals during transmission and storage. It can protect corporate confidential information from external threats. Strong access control mechanisms allow for strict rights management of data access, ensuring that only authorized users and systems can access specific data. Anonymization techniques, through de-identification and pseudo-anonymization, ensure data privacy while not compromising the model's training and inference performance. Additionally, in scenarios involving multiple production facilities or partners, federated learning, as an emerging distributed learning technology, provides a highly effective solution. Federated learning allows participants to retain data locally, eliminating the need to centralize data in the cloud and on a central server for processing. This not only effectively avoids the risk of data privacy breaches but also ensures data sovereignty. The participants only need to share the updated parameters of the model, not the raw data, so that model training and optimization can still be performed with all the knowledge while ensuring data privacy. In this way, different organizations or businesses can benefit from collective intelligence together without having to expose their sensitive data. This is especially important for manufacturing companies with global operations. Additionally, organizations should not only focus on the technical aspects of security when deploying LLM, but also ensure compliance with standards such as ISO 27001, GDPR, CCPA and other requirements. For example, the European GDPR (General Data Protection Regulation) requires companies to explicitly obtain user consent when processing personal data and ensure transparent use of the data. In the U.S., certain sectors (such as healthcare and finance) have more stringent legal regulations on data usage and sharing. Companies must establish a compliance-oriented governance framework and conduct regular reviews of data processing workflows to ensure ongoing adherence to applicable regulations. Overall, data privacy is a key factor that needs to be prioritized for LLM in industrial applications.

7.3.3. Interpretability and transparent

Since LLM is based on the transformer architecture, LLMs usually operate as a "black box" model and lack sufficient interpretability. This lack of interpretability poses a significant challenge for critical application areas, especially in tasks such as predictive maintenance and quality control. In these areas, understanding LLM's decision-making process is critical because engineers and operators need to ensure that the model's predictions and decisions are reasonable and reliable. Especially in predictive maintenance, LLMs are often used to analyze equipment status, predict failures, and provide maintenance recommendations. So, understanding how the model arrives at these conclusions is critical for ensuring system stability and avoiding misjudgments. If the decision-making process is not transparent, the operator may not be able to judge whether the model has adequately considered all relevant factors, which in turn affects the quality of the final decision. In addition, the non-interpretability of LLMs may raise a number of issues such as trust, accountability, and validation. In smart manufacturing and process applications, especially in scenarios involving safety, compliance, and responsibility for human lives or high-value assets, the "black-box" nature of the model can lead to untraceable and unverifiable decision-making processes. In such cases, the lack of transparent explanations will make it difficult to pursue accountability in the event of errors or non-compliant behavior in the system. For example, in an automated production process, if the quality control decisions made by the LLM are flawed, the production line may produce defective or substandard products. Without a clear understanding of how the model draws its conclusions, managers will have difficulty identifying the root cause of the problem and thus will not be able to effectively make corrections and improvements. To address these challenges, future research must focus more on improving the interpretability of LLM. This involves not only the development of new model architectures and algorithms but also the introduction of explainable AI (XAI) techniques. Addressing this issue will not only help drive the widespread deployment of LLM in industry but will also ensure the safety and fairness of its output at LLM. To address this challenge, explainable AI techniques, as well as the proposal of novel algorithmic architectures, are important directions for solving the explainability of LLM in the future. For example, combining physics-informed neural networks (PINNs) with LLM may provide new ideas for improving the transparency and verifiability of models. PINNs can integrate physical laws and constraints into neural networks and can provide more physics-based interpretability, which is especially important in engineering and manufacturing. In the context of predictive maintenance and quality control, by introducing PINN, LLM's predictions can more easily align with the physical state of the equipment, fault patterns, and engineering background, thereby enhancing the traceability and verifiability of LLM's decisions.

7.3.4. Human–technology balance

LLMs can process vast amounts of customer data, product information, and market trends to provide real-time guidance and accurate customer demand forecasting for sales teams [201]. They can not only automatically generate personalized marketing content and optimize sales strategies, but also adjust sales plans by analyzing customer feedback in real time, thereby supporting personalized customization and flexible production scheduling. While LLMs offer significant advantages in sales processes within the context of Industry 5.0, they also have some limitations. Firstly, in the future of highly personalized industrial products and services, customer demands and market trends will change rapidly and be highly complex. This requires LLMs to

possess strong adaptability and a deep understanding of the market. However, current LLMs often lack a thorough understanding of this area. Secondly, especially in the personalized customization scenarios of Industry 5.0, salespeople often need to flexibly adjust and negotiate based on the unique needs of the customer, technical requirements, and production capabilities. These decision-making processes involve a large amount of unstructured data and emotional factors that are difficult for LLM to fully simulate and control. So while LLMs can provide data-driven analytics and recommendations, they cannot completely replace a salesperson's ability to make judgments and interact emotionally in complex scenarios. Finally, excessive reliance on LLMs may lead to the gradual decline of core skills within the sales team. Industry 5.0 emphasizes the deep collaboration between humans and machines, and salespeople not only need to exchange information with machines efficiently, but also need to judge customer needs and market trends through their own experience and intuition. Excessive reliance on LLMs for routine sales tasks may cause team members to lose their ability to engage in in-depth communication and negotiation with clients, leading to a decrease in salespeople's adaptability and flexibility, thereby impacting the team's performance in complex business decisions. In summary, despite the immense potential of LLMs in sales processes within Industry 5.0, their integration with human users must be balanced, finding the right boundary between relying on technology for data support and maintaining human judgment. LLMs can provide data-driven suggestions and trend forecasts for salespeople, but the final decision should still be made by the sales team based on the actual situation and business objectives. As a result, the sales team needs to regularly evaluate LLM's proposal to ensure that it remains aligned with the company's goals, customer needs, and market trends in order to avoid the potential risks associated with technological over-reliance.

7.3.5. Worker safety and well-being

Industry 5.0 builds on the emphasis on automation and connectivity in Industry 4.0, more deeply integrating the human factor and prioritizing worker well-being and sustainability. While Industry 4.0 has led to a significant increase in automation and intelligence, it has also brought new security risks and psychological challenges. First, the close collaboration of people with equipment, robots, and other intelligent systems increases potential physical safety hazards. Second, as AI systems become more widely used in industry, workers may face technological dependence, lack of trust in automated decision-making, and even feel job stress and psychological anxiety. Therefore, it is particularly important to create a work environment that supports workers' physical and mental health, including optimizing work processes to reduce worker fatigue, providing training and psychological support for new technologies, and so on. In recent years, the development of human digital twin technology has provided new means for safety and physical and mental health. By constructing virtual digital models of workers and monitoring their physiological state, emotional changes, and work environment risks in real time, potential safety hazards and health problems can be detected in time, and personalized risk warning and intervention can be realized. In addition, the rise of wearable and sensors has supported this process with more real-time and accurate data. By collecting real-time physiological data (e.g., heart rate, body temperature, exercise status, etc.) and environmental data (e.g., temperature, humidity, gas concentration, etc.) from workers, these devices provide real-time feedback on the health status of workers. Such technologies not only help to improve the safety of workers at work, but also provide strong support for the management of physical health, driving the sustainable development of human–machine collaboration in Industry 5.0. The combined use of LLMs, human digital twins, and wearable devices can more comprehensively safeguard the safety and well-being of workers in smart manufacturing environments, thereby realizing the vision of human-centered manufacturing in Industry 5.0.

7.3.6. Job transformation and workforce impact

In today's world of rapid technological change, the rise of generative artificial intelligence (AI) techniques such as Large Language Models (LLMs) promises dramatic change in the future. These technologies have the potential not only to disrupt the existing labor market but also to change the nature of work, with some low-skilled jobs in particular likely to be replaced by AI systems. However, historical experience shows that while technological advances may cause some jobs to disappear, they will also create new jobs. These jobs include not only positions focused on managing and developing AI technology, but also roles such as training, fine-tuning, maintaining, and overseeing AI systems. This transformation of the labor market has undoubtedly created new challenges and opportunities for individuals and organizations. The ability to continue to learn will be the key to individuals adapting to new changes. At the same time, Industry 5.0 emphasizes the universality of technology, asserting that workers of all levels and backgrounds can benefit from technological advances. This means that while promoting the popularization and application of AI technology, measures must be taken to avoid social inequalities resulting from the technological divide. In particular, against the backdrop of rapid technological development, how to ensure that low-skilled workers, marginalized groups, and the workforce in different regions have equal access to new technologies and enjoy the opportunities brought about by the transformation is a topic that needs to be focused on by all sectors of society. To effectively address these challenges, regulators, policymakers, researchers, and industry leaders need to work together to ensure that the principles of responsibility are followed in the development and deployment of AI. It is important to guard against potential social inequities while fully unlocking the economic benefits of the technology.

7.4. Green manufacturing

Energy efficiency is also a key issue for enterprises in the application of LLMs, especially in the context of global efforts to promote sustainable development and green manufacturing programs [202]. Due to their large number of parameters and computational demands, LLMs are typically seen as energy-intensive applications, which not only result in higher operational costs for businesses but also potentially have a greater negative impact on the environment. For example, LLMs require a large amount of computational resources and storage space during the training and inference process, which directly leads to large power consumption and high carbon emissions. Therefore, how to optimize the energy use of LLM and reduce its impact on the environment has become a major challenge in the application of AI systems. To mitigate this problem, there are a variety of strategies that manufacturers can employ to improve the energy efficiency of LLMs. First, the development of energy-efficient algorithms is an important direction. By optimizing the model training algorithm and the computation during inference, energy consumption can be reduced without sacrificing performance. For example, model compression and quantization techniques can reduce computation and storage requirements while preserving model accuracy. Additionally, the design of adaptive model architectures is one of the effective ways to address the issue of reducing energy consumption in LLMs. By dynamically adjusting the model size and computational complexity according to task requirements, excessive computational resources can be avoided when processing simple tasks. These methods help reduce unnecessary energy consumption. At the data center level, the application of renewable energy and the optimization of cooling systems are also key factors in enhancing the sustainability of LLM deployment and industrial production. For example, using renewable energy sources such as solar and wind power to supply electricity to data centers. At the same time, innovative cooling technologies, such as liquid-cooling systems or intelligent temperature control technologies, can significantly improve the energy efficiency of data centers and reduce the

additional energy consumption caused by traditional air conditioning cooling methods. These energy-saving measures can greatly reduce the overall energy consumption of LLM and promote the green and sustainable development of AI technology.

7.5. Talent development

Companies or enterprises that integrate LLMs into industrial smart manufacturing systems early will gain significant first-mover advantages. However, while the introduction of LLM technology can lead to long-term efficiency improvements, it also faces potential cost constraints, such as significant initial investments. This includes not only the costs of introducing the technology itself and system integration but also the training costs for employees and potential internal resistance within the organization. Therefore, business managers need to carefully balance short-term costs and long-term benefits when making decisions, ensuring that they can effectively manage risks during the transition while driving technological advancement. It is important to note that in the era of Industry 5.0, LLM technology should not be exclusive to large companies or wealthy enterprises. As core participants in Industry 5.0, SMEs [203] also face the challenge of improving production efficiency and competitiveness, and should have the opportunity to fully take advantage of these advanced technologies to improve their competitiveness. To this end, SMEs need to obtain the help of external experts, especially during the introduction and application of LLM technology. At the national level, the government should also encourage the development of SMEs and provide more resources, such as tax incentives, technology subsidies, and low-interest loans, to help them overcome the obstacles encountered during technology adoption. At the same time, companies should also recognize that the introduction of Industry 5.0 and LLM technology is not meant to replace workers, but to organically integrate artificial intelligence with human operations to enhance human capabilities. In this process, the role of LLM is to enhance human decision-making and work efficiency, not to replace human labor. In fact, machines and robots can never fully replace the role of humans. Because humans have creativity, judgment, and flexibility that no automated system can have. Without human involvement, automation and digitalization processes cannot proceed smoothly. Humans can improve the system's fault tolerance and ensure the stable operation of intelligent manufacturing systems in complex environments. Therefore, enterprises should increase investments in continuous learning and skill development for employees. Ensure that every employee can adapt to and fully utilize new technologies, enabling them to stand out in the wave of smart manufacturing. By prioritizing talent development, companies can not only build a workforce that adapts to future growth but can also gain a sustainable competitive advantage in the era of Industry 5.0. In this environment, companies will maintain flexibility and foresight in advancing smart manufacturing, enhancing production efficiency, and fostering innovation, ultimately achieving long-term sustainable development.

8. Conclusion

With the continuous development of industrial technology, global manufacturing is transitioning from Industry 4.0 to Industry 5.0. In contrast to Industry 4.0, Industry 5.0 serves as the guiding principle for the next generation of intelligent manufacturing, focusing on human-centric intelligent manufacturing. This transformation focuses not only on highly automated production processes but also on the deep integration of humans and artificial intelligence. The purpose of the transformation combines the technological advantages of AI with human intelligence to realize the mutually beneficial co-existence of workers and technology. In this process, LLMs, as advanced artificial intelligence technologies, are playing a key role in achieving Industry 5.0. LLMs not only bring new transformative opportunities to industrial production, but also change the way humans interact with machines. This paper provides an overview of how the combination of LLMs with Industry 5.0 enabling technologies shows how this fusion drives industrial innovation, optimizes production workflows, and improves the efficiency of human–machine collaboration. In addition, this study highlights the vast potential of LLMs in transforming various aspects of smart manufacturing. By thoroughly exploring the application of LLMs in six areas—industrial design, process planning, manufacturing execution, product quality control, and others, we demonstrate the potential of LLMs in advancing human-centric smart manufacturing in Industry 5.0. Finally, we discuss the challenges and future research directions for LLM-based solutions in the context of Industry 5.0. Looking ahead, with the continued advances in smart manufacturing and artificial intelligence technologies, LLMs will play an increasingly important role in the future. By further advancing the integration of LLMs with industrial systems, Industry 5.0 will move beyond human-centric manufacturing and evolve into humanity-centered manufacturing. Humanity-centered manufacturing is not just about "how to use technology better", but about "what kind of future society is needed". It emphasizes that the transformation of the manufacturing industry should not be at the expense of the majority of people, but rather, it is necessary to find a path for the co-development of technology and human beings. We hope that this work will inspire more researchers to contribute to the exploration of this exciting and evolving field.

CRediT authorship contribution statement

Yunfei Ma: Writing – review & editing, Writing – original draft. **Shuai Zheng:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Zheng Yang:** Resources, Methodology. **Pai Zheng:** Validation, Supervision. **Jiewu Leng:** Writing – review & editing, Formal analysis. **Jun Hong:** Resources, Project administration.

Declaration of competing 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.

References

[1] Leng J, Sha W, Wang B, Zheng P, Zhuang C, Liu Q, Wuest T, Mourtzis D, Wang L. Industry 5.0: Prospect and retrospect. J Manuf Syst 2022;65:279–95.
[2] Xu X, Lu Y, Vogel-Heuser B, Wang L. Industry 4.0 and industry 5.0—Inception, conception and perception. J Manuf Syst 2021;61:530–5. http://dx.doi.org/10.1016/j.jmsy.2021.10.006.
[3] Ordieres-Meré J, Gutierrez M, Villalba-Díez J. Toward the industry 5.0 paradigm: Increasing value creation through the robust integration of humans and machines. Comput Ind 2023;150:103947. http://dx.doi.org/10.1016/j.compind.2023.103947.
[4] Wan PK, Leirmo TL. Human-centric zero-defect manufacturing: State-of-the-art review, perspectives, and challenges. Comput Ind 2023;144:103792. http://dx.doi.org/10.1016/j.compind.2022.103792.
[5] Lee J, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. 2018, arXiv preprint arXiv:1810.04805. 3 (8).
[6] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21(140):1–67.
[7] Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, Xia H, Xu J, Wu Z, Liu T, et al. A survey on in-context learning. 2022, arXiv preprint arXiv:2301.00234.
[8] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, et al. Emergent abilities of large language models. 2022, arXiv preprint arXiv:2206.07682.
[9] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, editors. Advances in neural information processing systems. vol. 33, Curran Associates, Inc.; 2020, p. 1877–901.
[10] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research, vol. 139, PMLR; 2021, p. 8821–31.

[11] Wermelinger M. Using GitHub copilot to solve simple programming problems. In: Proceedings of the 54th ACM technical symposium on computer science education v. 1. SIGCSE 2023, New York, NY, USA: Association for Computing Machinery; 2023, p. 172–8. http://dx.doi.org/10.1145/3545945.3569830.
[12] Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, et al. Tacotron: Towards end-to-end speech synthesis. 2017, arXiv preprint arXiv:1703.10135.
[13] Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, Yuan Z, Huang Y, Sun H, Gao J, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. 2024, arXiv preprint arXiv:2402.17177.
[14] Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science 2023;381(6654):187–92. http://dx.doi.org/10.1126/science.adh2586, arXiv:https://www.science.org/doi/pdf/10.1126/science.adh2586.
[15] Brynjolfsson E, Li D, Raymond L. Generative AI at work. Q J Econ 2025;qjae044.
[16] Liao Y, Deschamps F, Loures EdFR, Ramos LFP. Past, present and future of industry 4.0-a systematic literature review and research agenda proposal. Int J Prod Res 2017;55(12):3609–29.
[17] Kagermann H. Change through digitization—Value creation in the age of industry 4.0. In: Management of permanent change. Springer; 2014, p. 23–45.
[18] Ivanov D. The industry 5.0 framework: viability-based integration of the resilience, sustainability, and human-centricity perspectives. Int J Prod Res 2023;61(5):1683–95.
[19] Breque M, De Nul L, Petridis A, et al. Industry 5.0: towards a sustainable, human-centric and resilient European industry. Luxemb LU: Eur Comm Dir-Gen Res Innov 2021;46.
[20] Frederico GF. From supply chain 4.0 to supply chain 5.0: Findings from a systematic literature review and research directions. Logistics 2021;5(3):49.
[21] Psarommatis F, May G, Azamfirei V. Envisioning maintenance 5.0: Insights from a systematic literature review of industry 4.0 and a proposed framework. J Manuf Syst 2023;68:376–99.
[22] Marinelli M. From industry 4.0 to construction 5.0: exploring the path towards human–robot collaboration in construction. Systems 2023;11(3):152.
[23] Symeonaki E, Maraveas C, Arvanitis KG. Recent advances in digital twins for agriculture 5.0: Applications and open issues in livestock production systems. Appl Sci 2024;14(2). http://dx.doi.org/10.3390/app14020686.
[24] Holzinger A, Schweier J, Gollob C, Nothdurft A, Hasenauer H, Kirisits T, Häggström C, Visser R, Cavalli R, Spinelli R, et al. From industry 5.0 to forestry 5.0: Bridging the gap with human-centered artificial intelligence. Curr For Rep 2024;10(6):442–55.
[25] Hassoun A, Jagtap S, Trollman H, Garcia-Garcia G, Duong LN, Saxena P, Bouzembrak Y, Treiblmaier H, Para-López C, Carmona-Torres C, et al. From food industry 4.0 to food industry 5.0: Identifying technological enablers and potential future applications in the food sector. Compr Rev Food Sci Food Saf 2024;23(6):e370040.
[26] Tandel V, Kumari A, Tanwar S, Singh A, Sharma R, Yamsani N. Intelligent wearable-assisted digital healthcare industry 5.0. Artif Intell Med 2024;157:103000. http://dx.doi.org/10.1016/j.artmed.2024.103000.
[27] Kour R, Karim R, Dersin P, Venkatesh N. Cybersecurity for industry 5.0: trends and gaps. Front Comput Sci 2024;6:1434436.
[28] Gladysz B, Tran T-a, Romero D, van Erp T, Abonyi J, Ruppert T. Current development on the operator 4.0 and transition towards the operator 5.0: A systematic literature review in light of industry 5.0. J Manuf Syst 2023;70:160–85.
[29] Leng J, Zhu X, Huang Z, Li X, Zheng P, Zhou X, Mourtzis D, Wang B, Qi Q, Shao H, Wan J, Chen X, Wang L, Liu Q. Unlocking the power of industrial artificial intelligence towards industry 5.0: Insights, pathways, and challenges. J Manuf Syst 2024;73:349–63. http://dx.doi.org/10.1016/j.jmsy.2024.02.010.
[30] Guo J, Leng J, Zhao JL, Zhou X, Yuan Y, Lu Y, Mourtzis D, Qi Q, Huang S, Song X, Liu Q, Wang L. Industrial metaverse towards industry 5.0: Connotation, architecture, enablers, and challenges. J Manuf Syst 2024;76:25–42. http://dx.doi.org/10.1016/j.jmsy.2024.07.007.
[31] Zafar MH, as EFL, Sanfilippo F. Exploring the synergies between collaborative robotics, digital twins, augmentation, and industry 5.0 for smart manufacturing: A state-of-the-art review. Robot Comput-Integr Manuf 2024;89:102769. http://dx.doi.org/10.1016/j.rcim.2024.102769.
[32] Wang B, Zhou H, Li X, Yang G, Zheng P, Song C, Yuan Y, Wuest T, Yang H, Wang L. Human digital twin in the context of industry 5.0. Robot Comput-Integr Manuf 2024;85:102626. http://dx.doi.org/10.1016/j.rcim.2023.102626.
[33] Fang W, Chen L, Zhang T, Chen C, Teng Z, Wang L. Head-mounted display augmented reality in manufacturing: A systematic review. Robot Comput-Integr Manuf 2023;83:102567. http://dx.doi.org/10.1016/j.rcim.2023.102567.
[34] Lv Z. Digital twins in industry 5.0. Research 2023;6:0071.
[35] del Real Torres A, Andreiana DS, Ojeda Roldán Á, Hernández Bustos A, Acevedo Galicia LE. A review of deep reinforcement learning approaches for smart manufacturing in industry 4.0 and 5.0 framework. Appl Sci 2022;12(23):12377.
[36] Chen H. Large knowledge model: Perspectives and challenges. 2023, arXiv preprint arXiv:2312.02706.
[37] Li J, Gao Y, Yang Y, Bai Y, Zhou X, Li Y, Sun H, Liu Y, Si X, Ye Y, Wu Y, Lin Y, Xu B, Ren B, Feng C, Huang H. Fundamental capabilities and applications of large language models: A survey. ACM Comput Surv 2025. http://dx.doi.org/10.1145/3735632.
[38] Lee S, Sim W, Shin D, Seo W, Park J, Lee S, Hwang S, Kim S, Kim S. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. ACM Trans Intell Syst Technol 2024.
[39] Lou R, Zhang K, Yin W. Large language model instruction following: A survey of progresses and challenges. Comput Linguist 2024;50(3):1053–95.
[40] Khamassi M, Nahon M, Chatila R. Strong and weak alignment of large language models with human values. Sci Rep 2024;14(1):19399.
[41] Pan H, Zhai Z, Yuan H, Lv Y, Fu R, Liu M, Wang Z, Qin B. Kwaiagents: Generalized information-seeking agent system with large language models. 2023, arXiv preprint arXiv:2312.04889.
[42] Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: Scoping review. J Med Internet Res 2024;26:e60501.
[43] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, et al. Gpt-4 technical report. 2023, arXiv preprint arXiv:2303.08774.
[44] Wei J, Wang X, Schuurmans D, Bosma M, ichter b, Xia F, Chi E, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in neural information processing systems. vol. 35, Curran Associates, Inc.; 2022, p. 24824–37.
[45] Sha Y, Gou S, Liu B, Faber J, Liu N, Schramm S, Stoecker H, Steckenreiter T, Vnucec D, Wetzstein N, Widl A, Zhou K. Hierarchical knowledge guided fault intensity diagnosis of complex industrial systems. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. KDD '24, New York, NY, USA: Association for Computing Machinery; 2024, p. 5657–68. http://dx.doi.org/10.1145/3637528.3671610.
[46] Zaken EB, Ravfogel S, Goldberg Y. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. 2021, arXiv preprint arXiv:2106.10199.
[47] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. Lora: Low-rank adaptation of large language models. 2021, arXiv preprint arXiv:2106.09685.
[48] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, editors. Advances in neural information processing systems. vol. 33, Curran Associates, Inc.; 2020, p. 9459–74.
[49] Kernan Freire S, Wang C, Foosherian M, Wellsandt S, Ruiz-Arenas S, Niforatos E. Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. Front Artif Intell 2024;7:1293084.
[50] Zhou Y, Liu Y, Li X, Jin J, Qian H, Liu Z, Li C, Dou Z, Ho T-Y, Yu PS. Trustworthiness in retrieval-augmented generation systems: A survey. 2024, arXiv preprint arXiv:2409.10102.
[51] Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E, et al. The rise and potential of large language model based agents: A survey. Sci China Inf Sci 2025;68(2):121101.
[52] Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, et al. A survey on large language model based autonomous agents. Front Comput Sci 2024;18(6):186345.
[53] Talebirad Y, Nadiri A. Multi-agent collaboration: Harnessing the power of intelligent llm agents. 2023, arXiv preprint arXiv:2306.03314.
[54] Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E. A survey on multimodal large language models. 2023, arXiv preprint arXiv:2306.13549.
[55] Fan J, Yin Y, Wang T, Dong W, Zheng P, Wang L. Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey. Front Eng Manag 2025;1–24.
[56] Mo F, Chaplin JC, Sanderson D, Ratchev S. Advancing capability matching in manufacturing reconfiguration with large language models. In: International conference on flexible automation and intelligent manufacturing. Springer; 2024, p. 215–22.
[57] Liu Z, Bahety A, Song S. REFLECT: Summarizing robot experiences for failure explanation and correction. 2023, arXiv preprint arXiv:2306.15724.
[58] Wang T, Fan J, Zheng P. An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing. J Manuf Syst 2024;75:299–305.
[59] Park S, Wang X, Menassa CC, Kamat VR, Chai JY. Natural language instructions for intuitive human interaction with robotic assistants in field construction work. Autom Constr 2024;161:105345.
[60] Ye Y, You H, Du J. Improved trust in human-robot collaboration with ChatGPT. IEEE Access 2023;11:55748–54. http://dx.doi.org/10.1109/ACCESS.2023.3282111.
[61] Farooq MU, Kang G, Seo J, Bae J, Kang S, Jang YJ. DAIM-HRI: A new human-robot integration technology for industries. In: 2024 IEEE international conference on advanced robotics and its social impacts. ARSO, 2024, p. 7–12. http://dx.doi.org/10.1109/ARSO60199.2024.10557811.

[62] Gkournelos C, Konstantinou C, Makris S. An LLM-based approach for enabling seamless human-robot collaboration in assembly. CIRP Ann 2024;73(1):9–12. http://dx.doi.org/10.1016/j.cirp.2024.04.002.
[63] Wang C, Hasler S, Tanneberg D, Ocker F, Joublin F, Ceravola A, Deigmoeller J, Gienger M. LaMI: Large language models for multi-modal human-robot interaction. In: Extended abstracts of the CHI conference on human factors in computing systems. CHI EA '24, New York, NY, USA: Association for Computing Machinery; 2024, http://dx.doi.org/10.1145/3613905.3651029.
[64] Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, Finn C, Fu C, Gopalakrishnan K, Hausman K, et al. Do as i can, not as i say: Grounding language in robotic affordances. 2022, arXiv preprint arXiv:2204.01691.
[65] Lykov A, Tsetserukou D. LLM-BRAIn: AI-driven fast generation of robot behaviour tree based on large language model. In: 2024 2nd international conference on foundation and large language models. FLLM, 2024, p. 392–7. http://dx.doi.org/10.1109/FLLM63129.2024.10852491.
[66] Lykov A, Konenkov M, Gbagbe KF, Litvinov M, Peter R, Davletshin D, Fedoseev A, Kobzarev O, Alabbas A, Alyounes O, et al. Cognitiveos: Large multimodal model based system to endow any type of robot with generative ai. 2024, arXiv preprint arXiv:2401.16205.
[67] An T, Zhou Y, Zou H, Yang J. Iot-llm: Enhancing real-world iot task reasoning with large language models. 2024, arXiv preprint arXiv:2410.02429.
[68] Mo S, Salakhutdinov R, Morency L-P, Liang PP. Iot-lm: Large multisensory language models for the internet of things. 2024, arXiv preprint arXiv:2407.09801.
[69] Zhong N, Wang Y, Xiong R, Zheng Y, Li Y, Ouyang M, Shen D, Zhu X. CASIT: Collective intelligent agent system for internet of things. IEEE Internet Things J 2024;11(11):19646–56. http://dx.doi.org/10.1109/JIOT.2024.3366906.
[70] Cui H, Du Y, Yang Q, Shao Y, Liew SC. LLMind: Orchestrating AI and IoT with LLM for complex task execution. IEEE Commun Mag 2024;1–7. http://dx.doi.org/10.1109/MCOM.002.2400106.
[71] Li D, Li H, Li J, Li H-W, Wang H, Minerva R, Crespi N, Li K-C. Blockchain-enabled large language models for prognostics and health management framework in industrial internet of things. In: International conference on blockchain and trustworthy systems. Springer; 2024, p. 3–16.
[72] Ferrag MA, Ndhlovu M, Tihanyi N, Cordeiro LC, Debbah M, Lestable T, Thandi NS. Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT devices. IEEE Access 2024;12:23733–50. http://dx.doi.org/10.1109/ACCESS.2024.3363469.
[73] Nakanishi H, Hisafuru K, Hasegawa K, Hidano S, Fukushima K, Hashimoto K, Togawa N. Initial seeds generation using LLM for IoT device fuzzing. In: 2024 11th international conference on internet of things: systems, management and security. IOTSMS, 2024, p. 5–10. http://dx.doi.org/10.1109/IOTSMS62296.2024.10710191.
[74] Wang T, Zhao Z, Wu K. Exploiting LLM embeddings for content-based IoT anomaly detection. In: 2024 IEEE Pacific rim conference on communications, computers and signal processing. PACRIM, 2024, p. 1–6. http://dx.doi.org/10.1109/PACRIM61180.2024.10690230.
[75] Mihai S, Yaqoob M, Hung DV, Davis W, Towakel P, Raza M, Karamanoglu M, Barn B, Shetve D, Prasad RV, Venkataraman H, Trestian R, Nguyen HX. Digital twins: A survey on enabling technologies, challenges, trends and future prospects. IEEE Commun Surv Tutorials 2022;24(4):2255–91. http://dx.doi.org/10.1109/COMST.2022.3208773.
[76] Sun Y, Zhang Q, Bao J, Lu Y, Liu S. Empowering digital twins with large language models for global temporal feature learning. J Manuf Syst 2024;74:83–99. http://dx.doi.org/10.1016/j.jmsy.2024.02.015.
[77] Li M, Wang R, Zhou X, Zhu Z, Wen Y, Tan R. ChatTwin: Toward automated digital twin generation for data center via large language models. In: Proceedings of the 10th ACM international conference on systems for energy-efficient buildings, cities, and transportation. BuildSys '23, New York, NY, USA: Association for Computing Machinery; 2023, p. 208–11. http://dx.doi.org/10.1145/3600100.3623719.
[78] Li X, He B, Wang Z, Zhou Y, Li G. Towards cognitive digital twin system of human-robot collaboration manipulation. Authorea Prepr 2024.
[79] Luo P, Parn E, Brilakis I. ChatTwin: Enabling natural language interactions with infrastructure digital twins. In: Proceedings of the 2024 European conference on computing in construction. Computing in construction, vol. 5, Chania, Greece: European Council on Computing in Construction; 2024, http://dx.doi.org/10.35490/EC3.2024.259.
[80] Šturm J, Zajec P, Škrjanc M, Mladenić D, Grobelnik M. Enhancing cognitive digital twin interaction using an LLM agent. In: 2024 47th MIPRO ICT and electronics convention. MIPRO, 2024, p. 103–7. http://dx.doi.org/10.1109/MIPRO60963.2024.10569919.
[81] Xia Y, Dittler D, Jazdi N, Chen H, Weyrich M. LLM experiments with simulation: Large language model multi-agent system for simulation model parametrization in digital twins. In: 2024 IEEE 29th international conference on emerging technologies and factory automation. ETFA, 2024, p. 1–4. http://dx.doi.org/10.1109/ETFA61755.2024.10710900.
[82] Hanqing Y, SIEW M, JOE-WONG C. An LLM-based digital twin for optimizing human-in-the loop systems [c]. In: 2024 IEEE international workshop on foundation models for cyber-physical systems and internet of things, Hong kong, China. 2024.
[83] Giunchi D, Numan N, Gatti E, Steed A. DreamCodeVR: Towards democratizing behavior design in virtual reality with speech-driven programming. In: 2024 IEEE conference virtual reality and 3D user interfaces. VR, 2024, p. 579–89. http://dx.doi.org/10.1109/VR58804.2024.00078.
[84] Yin Z, Wang Y, Papatheodorou T, Hui P. Text2VRScene: Exploring the framework of automated text-driven generation system for VR experience. In: 2024 IEEE conference virtual reality and 3D user interfaces. VR, 2024, p. 701–11. http://dx.doi.org/10.1109/VR58804.2024.00090.
[85] De La Torre F, Fang CM, Huang H, Banburski-Fahey A, Amores Fernandez J, Lanier J. LLMR: Real-time prompting of interactive worlds using large language models. In: Proceedings of the 2024 CHI conference on human factors in computing systems. CHI '24, New York, NY, USA: Association for Computing Machinery; 2024, http://dx.doi.org/10.1145/3613904.3642579.
[86] Xu F, Nguyen T, Du J. Augmented reality for maintenance tasks with ChatGPT for automated text-to-action. J Constr Eng Manag 2024;150(4):04024015. http://dx.doi.org/10.1061/JCEMD4.COENG-14142, arXiv:https://ascelibrary.org/doi/pdf/10.1061/JCEMD4.COENG-14142.
[87] Zhao Z, Lou S, Tan R, Lv C. An AR-assisted human-robot interaction system for improving LLM-based robot control. In: 2024 IEEE international conference on cybernetics and intelligent systems (CIS) and IEEE international conference on robotics, automation and mechatronics. RAM, 2024, p. 144–9. http://dx.doi.org/10.1109/CIS-RAM61939.2024.10673005.
[88] Chen H, Hou L, Wu S, Zhang G, Zou Y, Moon S, Bhuiyan M. Augmented reality, deep learning and vision-language query system for construction worker safety. Autom Constr 2024;157:105158. http://dx.doi.org/10.1016/j.autcon.2023.105158.
[89] Fan H, Zhang H, Ma C, Wu T, Fuh JYH, Li B. Enhancing metal additive manufacturing training with the advanced vision language model: A pathway to immersive augmented reality training for non-experts. J Manuf Syst 2024;75:257–69. http://dx.doi.org/10.1016/j.jmsy.2024.06.007.
[90] Hang C-N, Yu P-D, Morabito R, Tan C-W. Large language models meet next-generation networking technologies: A review. Futur Internet 2024;16(10). http://dx.doi.org/10.3390/fi16100365, URL https://www.mdpi.com/1999-5903/16/10/365.
[91] Kan KB, Mun H, Cao G, Lee Y. Mobile-LLaMA: Instruction fine-tuning open-source LLM for network analysis in 5G networks. IEEE Netw 2024;38(5):76–83. http://dx.doi.org/10.1109/MNET.2024.3421306.
[92] Xiao Z, Ye C, Hu Y, Yuan H, Huang Y, Feng Y, Cai L, Chang J. LLM agents as 6G orchestrator: A paradigm for task-oriented physical-layer automation. 2024, arXiv preprint arXiv:2410.03688.
[93] Wang J, Zhang L, Yang Y, Zhuang Z, Qi Q, Sun H, Lu L, Feng J, Liao J. Network meets ChatGPT: Intent autonomous management, control and operation. J Commun Inf Netw 2023;8(3):239–55. http://dx.doi.org/10.23919/JCIN.2023.10272352.
[94] Mekrache A, Ksentini A, Verikoukis C. Intent-based management of next-generation networks: an LLM-centric approach. IEEE Netw 2024;38(5):29–36. http://dx.doi.org/10.1109/MNET.2024.3420120.
[95] Shen Y, Shao J, Zhang X, Lin Z, Pan H, Li D, Zhang J, Letaief KB. Large language models empowered autonomous edge AI for connected intelligence. IEEE Commun Mag 2024;62(10):140–6. http://dx.doi.org/10.1109/MCOM.001.2300550.
[96] Ali T, Kostakos P. HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs). 2023, arXiv preprint arXiv:2309.16021.
[97] Kaheh M, Kholgh DK, Kostakos P. Cyber sentinel: Exploring conversational agents in streamlining security tasks with gpt-4. 2023, arXiv preprint arXiv:2309.16422.
[98] He Z, Li Z, Yang S, Qiao A, Zhang X, Luo X, Chen T. Large language models for blockchain security: A systematic literature review. 2024, arXiv preprint arXiv:2403.14280.
[99] Zheng Z, Xie S, Dai H-N, Chen W, Chen X, Weng J, Imran M. An overview on smart contracts: Challenges, advances and platforms. Future Gener Comput Syst 2020;105:475–91. http://dx.doi.org/10.1016/j.future.2019.12.019.
[100] Petrović N, Al-Azzoni I. Model-driven smart contract generation leveraging ChatGPT. In: International conference on systems engineering. Springer; 2023, p. 387–96.
[101] Ortu M, Ibba G, Conversano C, Tonelli R, Destefanis G. Identifying and fixing vulnerable patterns in ethereum smart contracts: A comparative study of fine-tuning and prompt engineering using large language models. 2023, Available at SSRN 4530467.
[102] Ma W, Wu D, Sun Y, Wang T, Liu S, Zhang J, Xue Y, Liu Y. Combining fine-tuning and LLM-based agents for intuitive smart contract auditing with justifications. 2024, arXiv preprint arXiv:2403.16073.
[103] Gai Y, Zhou L, Qin K, Song D, Gervais A. Blockchain large language models. 2023, arXiv preprint arXiv:2304.12749.
[104] Mbula Mboma JG, Tshipata OT, Kambale WV, Kyamakya K. Assessing how large language models can be integrated with or used for blockchain technology: Overview and illustrative case study. In: 2023 27th international conference on circuits, systems, communications and computers. CSCC, 2023, p. 59–70. http://dx.doi.org/10.1109/CSCC58962.2023.00018.

[105] Benzinho J, Ferreira J, Batista J, Pereira L, Maximiano M, Távora V, Gomes R, Remédios O. LLM based chatbot for farm-to-fork blockchain traceability platform. Appl Sci 2024;14(19). http://dx.doi.org/10.3390/app14198856.
[106] Luo H, Luo J, Vasilakos AV. BC4LLM: A perspective of trusted artificial intelligence when blockchain meets large language models. Neurocomputing 2024;599:128089. http://dx.doi.org/10.1016/j.neucom.2024.128089.
[107] Zuo X, Wang M, Zhu T, Zhang L, Ye D, Yu S, Zhou W. Federated TrustChain: Blockchain-enhanced LLM training and unlearning. 2024, arXiv preprint arXiv:2406.04076.
[108] Xiao Y, Zheng S, Shi J, Du X, Hong J. Knowledge graph-based manufacturing process planning: A state-of-the-art review. J Manuf Syst 2023;70:417–35.
[109] Wen Y, Wang Z, Sun J. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. 2023, arXiv preprint arXiv:2308.09729.
[110] Wei X, Cui X, Cheng N, Wang X, Zhang X, Huang S, Xie P, Xu J, Chen Y, Zhang M, et al. Zero-shot information extraction via chatting with chatgpt. 2023, arXiv preprint arXiv:2302.10205.
[111] Gui H, Yuan L, Ye H, Zhang N, Sun M, Liang L, Chen H. Iepile: Unearthing large-scale schema-based information extraction corpus. 2024, arXiv preprint arXiv:2402.14710.
[112] Zhang K, Gutiérrez BJ, Su Y. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. 2023, arXiv preprint arXiv:2305.11159.
[113] Zhang R, Su Y, Trisedya BD, Zhao X, Yang M, Cheng H, Qi J. AutoAlign: Fully automatic and effective knowledge graph alignment enabled by large language models. IEEE Trans Knowl Data Eng 2024;36(6):2357–71. http://dx.doi.org/10.1109/TKDE.2023.3325484.
[114] Zhou B, Li X, Liu T, Xu K, Liu W, Bao J. CausalKGPT: Industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. Adv Eng Inform 2024;59:102333. http://dx.doi.org/10.1016/j.aei.2023.102333.
[115] Liu P, Qian L, Zhao X, Tao B. Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly. IEEE Trans Ind Inform 2024;20(6):8160–9. http://dx.doi.org/10.1109/TII.2024.3366977.
[116] Qi Z, Yu Y, Tu M, Tan J, Huang Y. Foodgpt: A large language model in food testing domain with incremental pre-training and knowledge graph prompt. 2023, arXiv preprint arXiv:2308.10173.
[117] Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: A roadmap. IEEE Trans Knowl Data Eng 2024;36(7):3580–99. http://dx.doi.org/10.1109/TKDE.2024.3352100.
[118] Wang B, Zheng L, Wang Y, Fang W, Wang L. Towards the industry 5.0 frontier: Review and prospect of XR in product assembly. J Manuf Syst 2024;74:777–811. http://dx.doi.org/10.1016/j.jmsy.2024.05.002.
[119] Alam MF, Ahmed F. Gencad: Image-conditioned computer-aided design generation with transformer-based contrastive representation and diffusion priors. 2024, arXiv preprint arXiv:2409.16294.
[120] You Y, Uy MA, Han J, Thomas R, Zhang H, You S, Guibas L. Img2cad: Reverse engineering 3d cad models from images through vlm-assisted conditional factorization. 2024, arXiv preprint arXiv:2408.01437.
[121] Badagabettu A, Yarlagadda SS, Farimani AB. Query2CAD: Generating CAD models using natural language queries. 2024, arXiv preprint arXiv:2406.00144.
[122] Xu J, Wang C, Zhao Z, Liu W, Ma Y, Gao S. CAD-MLLM: Unifying multimodality-conditioned CAD generation with MLLM. 2024, arXiv preprint arXiv:2411.04954.
[123] Deng H, Khan S, Erkoyuncu JA. An investigation on utilizing large language model for industrial computer-aided design automation. Procedia CIRP 2024;128:221–6. http://dx.doi.org/10.1016/j.procir.2024.07.049, 34th CIRP Design Conference.
[124] Kapsalis T. CADgpt: Harnessing natural language processing for 3D modelling to enhance computer-aided design workflows. 2024, arXiv preprint arXiv:2401.05476.
[125] Wu H, He Z, Zhang X, Yao X, Zheng S, Zheng H, Yu B. ChatEDA: A large language model powered autonomous agent for EDA. IEEE Trans Comput-Aided Des Integr Circuits Syst 2024;43(10):3184–97. http://dx.doi.org/10.1109/TCAD.2024.3383347.
[126] Shi L, Kazda M, Sears B, Shropshire N, Puri R. Ask-EDA: A design assistant empowered by LLM, hybrid RAG and abbreviation de-hallucination. 2024, arXiv preprint arXiv:2406.06575.
[127] Xu K, Sun J, Hu Y, Fang X, Shan W, Wang X, Jiang Z. MEIC: Re-thinking RTL debug automation using LLMs. 2024, arXiv preprint arXiv:2405.06840.
[128] Thakur S, Blocklove J, Pearce H, Tan B, Garg S, Karri R. Autochip: Automating hdl generation using llm feedback. 2023, arXiv preprint arXiv:2311.04887.
[129] Sun Y, Li X, Liu C, Deng X, Zhang W, Wang J, Zhang Z, Wen T, Song T, Ju D. Development of an intelligent design and simulation aid system for heat treatment processes based on LLM. Mater Des 2024;248:113506. http://dx.doi.org/10.1016/j.matdes.2024.113506.
[130] Kumar V, Gleyzer L, Kahana A, Shukla K, Karniadakis GE. Mycrunchgpt: A LLM assisted framework for scientific machine learning. J Mach Learn Model Comput 2023;4(4):41–72.
[131] Chen Y, Zhu X, Zhou H, Ren Z. MetaOpenFOAM: an LLM-based multi-agent framework for CFD. 2024, arXiv preprint arXiv:2407.21320.
[132] Xu Q, Zhou G, Zhang C, Chang F, Cao Y, Zhao D. Generative AI and digital twin integrated intelligent process planning: A conceptual framework. 2023, PREPRINT (Version 1) available at Research Square.
[133] Lee J, Su H. A unified industrial large knowledge model framework in smart manufacturing. 2023, arXiv preprint arXiv:2312.14428.
[134] Ni M, Wang T, Leng J, Chen C, Cheng L. A large language model-based manufacturing process planning approach under industry 5.0. Int J Prod Res 2025;1–20.
[135] Potthoff L, Naussedat R, Gunnemann L. Exploring generative ai's role in manual assembly: Application potentials and use concepts. Procedia CIRP 2024;130:194–9. http://dx.doi.org/10.1016/j.procir.2024.10.075, 57th CIRP Conference on Manufacturing Systems 2024 (CMS 2024).
[136] Mandvikar S, Achanta A. Process automation 2.0 with generative AI framework. Int J Sci Res(Raipur) 2023;12(10):1614–9.
[137] Meyer F, Freitag L, Hinrichsen S, Niggemann O. Potentials of large language models for generating assembly instructions. In: 2024 IEEE 29th international conference on emerging technologies and factory automation. ETFA, 2024, p. 1–8. http://dx.doi.org/10.1109/ETFA61755.2024.10710806.
[138] Azab A, Osman H, Baki F. CAPP-GPT: A computer-aided process planning-generative pretrained transformer framework for smart manufacturing. Manuf Lett 2024;41:51–62. http://dx.doi.org/10.1016/j.mfglet.2024.09.009, 52nd SME North American Manufacturing Research Conference (NAMRC 52).
[139] van Erp T, Goncalves R, Rytter NGM. Design of matrix production systems: A skill-based systems engineering approach. Procedia CIRP 2023;120:1173–8. http://dx.doi.org/10.1016/j.procir.2023.09.144, 56th CIRP International Conference on Manufacturing Systems 2023.
[140] Holland M, Chaudhari K. Large language model based agent for process planning of fiber composite structures. Manuf Lett 2024;40:100–3. http://dx.doi.org/10.1016/j.mfglet.2024.03.010.
[141] Xu Q, Zhou G, Zhang C, Chang F, Cao Y, Zhao D. Generative AI and DT integrated intelligent process planning: a conceptual framework. Int J Adv Manuf Technol 2024;133(5):2461–85.
[142] Zhang C, Xu Q, Yu Y, Zhou G, Zeng K, Chang F, Ding K. A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing. Robot Comput-Integr Manuf 2025;92:102883.
[143] Wang T, Fan J, Zheng P. An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing. J Manuf Syst 2024;75:299–305. http://dx.doi.org/10.1016/j.jmsy.2024.04.020.
[144] Li C, Chrysostomou D, Yang H. A speech-enabled virtual assistant for efficient human–robot interaction in industrial environments. J Syst Softw 2023;205:111818. http://dx.doi.org/10.1016/j.jss.2023.111818.
[145] Joglekar O, Kozlovsky S, Lancewicki T, Tchuiev V, Feldman Z, Di Castro D. Towards natural language-driven industrial assembly using foundation models. In: ICLR 2024 workshop on large language model (LLM) agents. 2024.
[146] Mathew JG, Monti F, Firmani D, Leotta F, Mandreoli F, Mecella M. Composing smart data services in shop floors through large language models. In: International conference on service-oriented computing. Springer; 2024, p. 287–96.
[147] Zeydan E, Mangues J, Arslan SS, Turk Y, Liyanage M. Generative artificial intelligence for intent-based industrial automation. IEEE Consum Electron Mag 2024;1–6. http://dx.doi.org/10.1109/MCE.2024.3490780.
[148] Ahn J, Yun S, Kwon J-W, Kim W-T. Literacy deep reinforcement learning-based federated digital twin scheduling for the software-defined factory. Electron (2079- 9292) 2024;13(22).
[149] Xia Y, Shenoy M, Jazdi N, Weyrich M. Towards autonomous system: flexible modular production system enhanced with large language model agents. In: 2023 IEEE 28th international conference on emerging technologies and factory automation. ETFA, 2023, p. 1–8. http://dx.doi.org/10.1109/ETFA54631.2023.10275362.
[150] Zhao Z, Tang D, Zhu H, Zhang Z, Chen K, Liu C, Ji Y. A large language model-based multi-agent manufacturing system for intelligent shopfloor. 2024, arXiv preprint arXiv:2405.16887.
[151] Gkournelos C, Konstantinou C, Makris S. An LLM-based approach for enabling seamless human-robot collaboration in assembly. CIRP Ann 2024;73(1):9–12.
[152] Tao L, Li S, Huang Q, Zhao Z, Su X, Jin K. Research of preventive maintenance plans for wind power equipment based on maintenance knowledge fusion large model. IFAC-Pap 2024;58(29):361–6.
[153] Chen Y, Liu C. Remaining useful life prediction: A study on multidimensional industrial signal processing and efficient transfer learning based on large language models. 2024, arXiv preprint arXiv:2410.03134.
[154] Tao L, Li S, Liu H, Huang Q, Ma L, Ning G, Chen Y, Wu Y, Li B, Zhang W, et al. An outline of prognostics and health management large model: Concepts, paradigms, and challenges. 2024, arXiv preprint arXiv:2407.03374.
[155] Lukens S, McCabe LH, Gen J, Ali A. Large language model agents as prognostics and health management copilots. In: Annual conference of the PHM society. vol. 16, 2024, 1.
[156] Paroha AD, Chotrani A. A comparative analysis of TimeGPT and time-LLM in predicting ESP maintenance needs in the oil and gas sector. Int J Comput Appl 2024;975:8887.

[157] Wang H, Li Y-F. Large language model empowered by domain-specific knowledge base for industrial equipment operation and maintenance. In: 2023 5th international conference on system reliability and safety engineering. SRSE, 2023, p. 474–9. http://dx.doi.org/10.1109/SRSE59585.2023.10336112.
[158] Cao X, Xu W, Zhao J, Duan Y, Yang X. Research on large language model for coal mine equipment maintenance based on multi-source text. Appl Sci 2024;14(7). http://dx.doi.org/10.3390/app14072946.
[159] Wang P, Karigiannis J, Gao RX. Ontology-integrated tuning of large language model for intelligent maintenance. CIRP Ann 2024;73(1):361–4. http://dx.doi.org/10.1016/j.cirp.2024.04.012.
[160] Gu Z, Zhu B, Zhu G, Chen Y, Tang M, Wang J. AnomalyGPT: Detecting industrial anomalies using large vision-language models. Proc AAAI Conf Artif Intell 2024;38(3):1932–40. http://dx.doi.org/10.1609/aaai.v38i3.27963.
[161] Wang H, Li C, Li Y-F, Tsung F. An intelligent industrial visual monitoring and maintenance framework empowered by large-scale visual and language models. IEEE Trans Ind Cyber-Phys Syst 2024;2:166–75. http://dx.doi.org/10.1109/TICPS.2024.3414292.
[162] Fu T, Liu S, Li P. Intelligent smelting process, management system: Efficient and intelligent management strategy by incorporating large language model. Front Eng Manag 2024;11(3):396–412.
[163] Chen H. How do people's attitudes towards AI in different national contexts affect the development of AI in the automobile industry?: A cross-country study of Germany and China [Master's thesis], University of Twente; 2024.
[164] Ma Y, Zheng S, Yang Z, Pan H, Hong J. A knowledge-graph enhanced large language model-based fault diagnostic reasoning and maintenance decision support pipeline towards industry 5.0. Int J Prod Res 2025;1–22.
[165] Tony Liu D, William Xu X. A review of web-based product data management systems. Comput Ind 2001;44(3):251–62. http://dx.doi.org/10.1016/S0166-3615(01)00072-0.
[166] Xue S, Jiang C, Shi W, Cheng F, Chen K, Yang H, Zhang Z, He J, Zhang H, Wei G, et al. Db-gpt: Empowering database interactions with private large language models. 2023, arXiv preprint arXiv:2312.17449.
[167] Li P, He Y, Yashar D, Cui W, Ge S, Zhang H, Fainman DR, Zhang D, Chaudhuri S. Table-gpt: Table-tuned gpt for diverse table tasks. 2023, arXiv preprint arXiv:2310.09263.
[168] Xiong G, Bao J, Zhao W. Interactive-kbqa: Multi-turn interactions for knowledge base question answering with large language models. 2024, arXiv preprint arXiv:2402.15131.
[169] Jiang J, Zhou K, Dong Z, Ye K, Zhao WX, Wen J-R. Structgpt: A general framework for large language model to reason over structured data. 2023, arXiv preprint arXiv:2305.09645.
[170] Du K, Yang B, Xie K, Dong N, Zhang Z, Wang S, Mo F. LLM-MANUF: An integrated framework of fine-tuning large language models for intelligent decision-making in manufacturing. Adv Eng Inform 2025;65:103263.
[171] Ren M, Fan J, Yu C, Zheng P. CockpitGemini: A personalized design framework for smart vehicle cockpits integrating generative model-based multi-agent systems and human digital twins. Int J AI Mater Des 2024;1(3):4–19.
[172] Sun C, Yang K, Reddy RG, Fung YR, Chan HP, Small K, Zhai C, Ji H. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. 2024, arXiv preprint arXiv:2402.11060.
[173] Shang F, Zhao F, Zhang M, Sun J, Shi J. Personalized recommendation systems powered by large language models: Integrating semantic understanding and user preferences. Int J Innov Res Eng Manag 2024;11(4):39–49.
[174] Xu S, Wei Y, Zheng P, Zhang J, Yu C. LLM enabled generative collaborative design in a mixed reality environment. J Manuf Syst 2024;74:703–15. http://dx.doi.org/10.1016/j.jmsy.2024.04.030.
[175] Luo J, Ouyang C, Jing Y, Fang H, Xiao Y, Zhang Q, Li R. Application of LLM techniques for data insights in DHP. In: 2024 IEEE 4th international conference on digital twins and parallel intelligence. DTPI, 2024, p. 656–61. http://dx.doi.org/10.1109/DTPI61353.2024.10778677.
[176] Gao J, Guo Y, Lim G, Zhang T, Zhang Z, Li TJ-J, Perrault ST. CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In: Proceedings of the 2024 CHI conference on human factors in computing systems. CHI '24, New York, NY, USA: Association for Computing Machinery; 2024, http://dx.doi.org/10.1145/3613904.3642002.
[177] Xue S, Jiang C, Shi W, Cheng F, Chen K, Yang H, Zhang Z, He J, Zhang H, Wei G, Zhao W, Zhou F, Qi D, Yi H, Liu S, Chen F. DB-GPT: Empowering database interactions with private large language models. 2023, arXiv preprint arXiv:2312.17449. URL https://arxiv.org/abs/2312.17449.
[178] Xue S, Qi D, Jiang C, Shi W, Cheng F, Chen K, Yang H, Zhang Z, He J, Zhang H, Wei G, Zhao W, Zhou F, Yi H, Liu S, Yang H, Chen F. Demonstration of DB-GPT: Next generation data interaction system empowered by large language models. In: Proceedings of the VLDB endowment. 2024, URL https://arxiv.org/abs/2404.10209.
[179] Zekhnini K, Cherrafi A, Bouhaddou I, Benghabrit Y, Garza-Reyes JA. Supply chain management 4.0: a literature review and research framework. Benchmarking: An Int J 2021;28(2):465–501.
[180] Richey Jr RG, Chowdhury S, Davis-Sramek B, Giannakis M, Dwivedi YK. Artificial intelligence in logistics and supply chain management: A primer and roadmap for research. J Bus Logist 2023;44(4):532–49.
[181] Skórnóg D, Kmiecik M. Supporting the inventory management in the manufacturing company by ChatGPT. Logforum 2023;19(4).
[182] Kumar A, Gupta N, Bapat G. Who is making the decisions? How retail managers can use the power of ChatGPT. J Bus Strat 2024;45(3):161–9.
[183] Li B, Mellou K, Zhang B, Pathuri J, Menache I. Large language models for supply chain optimization. 2023, arXiv preprint arXiv:2307.03875.
[184] Lin MS, Tang CGY, Kom XJ, Eyu JY, Xu C. Building a natural language processing model to extract order information from customer orders for interpretative order management. In: 2022 IEEE international conference on industrial engineering and engineering management. IEEM, 2022, p. 0081–6. http://dx.doi.org/10.1109/IEEM55944.2022.9989801.
[185] Duman Keles F, Wijewardena PM, Hegde C. On the computational complexity of self-attention. In: Agrawal S, Orabona F, editors. Proceedings of the 34th international conference on algorithmic learning theory. Proceedings of machine learning research, vol. 201, PMLR; 2023, p. 597–619, URL https://proceedings.mlr.press/v201/duman-keles23a.html.
[186] Van de Ven GM, Tuytelaars T, Tolias AS. Three types of incremental learning. Nat Mach Intell 2022;4(12):1185–97.
[187] Zhu Z, Wang X, Zhao W, Min C, Deng N, Dou M, Wang Y, Shi B, Wang K, Zhang C, et al. Is sora a world simulator? A comprehensive survey on general world models and beyond. 2024, arXiv preprint arXiv:2405.03520.
[188] Sun C, Huang S, Pompili D. LLM-based multi-agent reinforcement learning: Current and future directions. 2024, arXiv preprint arXiv:2405.11106.
[189] Wang Y, Gu Z, Zhang S, Zheng S, Wang T, Li T, Feng H, Xiao Y. LLM-GAN: Construct generative adversarial network through large language models for explainable fake news detection. 2024, arXiv preprint arXiv:2409.01787.
[190] Kurkute MV, Namperumal G, Selvaraj A. Scalable development and deployment of LLMs in manufacturing: Leveraging AI to enhance predictive maintenance, quality control, and process automation. Aust J Mach Learn Res Amp; Appl 2023;3(2):381–430.
[191] Li Z, He Y, Yu H, Kang J, Li X, Xu Z, Niyato D. Data heterogeneity-robust federated learning via group client selection in industrial IoT. IEEE Internet Things J 2022;9(18):17844–57. http://dx.doi.org/10.1109/JIOT.2022.3161943.
[192] Lyon RF. Machine hearing: An emerging field [exploratory DSP]. IEEE Signal Process Mag 2010;27(5):131–9. http://dx.doi.org/10.1109/MSP.2010.937498.
[193] Al-Shaaby A, Aljamaan H, Alshayeb M. Bad smell detection using machine learning techniques: a systematic literature review. Arab J Sci Eng 2020;45(4):2341–69.
[194] Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: A survey. Int J Comput Vis 2021;129(6):1789–819.
[195] Kethireddy RR. Secure model distribution and deployment for LLMs. J Recent Trends Comput Sci Eng (JRTCSE) 2024;12(4):1–14.
[196] John MM, Holmström Olsson H, Bosch J. Architecting AI deployment: A systematic review of state-of-the-art and state-of-practice literature. In: Software business: 11th international conference, ICSOB 2020, Karlskrona, Sweden, November 16–18, 2020, proceedings 11. Springer; 2021, p. 14–29.
[197] Rietsche R, Dremel C, Bosch S, Steinacker L, Meckel M, Leimeister J-M. Quantum computing. Electron Mark 2022;32(4):2525–36.
[198] Sakib SK, Bijoy Das A. Challenging fairness: A comprehensive exploration of bias in LLM-based recommendations. In: 2024 IEEE international conference on big data. BigData, 2024, p. 1585–92. http://dx.doi.org/10.1109/BigData62323.2024.10825082.
[199] Heersmink R, de Rooij B, Clavel Vázquez MJ, Colombo M. A phenomenology and epistemology of large language models: transparency, trust, and trustworthiness. Ethics Inf Technol 2024;26(3):41.
[200] Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confid Comput 2024;4(2):100211. http://dx.doi.org/10.1016/j.hcc.2024.100211.
[201] Wang W, Zhang P, Sun C, Feng D. Smart customer service in unmanned retail store enhanced by large language model. Sci Rep 2024;14(1):19838.
[202] Jiang P, Sonne C, Li W, You F, You S. Preventing the immense increase in the life-cycle energy and carbon footprints of LLM-powered intelligent chatbots. Engineering 2024;40:202–10. http://dx.doi.org/10.1016/j.eng.2024.04.002.
[203] Moeuf A, Lamouri S, Pellerin R, Tamayo-Giraldo S, Tobon-Valencia E, Eburdy R. Identification of critical success factors, risks and opportunities of industry 4.0 in SMEs. Int J Prod Res 2020;58(5):1384–400. http://dx.doi.org/10.1080/00207543.2019.1636323, arXiv:https://doi.org/10.1080/00207543.2019.1636323.