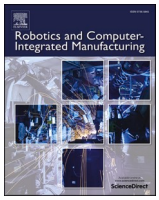
Robotics and Computer–Integrated Manufacturing 92 (2025) 102883

Contents lists available at ScienceDirect

Robotics and Computer-Integrated Manufacturing

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



Review

A survey on potentials, pathways and challenges of large language models in new-generation intelligent manufacturing

Chao Zhang [a,b], Qingfeng Xu [a,b], Yongrui Yu [a,b], Guanghui Zhou [a,b,*], Keyan Zeng [a,b], Fengtian Chang [c], Kai Ding [c]

[a] State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710054, China
[b] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[c] School of Construction Machinery, Chang'an University, Xi'an 710064, China

ARTICLE INFO

Keywords:
Large language models
Intelligent manufacturing
New-generation artificial intelligence
LLM applications
Industry 5.0

ABSTRACT

Nowadays, Industry 5.0 starts to gain attention, which advocates that intelligent manufacturing should adequately consider the roles and needs of humans. In this context, how to enhance human capabilities or even liberate humans from the processes of perception, learning, decision-making, and execution has been one of the key issues to be addressed in intelligent manufacturing. Large language models (LLMs), as the breakthrough in new-generation artificial intelligence, could provide human-like interaction, reasoning, and replies suitable for various application scenarios, thus demonstrating significant potential to address the above issues by providing aid or becoming partners for humans in perception, learning, decision-making, and execution in intelligent manufacturing. The combination of LLMs and intelligent manufacturing has inherent advantages and is expected to become the next research hotspot. Hence, this paper primarily conducts a systematic literature review on the application of LLMs in intelligent manufacturing to identify the promising research topics with high potential for further investigations. Firstly, this paper reveals the concept, connotation, and foundational architecture of LLMs. Then, several typical and trending interdisciplinary LLM applications, such as healthcare, drug discovery, social & economic, education, and software development, are summarized, on which an LLM-enabled intelligent manufacturing architecture is designed to provide a reference for applying LLMs in intelligent manufacturing. Thirdly, the specific pathways for applying LLMs in intelligent manufacturing are explored from the perspectives of design, production, and service. Finally, this paper identifies the limitations, barriers, and challenges that will be encountered during the research and application of LLMs in intelligent manufacturing, while providing potential research directions to address these limitations, barriers, and challenges.

1. Introduction

Since the European Commission announced its Policy Brief on the Fifth Industrial Revolution in 2021, Industry 5.0 has been gaining more attention in recent years [1–3]. Industry 5.0 complements the existing Industry 4.0 paradigm by highlighting research and innovation as drivers for a transition to a sustainable, human-centric, and resilient industry. The evolution of the industrial revolution has driven transformative developments in manufacturing systems [4], typically from the machine-centric traditional manufacturing to system-centric intelligent manufacturing (IM), till new-generation intelligent manufacturing (New-IM) [5,6]. As shown in Fig. 1, with the evolution of manufacturing systems, the roles of humans have evolved from the coexistence of human-machines in traditional manufacturing to the cooperation and collaboration in current IM, and will eventually evolve into compassion and coevolution with machines in New-IM. Meanwhile, human needs have also evolved from the basic safety and health in traditional manufacturing to the belonging in current IM, and will eventually evolve into the esteem and self-actualization in New-IM [7,8]. In this context, how to enhance human capabilities or even liberate humans from the processes of perception, learning, decision-making, and execution has been one of the key issues to be solved in New-IM.

New-generation artificial intelligence has achieved enormous progress in recent years, propelling the technological world forward [9]. The most recent advancement in this field is LLM [10], which has achieved significant success across various natural language processing (NLP)

* Corresponding author at: School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China.
E-mail address: ghzhou@mail.xjtu.edu.cn (G. Zhou).

https://doi.org/10.1016/j.rcim.2024.102883
Received 1 June 2024; Received in revised form 19 August 2024; Accepted 19 September 2024
Available online 26 September 2024

tasks, encompassing question-answering, summarization, and machine translation, among others. LLMs could provide human-like interaction, reasoning, and replies suitable for various application scenarios, thus providing great potential to maximize the basic interests of related humans and address the above issues by providing aids or becoming a partner in the processes of perception, learning, decision-making, and execution. In fact, as shown in Fig. 2, the trend of integration between LLM and New-IM has emerged, where LLM has been successfully used in several manufacturing scenarios, such as information retrieval [11], human-centric smart manufacturing [12,13], and intelligent operation and maintenance [14–16]. Nevertheless, the applications of LLM in New-IM is still in its initial stages. The potential, pathways, and challenges of LLM in New-IM remain under exploration.

Motivated by the above observations, the current state of LLMs is summarized, and their potentials, pathways, and challenges for improving humans in New-IM towards Industry 5.0 are explored in this paper. We searched within the topic words with "Intelligent manufacturing", "Large language model", "Industry 5.0" in the Web of Science database in recent years, using "OR" as a retrieval method between the above topic words. Based on the retrieval results, we focused on a total of 157 excellent and representative papers, such as "review papers", "hot papers", and "highly cited papers", etc., to explore the potential architecture, technological pathways, and challenges of applying LLMs in New-IM. Briefly, the main contributions of this paper are as follows:

- Revealing the concept, connotation, and foundation architecture of LLM: The concept and connotation of LLM are summarized according to its rise and evolutionary trajectory. Then, several typical foundational architectures of LLM are summarized, and their pros and cons are compared.
- Designing a New-IM-LLM architecture in the context of Industry 5.0, inspired by the interdisciplinary applications of LLM: Several typical and hotspot interdisciplinary LLM applications, such as healthcare, drug discovery, social & economic, education, and software development, are summarized, providing insights into the application of LLM in New-IM. Then, we design a New-IM-LLM architecture in the context of Industry 5.0, and discuss the ways in which LLM empowers New-IM considering its technological development and maturity.
- Exploring the pathways of LLMs when applying in design, production, and service: Based on the New-IM-LLM architecture, this paper further explores pathways for applying LLMs in New-IM-LLM from the perspectives of design, production, and service. Additionally, several impressive application examples are analyzed based on the current published works.
- Identification of challenges and future research directions towards New-IM-LLM: The application of LLM in New-IM is still in its embryonic stage, where tremendous challenges will be encountered during the applications of LLMs in this field. This paper summarizes these limitations, barriers, and challenges, and provides valuable research directions as potential solutions to address the above issues.

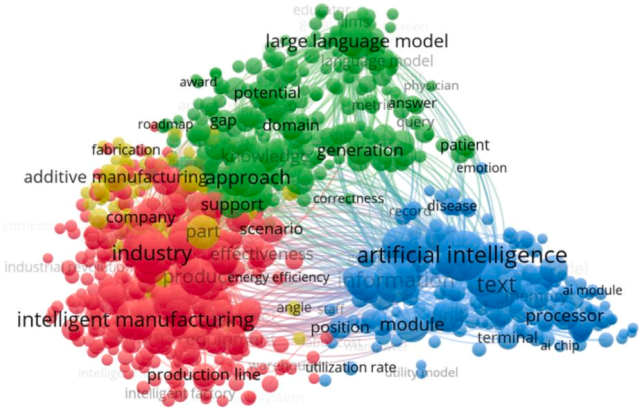


Fig. 2. Integration trends between LLM and New-IM.

The rest of the paper are structured as in Fig. 3. Section 2 reveals the concept, connotation, and foundation architecture of LLM, which provides a theoretical basis to understand the pros and cons in the application of LLM in New-IM. Inspired by the interdisciplinary applications of LLM in several research hotspots, Section 3 designs a New-IM-LLM architecture in the context of Industry 5.0, to provide a technical

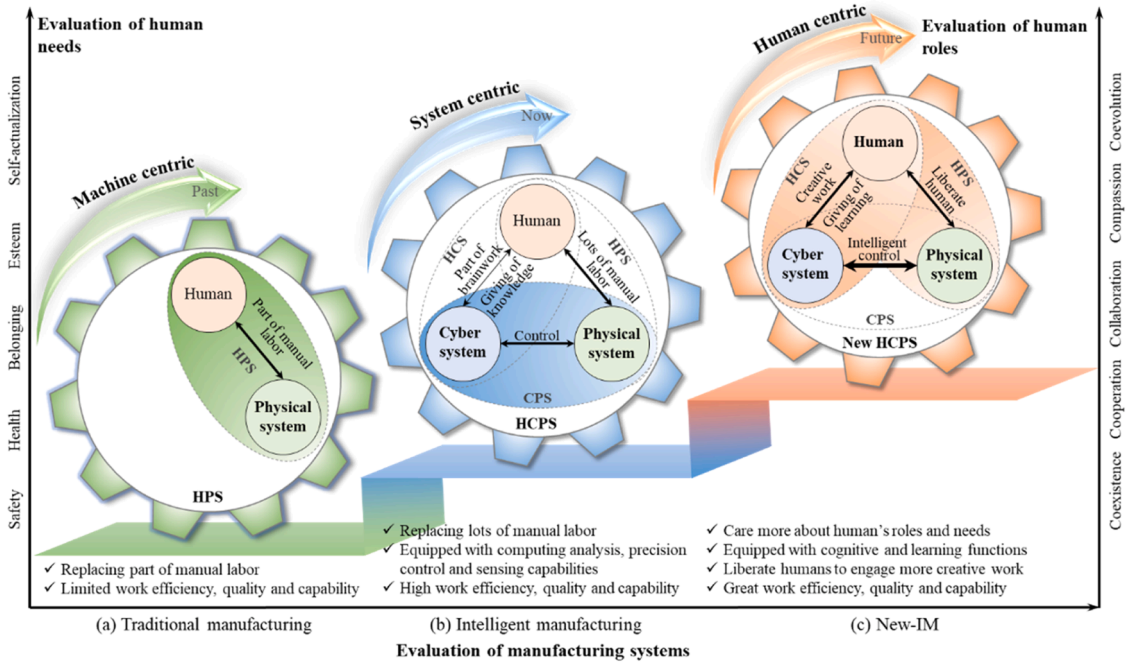


Fig. 1. Evolution of human roles and needs in manufacturing [4,5,7].

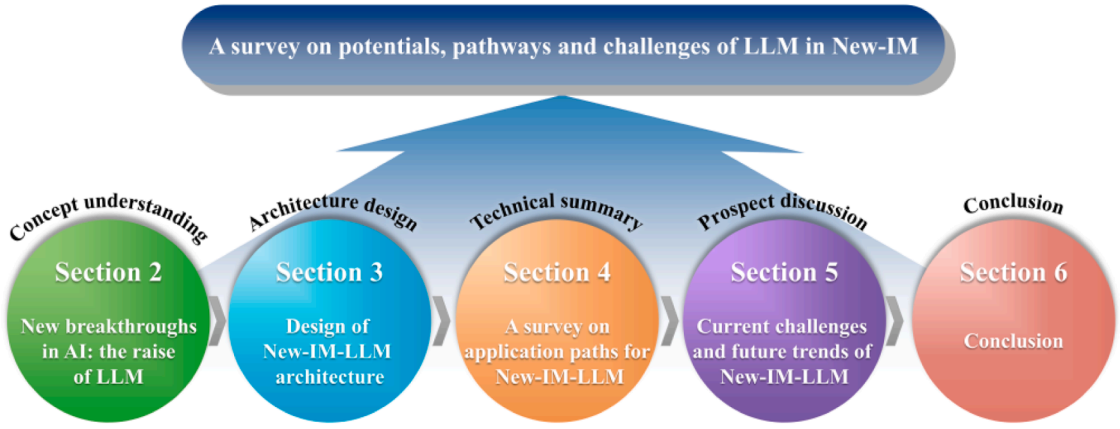


Fig. 3. The overall structure of the paper.

reference for applying LLMs in New-IM. Based on the framework, Section 4 explores pathways for applying LLMs in the design, production and service of New-IM. Section 5 summarizes the current challenges and provides future research directions for the application of LLM in New-IM. Finally, in Section 6, the contributions of this paper are concluded.

2. New breakthroughs in AI: the raise of LLM

The concept and connotation of LLM are revealed according to its rise and evolutionary vein. Then, several typical foundation architectures of LLM are summarized, where their pros and cons are compared.

2.1. The raise and evolutionary vein of LLM

In recent years, artificial intelligence has achieved significant advancements, especially in the development of LLMs, which have led the forefront of technological innovation. LLMs have the capacity to understand complex language patterns and generate expert-level text responses, enabling them to perform NLP tasks such as question answering, summarization, and machine translation with flexibility. Since the release of ChatGPT in 2022, LLMs have quickly become a central focus of research and interest in both industry and academia [17–19]. To better understand the concept and connotation of LLMs, this section summarizes the rise and evolutionary vein of LLMs, which starts from rule-based language model (LM), to statistical LM, neural LM, and pre-trained LM, till LLM, as shown in Fig. 4.

The development of LLMs can be traced back to the 1950s with the introduction of rule-based language models, which relied on manually crafted rules for language processing. Although these models provided good interpretability, they were limited in handling complex natural language texts. In the 1990s, the advent of hidden Markov models (HMM) [20] and Gaussian mixture models (GMM) [21] facilitated the widespread adoption of statistical LMs, such as n-gram models [22–24]. Despite improving task performance, these models struggled to capture complex contextual relationships [25,26]. The emergence of neural LMs [27–29] in 2010 marked a significant breakthrough, with models like Recurrent Neural Network (RNN) [28], Long short-term memory (LSTM) [30], and Gated Recurrent Units (GRU) [31] being better equipped to manage long-distance dependencies [32]. In 2013, word2vec [33,34] introduced a novel method for word vector representation using neural networks, effectively capturing semantic and syntactic relationships in vector space. However, challenges in managing long-distance dependencies persisted. In 2018, pre-trained LMs based on the Transformer architecture [35], such as BERT [36], successfully addressed these challenges, significantly enhancing the performance of NLP tasks.

This breakthrough spurred extensive research on generative models, such as GPT-2 [37] and BART [38], and gradually established the "pre-training and fine-tuning" learning paradigm. Researchers found that model size significantly impacts the performance of downstream tasks. Since 2020, large-scale pre-trained LMs, such as GPT-3 (175 billion parameters) [39], have demonstrated capabilities far surpassing those of smaller models. Thus, the term "large language models" emerged and quickly garnered widespread attention.

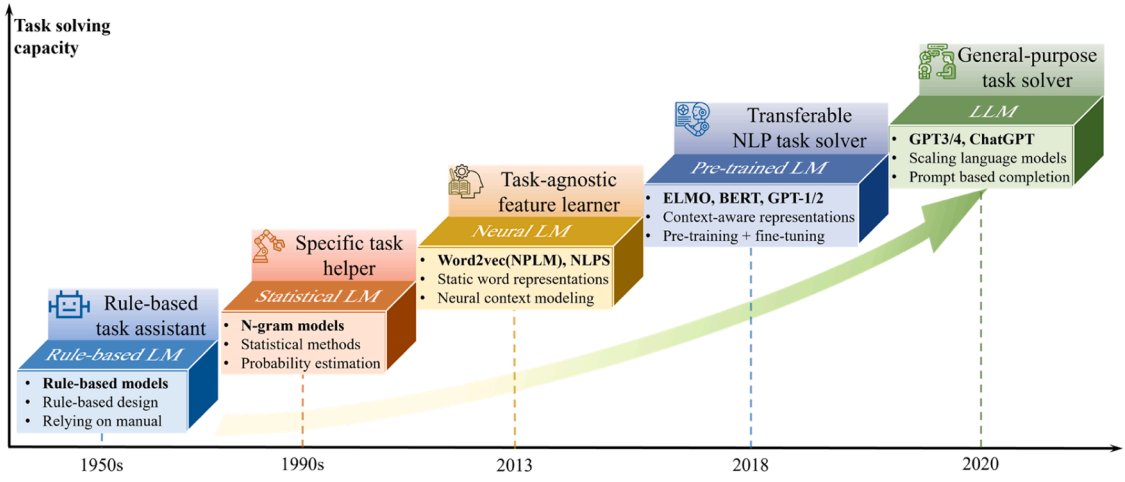


Fig. 4. Raise and evaluation vein of LLM.

2.2. Foundation architecture of LLM

As discussed in Section 2.1, the Transformer architecture was introduced to address the limitations of neural LMs, such as RNNs, in handling variable-length sequences and context awareness. As shown in Fig. 5, this architecture consists of an encoder and a decoder, with core components including multi-head attention mechanisms and feedforward neural networks that enhance its ability to manage long-term dependencies and support high parallelism [40]. These features make the Transformer highly effective across a range of NLP tasks, particularly well-suited for large-scale pre-training. Consequently, it has become the backbone architecture for mainstream models such as T5 [41], BERT, and the GPTs [37,39,42]. This section introduces the architecture and characteristics of these models, while Table 1 details the configurations of models with different architectures, including their pros and cons, applicable NLP tasks (AT in NLP), and links to their respective open-source code repositories.

Encoder-Decoder. The encoder-decoder architecture captures contextual information using a bidirectional self-attention mechanism in the encoder and generates high-quality text in the decoder through a cross-attention mechanism. As shown in Fig. 6(a), this architecture comprises an encoder and a decoder. The encoder processes input vectors through an embedding layer and multiple Transformer modules, producing intermediate representations. The decoder then links these input vectors with the encoder's output via cross-attention layers to generate the final output. This architecture is particularly well-suited for tasks requiring input comprehension and output generation, such as machine translation and text summarization. Its primary strengths are its robust bidirectional information processing and flexibility. However, because it includes both an encoder and a decoder, it requires more computational resources and longer training times. Typical large language models based on this architecture include T5 and BART.

Encoder. The encoder processes input data using self-attention layers and feedforward neural networks to extract rich feature representations. As shown in Fig. 6(b), the encoder comprises multiple encoding blocks, each containing a Trm module that captures global contextual information through self-attention layers. The encoder's strength lies in its robust contextual understanding and efficient feature extraction, making it well-suited for tasks such as text similarity analysis, text classification, and named entity recognition. However, without an autoregressive decoder, these models are not suitable for text generation tasks. Furthermore, although they perform exceptionally well across various tasks, they require fine-tuning to achieve optimal results, limiting their adaptability when annotated data is scarce. Typical encoder-based large language models include BERT and ERNIE 3.0 [43].

Decoder. The decoder processes input vectors using unidirectional self-attention and cross-attention layers, integrating the encoder's output information. As shown in Fig. 6(c), the decoder consists of multiple decoding blocks, each containing a Trm module that captures dependencies in the generated sequence through unidirectional self-attention layers. The decoder excels in tasks such as text generation, auto-completion, and dialogue generation, making it particularly suitable for real-time interactive applications. Unlike encoder-decoder models, decoder-only models generate responses immediately upon receiving input, making them ideal for applications like chatbots. These models excel in creative tasks, such as generating artistic texts, composing music, or writing code. However, their comprehension of complex texts and ability to handle detailed queries or information

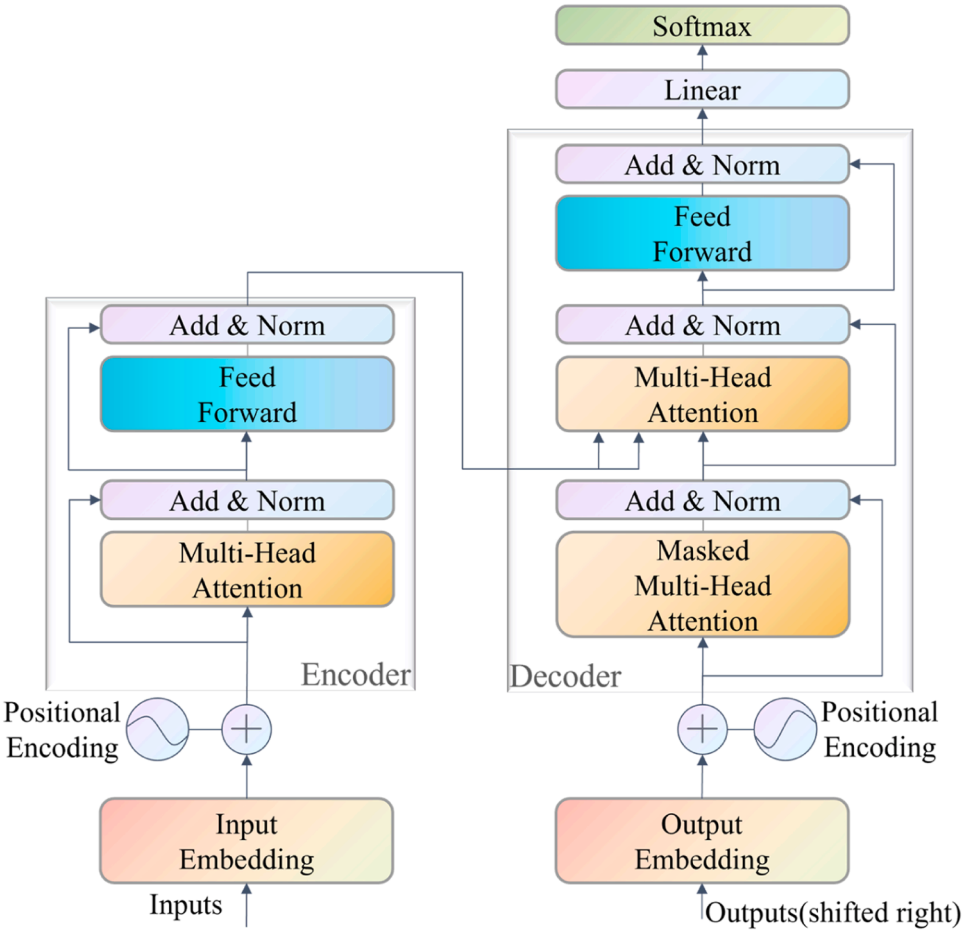


Fig. 5. Architecture of Transformer.

Table 1
Comparative analysis of mainstream LLM across different architectures.

Architecture		Pros	Cons	AT in NLP	Typical LLM	Year	Language	Size	code
Encoder-Decoder		• Strong bidirectional information processing capabilities; • Flexible and efficient.	• High computational resource Requirements; • challenging model Training.	• Text summarization; • Machine translation; • Question answering systems; • Data augmentation.	T5 [41]	2019	EN	13B	https://github.com/google-research/text-to-text-transfer-transformer
					BART [38]	2020	EN	6.5B	https://github.com/huggingface/transformers
Encoder		• Strong contextual understanding capability; • Efficient feature extraction.	• Weak text generation capability; • Limited flexibility.	• Text similarity analysis; • Text classification; • Named entity recognition.	BERT [36]	2018	EN	0.34B	https://github.com/google-research/bert
					ERNIE 3.0 [43]	2021	ZH	260B	https://github.com/PaddlePaddle/ERNIE
Decoder	Causal Decoder	• Strong generative capability; • Flexible and efficient; • Dynamic and innovative expansion; • Real-time interactive capability.	• Weaker context comprehension; • Amplifies biased information; • "Hallucination" issue.	• Text generation; • Text auto-completion; • Dialogue generation.	GPT-4 [52]	2023	EN	175B	/
					PaLM [53]	2020	Multi	540B	https://github.com/lucidrains/PaLM-pytorch
					Gopher [54]	2021	EN	280B	/
					LaMDA [55]	2022	EN	137B	https://github.com/conceptofmind/LaMDA-rlhf-pytorch
					OPT [56]	2022	EN	175B	https://github.com/facebookresearch/metaseq
					BLOOM [57]	2022	Multi	176B	https://github.com/bigscience-workshop/bigscience
					MT-NLG [58]	2022	EN	530B	https://github.com/microsoft/DeepSpeed
					LLaMA2 [59]	2023	Multi	70B	https://huggingface.co/meta-llama/Llama-2–70b-chat-hf
	Prefix Decoder				ChatGLM-130B [60]	2022	EN, ZH	130B	https://github.com/THUDM/GLM-130B

Note: All the tasks listed in the 'AT in NLP' column can be performed by all three models, but there are differences in their relative applicability to specific tasks.; The "Language" column in the table refers to the primary language used during the model's training; The "Size" column in the table indicates the number of parameters in the model, usually measured in billions (B).

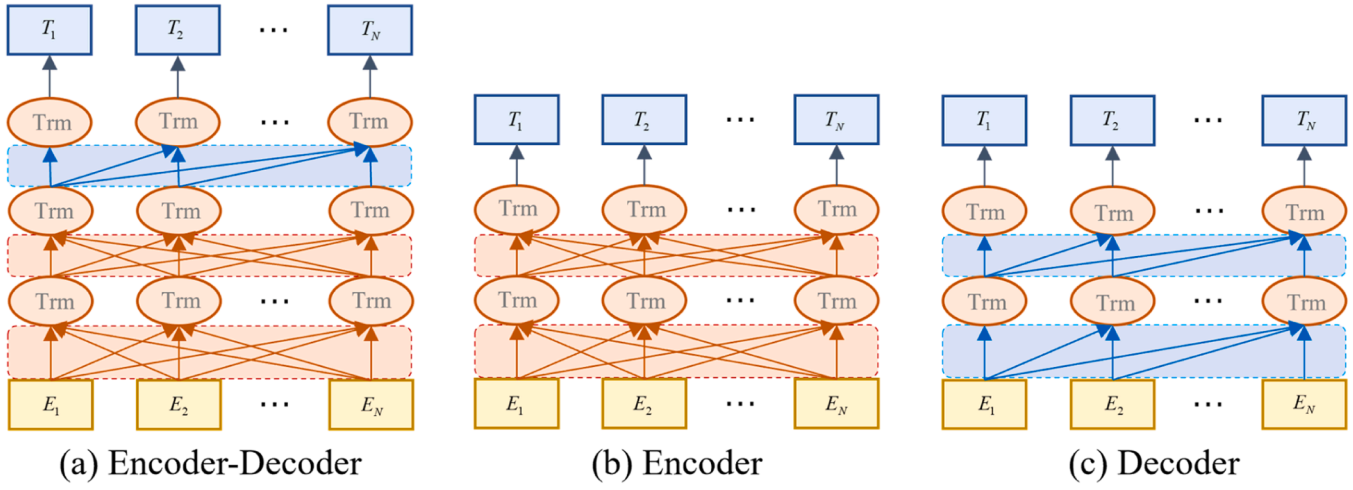


Fig. 6. Architecture categories of LLMs.

extraction is weaker compared to encoder or encoder-decoder models [44,45]. Long-form generation may lead to "hallucinations" or topic drift, and biases in the training data may be amplified. Most current models, such as the GPT series, use decoder architectures, including causal decoders and prefix decoders, with Chat General Language Model (ChatGLM) [46] being a notable example.

In the context of Industry 5.0, New-IM is steadily progressing toward deep collaboration between humans and intelligent systems. LLMs are pivotal in driving this transformation. From knowledge comprehension [47] and intelligent interaction [48] to customized production [49] and intelligent maintenance [50,51], LLM architectures offer robust support for the future development of the manufacturing industry. Leveraging their advanced natural language processing and generation capabilities, these models enable manufacturing systems to efficiently interpret

complex data and requirements, generating solutions that drive intelligent automation in production processes. As LLM technology continues to evolve and be applied, the manufacturing sector will be better positioned to respond to market demands, achieve deeper collaboration between humans and intelligent systems, and elevate the industry to new heights.

3. Design of new-IM-LLM architecture inspired by cross-disciplinary insights

Several typical and hot-spot interdisciplinary LLMs applications, such as healthcare, drug discovery, social and economic fields, education, and software development, are summarized in this section. Inspired by the above interdisciplinary LLMs applications, a New-IM-LLM architecture in the context of the Industry 5.0 framework is designed.

3.1. Dissecting LLM applications: insights and competences

Research assessing the capabilities of LLMs across diverse science and engineering domains revealed consistent performance in understanding, reasoning, and generation, irrespective of specialization [61]. This consistent performance implies that findings and insights derived from other specialized domains can be extrapolated to the manufacturing domain. Consequently, insights gained from diverse disciplines can be capitalized on to expedite the development of LLMs applications in manufacturing through information sharing across disciplines. This section conducts a comprehensive analysis of LLMs applications across various domains and distills its core competences, thereby informing the design of the New-IM-LLM architecture.

3.1.1. A cross-disciplinary comparison of LLM applications

The analysis primarily centered on stakeholder-centric applications within or closely associated with engineering disciplines. This approach facilitated a manageable scope, directing attention to applications aimed at specific deliverables, unlike artistic creation or basic science pursuits, which may prioritize exploration or discovery. As indicated above, this section presents a consolidated overview derived from published studies across five diverse domains in science and engineering: Healthcare, Drug Discovery, Social & Economic, Education, and Software Development. As shown in Fig. 7, through the cross-disciplinary comparison of the five domains mentioned above, we have gained insights into LLMs applications and identified several paradigms for implementing LLM in intelligent manufacturing.

LLM in healthcare. The potential applications of LLM in healthcare for both patients and providers are numerous [62], with case studies in medical disciplines such as ophthalmology [63], cardiology [64], and obstetrics and gynecology [65]. Specifically, LLM can serve as a valuable tool for remote triage of patients: ChatGPT can already deliver higher quality responses to patient queries without the need for fine-tuning [66, 67]. Although there are still limitations in terms of feedback efficiency and expertise depth, this progress greatly reduces the threshold for manufacturing fields with strong and broad professional characteristics. In addition, the performance of LLMs in medical applications can be further enhanced by fine-tuning and integrating multimodal data [68], indicating that LLMs specialized in the manufacturing field can also prove useful by providing new functionalities. Another enlightening example is the application of LLM in electronic medical records (EMR), addressing the challenge of processing heterogeneous medical data [69]. Similarly, manufacturing process sheets, which encompass structured and unstructured text, images, and experimental data akin to EMRs, are ripe for integrating LLMs for further applications. Furthermore, researchers have successfully empowered LLMs to autonomously develop machine learning models for clinical outcome prediction [70]. This underscores the potential of LLM in advanced data analysis and enhances the feasibility of facilitating its wider application in professional practice through specialized tuning.

LLM in drug discovery. The emergence of LLM has catalyzed a paradigm shift in the field of drug discovery and development [71]. Scientists have recognized LLM's utility in connecting and representing data in diverse formats, facilitating the transformation of discrete knowledge into well-structured mechanistic assertions and the generation of knowledge graphs (KGs) [72]. Leveraging LLMs, these KGs can be enriched to expedite the identification of potential drug targets [73]. For closely related manufacturing knowledge, such as manufacturing processes and failure information, LLM-empowered KGs also have extremely high application value. Additionally, LLM aids in predicting various drug features and provides a blueprint for drug compounds with novel structures [74]. This aligns with the objectives of manufacturing process planning for products. Moreover, researchers have harnessed

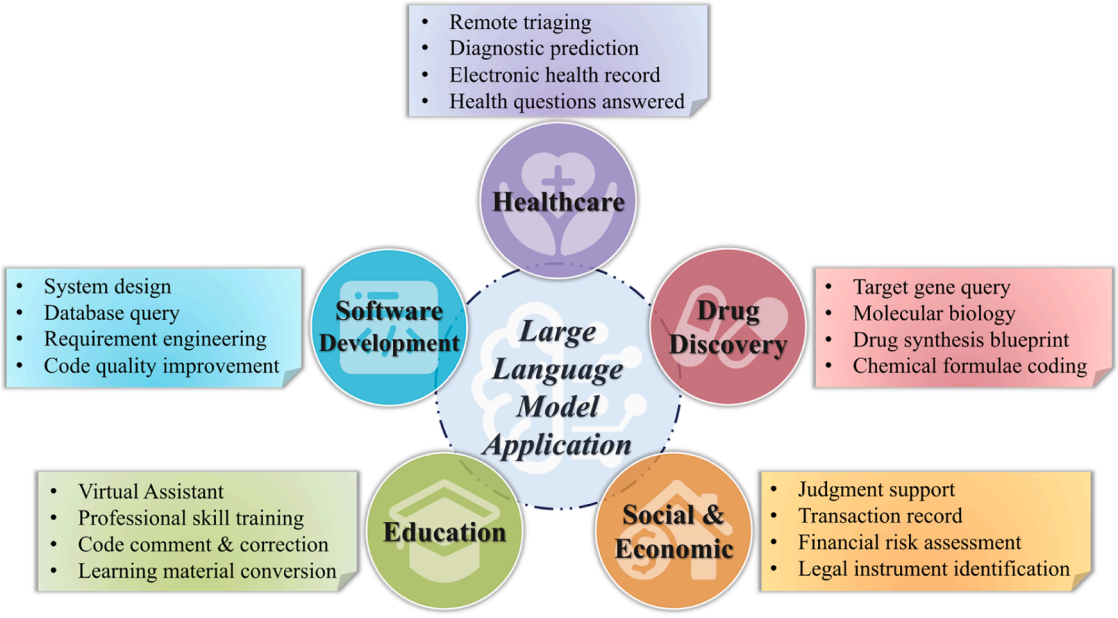


Fig. 7. Cross-disciplinary comparison of LLM applications.

LLMs to predict the synergistic effects of drug combinations in rare tissues lacking structured data and features [75]. This underscores LLM's potential as a viable alternative for inference in scenarios characterized by limited structured data and sample size, presenting promising implications for the manufacturing sector facing analogous challenges.

LLM in social and economic fields. The domain of social and economic is characterized by vast amounts of textual data necessitating analysis and decision-making. LLMs demonstrate prowess in conducting comprehensive data analysis to assist relevant organizations in making more informed decisions [76]. Particularly noteworthy is the exceptional performance of LLM in financial domain research [77], investment literacy [78], and dealing with legal texts [79] that are notorious for their complexity, formal language, and specialized terminology, which is achieved through fine-tuning the output quality by incorporating financial expertise and iterative refinement based on researchers' insights. This capability can play a similar role in manufacturing domains such as engineering and technical manuals, which can be tailored through fine-tuning to enable reasoning based on specialized domain corpora and interaction with technicians through question-and-answer systems [80]. Moreover, numerous e-commerce companies have deployed LLM applications to streamline operations, offering cost-effective and personalized products and interactions that reduce the cognitive burden associated with decision-making for consumers [81]. This decision-support capability holds significant promise in aiding technicians faced with intricate manufacturing scenarios characterized by multidisciplinary interconnections for process design and implementation.

LLM in education. The utilization of LLM in education emerges as a potential area of interest, addressing challenges inherent in traditional educational paradigms. Firstly, LLM's versatility across diverse domains facilitates the seamless integration of various digital education applications into a unified framework, thereby broadening the horizons of educational possibilities and experiences [82]. This is a positive sign for the entire manufacturing chain, which involves applications in multiple fields. In contrast, through prompt engineering, LLM also enables users to delineate the intended role explicitly in conversations, allowing for the differentiation of LLMs into distinct intelligent agents serving various functions (e.g., teacher agent, classmate agent), fostering collaborative efforts to maximize pedagogical effectiveness [83]. This underscores the potential of multi-agents in manufacturing scenarios which are characterized by complex tasks and roles. Analogous to its applications in other fields, the decision-making capabilities of LLM empower educators to craft personalized plans for individual students [84]. Moreover, it adeptly assimilates diverse learning materials provided by instructors, converting unstructured data into structured formats and generating appropriate content for student interaction [85]. Its user-friendly interface and compatibility engender a high receptivity among students in higher education settings [86], which is also a feature advantageous for its integration within the manufacturing industry, effectively reducing deployment costs for companies and lowering the knowledge threshold for workers.

LLM in software development. In software development field, LLM also exhibits significant potential in code completion and the generation of synthetic code from natural language, presenting a compelling opportunity for manufacturing fields with extensive programming requirements, such as robot operating systems and computer numerical control machine tools. In particular, LLMs can fulfill multiple roles, serving as development, testing, and verification experts [87], assisting human developers across all phases of the software development lifecycle through the collaboration of distributed autonomous agents [88]. Researchers have successfully implemented automated customization of algorithms for program generation using LLMs, thereby surmounting mathematical challenges [89]. Additionally, LLMs are employed in automating code review, enhancing code quality, and querying databases to refine software requirements and address technical challenges effectively [90]. Furthermore, LLMs facilitate code transformation across programming languages [91], a critical capability in manufacturing domains with diverse code usage scenarios. Moreover, LLMs' cross-programming language generalization capability proves invaluable for modeling needs in manufacturing domains with low programming dependency, such as CAD modeling and finite element simulation modeling [92].

3.1.2. Versatile competences of LLM

Application scenarios in other domains provide only superficial insights into LLM's potential in the manufacturing domain. To truly comprehend its capabilities, a deeper exploration of its intrinsic properties is required. Therefore, applications across various domains have been summarized, and an analysis of their corresponding LLMs competencies has been conducted.

The analysis presented in Table 2 reveals five remarkable competencies of LLMs in addressing distinct challenges across various domains. These competencies—natural language processing, professional perspective, data analysis, code generation, and human-machine interaction—enable LLMs to effectively tackle a wide range of problems. Many applications, of course, require the cooperation of multiple capabilities to accomplish their objectives. To streamline understanding and presentation, the most representative competencies have been chosen for correspondence. It should be noted that, since the table is summarized based on relevant literature in each domain, although we have tried to collect as much representative literature as possible, some application scenarios may have incomplete statistics. The empty entities in the table mean that, to the best of our knowledge, there is no application need for the corresponding competence in the field.

Firstly, LLMs adeptly processes natural language text, correlates context, and generates organized content based on specified requirements after fully understanding and analyzing it. Secondly, they offer professional insights and knowledge, catering to both novice users and addressing complex technical queries through fine-tuning. Thirdly, LLMs demonstrate proficiency in managing, analyzing, and deriving insights from large-scale, multimodal, and unstructured data sets. Additionally, LLMs efficiently generates functions and code based on natural language descriptions, streamlining software development and automation efforts. Lastly, LLMs excel in human-machine interaction with its multimodal perception capabilities, engaging users through question-answering, suggestions, and dialogue, which effectively improves the efficiency of the engineers involved.

These competences underscore the versatility and potential of LLMs to revolutionize problem-solving approaches and information dissemination across various domains, especially manufacturing.

Actually, the competences of LLMs have already led to noteworthy research and initial applications in various engineering fields beyond manufacturing, including materials engineering [92], chemical engineering [93], civil engineering [94], and electrical engineering [95]. The shared background and foundation among engineering disciplines provide valuable references for the manufacturing sector. If a problem arises that is beyond the technician's knowledge or is time-consuming and complex, LLMs can rapidly provide solutions and strategies relevant to the specialized area. By integrating multiple automation technologies with LLMs, intelligent agents can be developed based on the collaboration of multiple LLMs for the autonomous design, planning, and execution of complex experiments and processes [96]. Furthermore, by assimilating high-quality multimodal data and domain knowledge, LLMs with robust data processing and contextual correlation capabilities can accumulate richer and higher-quality knowledge reserves to better assist technicians in addressing technical challenges [97]. Additionally, LLMs can undertake technical specification reviews [98] and schedule planning [99] for engineering projects, mirroring application scenarios in the manufacturing field. It should be noted that while fine-tuning LLMs using specialized knowledge can enhance performance in most engineering application scenarios, independently collecting and fine-tuning expertise for each one in multi-agent systems is impractical.

Table 2
The applications and competence of LLMs in various fields.

Competence Field of application	Natural Language Processing	Professional Perspective	Data Analysis	Code Generation	Human-Machine Interaction
Healthcare	• Electronic health record [69]; • Translation and summarization [66].	• Supplementary diagnosis [62]; • Health questions answered [67].	• Clinical data analysis [63]; • Diagnostic prediction [70].	/	• Remote triaging [62]; • Personalizing patient visit [63].
Drug discovery	• Knowledge organization [72]; • Drug synthesis blueprint [74].	• Molecular biology [73].	• Drug characterization prediction [75].	• Chemical formulae code [71].	• Target gene query [73].
Social & Economic	• Financial transaction record [77]; • Legal instrument identification [79].	• Judgment basis support [79]; • Financial investment literacy [78].	• Financial risk assessment [76].	/	• Personalized product recommendation [81].
Education	• Learning material conversion [85]; • Lesson plan [84].	• Professional skill training [82].	• Learning outcome assessment [83].	• Code quality improvement [82].	• Virtual Assistant [86].
Software development	• Requirement engineering [90]; • Use case design [87].	• Solution generation [88]; • System design [89].	• Data modeling [92]; • Database query [87].	• Code quality improvement [90]; • Code conversion [91].	• Conversational software engineering [88].

However, LLMs' general competences are sufficient for simple tasks that do not require high accuracy, and the interaction with technicians through prompt engineering effectively bridges the performance gap between them.

The versatile competencies of LLMs demonstrate their substantial potential in intelligent manufacturing by facilitating accelerated design, production, and service optimization. Integrating LLMs can significantly improve manufacturing efficiency, foster technological innovation, and deliver strong decision-making support and automation throughout various stages of intelligent manufacturing. As LLMs become more deeply embedded in manufacturing environments, they are expected to streamline operations, enhance flexibility, and effectively tackle

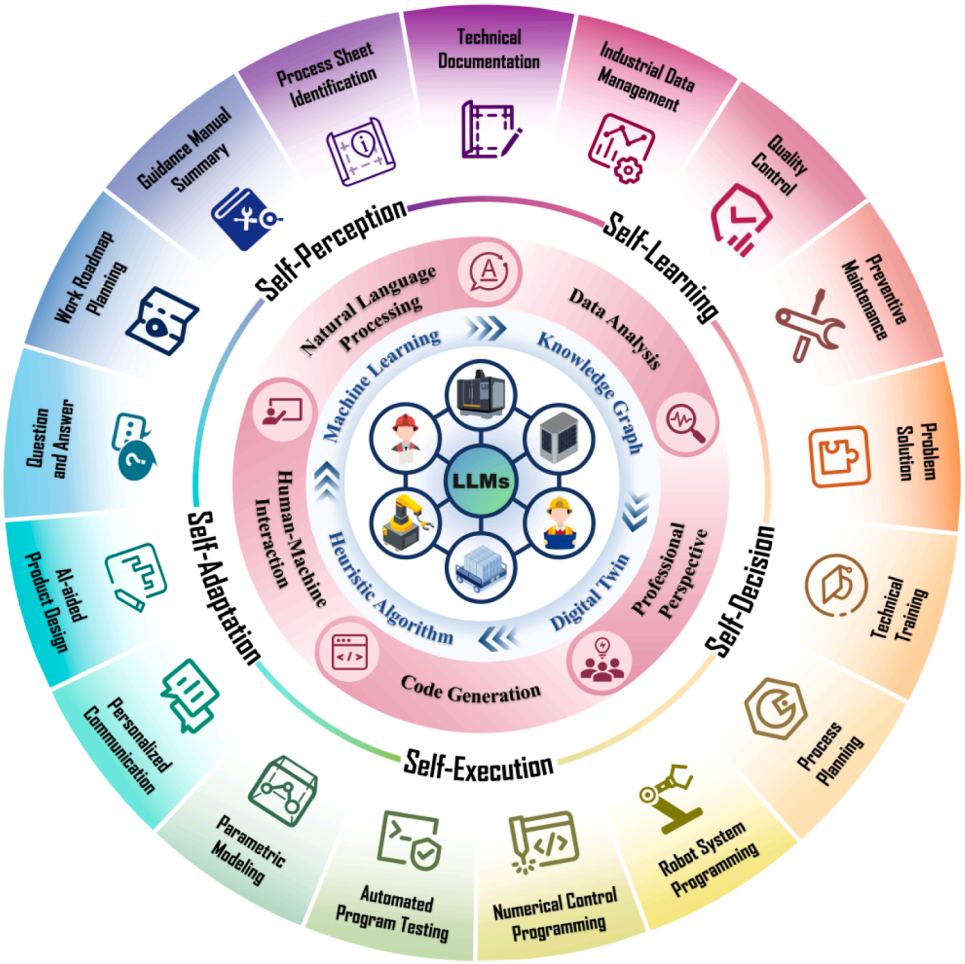


Fig. 8. Reference architecture of New-IM-LLM.

complex challenges, advancing the industry toward greater intelligence and efficiency. The following sections will examine specific application paradigms of LLMs in manufacturing, underscoring their potential to transform the industry and drive technological advancements.

3.2. Reference architecture of new-IM-LLM

Recent advancements in LLMs have profoundly influenced diverse domains, revolutionizing data interpretation and problem-solving methodologies, and driving industry advancement. Through a meta-analysis of existing literature, we identified paradigms and opportunities for LLMs applications in the New-IM domain, pinpointing the different stages for LLM integration.

3.2.1. The design of new-IM-LLM architecture

LLMs are currently positioned at the forefront of AI in manufacturing, showcasing great potential across all stages of the product lifecycle: design, production, and service. High-quality research is essential in delineating the strengths and limitations of emerging technologies. While numerous use cases of LLM technologies have been reported, most merely introduce LLMs at specific technology nodes as an auxiliary. There is a dearth of comprehensive, well-designed, and pragmatic architectural paradigms that systematically assess the utility of implementing innovative LLM-based tools in design, production, or service scenarios.

Therefore, an architectural paradigm for LLMs applications in the manufacturing domain is proposed, as illustrated in Fig. 8, built upon the preceding competence summary. The paradigm is intended to guide the industry in targeting LLM deployments, facilitating the comprehensive development of future LLM applications. Notably, the application of LLMs in manufacturing is not standalone but rather integrated with existing advanced technologies such as machine learning, KG, digital twins (DT), and heuristic algorithms. This integration enables collaborative technical support throughout the entire product lifecycle.

The capabilities of LLMs align, to some extent, with the functions of the New-IM paradigm. Firstly, excellent NLP capabilities make LLMs highly valuable in guidance manual summarization and process sheet identification and can provide reliable assistance in technical documentation and work roadmap planning. Leveraging NLP capabilities, LLMs extract product-related information from instruction manuals, process sheets, and technical documents, combined with extensive data collected from industrial sites, enabling self-perception of the manufacturing process. Secondly, LLMs' data analytics capabilities enable them to efficiently process massive amounts of historical or real-time data from manufacturing processes, making it easy to perform data-related tasks such as industrial data management, quality control, and predictive maintenance. Also, their data analysis capabilities organizes and analyzes vast amounts of text and data, identifying potential logical relationships within heterogeneous data sets through powerful contextual correlation. This facilitates the self-learning and expansion of professional expertise. Subsequently, the LLMs, fine-tuned by their expertise, acquire manufacturing domain professional perspectives that can effectively improve technical training and assist manufacturing workers in production to achieve solutions to specialized questions and planning of manufacturing processes. Further, upon acquiring task requirements, LLMs employ accumulated professional perspectives to reason about tasks and realize self-decision regarding task solutions. Lastly, LLMs' code generation competence can be used extensively in parametric modeling, programming of CNC and ROS systems, and final program testing, which improves the efficiency of process solution implementation. Therefore, utilizing their code generation capabilities, LLMs can invoke external interfaces to control devices, enabling self-execution. Despite the eventual aspiration for New-IM to detach completely from human intervention, current human participation remains indispensable. LLMs' excellent human-computer interaction capabilities allow it to communicate with humans on a personalized basis and answer questions in a multimodal manner, providing assisted interaction in areas such as product design. Hence, the high human-machine interaction capability of LLMs effectively underscores the demand for New-IM self-adaptation.

Overall, the reference architecture of New-IM-LLM aligns the five basic competences of LLMs with the five processes of self-perception, self-learning, self-decision, self-execution, and self-adaptation within the smart manufacturing paradigm. This architecture spans the entire manufacturing chain, fully exploits the application potential of LLM, and lays the foundation for the subsequent gradual integration of LLM into the manufacturing site.

3.2.2. The integration levels of LLM in IM

The integrated application of LLMs in real-world scenarios is a gradual process, divided into distinct stages. Furthermore, progress across these stages is not consistently linear and requires manual oversight to ensure the safety of LLMs at higher integration stages for deployment. Drawing upon the previously summarized competencies of LLMs and their potential applications in manufacturing, the integration of LLMs in manufacturing can be conceptualized as evolving through a staged process, from assistive LLMs to systematized LLMs, as shown in Fig. 9.

Stage 1: Assistive LLMs serves as an auxiliary tool to help design technicians and implementation workers achieve their objectives. In the initial stage of integration, task comprehension and execution remain predominantly human-driven. In the context of Industry 5.0, LLMs do not directly engage with core tasks but leverage humans as intermediaries to access auxiliary tasks, which are low-level, specific, and well-defined, thus posing a low risk. Examples include assisting in gathering and summarizing historical data on similar products, offering basic operating instructions to workers, providing editing suggestions to technicians involved in processes, and summarizing worksheets.

Stage 2: Collaborative LLMs functioning as a co-worker and undertaking small tasks that can be offloaded from complex tasks. Aligned with Industry 5.0 principles, in this stage, the LLM assumes a more proactive role, but needs humans to select or adjust the completion of tasks based on professional judgment. Examples include providing or suggesting alternatives for product design and various process planning scenarios, as well as acquiring real-time equipment operational data, analyzing it, and issuing alerts. A more advanced application of LLMs in the collaborative stage might involve operating in a semi-independent manner, such as acting as a programming engineer to provide device operation programs and conducting simulations and evaluations. In this scenario, humans monitor and oversee the results of evaluations, intervening when necessary to assume control.

Stage 3: Autonomous LLMs become an intelligent agent to execute a full spectrum of specific tasks in an integrated manner. In the autonomous stage, LLMs will achieve maximum range and autonomy, possessing comprehensive information processing and logical reasoning competence to execute a full spectrum of specific tasks, requiring only indirect human supervision. However, LLMs at this stage will not yet possess the capacity to independently perform complex tasks. Instead, they will target specific segments to achieve defined goals. For instance, an application at this stage could, theoretically, conduct a comprehensive product evaluation, select the appropriate process for planning, and provide a complete CNC program without human intervention. Nevertheless, LLMs in this stage will still rely on humans to achieve predefined objectives for completing complex tasks under the framework of Industry 5.0.

Stage 4: Systematized LLMs achieve a division of labor in multi-agent teams to accomplish a complete complex task in a selected scenario. Autonomous LLMs demonstrate proficiency in solving specified tasks but encounter logical inconsistencies when confronted with more complex tasks. This inconsistency arises from variations in inherent logic across different subdivided tasks, potentially leading to hallucination cascades with the simplistic reuse of LLMs. To address this challenge,

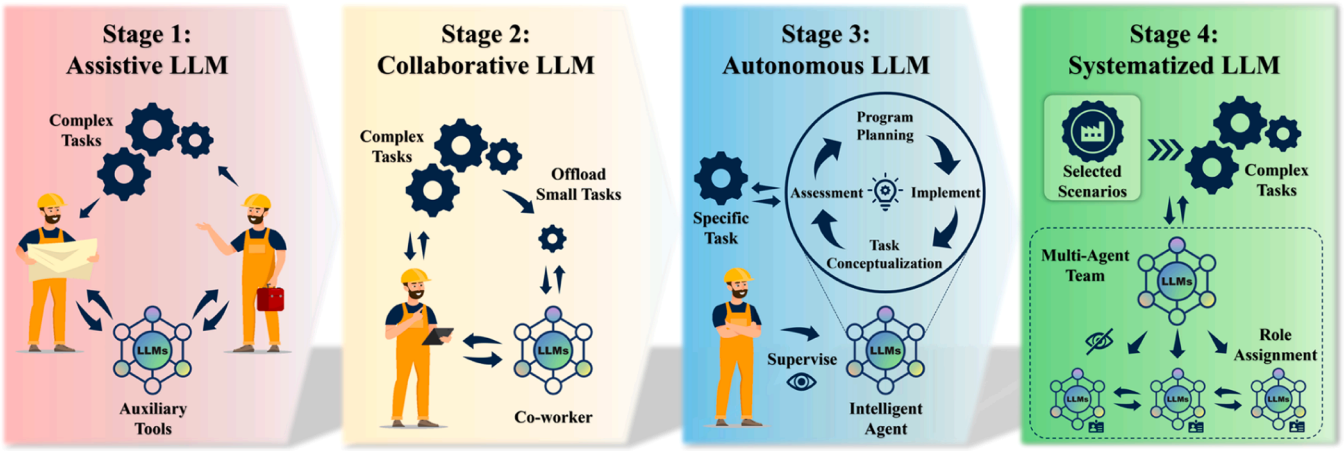


Fig. 9. Stages of LLM integration in IM.

systematic LLMs effectively decompose complex tasks into sub-tasks involving multiple agents working collaboratively. This is achieved through role assignment, wherein different agents are assigned distinct roles, enhancing task completion efficiency and effectiveness through cooperation within the multi-agent team. Reflecting the goals of Industry 5.0, supervision can also be realized by assigning the corresponding roles during the cooperation, obviating the need for direct human supervision before the final implementation stage.

4. A survey on application paths for new-IM-LLM

Based on the New-IM-LLM architecture, this section further explores pathways for applying LLMs in New-IM-LLM from the perspectives of design, production and service. Additionally, several impressive application examples are analyzed based on current works.

As illustrated in Fig. 10, New-IM-LLM requires domain-specific data and related technologies for support. On one hand, manufacturing enterprises accumulate terabytes (TB) or even petabytes (PB) of production

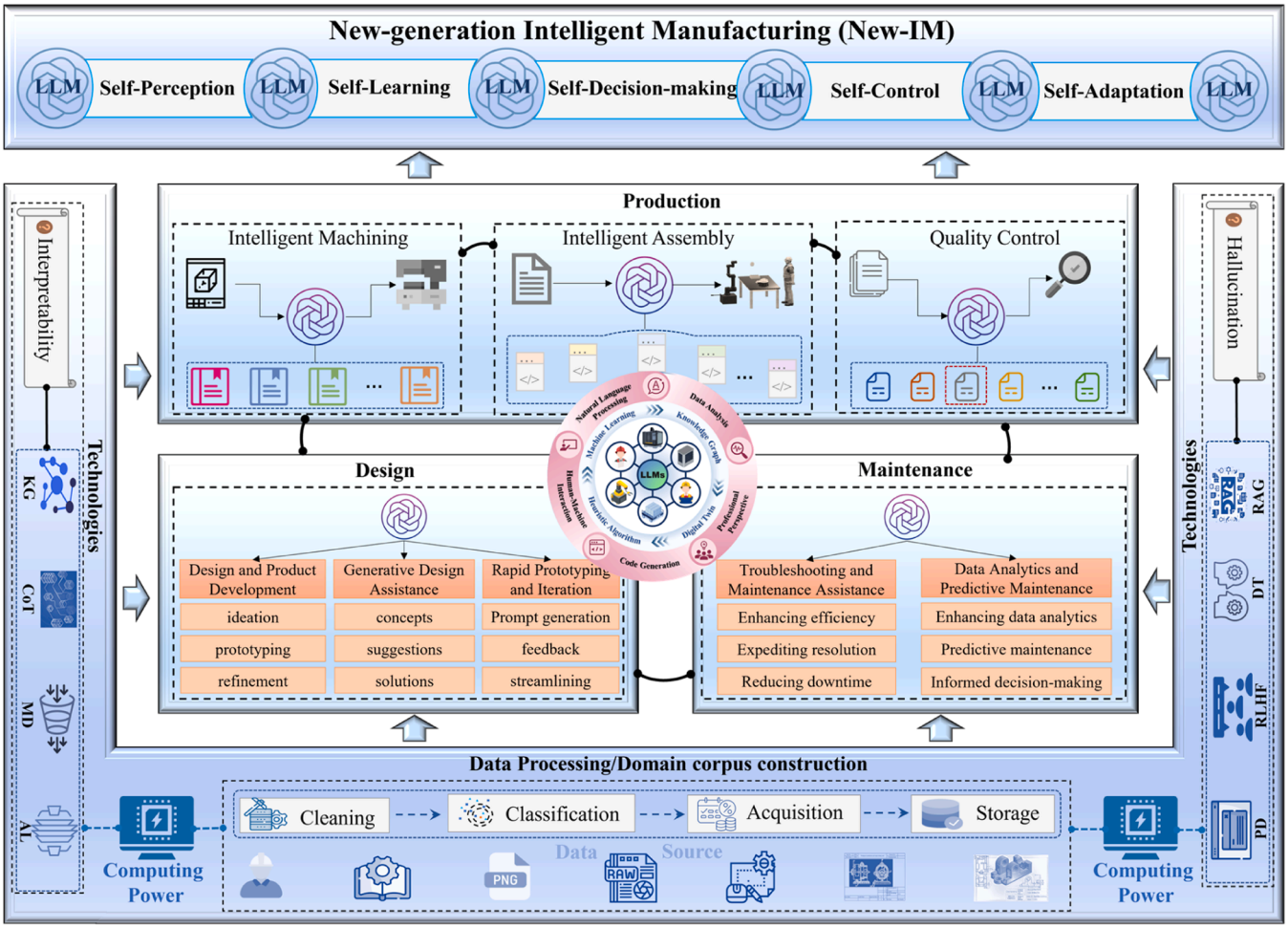


Fig. 10. Application architecture for New-IM-LLM.

data during the manufacturing process. This data is complex and diverse in form, necessitating relevant data processing technologies to clean, classify, retrieve, and store the production data to support LLMs training and enhance model training effectiveness. These data processing technologies include data mining, k-nearest neighbors/support vector machine (KNN/SVM), regularization, and databases such as MySQL, Redis, and Neo4j. On the other hand, LLMs, as generative AI models developed from deep learning, are essentially black-box models with poor interpretability of their decision-making processes. Additionally, LLMs suffer from the "hallucination" problem. In the NLP field, hallucination refers to a phenomenon where the generated content appears correct but is meaningless or incorrect [100]. For manufacturing enterprises, transparency and reliability of the model are crucial for production safety, product quality, and user acceptance. Therefore, the application of LLMs in New-IM-LLM necessitates certain key technologies to address these issues. Fortunately, extensive research has been conducted domestically and internationally on the interpretability and hallucination problems of LLM, yielding significant results. Techniques addressing interpretability issues include KG [101], chain of thought (CoT) [102], model distillation (MD) [103], and adversarial learning (AL) [104]. Techniques addressing hallucination issues include retrieval-augmented generation (RAG) [105], DT [106], reinforcement learning from human feedback (RLHF) [107], and prompt design (PD) [108]. New-IM-LLM will leverage these data and technologies in design, production, and service to support the self-perception, self-learning, self-decision-making, self-control, and self-adaptation of the New-IM.

4.1. LLM in design

Design involves not only the appearance and functionality of a product but also the entire process from concept to production, including design optimization, manufacturability analysis, and production process planning. It is characterized by multidisciplinary collaboration, innovation, personalized needs, complex decision-making processes, a focus on user experience, and efficient design and manufacturing workflows. LLMs significantly enhance design efficiency through automated generation and optimization, enabling the rapid creation of solutions that meet manufacturing requirements and reducing design iteration time. The integration of LLMs into the production design process increases efficiency, flexibility, and human-centeredness, providing robust support for tackling complex design challenges.

In Industry 5.0, the design methodologies augmented by LLM comprehensively address the multifaceted needs of designers, producers, and consumers. These LLMs, grounded in domain expertise, real-time data, and specialized software comprehension, significantly enhance decision-making capabilities in design under the guidance and support of domain experts. Compared to conventional design tools, LLMs demonstrate an exceptionally advanced capacity for discerning human intentions, thereby reinforcing the human-centered design ethos in engineering practices. Makatura et al. [109] presented the application of LLM in computational design and manufacturing (CDaM). They proposed that these tools can eliminate process obstacles by providing intuitive, unified, and user-friendly interfaces. Their study demonstrated that LLMs can accelerate the CDaM process and facilitate the development of new designs and manufacturing methods. Hu et al. [110] investigated the application of artificial intelligence-generated content (AIGC) technology in design, emphasizing its role in fostering design innovation. They focused on integrating the representative AIGC tool "Midjourney," proposed an AIGC-based product design pathway, and developed a supporting toolset named AMP-Cards. Wu et al. [111] applied generative AI to product color-matching design, achieving optimal solutions by integrating the functionalities of ChatGPT and Midjourney. This study highlights that generative AI, including LLMs, has the potential to collaborate with traditional CAD tools and revolutionize product design. Yang et al. [112] applied GPT-4 to hardware development, proposing a novel method for hardware design using natural language. By comparing it with handwritten hardware description language (HDL) designs, they highlighted the limitations of GPT, particularly in AI accelerators. Xu et al. [113] proposed an AI-enhanced multimodal collaborative design (AI-MCD) framework to address communication barriers and inefficiencies resulting from the diverse backgrounds of stakeholders in collaborative design processes. This framework integrates LLM with mixed reality (MR) technology, creating an interactive and immersive design environment.

Example I: LLM-enhanced design for manufacturing. Design for Manufacturing (DFM) is a methodology that streamlines the manufacturing process and enhances production efficiency. Leveraging its robust reasoning capabilities, LLMs comprehensively considers geometric features, material properties, and manufacturing process requirements based on input multimodal data. This approach yields more accurate and optimized design recommendations, thereby reducing human errors and enhancing design efficiency. By incorporating manufacturing considerations during the design phase, DFM mitigates potential production issues, reduces costs, and improves product quality and consistency. For the applications of LLMs in the domain of DFM, Makatura et al. [109] suggested that LLMs can significantly enhance the design and production of parts. Their findings are summarized in this example. Serving as a knowledge repository or database for the manufacturing field, LLMs can fully utilize their pattern recognition and language interpretation capabilities during the design phase. This case study employs GPT-4 to develop an LLM for DFM, named DesignGPT, as shown in Fig. 11(a). DesignGPT can be integrated into mobile devices, forming a design and manufacturing Q&A system, enabling designers to interact at any time. Initially, designers upload design files or describe the requirements. Upon receiving the input, DesignGPT preprocesses the design files (e.g., through voxelization) to prepare for analysis, as depicted in Fig. 11(b). DesignGPT then conducts a manufacturability analysis based on the preprocessed models, identifies non-manufacturable areas, marks them, and generates an analysis report, as shown in Fig. 11 (c, left). The analysis process primarily includes geometric feature detection, material property analysis, and manufacturing process requirement matching. Subsequently, DesignGPT searches the model database for manufacturable origin models similar to the design. Through optimization, DesignGPT generates a manufacturable design model and compares it with the reference model, as shown in Fig. 11 (c, right). Finally, designers can review detailed model information and improvement suggestions to optimize the design and obtain an improved model. Concurrently, the design undergoes further iterative optimization using incomplete models derived from manufacturability analysis and complete models built with manufacturability support, ensuring that the product design meets manufacturing requirements. In this context, incomplete models refer to designs identified as problematic in the initial analysis, while complete models are those optimized to meet manufacturability standards.

4.2. LLM in production

Intelligent production encompasses not only automation and digitalization but also the integration of advanced artificial intelligence technologies to achieve more efficient, flexible, and personalized production processes. Its features include efficient data processing, intelligent process optimization, adaptability, personalized production, smart assembly and quality control, and human-machine collaboration. LLMs, with their robust data processing, analysis, and generation capabilities, enhance the efficiency, flexibility, and human-centered nature of intelligent production systems. These capabilities provide LLMs with broad application prospects in intelligent production, further advancing the digital and intelligent transformation of the manufacturing.

Intelligent production serves as the central theme in New-IM [4], with intelligent production lines, intelligent workshops, and intelligent factories being the primary carriers of intelligent production [114–116].

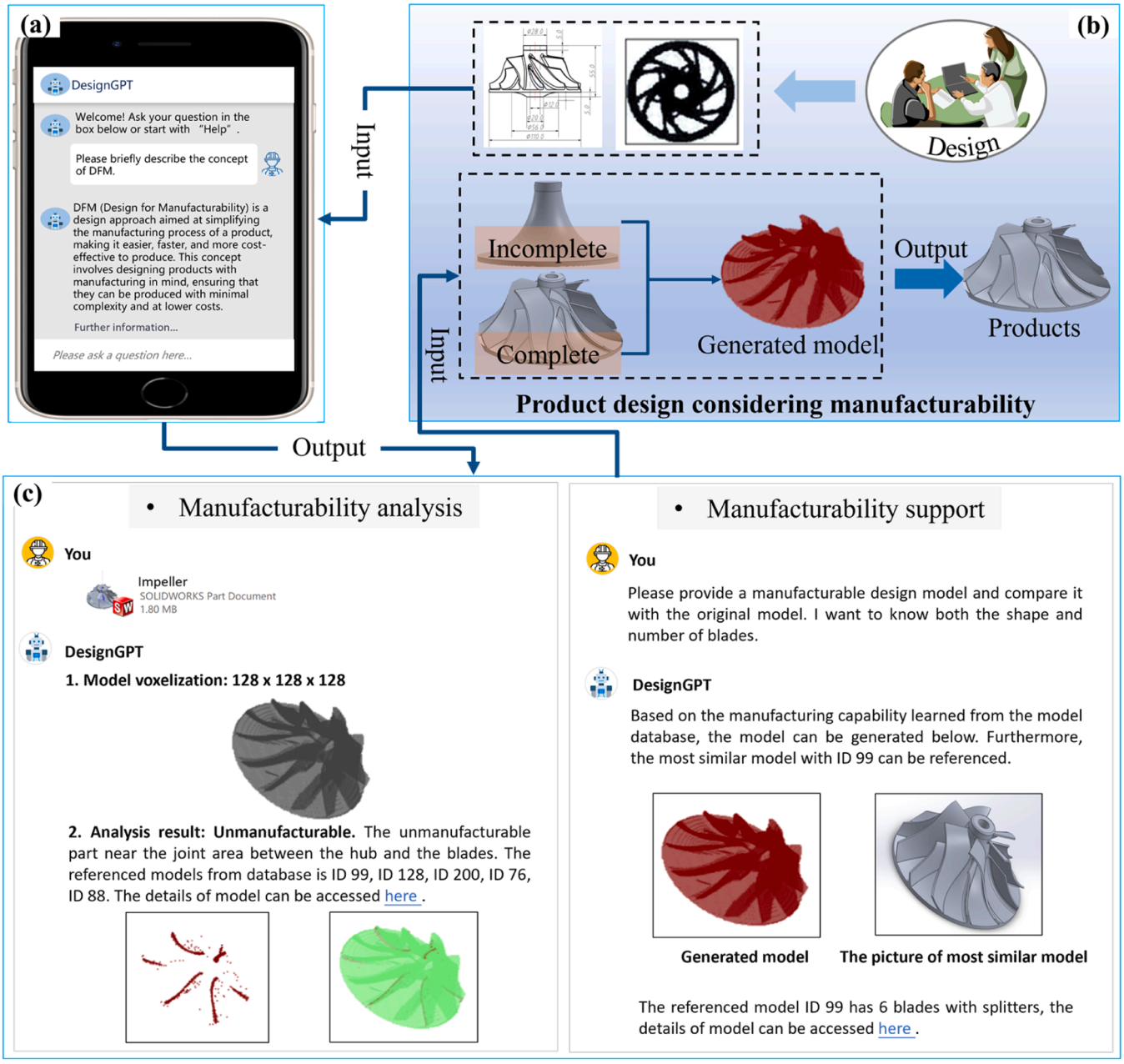


Fig. 11. An example of LLM enhanced DFM [109].

Intelligent machining, intelligent assembly, and quality control are key elements in the operation of these carriers. Therefore, LLMs in production refers to leveraging the powerful potential of LLMs in achieving human-like expert intelligence, enhancing critical aspects of production, and enabling enterprises to achieve more comprehensive, efficient, and personalized intelligent production.

Based on the support of the aforementioned data and related technologies, this section illustrates the application scenarios of LLM-empowered production in three key aspects: intelligent processing, intelligent assembly, and quality control.

LLM in Intelligent Machining. LLMs can extract information from machining technical documents and operation manuals through information extraction tasks such as named entity recognition and relation extraction. This information is used to construct a machining process knowledge base, thereby supporting process decision-making in part machining. Additionally, LLMs can address the current limitations of CAPP systems, including high entry barriers, poor human-computer interaction, and lack of flexibility. Guo et al. proposed a BERT-based process knowledge extraction model for automatically extracting process text knowledge, which improves the accuracy of knowledge extraction compared to non-BERT models [117]. LLMs can be combined with KGs to enable knowledge reasoning, and there has been considerable research in this area of AI. Xiao et al. [118] suggested that LLMs, when integrated with process knowledge graphs (PKGs), can be used to reason new process routes. Although research in this area is still in its infancy, it holds great promise.

LLM in Intelligent Assembly. LLMs enhance the assembly process by improving the comprehension and decision-making capabilities of assembly robots. LLMs can be used to interpret instructions from assembly workers to robots and help robots quickly adapt to new assembly requirements. In assembly tasks, non-standard situations or special requirements frequently arise. For example, if the assembly instructions for a component suddenly change, traditional robot programs may struggle to cope. By utilizing LLMs, robots can quickly understand text

prompts or instructions related to specific contexts and dynamically adjust their behavior, thereby helping robots swiftly adapt to new assembly requirements. You et al. [119] applied LLM to the automation of construction task sequence planning and proposed a new robotic system called RoboGPT. Using ChatGPT for construction sequence planning, it can handle complex construction operations and adapt to dynamic changes. Fan et al. [120] explored the potential of LLM agents in industrial robots and proposed specific approaches such as visual-semantic control and real-time feedback loops. Gkournelos et al. [121] proposed an LLM-based manufacturing execution system that enhances human-robot collaboration (HRC) in smart manufacturing. The system provides a natural language interface, integrates DT, and employs robotic behavior control. It demonstrated significant improvements in collaboration and efficiency in two HRC assembly case studies.

LLM in Quality Control. Quality control is crucial in enterprise manufacturing, directly affecting product reliability and customer satisfaction [122]. LLM-enhanced quality control refers to the use of LLMs to detect and optimize product quality defects. Zhou et al. [123] developed an enhanced LLM named CausalKGPT, integrating causal science to analyze and respond to quality defects in the manufacturing process. Rane et al. [124] proposed ChatGPT-driven intelligent manufacturing, which can analyze textual and visual data to help identify defects or deviations in quality standards. Furthermore, this approach allows for real-time monitoring and analysis of the manufacturing process, enabling timely intervention when anomalies are detected.

Example II: LLM-enhanced intelligent process planning. Systematic understanding and addressing of complex process requirements enable LLMs to enhance the efficiency and consistency of process planning. Additionally, human-computer interaction increases flexibility, facilitating easier access to and application of process knowledge by planners, thus improving overall design and production efficiency. In intelligent machining, Xu et al. [125] proposed a generative intelligent process planning (GIPP) framework that integrates generative AI with DT. Their research outcomes are encapsulated in this example. This framework leverages the capabilities of LLMs in understanding human intentions, generating efficient content, and facilitating human-computer interaction. Combined with the real-time monitoring and verification abilities of DT, it enables the generation of high-efficiency and highly reliable process knowledge and process plans, as illustrated in Fig. 12. To this end, the GPT model was fine-tuned to develop ProcessGPT, an LLM for

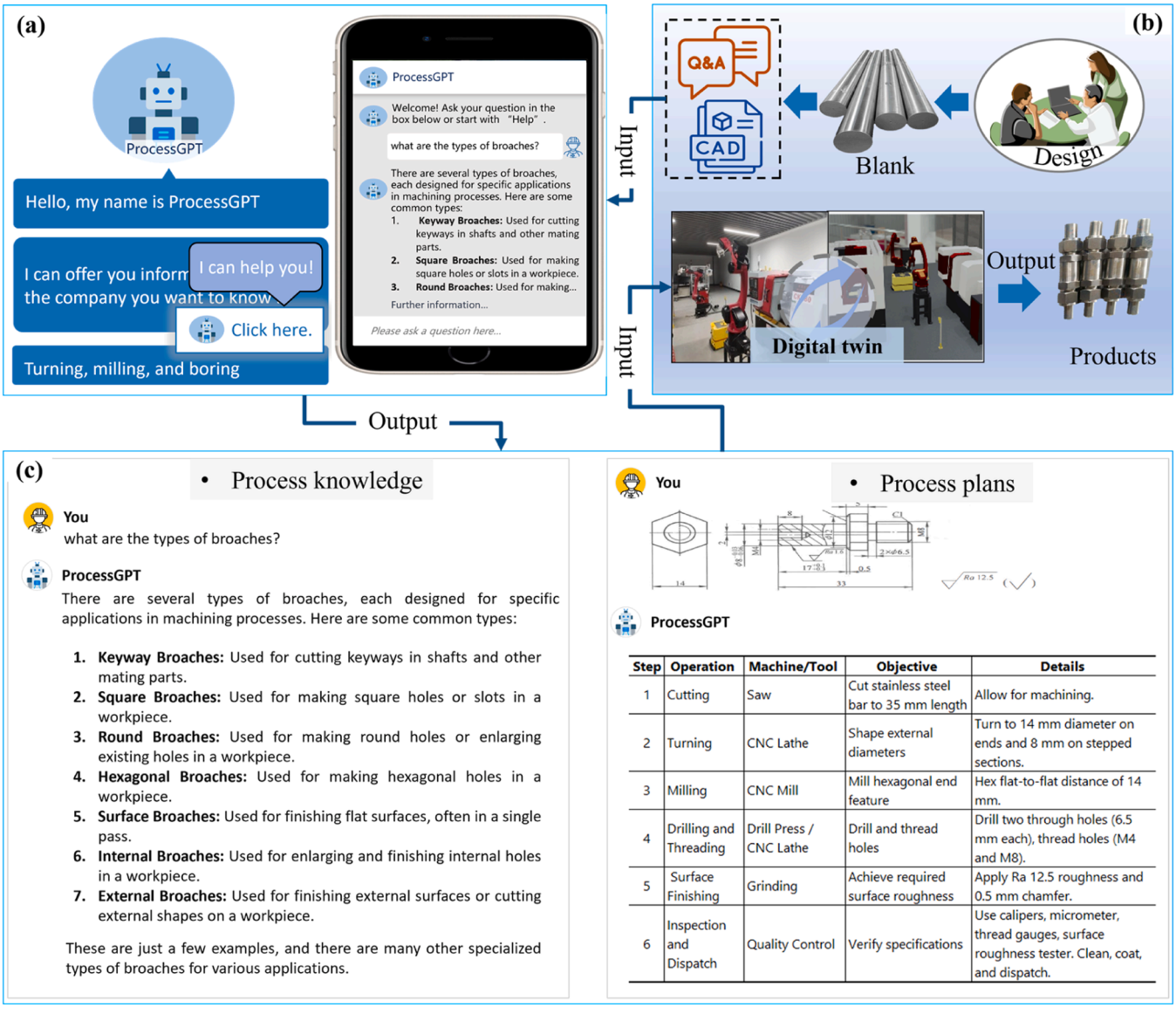


Fig. 12. An example of LLM enhanced intelligent process planning [125].

process planning. Like Example I, ProcessGPT can be integrated into mobile devices to create a process questions and answers (Q&A) system that process planners can access and learn from anytime, anywhere, as shown in Fig. 12(a). Initially, the designer specifies the technical requirements based on product specifications, including raw material and geometric parameters. This information can be input into ProcessGPT via process queries or 2D CAD models, as shown in Fig. 12(b). Subsequently, ProcessGPT generates the corresponding process knowledge and process plans, as depicted in Fig. 12(c). Ultimately, the obtained process plans are input into a DT-based computer numerical control (CNC) system, where the final products are produced through online simulation, monitoring, evaluation, and verification.

4.3. LLM in service

Product service is a vital component of manufacturing, playing a key role in enhancing customer experience, strengthening enterprise competitiveness, and promoting long-term growth. It encompasses a range of activities, from product support to maintenance, repair, and operations (MRO), with the goals of improving customer satisfaction, increasing product reliability, extending product lifespan, and fostering sustainable business development. Key characteristics include the complexity of MRO, the real-time and precise nature of services, knowledge accumulation and transfer, personalized interaction, and fault prediction with preventive maintenance. In product service scenarios, LLMs, with their robust data processing, real-time analysis, knowledge integration, and personalized service capabilities, can significantly enhance maintenance efficiency, accuracy, and customer satisfaction. This enables intelligent and personalized service optimization, substantially improving the efficiency and quality of product services, and advancing the manufacturing industry toward greater intelligence, personalization, and efficiency.

MRO ensures the normal, safe, and reliable operation of manufacturing equipment, encompassing real-time condition monitoring, equipment data analysis, troubleshooting, maintenance, and system optimization. While existing AI technologies have made significant strides in predictive maintenance, their application scenarios remain relatively limited, often necessitating engineer involvement for assistance [126].

In contrast, LLMs offer comprehensive applicability across various MRO processes, enhancing automation and intelligence. LLM-based MRO systems can furnish engineers with real-time maintenance recommendations and solutions through the analysis of equipment status, extensive historical data, and best practices. By providing operational recommendations, optimizing parameter settings, and guiding operational processes in an interactive manner, LLMs elevate maintenance efficiency and accuracy, surpassing previous AI technologies' capabilities [127].

The NLP competence forms the bedrock of LLMs, facilitating the summarization and comprehension of specialized knowledge content by parsing multimodal and unstructured maintenance records and technical manuals. This competence aids engineers in drafting maintenance plans, compiling reports, and conveying technical information to non-technical stakeholders. For instance, LLMs extract rich historical information regarding equipment operating statuses from power equipment operation and maintenance records for equipment diagnosis [128]. Additionally, it sifts through plant repair and maintenance reports to extract relevant information, optimizing predictive maintenance processes [129]. In the domain of aircraft maintenance, LLM extracts maintenance experience information from maintenance records to pinpoint fault causes and guide maintenance personnel in troubleshooting efforts [130]. Furthermore, the integration with KG enables LLM to mine compressor fault text data, facilitating knowledge extraction and analysis [131].

Furthermore, leveraging learned expertise, LLMs can suggest potential causes of equipment failure based on described features or failure modes. Insights on best practices, potential risks, cost-effectiveness, and the impact of different maintenance strategies associated with the underlying problem are also provided. Real-time analysis of equipment operational data enables the identification of potential signs of equipment degradation or failure, along with recommendations for appropriate interventions. For instance, a domain knowledge base has been developed using manual documents containing troubleshooting, operation, and maintenance expertise as the primary knowledge source, facilitating fault identification and guidance for wind turbines [132]. In another study, ChatGPT, fine-tuning on 3D printing codes and parameter data, provided logical responses based on equipment faults and product quality, thereby achieving fault analysis and optimization of additive manufacturing processes [133]. Moreover, a novel multisource text-based equipment maintenance LLM was proposed to effectively manage and utilize massive coal mine equipment maintenance knowledge, enabling professional consultation and maintenance decision analysis [134]. Data augmentation can also be realized based on generative AI, followed by sequential training of fault diagnosis models to achieve high-precision detection of transformer faults [135]. Additionally, LLMs can facilitate the identification and management of faulty facilities even with limited samples. A combination of temporal association rule mining and semantic similarity concepts derived from LLM was employed to apply association rule mining to maintenance requests [15].

Example III: LLM-enhanced fault diagnosis. Integrating the reasoning capabilities of LLMs with the structured information of knowledge graphs enables systematic processing of complex fault information, resulting in accurate diagnostic outcomes. Furthermore, human-computer interaction enhances flexibility, improving diagnostic efficiency and reliability, and thereby reducing downtime and economic losses. In service applications, Liu et al. [136] proposed a knowledge-enhanced joint model that integrates an aviation assembly KG into an LLM. Their research results are exemplified in the following case. The knowledge supporting this process is derived from professional guide manuals and is assimilated into the joint knowledge augmentation model through prefix-tuning training. As shown in Fig. 13, the LLM, through knowledge graph-based reasoning, can identify the root causes of faults observed in industrial environments and provide tailored troubleshooting solutions. Upon a fault occurrence, the LLM receives inputs regarding the fault phenomenon or fault number via speech or text, as illustrated in Fig. 13(a). These inputs are processed by the LLM, which then generates visual schematic diagrams of faulty parts, introductory information, and sub-knowledge graphs of key component entities, enhancing the interpretability of the knowledge-based reasoning outcomes, as depicted in Fig. 13(b). Finally, leveraging the LLM, workers can access additional technical support through problem queries, including fault symptom analysis, fault cause analysis, and troubleshooting plans, as shown in Fig. 13(c). Fault symptom analysis involves the LLM providing detailed descriptions of the symptoms when users inquire about the causes of faults. Fault cause analysis entails the LLM explaining the specific reasons for the faults, such as a fuel truck return switch not being tightly closed, which leads to hydraulic system malfunction. The troubleshooting plan includes detailed steps to resolve the fault, such as turning off the fuel truck's return oil switch, with references to specific manual sections for guidance.

5. Discussion on current challenges and future trends of new-IM-LLM

The limitations, barriers, and challenges of LLMs in New-IM are summarized in this section, and valuable research directions to continuously improve New-IM-LLM are provided.

5.1. Challenge ahead

Although LLMs hold significant potential for applications in New-IM

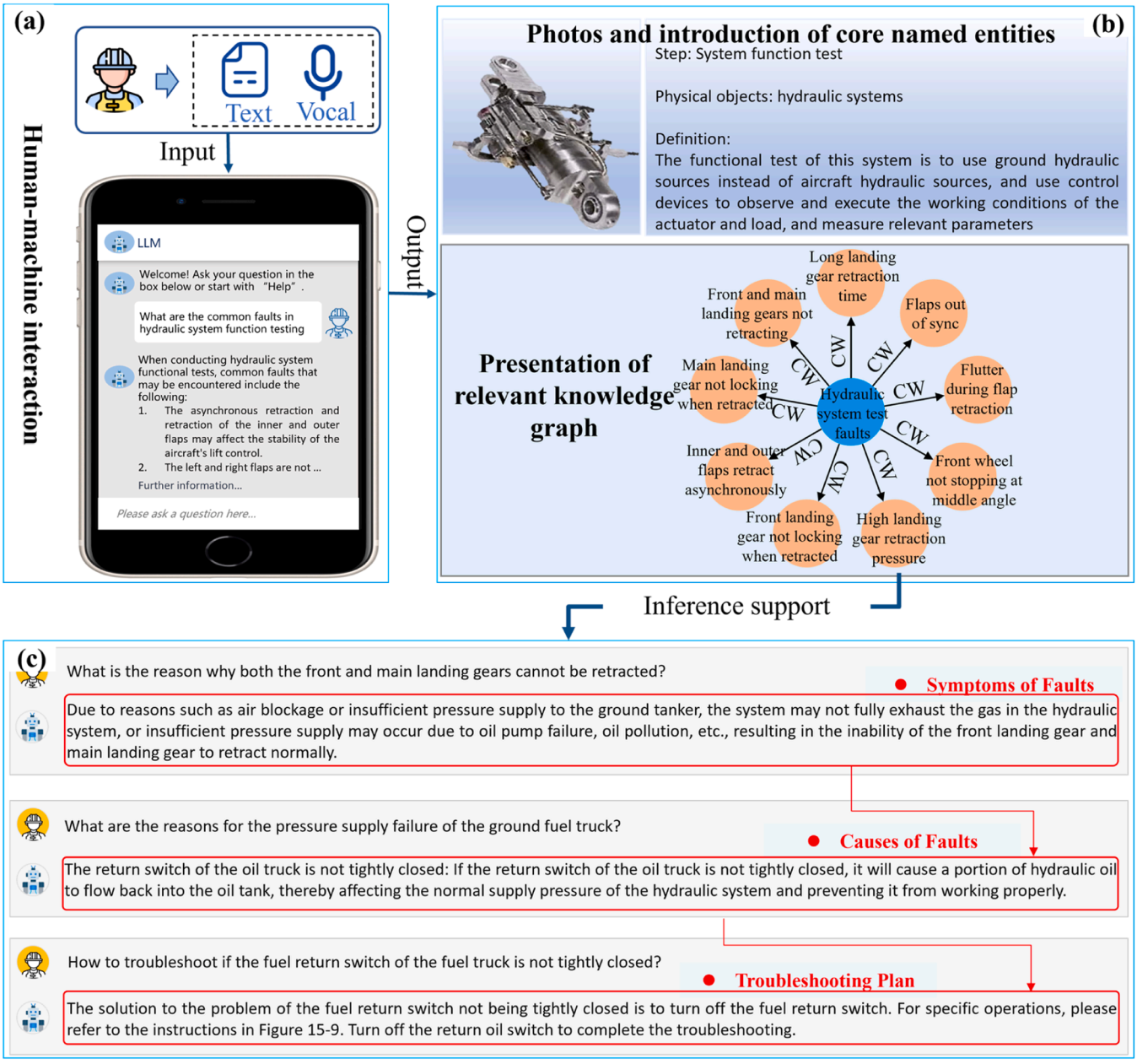


Fig. 13. An example of combing LLM and KG for assembly fault diagnosis [136].

in the context of Industry 5.0, they still faces several challenges that require further research and resolution. This section summarizes these challenges from four aspects, including limited LLM data, expensive LLM computing resources, uncertain LLM application performance, and ethical and privacy concerns.

Limited LLM data: Extensive and effective domain data is critical for training a high-performance LLM in that specific domain [137]. However, unlike several typical and hotspot interdisciplinary LLM application domains that have comprehensive open-source databases, such as PubMed [138], New-IM lacks similar resources, making it challenging to develop high-performance LLMs for manufacturing applications. Additionally, manufacturing data has multi-modal characteristics, typically including text, images, models, sensor data, etc. [139]. LLMs must effectively integrate these multimodal data to provide comprehensive and accurate analyses. However, the lack of specialized tools for multimodal data integration in New-IM further complicates this task.

Expensive LLM computing resources: Both the training and use of LLMs require significant computational resources to process large volumes of data and handle substantial model sizes, which are too expensive to limit the widespread application of LLMs in manufacturing enterprises. Although advances in model compression techniques such as pruning [140], quantization [141], and knowledge distillation [142] have enabled the creation of more compact LLMs, balancing training effectiveness and model lightweighting remains a major challenge.

Uncertain LLM application performance: The uncertainty of LLMs, such as unexplainable outputs and hallucination problems, will greatly limit its application in specific manufacturing scenarios. Explainable LLMs output are critical for specific manufacturing scenarios in New-IM, as it usually demands highly transparent decision-making to ensure trust and reliability [143]. However, most current LLMs cannot provide clear and interpretable decision pathways making it difficult for decision-makers to understand and trust their outputs [144]. Additionally, the inherent problem-"hallucination" of LLM itself might cause LLM to generate nonexistent or infeasible manufacturing processes, misleading production [145]. These challenges directly affect the trust that enterprises place in LLM technology, complicating its application in

manufacturing.

Ethical and privacy concerns: Regarding LLM ethics, it is essential to determine the responsibility for decision errors or accidents when using LLM for decision-making. This involves delineating responsibilities among model developers, data providers, and users [146]. As for data privacy, the substantial data generated in manufacturing might include personal information about employees, such as job performance and health data [147]. Improper handling or disclosure of this data could violate employees' privacy rights.

5.2. Future trends

Given the limitations, barriers, and challenges of LLMs in New-IM, several research directions have been identified as future trends for their application: the development of multi-modal LLM and efficient fine-tuning techniques to reduce dependence on large-scale datasets; the creation of lightweight LLMs to minimize computing costs; the integration of KG and DT with LLM to address interpretability and hallucination issues; and the establishment of ethical and legal regulatory frameworks to tackle ethical and privacy concerns.

Developing multimodal LLM and efficient fine-tuning techniques to reduce the dependence on large-scale datasets. Multimodal LLMs can handle various types of data simultaneously (such as text, images, audio, etc.), with these different modalities complementing each other [148]. For instance, in New-IM, text data can describe equipment operation manuals and fault records, while image data can capture the actual state of the equipment and fault phenomena. By integrating these data types, model can learn from diverse information sources, thereby reducing the dependency on a single data modality. Moreover, multimodal pre-trained models can leverage transfer learning by pre-training on extensive non-specific domain data and then fine-tuning with a small amount of domain-specific data [149]. This method utilizes the rich information from large-scale non-specific data, significantly decreasing the demand for extensive domain-specific data.

Implementing lightweight LLMs to minimize the computing costs. Lightweight methods for LLM, such as pruning, quantization, knowledge distillation, and distributed computing architectures, can decrease the number of LLM parameters, computational complexity, and storage requirements, thereby reducing computational overhead. Additionally, the application of model compression algorithms [150], including low-rank decomposition and sparse coding, ensures that model performance is minimally affected through meticulous fine-tuning during the compression process.

Combining KGs and DT with LLMs to address the interpretability and hallucination issues in LLM. KGs can enable LLM with structured contextual information, ensuring that the outputs of LLM have clear sources and logical foundations [151,152]. When an LLM reasons based on a KG, each decision step can be traced back to specific nodes and edges in the graph. This transparency allows engineers to clearly understand how the model arrives at its conclusions, thereby enhancing interpretability. DT provides real-time feedback on the operational data of physical devices, enabling LLMs to dynamically adjust and self-learn, progressively improving and optimizing predictive capabilities while reducing the generation of erroneous information [153–155]. Additionally, DT can simulate and validate the process routes and operational steps generated by LLMs, ensuring their feasibility in a virtual environment. Therefore, DT can help LLM to identify and correct hallucination issues in advance, thereby preventing problems in actual production.

Establishing ethical and legal regulatory frameworks and employing federated learning to address the ethical and privacy concerns associated with deploying LLMs. Ethical and legal regulatory frameworks serve as guiding principles for the responsible use of LLM, ensuring compliance with data protection and privacy regulations, thus effectively alleviating corporate concerns regarding ethical AI deployment [156]. Federated learning [157] allows models to be trained locally on multiple nodes without the need to centralize data, thereby safeguarding data privacy.

6. Conclusion

In the context of Industry 5.0, a comprehensive survey on the potentials, pathways, and challenges for applying LLMs in New-IM is provided in this paper. Firstly, the concept, connotation, and foundational architectures of LLMs are summarized. Then, a New-IM-LLM architecture is designed based on several typical and hotspot interdisciplinary LLMs applications, which could aid, cooperate, or partially replace humans during perception, learning, decision-making, and execution processes in New-IM, thus liberating people from heavy mental labor. Thirdly, the pathways for applying LLMs in New-IM-LLM are explored from the perspective of design, production and service, providing insight into the research and application of LLM in New-IM. Finally, limited LLMs data, expensive LLMs computing resources, uncertain LLMs application performance, and ethical and privacy concerns are identified as limitations, barriers, and challenges for developing New-IM-LLM, while several potential solutions are provided as valuable research directions to continuously improve New-IM-LLM.

CRediT authorship contribution statement

Chao Zhang: Writing – original draft, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Qingfeng Xu:** Writing – original draft, Methodology, Investigation. **Yongrui Yu:** Writing – original draft, Investigation, Formal analysis. **Guanghui Zhou:** Writing – review & editing, Supervision, Resources, Project administration. **Keyan Zeng:** Writing – original draft, Investigation. **Fengtian Chang:** Writing – review & editing, Investigation. **Kai Ding:** Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the National Natural Science Foundation of China [grant numbers 52105530, 52375511, 52475534]; the China National Postdoctoral Program for Innovative Talents [grant number BX2021244]; the Key Research and Development Program of Shaanxi Province, China [grant number 2023-ZDLNY-71]; and the Fundamental Research Funds for the Central Universities (grant number xzy012022053).

References

[1] J. Leng, W. Sha, B. Wang, P. Zheng, C. Zhuang, Q. Liu, T. Wuest, D. Mourtzis, L. Wang, Industry 5.0: prospect and retrospect, J. Manuf. Syst. 65 (2022) 279–295, https://doi.org/10.1016/j.jmsy.2022.09.017.

[2] C. Zhang, Z. Wang, G. Zhou, F. Chang, D. Ma, Y. Jing, W. Cheng, K. Ding, D. Zhao, Towards new-generation human-centric smart manufacturing in Industry 5.0: a systematic review, Adv. Eng. Inf. 57 (2023) 102121, https://doi.org/10.1016/j.aei.2023.102121.

[3] J. Leng, Y. Zhong, Z. Lin, K. Xu, D. Mourtzis, X. Zhou, P. Zheng, Q. Liu, J.L. Zhao, W. Shen, Towards resilience in Industry 5.0: a decentralized autonomous manufacturing paradigm, J. Manuf. Syst. 71 (2023) 95–114, https://doi.org/10.1016/j.jmsy.2023.08.023.

[4] J. Zhou, P. Li, Y. Zhou, B. Wang, J. Zang, L. Meng, Toward new-generation intelligent manufacturing, engineering 4 (2018) 11–20. https://doi.org/10.1016/j.eng.2018.01.002.

[5] G. Zhou, C. Zhang, Z. Li, K. Ding, C. Wang, Knowledge-driven digital twin manufacturing cell towards intelligent manufacturing, Int. J. Prod. Res. 58 (2020) 1034–1051, https://doi.org/10.1080/00207543.2019.1607978.

[6] C. Zhang, G. Zhou, J. Li, F. Chang, K. Ding, D. Ma, A multi-access edge computing enabled framework for the construction of a knowledge-sharing intelligent machine tool swarm in Industry 4.0, J. Manuf. Syst. 66 (2023) 56–70, https://doi.org/10.1016/j.jmsy.2022.11.015.

[7] Y. Lu, H. Zheng, S. Chand, W. Xia, Z. Liu, X. Xu, L. Wang, Z. Qin, J. Bao, Outlook on human-centric manufacturing towards Industry 5.0, J. Manuf. Syst. 62 (2022) 612–627, https://doi.org/10.1016/j.jmsy.2022.02.001.

[8] B. Wang, H. Zhou, X. Li, G. Yang, P. Zheng, C. Song, Y. Yuan, T. Wuest, H. Yang, L. Wang, Human digital twin in the context of Industry 5.0, Robot. Comput. Integr. Manuf. 85 (2024) 102626, https://doi.org/10.1016/j.rcim.2023.102626.

[9] A.V. Samsonovich, S.A. Shumsky, V.E. Karpov, A.A. Kotov, A.G. Kolonin, Key advanced research initiative: a manifesto for the new-generation artificial intelligence, Procedia Comput. Sci. 213 (2022) 824–831, https://doi.org/10.1016/j.procs.2022.11.140.

[10] L. Hu, Z. Liu, Z. Zhao, L. Hou, L. Nie, J. Li, A survey of knowledge enhanced pre-trained language models, IEEE Trans. Knowl. Data Eng. 36 (2024) 1413–1430, https://doi.org/10.1109/TKDE.2023.3310002.

[11] L. Xia, C. Li, C. Zhang, S. Liu, P. Zheng, Leveraging error-assisted fine-tuning large language models for manufacturing excellence, Robot. Comput. Integr. Manuf. 88 (2024) 102728, https://doi.org/10.1016/j.rcim.2024.102728.

[12] T. Wang, J. Fan, P. Zheng, An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing, J. Manuf. Syst. (2024), https://doi.org/10.1016/j.jmsy.2024.04.020. S0278612524000864.

[13] S. Lou, Y. Zhang, R. Tan, C. Lv, A human-cyber-physical system enabled sequential disassembly planning approach for a human-robot collaboration cell in Industry 5.0, Robot. Comput.-Integrat. Manuf. 87 (2024) 102706, https://doi.org/10.1016/j.rcim.2023.102706.

[14] P.Y. Abijith, P. Patidar, G. Nair, R. Pandya, Large language models trained on equipment maintenance text, in: 2023: p. D021S065R003. https://doi.org/10.2118/216336-MS.

[15] M. Lowin, A text-based predictive maintenance approach for facility management requests utilizing association rule mining and large language models, MAKE 6 (2024) 233–258, https://doi.org/10.3390/make6010013.

[16] S.M.R. Naqvi, M. Ghufran, C. Varnier, J.-M. Nicod, K. Javed, N. Zerhouni, Unlocking maintenance insights in industrial text through semantic search, Comput. Ind. 157–158 (2024) 104083, https://doi.org/10.1016/j.compind.2024.104083.

[17] R. Qureshi, M. Irfan, H. Ali, A. Khan, A.S. Nittala, S. Ali, A. Shah, T.M. Gondal, F. Sadak, Z. Shah, M.U. Hadi, S. Khan, Q. Al-Tashi, J. Wu, A. Bermak, T. Alam, Artificial intelligence and biosensors in healthcare and its clinical relevance: a review, IEEe Access. 11 (2023) 61600–61620, https://doi.org/10.1109/ACCESS.2023.3285596.

[18] H. Wang, M. Liu, W. Shen, Industrial-generative pre-trained transformer for intelligent manufacturing systems, IET Collab. Intell. Manufact. 5 (2023) e12078, https://doi.org/10.1049/cim2.12078.

[19] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W.X. Zhao, Z. Wei, J. Wen, A survey on large language model based autonomous agents, Front. Comput. Sci. 18 (2024) 186345, https://doi.org/10.1007/s11704-024-40231-1.

[20] K. Knill, S. Young, Hidden markov models in speech and language processing, in: S. Young, G. Bloothooft (Eds.), Corpus-Based Methods in Language and Speech Processing, Springer, Netherlands, Dordrecht, 1997: pp. 27–68. https://doi.org/10.1007/978-94-017-1183-8_2.

[21] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted gaussian mixture models, Digit. Signal Process. 10 (2000) 19–41, https://doi.org/10.1006/dspr.1999.0361.

[22] S.M. Thede, M.P. Harper, A second-order Hidden Markov Model for part-of-speech tagging, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, College Park, Maryland, 1999, pp. 175–182, https://doi.org/10.3115/1034678.1034712.

[23] L.R. Bahl, P.F. Brown, P.V. De Souza, R.L. Mercer, A tree-based statistical language model for natural language speech recognition, IEEE Trans. Acoust., Speech, Signal Processing 37 (1989) 1001–1008, https://doi.org/10.1109/29.32278.

[24] T. Brants, A.C. Popat, P. Xu, F.J. Och, J. Dean, Large language models in machine translation, (2007) 858–867.

[25] S. Katz, Estimation of probabilities from sparse data for the language model component of a speech recognizer, IEEE Trans. Acoust. 35 (1987) 400–401, https://doi.org/10.1109/TASSP.1987.1165125.

[26] W.A. Gale, G. Sampson, Good-turing frequency estimation without tears*, J. Quant. Linguist. 2 (1995) 217–237, https://doi.org/10.1080/09296179508590051.

[27] Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model, in: T. Leen, T. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems, MIT Press, 2000, in: https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf.

[28] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, S. Khudanpur, Recurrent neural network based language model, in: Interspeech, Makuhari, 2010: pp. 1045–1048.

[29] S. Kombrink, T. Mikolov, M. Karafiát, L. Burget, Recurrent neural network based language modeling in meeting recognition. Interspeech 2011, ISCA, 2011, pp. 2877–2880, https://doi.org/10.21437/Interspeech.2011-720.

[30] A. Graves, Long short-term memory, in: A. Graves (Ed.), Supervised Sequence Labelling with Recurrent Neural Networks, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 37–45. https://doi.org/10.1007/978-3-642-24797-2_4.

[31] R. Dey, F.M. Salem, Gate-variants of Gated Recurrent Unit (GRU) neural networks, in: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), IEEE, Boston, MA, 2017, pp. 1597–1600, https://doi.org/10.1109/MWSCAS.2017.8053243.

[32] U. Khandelwal, H. He, P. Qi, D. Jurafsky, Sharp nearby, fuzzy far away: how neural language models use context, arXiv Preprint arXiv:1805.04623 (2018). https://doi.org/10.48550/arXiv.1805.04623.

[33] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2013.https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv Preprint arXiv:1301.3781 (2013). https://doi.org/10.48550/arXiv.1301.3781.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. ukasz Kaiser, I. Polosukhin, Attention is All you Need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[36] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv Preprint arXiv:1810.04805 (2018).

[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (2019) 9.

[38] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv Preprint arXiv:1910.13461 (2019). https://doi.org/10.48550/arXiv.1910.13461.

[39] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[40] H. Zhu, A graph neural network-enhanced knowledge graph framework for intelligent analysis of policing cases, MBE 20 (2023) 11585–11604, https://doi.org/10.3934/mbe.2023514.

[41] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 5485–5551.

[42] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, (2018).

[43] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, others, Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation, arXiv Preprint arXiv:2107.02137 (2021). https://doi.org/10.48550/arXiv.2107.02137.

[44] P. Cai, Y. Fan, F. Leu, Compare encoder-decoder, encoder-only, and decoder-only architectures for text generation on low-resource datasets. In: Barolli, L. (eds) Advances on Broad-Band Wireless Computing, Communication and Applications. BWCCA 2021. Lecture Notes in Networks and Systems, vol 346. Springer, Cham. (2020) 119. https://doi.org/10.1038/s41746-020-00323-1.

[45] A. Benayas, M. Sicilia, M. Mora-Cantallops, A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: navigating the trade-offs in model size and performance. (2024) PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.3.rs-3865391/v1.

[46] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, J. Tang, Glm: general language model pretraining with autoregressive blank infilling, arXiv Preprint arXiv:2103.10360 (2021). https://doi.org/10.48550/arXiv.2103.10360.

[47] J. Moon, G. Park, M. Yang, J. Jeong, Design and verification of process discovery based on NLP approach and visualization for manufacturing industry, Sustainability. 14 (2022) 1103, https://doi.org/10.3390/su14031103.

[48] J. Lim, S. Patel, A. Evans, J. Pimley, Y. Li, I. Kovalenko, Enhancing human-robot collaborative assembly in manufacturing systems using large language models, arXiv preprint arXiv:2406.01915 (2024). https://doi.org/10.48550/arXiv.2406.01915.

[49] J. An, T. Liu, Y. Chen, Advancing mass customization through gpt language models: a multidimensional analysis of market, technological, and managerial innovations, in: Y. Zhong (Ed.), International Conference on Mechatronics and Intelligent Robotics 845, Springer, Singapore, 2023, pp. 27–40, https://doi.org/10.1007/978-981-99-8498-5_3. Lecture Notes in Mechanical Engineering.

[50] J. Myöhänen, Improving industrial performance with language models: a review of predictive maintenance and process optimization, (2023). https://urn.fi/URN:NBN:fi-fe2023053150826.

[51] M. Lowin, A text-based predictive maintenance approach for facility management requests utilizing association rule mining and large language models, Mach. Learn. Knowl. Extract. 6 (2024) 233–258, https://doi.org/10.3390/make6010013.

[52] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, others, Gpt-4 technical report, arXiv Preprint arXiv:2303.08774 (2023). https://doi.org/10.48550/arXiv.2303.08774.

[53] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A.M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, PaLM: scaling Language Modeling with Pathways, Journal of Machine Learning Research 24 (2023) 1–113.

[54] J.W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, others, Scaling language models: methods, analysis & insights from training gopher, arXiv Preprint arXiv:2112.11446 (2022). https://doi.org/10.48550/arXiv.2112.11446.

[55] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, others, Lamda: language models for dialog applications, arXiv Preprint arXiv:2201.08239 (2022). https://doi.org/10.48550/arXiv.2201.08239.

[56] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X.V. Lin, others, Opt: open pre-trained transformer language models, arXiv Preprint arXiv:2205.01068 (2022). https://doi.org/10.48550/arXiv.2205.01068.

[57] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A.S. Luccioni, F. Yvon, M. Gallé, others, Bloom: a 176b-parameter open-access multilingual language model, (2023). https://inria.hal.science/hal-03850124.

[58] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, others, Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model, arXiv Preprint arXiv:2201.11990 (2022). https://doi.org/10.48550/arXiv.2201.11990.

[59] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, others, Llama 2: open foundation and fine-tuned chat models, arXiv Preprint arXiv:2307.09288 (2023). https://doi.org/10.48550/arXiv.2307.09288.

[60] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W.L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, J. Tang, GLM-130B: an open bilingual pre-trained model, (2023). http://arxiv.org/abs/2210.02414 (accessed May 24, 2024).

[61] L. Schulze Balhorn, J.M. Weber, S. Buijsman, J.R. Hildebrandt, M. Ziefle, A. M. Schweidtmann, Empirical assessment of ChatGPT's answering capabilities in natural science and engineering, Sci. Rep. 14 (2024) 4998, https://doi.org/10.1038/s41598-024-54936-7.

[62] J. Clusmann, F.R. Kolbinger, H.S. Muti, Z.I. Carrero, J.-N. Eckardt, N.G. Laleh, C. M.L. Löffler, S.-C. Schwarzkopf, M. Unger, G.P. Veldhuizen, S.J. Wagner, J. N. Kather, The future landscape of large language models in medicine, Commun. Med. 3 (2023) 141, https://doi.org/10.1038/s43856-023-00370-1.

[63] B.K. Betzler, H. Chen, C.-Y. Cheng, C.S. Lee, G. Ning, S.J. Song, A.Y. Lee, R. Kawasaki, P. Van Wijngaarden, A. Grzybowski, M. He, D. Li, A. Ran Ran, D.S. W. Ting, K. Teo, P. Ruamviboonsuk, S. Sivaprasad, V. Chaudhary, R. Tadayoni, X. Wang, C.Y. Cheung, Y. Zheng, Y.X. Wang, Y.C. Tham, T.Y. Wong, Large language models and their impact in ophthalmology, Lancet Digit. Health 5 (2023) e917–e924, https://doi.org/10.1016/S2589-7500(23)00201-7.

[64] M.J. Boonstra, D. Weissenbacher, J.H. Moore, G. Gonzalez-Hernandez, F. W. Asselbergs, Artificial intelligence: revolutionizing cardiology with large language models, Eur. Heart J. 45 (2024) 332–345, https://doi.org/10.1093/eurheartj/ehad838.

[65] M.R. Chavez, T.S. Butler, P. Rekawek, H. Heo, W.L. Kinzler, Chat generative pre-trained transformer: why we should embrace this technology, Am. J. Obstet. Gynecol. 228 (2023) 706–711, https://doi.org/10.1016/j.ajog.2023.03.010.

[66] J.W. Ayers, A. Poliak, M. Dredze, E.C. Leas, Z. Zhu, J.B. Kelley, D.J. Faix, A. M. Goodman, C.A. Longhurst, M. Hogarth, D.M. Smith, Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum, JAMa Intern. Med. 183 (2023) 589, https://doi.org/10.1001/jamainternmed.2023.1838.

[67] I.A. Bernstein, Y. (Victor) Zhang, D. Govil, I. Majid, R.T. Chang, Y. Sun, A. Shue, J. C. Chou, E. Schehlein, K.L. Christopher, S.L. Groth, C. Ludwig, S.Y. Wang, Comparison of ophthalmologist and large language model Chatbot responses to online patient eye care questions, JAMa Netw. Open. 6 (2023) e2330320, https://doi.org/10.1001/jamanetworkopen.2023.30320.

[68] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S. W. Ting, Large language models in medicine, Nat. Med. 29 (2023) 1930–1940, https://doi.org/10.1038/s41591-023-02448-8.

[69] S.B. Patel, K. Lam, ChatGPT: the future of discharge summaries? Lancet Digit. Health 5 (2023) e107–e108, https://doi.org/10.1016/S2589-7500(23)00021-3.

[70] S. Tayebi Arasteh, T. Han, M. Lotfinia, C. Kuhl, J.N. Kather, D. Truhn, S. Nebelung, Large language models streamline automated machine learning for clinical studies, Nat. Commun. 15 (2024) 1603, https://doi.org/10.1038/s41467-024-45879-8.

[71] K. Swanson, G. Liu, D.B. Catacutan, A. Arnold, J. Zou, J.M. Stokes, Generative AI for designing and validating easily synthesizable and structurally novel antibiotics, Nat. Mach. Intell. 6 (2024) 338–353, https://doi.org/10.1038/s42256-024-00809-7.

[72] N. Savage, Drug discovery companies are customizing ChatGPT: here's how, Nat. Biotechnol. 41 (2023) 585–586, https://doi.org/10.1038/s41587-023-01788-7.

[73] C. Chakraborty, M. Bhattacharya, S.-S. Lee, Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development, Mol. Therapy - Nucl. Acids 33 (2023) 866–868, https://doi.org/10.1016/j.omtn.2023.08.009.

[74] A. Zhao, Y. Wu, Future implications of ChatGPT in pharmaceutical industry: drug discovery and development, Front. Pharmacol. 14 (2023) 1194216, https://doi.org/10.3389/fphar.2023.1194216.

[75] T. Li, S. Shetty, A. Kamath, A. Jaiswal, X. Jiang, Y. Ding, Y. Kim, CancerGPT for few shot drug pair synergy prediction using large pretrained language models, Npj Digit. Med. 7 (2024) 40, https://doi.org/10.1038/s41746-024-01024-9.

[76] M. Zhou, W. Chen, S. Zhu, T. Cai, J. Yu, G. Dai, Application of large language models in professional fields, in: 2023 11th International Conference on Information Systems and Computing Technology (ISCTech), IEEE, Qingdao, China, 2023, pp. 142–146, https://doi.org/10.1109/ISCTech60480.2023.00033.

[77] M. Dowling, B. Lucey, ChatGPT for (Finance) research: the Bananarama conjecture, Financ. Res. Lett. 53 (2023) 103662, https://doi.org/10.1016/j.frl.2023.103662.

[78] P. Niszczota, S. Abbas, GPT has become financially literate: insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice, Financ. Res. Lett. 58 (2023) 104333, https://doi.org/10.1016/j.frl.2023.104333.

[79] D.H. Anh, D.-T. Do, V. Tran, N.L. Minh, The impact of large language modeling on natural language processing in legal texts: a comprehensive survey, in: 2023 15th International Conference on Knowledge and Systems Engineering (KSE), IEEE, Hanoi, Vietnam, 2023, pp. 1–7, https://doi.org/10.1109/KSE59128.2023.10299488.

[80] X. Yang, Z. Wang, Q. Wang, K. Wei, K. Zhang, J. Shi, Large language models for automated Q&A involving legal documents: a survey on algorithms, frameworks and applications, IJWIS (2024), https://doi.org/10.1108/IJWIS-12-2023-0256.

[81] N. Kshetri, Generative artificial intelligence and E-commerce, Computer 57 (2024) 125–128. https://doi.org/10.1109/MC.2023.3340772.

[82] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Differ. 103 (2023) 102274, https://doi.org/10.1016/j.lindif.2023.102274.

[83] J. Jeon, S. Lee, Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT, Educ. Inf. Technol. 28 (2023) 15873–15892, https://doi.org/10.1007/s10639-023-11834-1.

[84] S. Murugesan, A.K. Cherukuri, The rise of generative artificial intelligence and its impact on education: the promises and perils, Computer. (Long. Beach. Calif) Computer. (Long. Beach. Calif) 56 (2023) 116–121, https://doi.org/10.1109/MC.2023.3253292.

[85] M. Javaid, A. Haleem, R.P. Singh, S. Khan, I.H. Khan, Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system, BenchCouncil Trans. Benchmarks, Standard. Eval. 3 (2023) 100115, https://doi.org/10.1016/j.tbench.2023.100115.

[86] M. Zaabi, W. Hariri, N. Smaoui, A review study of ChatGPT applications in education, in: 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA), IEEE, Hammamet, Tunisia, 2023, pp. 1–5, https://doi.org/10.1109/INISTA59065.2023.10310439.

[87] L. Belzner, T. Gabor, M. Wirsing, Large language model assisted software engineering: prospects, challenges, and a case study, in: B. Steffen (Ed.), Bridging the Gap Between AI and Reality, Springer Nature Switzerland, Cham, 2024: pp. 355–374. https://doi.org/10.1007/978-3-031-46002-9_23.

[88] S. Suri, S.N. Das, K. Singi, K. Dey, V.S. Sharma, V. Kaulgud, Software engineering using autonomous agents: are we there yet?, in: 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE) IEEE, Luxembourg, 2023, pp. 1855–1857, https://doi.org/10.1109/ASE56229.2023.00174.

[89] B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M.P. Kumar, E. Dupont, F.J.R. Ruiz, J.S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, A. Fawzi, Mathematical discoveries from program search with large language models, Nature 625 (2024) 468–475, https://doi.org/10.1038/s41586-023-06924-6.

[90] L. Layman, R. Vetter, Generative artificial intelligence and the future of software testing, Computer. (Long. Beach. Calif) 57 (2024) 27–32, https://doi.org/10.1109/MC.2023.3306998.

[91] L.C. Fernandes, Programming computational electromagnetic applications assisted by large language models [Em Programmer's Notebook], IEEE Antennas Propag. Mag. 66 (2024) 63–71, https://doi.org/10.1109/MAP.2023.3336708.

[92] Z.-Y. Chen, F.-K. Xie, M. Wan, Y. Yuan, M. Liu, Z.-G. Wang, S. Meng, Y.-G. Wang, MatChat: a large language model and application service platform for materials science, Chinese Phys. B 32 (2023) 118104, https://doi.org/10.1088/1674-1056/ad04cb.

[93] A.M. Schweidtmann, Generative artificial intelligence in chemical engineering, Nature Chem. Eng. 1 (2024) 193, https://doi.org/10.1038/s44286-024-00041-5. –193.

[94] V. Dudhee, V. Vukovic, How large language models and artificial intelligence are transforming civil engineering, Proc. Institut. Civil Eng. - Civil Eng. 176 (2023) 150, https://doi.org/10.1680/jcien.2023.176.4.150. –150.

[95] R.S. Bonadia, F.C.L. Trindade, W. Freitas, B. Venkatesh, On the potential of ChatGPT to generate distribution systems for load flow studies using OpenDSS, IEEE Trans. Power Syst. 38 (2023) 5965–5968, https://doi.org/10.1109/TPWRS.2023.3315543.

[96] D.A. Boiko, R. MacKnight, B. Kline, G. Gomes, Autonomous chemical research with large language models, Nature 624 (2023) 570–578, https://doi.org/10.1038/s41586-023-06792-0.
[97] Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, S. Shi, Generative artificial intelligence and its applications in materials science: current situation and future perspectives, J. Materiomics 9 (2023) 798–816, https://doi.org/10.1016/j.jmat.2023.05.001.
[98] J. Lee, W. Jung, S. Baek, In-house knowledge management using a large language model: focusing on technical specification documents review, Appl. Sci. 14 (2024) 2096, https://doi.org/10.3390/app14052096.
[99] S.A. Prieto, E.T. Mengiste, B. García De Soto, Investigating the Use of ChatGPT for the scheduling of construction projects, Buildings 13 (2023) 857, https://doi.org/10.3390/buildings13040857.
[100] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, others, A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions, arXiv Preprint arXiv:2311.05232 (2023). https://doi.org/10.48550/arXiv.2311.05232.
[101] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: a roadmap, IEEe Trans. Knowl. Data Eng. (2024) 1–20, https://doi.org/10.1109/TKDE.2024.3352100.
[102] J. Wei, X. Wang, D. Schuurmans, M. Bosma, Brian Ichter, F. Xia, E. Chi, Q.V. Le, D. Zhou, Chain-of-Thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2022: pp. 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
[103] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, T. Zhou, A survey on knowledge distillation of large language models, arXiv Preprint arXiv:2402.13116 (2024). https://doi.org/10.48550/arXiv.2402.13116.
[104] D. Wang, C. Gong, Q. Liu, Improving neural language modeling via adversarial training, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 6555–6565, in: https://proceedings.mlr.press/v97/wang19f.html.
[105] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020: pp. 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (accessed January 25, 2024).
[106] F. Tao, H. Zhang, A. Liu, A.Y.C. Nee, Digital twin in industry: state-of-the-Art, IEEE Trans. Ind. Inf. 15 (2019) 2405–2415, https://doi.org/10.1109/TII.2018.2873186.
[107] H. Lee, S. Phatale, H. Mansoor, K. Lu, T. Mesnard, C. Bishop, V. Carbune, A. Rastogi, Rlaif: scaling reinforcement learning from human feedback with ai feedback, arXiv Preprint arXiv:2309.00267 (2023). https://doi.org/10.48550/arXiv.2309.00267.
[108] B. Meskó, Prompt engineering as an important emerging skill for medical professionals: tutorial, J. Med. Internet. Res. 25 (2023) e50638, https://doi.org/10.2196/50638.
[109] L. Makatura, M. Foshey, B. Wang, F. HähnLein, P. Ma, B. Deng, M. Tjandrasuwita, A. Spielberg, C.E. Owens, P.Y. Chen, How can large language models help humans in design and manufacturing? arXiv Preprint arXiv:2307.14377 (2023). https://doi.org/10.48550/arXiv.2307.14377.
[110] H. Yin, Z. Zhang, Y. Liu, The exploration of integrating the midjourney artificial intelligence generated content tool into design systems to direct designers towards future-oriented innovation, Systems. (Basel) 11 (2023) 566, https://doi.org/10.3390/systems11120566.
[111] F. Wu, S.-W. Hsiao, P. Lu, An AIGC-empowered methodology to product color matching design, Displays 81 (2024) 102623, https://doi.org/10.1016/j.displa.2023.102623.
[112] K. Yang, H. Liu, Y. Zhao, T. Deng, A new design approach of hardware implementation through natural language entry, IET Collab. Intel. Manufact. 5 (2023) e12087, https://doi.org/10.1049/cim2.12087.
[113] S. Xu, Y. Wei, P. Zheng, J. Zhang, C. Yu, LLM enabled generative collaborative design in a mixed reality environment, J. Manuf. Syst. 74 (2024) 703–715, https://doi.org/10.1016/j.jmsy.2024.04.030.
[114] R. Jardim-Goncalves, D. Romero, A. Grilo, Factories of the future: challenges and leading innovations in intelligent manufacturing, Int. J. Comput. Integr. Manuf. 30 (2017) 4–14, https://doi.org/10.1080/0951192X.2016.1258120.
[115] S. Wang, J. Wan, D. Li, C. Zhang, Implementing smart factory of Industrie 4.0: an outlook, Int. J. Distrib. Sens. Netw. 12 (2016) 3159805, https://doi.org/10.1155/2016/3159805.
[116] G.-J. Cheng, L.-T. Liu, X.-J. Qiang, Y. Liu, Industry 4.0 development and application of intelligent manufacturing, in: 2016 International Conference on Information System and Artificial Intelligence (ISAI), IEEE, Hong Kong, China, 2016, pp. 407–410, https://doi.org/10.1109/ISAI.2016.0092.
[117] L. Guo, F. Yan, T. Li, T. Yang, Y. Lu, An automatic method for constructing machining process knowledge base from knowledge graph, Robot. Comput. Integr. Manuf. 73 (2022) 102222, https://doi.org/10.1016/j.rcim.2021.102222.
[118] Y. Xiao, S. Zheng, J. Shi, X. Du, J. Hong, Knowledge graph-based manufacturing process planning: a state-of-the-art review, J. Manuf. Syst. 70 (2023) 417–435, https://doi.org/10.1016/j.jmsy.2023.08.006.
[119] H. You, Y. Ye, T. Zhou, Q. Zhu, J. Du, Robot-enabled construction assembly with automated sequence planning based on ChatGPT: roboGPT, Buildings 13 (2023) 1772, https://doi.org/10.3390/buildings13071772.
[120] H. Fan, X. Liu, J.Y.H. Fuh, W.F. Lu, B. Li, Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics, J. Intell. Manuf. (2024), https://doi.org/10.1007/s10845-023-02294-y.
[121] C. Gkournelos, C. Konstantinou, S. Makris, An LLM-based approach for enabling seamless Human-Robot collaboration in assembly, CIRP Annals (2024), https://doi.org/10.1016/j.cirp.2024.04.002. S000785062400012X.
[122] V. Shivajee, R.K. Singh, S. Rastogi, Manufacturing conversion cost reduction using quality control tools and digitization of real-time data, J. Clean. Prod. 237 (2019) 117678, https://doi.org/10.1016/j.jclepro.2019.117678.
[123] B. Zhou, X. Li, T. Liu, K. Xu, W. Liu, J. Bao, CausalKGPT: industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing, Adv. Eng. Inf. 59 (2024) 102333, https://doi.org/10.1016/j.aei.2023.102333.
[124] N. Rane, S. Choudhary, J. Rane, Intelligent manufacturing through generative artificial intelligence, such as ChatGPT or Bard, SSRN Journal (2024), https://doi.org/10.2139/ssrn.4681747.
[125] Q. Xu, G. Zhou, C. Zhang, F. Chang, Y. Cao, D. Zhao, Generative AI and DT integrated intelligent process planning: a conceptual framework, Int. J. Adv. Manuf. Technol. 133 (2024) 2461–2485, https://doi.org/10.1007/s00170-024-13861-9.
[126] A. Ucar, M. Karakose, N. Kırımça, Artificial intelligence for predictive maintenance applications: key components, trustworthiness, and future trends, App. Sci. 14 (2024) 898, https://doi.org/10.3390/app14020898.
[127] X. Liu, J. Vatn, S. Yin, V. Maithani, Performance of ChatGPT on CMRP: potential for assisting maintenance and reliability professionals using large language models, in: IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society, IEEE, Singapore, Singapore, 2023, pp. 1–7, https://doi.org/10.1109/IECON51785.2023.10311736.
[128] J. Jia, H. Fu, Z. Zhang, J. Yang, Diagnosis of power operation and maintenance records based on pre-training model and prompt learning, in: 2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), IEEE, Chizhou, China, 2022, pp. 58–61, https://doi.org/10.1109/DCABES57229.2022.00029.
[129] E. Öztürk, A. Solak, D. Bäcker, L. Weiss, K. Wegener, Analysis and relevance of service reports to extend predictive maintenance of large-scale plants, Procedia CIRP. 107 (2022) 1551–1558, https://doi.org/10.1016/j.procir.2022.05.190.
[130] F. Jiang, B. Jia, J. Wang, G. Zheng, Research on failure cause analysis method based on aircraft maintenance records, Chinese Society of Aeronautics and Astronautics, in: Proceedings of the 6th China Aeronautical Science and Technology Conference, Springer Nature Singapore, Singapore, 2024, pp. 374–388, https://doi.org/10.1007/978-981-99-8864-8_36.
[131] X. Qin, Y. He, J. Ma, W. Peng, E. Zio, H. Su, An effective knowledge mining method for compressor fault text data based on large language model, in: 2023 International Conference on Computer Science and Automation Technology (CSAT), IEEE, Shanghai, China, 2023, pp. 44–48, https://doi.org/10.1109/CSAT61646.2023.00024.
[132] H. Wang, Y.-F. Li, 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), IEEE, Beijing, China, 2023, pp. 474–479, https://doi.org/10.1109/SRSE59585.2023.10336112.
[133] S. Badini, S. Regondi, E. Frontoni, R. Pugliese, Assessing the capabilities of ChatGPT to improve additive manufacturing troubleshooting, Adv. Ind. Eng. Polym. Res. 6 (2023) 278–287, https://doi.org/10.1016/j.aiepr.2023.03.003.
[134] X. Cao, W. Xu, J. Zhao, Y. Duan, X. Yang, Research on large language model for coal mine equipment maintenance based on multi-source text, Appl. Sci. 14 (2024) 2946, https://doi.org/10.3390/app14072946.
[135] Y. Chen, Z. Zhao, J. Liu, S. Tan, C. Liu, Application of generative AI-based data augmentation technique in transformer winding deformation fault diagnosis, Eng. Fail Anal. 159 (2024) 108115, https://doi.org/10.1016/j.engfailanal.2024.108115.
[136] P. Liu, L. Qian, X. Zhao, B. Tao, Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly, IEEE Trans. Ind. Inf. (2024) 1–10, https://doi.org/10.1109/TII.2024.3366977.
[137] J. Guo, V. Mohanty, J.P. Ono, H. Hao, L. Gou, L. Ren, Investigating interaction modes and user agency in human-LLM collaboration for domain-specific data analysis, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024: pp. 1–9. https://doi.org/10.1145/3613905.3651042.
[138] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, BioGPT: generative pre-trained transformer for biomedical text generation and mining, (2023). https://doi.org/10.1093/bib/bbac409.
[139] W. Tao, M.C. Leu, Z. Yin, Multi-modal recognition of worker activity for human-centered intelligent manufacturing, Eng. Appl. Artif. Intell. 95 (2020) 103868, https://doi.org/10.1016/j.engappai.2020.103868.
[140] X. Ma, G. Fang, X. Wang, LLM-pruner: on the structural pruning of large language models, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2023: pp. 21702–21720. https://proceedings.neurips.cc/paper_files/paper/2023/file/44956951349095f74492a5471128a7e0-Paper-Conference.pdf.
[141] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, S. Han, Awq: activation-aware weight quantization for llm compression and acceleration, arXiv Preprint arXiv:2306.00978 (2023). https://doi.org/10.48550/arXiv.2306.00978.
[142] Y. Gu, L. Dong, F. Wei, M. Huang, MiniLLM: knowledge distillation of large language models, in: The Twelfth International Conference on Learning Representations, 2024. https://openreview.net/forum?id=5h0qf7IBZZ.

[143] C. Singh, J.X. Morris, J. Aneja, A.M. Rush, J. Gao, Explaining patterns in data with language models via interpretable autoprompting, (2023). https://openreview.net/forum?id=GvMuB-YsiK6.

[144] T. Saha, D. Ganguly, S. Saha, P. Mitra, Workshop on large language Models' interpretability and trustworthiness (LLMIT), in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, ACM, Birmingham United Kingdom, 2023, pp. 5290–5293, https://doi.org/10.1145/3583780.3615311.

[145] N. Chakraborty, M. Ornik, K. Driggs-Campbell, Hallucination detection in foundation models for decision-making: a flexible definition and review of the state of the art, arXiv Preprint arXiv:2403.16527 (2024). https://doi.org/10.48550/arXiv.2403.16527.

[146] S. Porsdam Mann, B.D. Earp, S. Nyholm, J. Danaher, N. Møller, H. Bowman-Smart, J. Hatherley, J. Koplin, M. Plozza, D. Rodger, P.V. Treit, G. Renard, J. McMillan, J. Savulescu, Generative AI entails a credit–blame asymmetry, Nat. Mach. Intell. 5 (2023) 472–475, https://doi.org/10.1038/s42256-023-00653-1.

[147] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, Y. Zhang, A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly, High-Confidence Comput. 4 (2024) 100211, https://doi.org/10.1016/j.hcc.2024.100211.

[148] Z. Ge, H. Huang, M. Zhou, J. Li, G. Wang, S. Tang, Y. Zhuang, WorldGPT: empowering LLM as multimodal world model, arXiv Preprint arXiv:2404.18202 (2024). https://doi.org/10.48550/arXiv.2404.18202.

[149] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, X. Huang, Z. Wang, L. Sheng, L. BAI, J. Shao, W. Ouyang, LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2023: pp. 26650–26685. https://proceedings.neurips.cc/paper_files/paper/2023/file/548a41b9cac6f50dccf7e63e9e1b1b9b-Paper-Datasets_and_Benchmarks.pdf.

[150] X. Zhu, J. Li, Y. Liu, C. Ma, W. Wang, A survey on model compression for large language models, arXiv Preprint arXiv:2308.07633 (2023). https://doi.org/10.48550/arXiv.2308.07633.

[151] C. Zhang, G. Zhou, J. Li, T. Qin, K. Ding, F. Chang, KAiPP: an interaction recommendation approach for knowledge aided intelligent process planning with reinforcement learning, Knowl. Based Syst. 258 (2022) 110009, https://doi.org/10.1016/j.knosys.2022.110009.

[152] A.H. Shariatmadari, S. Guo, S. Srinivasan, A. Zhang, Harnessing the power of knowledge graphs to enhance LLM explainability in the biomedical domain, (2024).

[153] C. Zhang, G. Zhou, Q. Xu, Z. Wei, C. Han, Z. Wang, A digital twin defined autonomous milling process towards the online optimal control of milling deformation for thin-walled parts, Int. J. Adv. Manuf. Technol. 124 (2023) 2847–2861, https://doi.org/10.1007/s00170-022-10667-5.

[154] J.A.H. Ali, B. Gaffinet, H. Panetto, Y. Naudet, Cognitive systems and interoperability in the enterprise: a systematic literature review, Annu. Rev. Control 57 (2024) 100954, https://doi.org/10.1016/j.arcontrol.2024.100954.

[155] J. Leng, D. Yan, Q. Liu, K. Xu, J.L. Zhao, R. Shi, L. Wei, D. Zhang, X. Chen, ManuChain: combining permissioned Blockchain with a holistic optimization model as Bi-level intelligence for smart manufacturing, IEEE Trans. Syst. Man Cybernetic.: Syst. 50 (2020) 182–192, https://doi.org/10.1109/TSMC.2019.2930418.

[156] U. Iqbal, T. Kohno, F. Roesner, LLM platform security: applying a systematic evaluation framework to OpenAI's ChatGPT plugins, arXiv Preprint arXiv:2309.10254 (2023). https://doi.org/10.48550/arXiv.2309.10254.

[157] N. Rieke, J. Hancox, W. Li, F. Milletarì, H.R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B.A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R.M. Summers, A. Trask, D. Xu, M. Baust, M.J. Cardoso, The future of digital health with federated learning, Npj Digit. Med. 3 (2020) 119, https://doi.org/10.1038/s41746-020-00323-1.