



A Review of Knowledge Base Construction Strategies for LLM-based Intelligent Decision Support System in Material Selection and Process Planning

Geonhwi Lee¹ · Solchan Kim¹ · Yulseok Byun¹ · Jiho Lee² · Hae-Jin Choi¹ 

Received: 30 November 2025 / Revised: 8 February 2026 / Accepted: 6 April 2026 / Published online: 21 May 2026
© The Author(s) 2026

Abstract

Material selection and process planning are core decision-making tasks in manufacturing, directly impacting cost, quality, productivity, and sustainability. However, diverse design alternatives, data scarcity, and qualitative requirements limit the effectiveness of existing data-driven approaches and expert systems, creating challenges in generalization and knowledge acquisition. Large language models (LLMs) have emerged as promising tools due to their natural language interaction and complex query-handling capabilities, yet hallucination and outdated knowledge pose significant risks in high-reliability manufacturing environments. To address these issues, this study proposes retrieval-augmented generation (RAG), which integrates reliable external knowledge, as the core architecture for an LLM-based intelligent decision support system (IDSS). The objective is to identify appropriate knowledge base (KB) structures and practical construction strategies for material selection and process design. This study reviews decision-making methodologies, research on material selection and computer-aided process planning (CAPP), and prior KB construction approaches, reframing them through a RAG-centric perspective. The analysis reveals that CAPP involves heterogeneous tasks across feature, operation, and system levels, requiring integrated use of rule-based reasoning, optimization, knowledge inference, and learning-based algorithms. Although knowledge graphs (KGs) effectively structure complex manufacturing knowledge, manual construction remains costly and time-consuming. Recent LLM-based techniques that automatically extract and refine knowledge from unstructured data have emerged as promising solutions to overcome these limitations and improve KG scalability. The combination of LLMs and KGs can alleviate the generalization constraints of traditional data models and the knowledge acquisition bottlenecks of expert systems, forming a foundation for next-generation intelligent manufacturing systems. This study outlines key future research directions: developing an upper-level integrated schema applicable across diverse manufacturing domains, establishing bottom-up methodologies for constructing reliable domain-specific KGs, and creating quantitative evaluation and management frameworks for the LLM–KG–RAG pipeline.

Keywords Large language model · Material selection · Process planning · Decision support system · Retrieval-augmented generation · Knowledge base

1 Introduction

In manufacturing, material selection and process planning directly affect manufacturing costs, quality, and productivity, and these decisions should be made in the early stages of product design [1–3]. In addition, these two decisions are not separate tasks but a coupled problem. This means that materials and manufacturing processes should be evaluated in an integrated manner by considering mutual constraints rather than being selected independently in a sequential manner. The properties and machinability of a material

✉ Hae-Jin Choi
hjchoi@cau.ac.kr

¹ School of Mechanical Engineering, Chung-Ang University,
84 Heukseok-ro, Seoul 06974, Dongjak-gu, Republic of
Korea

² Department of Mechanical Engineering, University of North
Texas, Denton, TX 76205, USA

restrict the range of applicable manufacturing processes, whereas the choice of manufacturing process and its operating conditions define the achievable ranges of performance and cost, thereby influencing the validity of material selection [4, 5]. However, due to numerous alternatives, diverse evaluation criteria, uncertain requirements, and complex constraint relationships between materials and processes, material and process selection is inherently difficult and complex [1, 3, 6]. Furthermore, process planning is also considered a highly challenging problem due to the diversity of manufacturing contexts, the complexity of decision-making logic, and the intricacy of process knowledge [7]. In particular, as sustainability has emerged as a core manufacturing value driven by increasingly stringent global environmental regulations and market demands, traditional cost–quality–productivity–centric decision-making has evolved into a complex multi-criteria decision-making problem that must also account for environmental factors such as energy efficiency, carbon emissions, and material recyclability [8, 9]. As a result, these inherent challenges have become even more intensified. In addition, the continuous development of new materials and manufacturing processes is expanding the range of alternatives that must be considered [10, 11]. Moreover, advances in digital technologies such as intelligent sensors and industrial IoT generate massive volumes of real-time manufacturing data, adding high-dimensional and dynamic characteristics to decision-making problems that are difficult to handle using traditional static analytical methods [12, 13].

A decision support system (DSS) is an interactive computer-based system that assists organizations or individuals in making rational and effective decisions by utilizing data, models, and knowledge when facing complex decision-making situations [14, 15]. Building on this concept, research over the past several decades has focused on developing systems that construct and leverage data, models, and knowledge bases to address material selection and process planning problems in manufacturing [1, 10, 16–19]. For example, multi-criteria decision-making (MCDM) techniques have been used to rank alternatives in material and process selection problems [1, 10, 16], and expert systems have been developed to support process planning by formalizing expert knowledge in the form of rule-based logic such as IF-THEN rules or knowledge graphs [17–19]. In addition, as rapid changes in manufacturing environments have increased the complexity and uncertainty of process parameters, various artificial intelligence (AI) techniques such as artificial neural networks, fuzzy logic, and genetic algorithms have been incorporated into intelligent decision support systems (IDSS) to address material and process selection and planning, aiming to overcome the limitations of traditional DSSs that struggle to handle subjective

and uncertain data [20, 21]. However, these systems often rely on constrained interfaces such as standardized input forms or predefined menu selections, which prevent them from sufficiently accommodating the diverse query styles that arise from differences in user background knowledge or expertise. Consequently, by forcing all users to follow the same input procedure, such systems suffer from reduced flexibility and usability.

As an alternative to overcome the rigid interfaces and low usability inherent in traditional systems, large language models (LLMs), which have rapidly advanced in recent years, have gained significant attention [22]. LLMs are being adopted across a wide range of domains owing to their generality and scalability, both of which derive from extensive training on massive datasets [22–25]. In particular, their user-friendly natural language interaction capabilities offer the potential to fundamentally address the limitations of conventional DSSs, which rely on rigid input structures and struggle to process diverse user queries. However, despite this promising potential, applying LLMs directly to critical decision-making tasks in manufacturing environments—where high reliability and accuracy are essential—entails substantial inherent risks. One of the most critical issues is the phenomenon of hallucination, in which an LLM generates plausible but factually incorrect information [26]. Such hallucinations may lead to severe consequences, including suggesting nonexistent or inappropriate materials and processes or recommending incorrect process sequences. In addition, LLMs are unable to incorporate dynamic knowledge such as newly introduced advanced materials or emerging process trends that appear after the time of training.

Retrieval-augmented generation (RAG) is a technique that enables an LLM to retrieve and utilize reliable external data and knowledge in real time when generating responses to user queries. It has emerged as a representative approach for mitigating hallucination issues in LLMs. The adoption of RAG has drawn considerable attention as a practical solution for designing systematic architecture for LLM-based IDSS. By referencing verified external knowledge during the response generation process, RAG can significantly reduce the risk of an LLM producing incorrect information or generating content that deviates from the intended context. In addition, RAG offers the strong advantage of incorporating up-to-date information by linking to continuously updated external knowledge sources. These characteristics play an essential role in material selection and process planning, fields in which rapid technological change and the availability of current information have a decisive impact on decision-making outcomes.

Because material selection and process planning are interrelated decisions [4, 5], practical decision support

should be able to handle evidence-grounded queries that explicitly connect material and process information, such as identifying materials that satisfy equipment or process constraints or determining appropriate process routes for a given material. Therefore, the knowledge source for RAG should support joint reasoning over material and process information rather than treating them as independent topics. To this end, the following two core research questions must be addressed.

1. What format should the knowledge source be constructed in to support RAG-based decision-making for material selection and process planning?
2. What domain knowledge should the knowledge source contain to support RAG-based decision-making for material selection and process planning?

To answer these research questions, this paper systematically analyzes existing studies on material selection and process planning from the perspective of constructing data and knowledge bases for RAG. It further discusses how knowledge base construction and algorithm development should evolve to address increasingly complex manufacturing decision-making problems. Prior work on material selection and process planning has accumulated mainly in a method-oriented manner, including MCDM, knowledge-based expert systems, and AI models, or in a domain- and case-specific manner [18, 27]. Meanwhile, in the manufacturing domain, document-based question answering using LLMs and RAG has also been proposed, with discussions focusing on pipeline design and knowledge graph-enhanced retrieval to mitigate limitations in retrieval precision and context selection [28]. However, existing reviews tend to organize material selection studies around MCDM and process planning studies around knowledge-based approaches. As a result, a systematic framework that integrates design principles for verifiable knowledge sources from the perspective of an IDSS based on RAG has not been systematically

synthesized across the material selection and process planning literature [18, 27, 28].

Accordingly, this paper makes the following contributions. First, it restructures the literature on material selection and process planning from the perspective of grounded generation based on RAG. Second, it analyzes the options and implications of knowledge source representations for RAG. Third, it reviews knowledge base construction and utilization cases from the perspectives of quality assurance and updateability. Fourth, it presents a research roadmap and design implications for developing next-generation IDSS based on LLMs. Figure 1 provides an overview of the overall structure and analytical perspective of the review proposed in this paper.

The structure of this paper is as follows. Section 2 reviews the background and existing research methodologies in the fields of material selection and process planning related to the topic of this study. Section 3 introduces the technical background of RAG and its applications in the manufacturing domain, which serve as key components of this work. Based on this foundation, Sect. 4 reviews and analyzes case studies on the construction and utilization of knowledge bases for RAG. Section 5 discusses the overall implications of the proposed approach and outlines future research directions, and Sect. 6 presents the conclusion.

2 Review of Material Selection and Process Planning

2.1 Material Selection

In product design, materials directly influence not only the final performance of a product but also various factors such as economic viability, sustainability, manufacturability, and supply stability, and the outcome of material selection constrains the range of manufacturing processes applicable in subsequent production stages [29–32]. Therefore, material selection must consider multiple aspects, including material properties, cost, manufacturability, environmental friendliness, and sustainability. However, the number of materials available for industrial application is vast, and these factors often conflict with or depend on one another, making material selection inherently a complex multi-criteria decision-making problem.

The material selection process typically consists of four steps: interpreting customer requirements and performance goals (establishing selection criteria), screening, comparison and ranking, and final selection [10, 33, 34]. In the screening step, materials that satisfy the defined criteria and constraints are filtered, and in the subsequent comparison and ranking step, performance scores are assigned for each

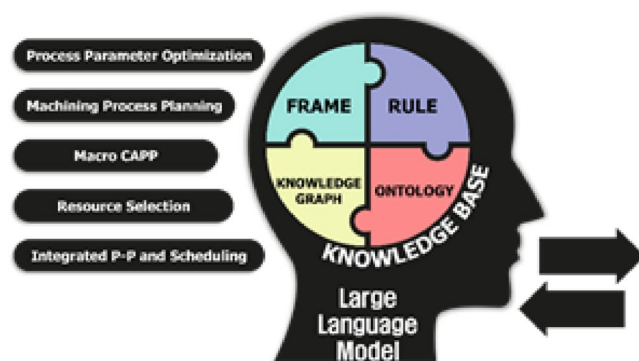


Fig. 1 Overview of the review scope and analytical direction adopted in this paper

material across multiple evaluation criteria to derive the final set of candidates. To support this process, prior research has applied chart-based screening methods, MCDM techniques, expert systems, and data-driven approaches to systematically address the complexity of material selection problems [10]. Accordingly, this subsection reviews the major approaches employed in existing material selection studies and analyzes the limitations inherent in each approach. Table 1 summarizes representative case studies on material selection according to each methodological approach and

outlines the key characteristics and research contents of these methods.

2.1.1 Chart-based Approaches

Chart-based approaches were among the earliest approaches used in material selection, with Ashby's chart being a representative example that visualizes two material properties simultaneously. Ashby charts are primarily employed in the material screening stage, as they allow performance indices to be overlaid on the chart for a large set of candidate

Table 1 Classification and summary of material selection studies by methodological approach

Approach	Method	Summary	Refs
Chart-based	Ashby's chart	Ashby chart-based, micro-electromechanical systems material selection	[35]
		Ashby approach-based bipolar plates material selection	[36]
MCDM	AHP	AHP based selection of green composite materials for orthopedic prosthesis selection	[40]
		AHP based selection of phase change materials (PCM) for solar energy storage selection	[41]
		AHP based materials and assembly method selection for railcar development selection	[42]
		Fuzzy TOPSIS method based eco-material selection for automobile body panels	[45]
	TOPSIS	TOPSIS and AHP methods based horizontal wind turbine blade material selection	[47]
		TOPSIS with various normalization methods based tribological coating material selection	[46]
		VIKOR	VIKOR method based brushed DC motor component material selection
	Entropy-VIKOR method based automotive piston component material selection		[51]
	ELECTRE	BBWM and ELECTRE III based honeycomb structural materials selection for electronics clean rooms selection	[54]
		Fermatean fuzzy ELECTRE based biomedical material selection for hip joint prosthesis selection	[55]
Expert System	Rule-based	The high reliability and consistency of ranking results provided by the PROMETHEE II technique in multi-criteria material selection problems.	[57]
		PSI and PROMETHEE method-based ranking of fiber reinforced composite materials	[58]
	Fuzzy rule-based	Expert system-based material selection for automotive parts and implants in manufacturing	[66]
		Rule-Based Expert System (KBES)-based material selection for racing bicycle forks	[67]
		Optimization of material selection in Additive Manufacturing using a Fuzzy Logic Rule-Based Expert Advisory System	[69]
Data-driven	Case-based reasoning	Case-Based Reasoning and Multi-Criteria Decision-Making (ARAMIS/AIR/CAIR) combined approach-based construction material selection in petrochemistry	[70]
	Surrogate models	A methodology is proposed that integrates topology optimization with IDEM-based design exploration for a vehicle antenna-supporting structure, enabling mass reduction while maintaining the required safety factor.	[74]
		A methodology for transparent GFRP that leverages fiber and polymer refractive-index and dispersion data, combined with a physics-informed LSTM-based ML model, to accurately predict visible-light transmittance and identify optimal fiber–polymer combinations.	[75]
		A methodology is presented that couples an ANN-based material model with FEA and thermal analysis to simultaneously optimize the alloy composition and geometry of A-USC superheater tubes, thereby reducing creep and enhancing heat transfer.	[76]
		The study proposes an AdaBoost-based forward and inverse machine learning framework that uses literature data to optimize the composition of cast heat-resistant aluminum alloys and accurately predict and improve their high-temperature tensile strength at 300–350 °C	[77]
	Generative AI	A framework that integrates a Wasserstein Generative Adversarial Nets WGAN-based generative model with Random Forest and other machine-learning property predictors to generate MPEA compositions and processing routes and design new alloys with targeted mechanical properties.	[78]
		The paper presents MatterGen, a diffusion-based generative model that jointly generates atom types, coordinates, and lattices to produce stable, novel inorganic crystal structures and can be fine-tuned to inverse-design materials that satisfy target chemical, symmetry, and property constraints.	[79]
		The paper proposes a generative deep-learning framework that combines a GAN with a boosted-tree surrogate model to learn composition–property mappings from limited data and to generate and inverse-design compositionally complex bulk metallic glasses	[80]

materials, making them useful for identifying groups of materials that can satisfy required functional specifications in conceptual design. Guisbiers and Wautelet performed material selection for micro-electromechanical systems based on Ashby charts [35], and de Oliveira et al. applied the Ashby approach to select materials for bipolar plates [36].

However, Ashby charts are particularly effective when two or three key criteria dominate the material selection process, but as the number of evaluation criteria increases, their two-dimensional visualization structure makes it difficult to reflect multi-criteria requirements simultaneously [10, 36]. In addition, Ashby charts are primarily intended to visually identify the feasibility region and screen candidate materials. Thus, they do not directly provide quantitative ranking among alternatives, which limits their use as a standalone method in the final material selection stage [10, 37]. Moreover, Ashby charts are limited in adequately addressing trade-offs across technical, economic, and environmental aspects. Reflecting these limitations, Risaliti et al. combined an Ashby chart method with MCDM approaches in the redesign of an electric-vehicle C-segment motor mounting bracket [37]. They first reduced the candidate pool using a Granta Selector-based Ashby chart and then applied the Vise Kriterijumska Optimisacija Kompromisno Resenje (VIKOR) method to rank the top 20 alternatives, ultimately identifying the combination of low-alloy steel AISI 9255 and press forming as the final solution. Compared to the baseline design, this resulted in approximately 44% mass reduction, 75% cost savings, and about a 60% reduction in carbon footprint. In summary, while chart-based screening is well suited for reducing the search space, it has limitations in final decision-making under conflicting criteria. Therefore, it should be integrated with methods that support quantitative ranking and compromise or trade-off decision-making.

2.1.2 Multi-Criteria Decision-Making Approaches

MCDM is a systematic methodology for determining the optimal alternative by simultaneously considering multiple conflicting criteria within a single decision-making process [38]. Material selection requires the simultaneous consideration of multiple, often conflicting criteria, which has led to the widespread use of MCDM techniques as systematic evaluation tools [1, 10, 16]. In particular, MCDM has played an important role in material selection because it structures diverse evaluation criteria and incorporates their relative importance, allowing material alternatives to be quantitatively compared and ranked. By organizing the evaluation criteria and assigning weights that reflect their relative importance, MCDM provides a structured basis for

quantitatively comparing and prioritizing material alternatives. Representative MCDM methods include Analytic Hierarchy Process (AHP), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), VIKOR, Elimination and Choice Expressing the Reality (ELECTRE), and Preference Ranking Organisation Method for Enrichment of Evaluations (PROMETHEE).

AHP is a technique that decomposes a decision problem into a hierarchical structure and derives criterion weights through pairwise comparisons, allowing qualitative and subjective factors to be incorporated in a systematic manner while offering simplicity and flexibility in its structure [39]. In addition, the process of deriving weights includes a mathematical consistency check to verify whether the decision maker's judgments are logically coherent, thereby ensuring the reliability of the results [39]. Based on these characteristics, AHP has been applied across various engineering domains, such as orthopedic prosthetics and bone fixation implants [40], solar energy storage devices [41], and railway vehicle structures [42], to evaluate and rank candidate materials and processes using multi-criteria assessment. However, because AHP is based on pairwise comparisons, the number of required comparisons increases sharply to $n(n-1)/2$ as the number of alternatives or criteria increases to n , which substantially increases computation time and the burden of user input [43]. Indeed, Mohamed et al. measured AHP computation time while incrementally increasing the number of requirements from 10 to 500. They found that computation time increased exponentially as the number of requirements grew, and the execution time for 500 requirements was reported as 161.709 min, which is approximately 64.65 times longer than that for 10 requirements (2.501 min) [43]. Therefore, AHP has limitations when applied alone to large-scale decision-making problems involving many alternatives and criteria. This calls for combining screening techniques to reduce the candidate set in advance, adopting other MCDM methods with relatively lower computational burden, or employing hybrid combinations with other MCDM methods.

TOPSIS is a technique that ranks alternatives by calculating their distances from the positive ideal solution and the negative ideal solution based on given criterion weights and each alternative's performance values [44]. It has the advantages of a straightforward procedure and a fixed sequence of computational steps, which makes it easy to apply even as the number of attributes increases [44]. However, because TOPSIS does not determine criterion weights by itself, practical applications commonly first assign criterion importance using a separate weighting method and then rank alternatives using TOPSIS [45, 46]. Mayyas et al. considered 21 candidate materials for automotive body-in-white (BIW) panels and applied fuzzy TOPSIS by

modeling environmental, mechanical, and functional attributes as fuzzy linguistic variables [45]. They evaluated each alternative's closeness to the ideal solution and ultimately selected high-strength steel-based materials as a superior eco-friendly alternative [45]. Okokpujie et al. addressed material selection for horizontal-axis wind turbine blades in low-wind-speed environments [46]. They derived criteria weights for cost, lightweight property, corrosion resistance, and durability using AHP and then applied TOPSIS to identify aluminum 6061-T9 alloy as the optimal alternative [46]. Yang et al. established 12 coating alternatives with different compositions and deposition conditions to select material and process conditions for boron-based tribological hard coatings, and they applied TOPSIS with AHP–entropy combined weights across seven performance indicators to identify the optimal coating alternative [47]. They also compared TOPSIS rankings obtained using seven representative normalization methods applied to the same dataset and found that the optimal alternative was not consistent across the normalization methods, demonstrating that the choice of normalization can significantly affect the relative closeness and the final ranking [47]. To integrate rankings derived from different normalization methods, they proposed a Final Rank Index and used it to determine the final optimal coating alternative [47].

VIKOR is similar to TOPSIS in that it calculates rankings based on relative performance with respect to the ideal solution [48]. However, while TOPSIS evaluates alternatives by their distance from the positive ideal solution and the negative ideal solution, VIKOR derives the final ranking from a compromise perspective that maximizes group utility and minimizes individual regret [48]. This characteristic makes VIKOR advantageous for deriving compromise solutions that reflect the trade-off between group utility and individual regret in multi-criteria decision-making problems with significant conflicts between criteria [49]. Indeed, Zulkafli et al. applied entropy-weighted VIKOR to comprehensively evaluate multi-criteria performance and derive optimal material combinations for a brushed DC motor, where conflicting requirements such as thermal performance, electromagnetic torque, cost, and mass coexist [50]. They set candidate materials for key components such as permanent magnets, casing, bearings, brushes, and slot insulation [50]. Furthermore, VIKOR has the advantage of enabling the assessment of the solution stability of compromise solutions under varying weights [49]. Indeed, Dev et al. established a candidate group of composites reinforced with recycled porcelain in LM-26 Al alloy, targeting automotive piston applications [51]. They ranked alternatives using entropy-weighted VIKOR to simultaneously consider multiple conflicting criteria: density reduction, strength and hardness improvement, and reduction in wear rate and

friction coefficient [51]. Furthermore, by performing sensitivity analysis assuming variations in criterion weights to examine changes in the ranking results, they confirmed that the 6 wt% porcelain-reinforced composition stably maintained its position as the optimal compromise from an overall performance perspective [51]. However, VIKOR has the limitation of excessively emphasizing individual regret due to its structure where the worst criterion, already included in the group utility calculation, is reflected again as individual regret during the compromise index calculation process [52]. In case studies of manufacturing process selection, it has been reported that this excessive penalty can lead to alternatives differing from those proposed as optimal in prior studies being derived as the optimal solution [52].

Meanwhile, ELECTRE and PROMETHEE perform decision-making based on outranking relationships among alternatives [8, 10]. ELECTRE compares all alternative pairs to establish outranking relationships, and in applications aiming to derive a final brief set of promising alternatives, such as ELECTRE I, it is used to justify decisions to exclude the remaining alternatives [53]. ELECTRE is useful for non-compensatory decision-making that does not allow trade-offs between criteria, and when a veto threshold is introduced, as in ELECTRE III, it can screen alternatives by blocking outranking so that a critical flaw in one criterion is not offset by advantages in other criteria [53, 54]. Indeed, Chen et al. applied the non-compensatory outranking characteristic of ELECTRE III to a multi-criteria comparison involving safety-related criteria such as fire and earthquake resistance in a case study on selecting honeycomb structural materials for electronics clean rooms [54]. Similarly, Kirişçi et al. addressed biomaterial selection for hip prostheses by incorporating criteria for which deficiencies are unacceptable from safety and suitability perspectives, such as toxicity, corrosion, and biocompatibility, into a Fermatean fuzzy ELECTRE-based decision-making process and derived the optimal alternative through outranking-based selection [55].

PROMETHEE ranks alternatives using preference functions and outranking flows, producing a partial ranking in PROMETHEE I and a complete ranking in PROMETHEE II [56]. Accordingly, PROMETHEE has also been applied in material selection studies that require clear and practical rankings of alternatives [57, 58]. For example, Mahajan et al. applied PROMETHEE II, TOPSIS, and VIKOR to the same set of alternatives and criteria in a case study on selecting natural fibers for sustainable composites, compared the resulting rankings using Spearman's rank correlation, and reported that PROMETHEE II showed a high average rank similarity to TOPSIS and VIKOR, suggesting it as a reliable method in terms of rank consistency [57]. Furthermore, Patnaik et al. defined mechanical properties and slurry wear-related characteristics as evaluation criteria

and jointly applied PSI and PROMETHEE to eight epoxy composite alternatives reinforced with polyester needle-punched nonwoven fabrics (PS1–PS4) and viscose fabrics (VS1–VS4) to derive the final ranking of fiber-reinforced composite alternatives [58].

Although various MCDM methods have effectively supported material selection problems, several structural limitations arise in practical application. First, the rank reversal problem, in which the existing ranking changes when alternatives are added or removed, undermines the consistency and reliability of decision-making [59]. Indeed, Mousavi-Nasab and Sotoudeh-Anvari reapplied MCDM methods such as TOPSIS, VIKOR, and Simple Additive Weighting (SAW) to the same data and weights used in various prior material selection studies [59]. They analyzed whether the relative rankings among the original alternatives changed by first calculating the rankings within the original alternative set, then removing the lowest-ranked alternative, and finally recalculating the rankings for the remaining alternatives [59]. The results confirmed that, in all methods except SAW, structural changes in the set of alternatives alone could cause rank reversal among top alternatives in some cases [59]. However, while Mousavi-Nasab and Sotoudeh-Anvari's reanalysis reported SAW as the only method in which rank reversal was not observed [59], Wang and Luo demonstrated through analysis and numerical examples that rank reversal can occur even in SAW under scenarios involving the addition or removal of alternatives [60]. In addition, because criterion weights heavily depend on experts' subjective judgments, uncertainty in weight estimation can arise and directly affect ranking changes and alternative selection outcomes [61, 62]. Furthermore, traditional MCDM proceeds under structural assumptions that criteria are independent and can be represented in a hierarchical structure and that aggregation is performed using weighted-sum aggregation or simple distance measures, which limits the ability to fully reflect interdependencies among criteria, feedback effects, and non-additive and nonlinear interactions [63].

2.1.3 Expert Systems

Expert systems are computer programs designed to replicate the intelligent behavior of human experts and provide decision-making capabilities for solving complex problems [10, 17, 18]. Among them, the rule-based expert system is the most representative type, and it represents information acquired from human experts in the form of IF–THEN rules [64]. Because rules are expressed in the way experts reason, this type of system has the advantage that the basis for its judgments can be easily traced and explained [65]. However, rule-based knowledge bases have a limitation in

that adding or modifying a small number of rules is relatively simple, whereas the knowledge base becomes highly complex as the number of rules increases [65]. Due to this characteristic, rule-based expert systems are effective in situations where explicit rules exist and explainability is critical. In a rule-based expert system for material selection, material property data and experts' experience and judgment are formalized into IF–THEN rules to build a rule-based knowledge base, and the inference engine then compares the problem information provided by the user against the rules in the knowledge base to perform inference and derive a conclusion [66, 67].

Fuzzy expert systems are developed based on fuzzy logic, enabling computers to mimic the general reasoning process of humans by operating in a less precise manner [64]. Accordingly, fuzzy expert systems can systematically represent and reason about ambiguous knowledge that is difficult to handle using probability theory and Boolean logic, while also enabling realistic decision support by mimicking the approximate reasoning style of human experts [64]. However, performance can deteriorate if appropriate rule structures and membership functions are not selected [68]. Furthermore, extracting both linguistic rules and fuzzy sets from experts is time-consuming, and system performance can be significantly affected by disagreements among experts [68]. Accordingly, fuzzy expert systems are effective in situations where quantitative data are scarce and reliance on expert experience is necessary. Indeed, Skrzek et al. proposed an expert advisory system that combines triangular and trapezoidal membership functions and IF–THEN rule-based fuzzy inference with criterion weights calculated by AHP to incorporate intuitive judgments and experience-based knowledge that are difficult to capture using AHP alone, and they reported that the recommendation results achieved approximately 85% agreement with user preferences in an additive manufacturing (AM) material selection case [69].

Case-based reasoning is a problem-solving method that stores past problem-solving cases in a database and applies solutions from retrieved similar cases to new problems when comparable issues arise [65]. Accordingly, CBR-based material selection studies have been explored to derive solutions from past material selection cases. For example, Berman et al. conducted material selection for petrochemical facilities using a hybrid expert system that retrieves similar cases via CBR and then validates and ranks the candidate materials from a multi-criteria perspective using MCDM techniques [70].

However, expert systems have a fundamental limitation known as the knowledge acquisition bottleneck, in which extracting and formalizing tacit knowledge, such as an expert's intuition or know-how, into explicit rules is

extremely difficult and costly [17]. Furthermore, as material datasets grow larger, the complexity of the rules that must be managed increases exponentially, which substantially reduces scalability and maintenance efficiency [18, 69]. In addition, because the system operates only within the scope of its initially encoded knowledge and rules, it is difficult to respond flexibly to environments with many exceptions or rapid changes, such as new material development. Indeed, Goel et al. explain that when new materials are added to the database during material selection, the number of output nodes must be increased, which in turn requires retraining the neural network [71]. This retraining requirement increases the operational burden under changing knowledge and resource configurations and can act as a practical constraint. Thus, the difficulty of knowledge acquisition, the management complexity associated with scaling, and the limited ability to respond to change can combine to create structural limitations in the application and operation of rule-based approaches.

2.1.4 Data-driven Approaches

Data-driven approaches have been widely applied in material selection to identify complex nonlinear relationships between material design variables and product performance within large design spaces, thereby enabling efficient exploration for optimal materials [72, 73]. By learning the relationship between material design variables (or material properties) and product performance, it is possible to identify ranges of material properties that satisfy target performance for screening suitable candidates or, conversely, to estimate design variables such as material composition and process parameters required to achieve the desired performance [74–77]. Indeed, Jang et al. constructed a Kriging surrogate model based on finite element analysis data to model the relationship between the geometry and material properties of antenna support structures and their mass and factor of safety. They then screened design–property combinations that satisfy the target constraints using the surrogate model and mapped them onto the property space of an Ashby plot to progressively narrow down the pool of realistically selectable material candidates [74]. Recently, research has also explored the use of generative models to predict the structure and composition of new materials, explore unexplored regions, and discover novel materials [78–80]. For example, Li et al. trained a Wasserstein WGAN-based generative model using a multicomponent alloy dataset to generate new alloy composition candidates. They combined a material property prediction model with Pareto-front-based multi-objective optimization to select promising compositions that satisfy trade-offs between target properties such as strength and elongation [78].

Generally, the validity domain of data-driven models is constrained by the training data, and prediction reliability may decline in extrapolation scenarios beyond the training distribution [81, 82]. Indeed, recent out-of-distribution (OOD) benchmark studies have reported that apparent OOD performance improvements often stem from interpolation within regions sufficiently covered by the training data, whereas prediction accuracy deteriorates markedly when test samples lie outside the training representation space [82]. To improve generalization performance, high-quality data encompassing diverse variables and conditions is essential, but real-world material data are frequently biased, imbalanced, and sparse, which limits the coverage of the training distribution and can reduce prediction reliability in OOD situations [81, 82]. Furthermore, these methods inherently rely on numerical data, creating structural limitations in directly integrating qualitative design requirements or experience-based expert heuristics that are difficult to quantify into decision-making.

2.2 Process Planning

Process planning is a technical decision-making process that systematically defines how a designed product will be manufactured, serving as a critical bridge between the design stage (CAD) and the actual production stage, including CAM and shop-floor operations [19, 83–85]. Rather than merely arranging machining or assembly steps, process planning interprets the design intent from the perspective of manufacturability and organizes all required procedures in a logical manner to transform the design into an executable manufacturing plan [85]. This process encompasses a wide range of tasks that constitute the manufacturing workflow, including the selection of machining and assembly methods, structuring of working steps, identification of required resources such as equipment, tools, and materials, determination of process parameters, definition of quality criteria, planning of inspection operations, and consideration of cost factors.

Computer-aided process planning (CAPP) is a technological concept developed to support or automate the intermediate link between design and manufacturing in a more consistent and efficient manner [19, 83–85]. CAPP covers a range of activities, from extracting manufacturing-related semantic information from design data to generating process routes, selecting resources and tools, determining process conditions, and planning inspection procedures required in process planning. Through this integration, CAPP strengthens the consistency of planning, reduces engineering effort, and ensures digital continuity across the entire manufacturing pipeline. In this section, we classify CAPP tasks into three levels that correspond to feature, operation, and

system. This classification is based on the scope and granularity of decision-making during the refinement of design information into an executable process plan. The feature-level represents interpretation and planning units through which design semantics are reflected in process planning. The operation-level represents decision and execution units that translate the plan into concrete working steps. The system-level represents integration units that consolidate multiple operations into a globally consistent process plan. This three-level hierarchy captures the expansion of decision scope from local interpretation to step-level specification and further to global integration, thereby providing a basis for structurally understanding CAPP according to the scale and characteristics of task types. Table 2 systematically categorizes representative case studies from the CAPP-related literature according to three task levels (feature, operation, and system) and five technology categories (rule-based, optimization, knowledge-driven reasoning, geometric deep learning, and reinforcement learning).

2.2.1 Multi-Level Structure of Decision-Making Tasks in CAPP

The feature-level represents the initial stage of process planning, where geometric data such as CAD models, B-Rep structures, or point clouds are interpreted from a manufacturing perspective to obtain meaningful information. Since subsequent tasks, including process planning, tool selection, and assembly structure derivation, directly rely on the information extracted at this stage, the accuracy and consistency of the feature-level serve as fundamental factors that determine the overall quality of the CAPP pipeline. Among its detailed tasks, feature recognition is a representative one, referring to the automatic identification of manufacturing features such as holes, pockets, slots, ribs, and chamfers from CAD geometry [86–93]. This process is not a simple geometric classification but a critical function that extracts semantic units required for assessing manufacturability, thereby linking CAD geometry with manufacturing knowledge. The feature-level also includes another independent task, manufacturing process identification, which interprets the identified features in a manufacturing context. Manufacturing process identification determines which manufacturing process can be applied to each recognized feature and requires a higher degree of manufacturing knowledge integration than feature recognition [94, 95].

The operation-level defines the actual manufacturing tasks based on the information obtained from the feature-level and represents the stage in which the greatest variety of decision-making activities occurs within the process planning framework. At this level, working steps are explicitly formulated, and the manner in which each task will be

executed, the resources required, and the conditions under which the task will be operated are determined in detail. Under this objective, the operation-level comprises several tasks with distinct functional roles, including machining process planning, tool and resource selection, and process parameter optimization. Machining process planning determines how and in what sequence the recognized features will be machined, encompassing a range of subproblems such as process selection, sequence generation, setup planning (fixture orientation and clamping strategy), and candidate path generation [96–103]. Subsequently, tool and resource selection addresses the choice of tools, machines, fixtures, and measurement devices required to execute the machining operations, playing an essential role in ensuring both feasibility and efficiency [100, 103–105]. Finally, process parameter optimization improves machining performance by optimizing cutting conditions such as speed, feed, and depth of cut, along with cooling parameters, machining time, and quality indicators [106–108].

The system-level defines the overarching manufacturing structure of the entire product, representing a high-level planning stage that considers not individual operations or working steps but the overall process flow, resource allocation, and system-wide constraints [84]. From this perspective, the system-level encompasses several upper-level tasks that determine the manufacturing framework of the product, including assembly process planning, Macro CAPP, and integrated process planning and scheduling (IPPS). Assembly process planning analyzes the structural composition of the product and determines the assembly sequence, assembly resources, and assembly constraints, thereby ensuring consistency and feasibility in the overall assembly flow [109–112]. Following this, Macro CAPP serves as a higher-level planning activity that determines the manufacturing route of the entire product, evaluates process alternatives, and configures equipment layouts and resource flows, supporting strategic decision-making at the manufacturing system-level [97, 105, 113–115]. Lastly, IPPS treats process planning and production scheduling as a unified problem, constituting a key approach for balancing plan executability with overall production efficiency [84, 116–118].

2.2.2 CAPP Spectrum of Technical Approaches in CAPP

The technical approaches applied in CAPP research vary depending on task characteristics and the form of decision-making required. Rule-based methods construct decision logic using explicit knowledge such as manufacturing rules, constraints, and geometric or process mappings, and they are primarily applied to tasks that require determining whether specific conditions are satisfied. Related literature has addressed tasks such as STEP-based manufacturing

Table 2 This table classifies the CAPP-related literature by three task levels and five technology categories. Filled circles (●) indicate technologies that constitute a major focus of each study, while hollow circles (○) indicate technologies used in a secondary or supporting capacity. The table provides an overview of which technologies are predominantly applied to each CAPP task type

CAPP task type	Feature-level		Operation-level			System-level		
	Task:	Feature recognition	Manufacturing process identification	Machining process planning	Tool and resource selection	Process parameter optimization	Assembly process planning	Macro CAPP
Task:	Feature of Task:	Automatically identifying and classifying manufacturable features from CAD/B-Rep / point-cloud geometry	Determining the appropriate machining process based on part geometry and feature characteristics	Generating machining working steps by selecting suitable operations and defining the required machining sequence	Selecting the optimal tools, equipment, and resources required to execute each working step under given machining constraints	Optimizing machining parameters to achieve desired quality, productivity, and cost	Determining feasible assembly sequences and required resources based on product structure and assembly constraints	Designing high-level manufacturing routes by selecting processes, machines, and operation flows for the entire part
								Jointly optimizing process planning and scheduling (IPPS)
Technology Category	Key Approaches							
		Using rule libraries and geometric-topological reasoning to derive feasible manufacturing decisions	-	● [102], [103]	○ [103]	-	● [112]	-
Optimization	Solving planning and scheduling using mathematical programming, heuristics, or ML-OR hybrids	-	-	-	● [104]	● [106], [108]	-	● [84], [116], [117], [118]
	Using graphs, ontologies, and rules to encode manufacturing knowledge and support constraint-based reasoning	○ [88], [89]	○ [94]	● [96], [99] ○ [97]	● [105] ○ [104]	-	● [110], [111]	● [97], [105], [113], [114], [115] ○ [116], [117]

Table 2 (continued)

CAPP task type		Feature-level		Operation-level		System-level		
Task:		Feature recognition	Manufacturing process identification	Machining process planning	Tool and resource selection	Process parameter optimization	Assembly process planning	Macro CAPP
CAPP task type	Feature of Task:	Automatically identifying and classifying manufacturable features from CAD/B-Rep / point-cloud geometry	Determining the appropriate machining process based on part geometry and feature characteristics	Generating machining working steps by selecting suitable operations and defining the required machining sequence	Selecting the optimal tools, equipment, and resources required to execute each working step under given machining constraints	Optimizing machining parameters to achieve desired quality, productivity, and cost	Determining feasible assembly sequences and required resources based on product structure and assembly constraints	Designing high-level manufacturing routes by selecting processes, machines, and operation flows for the entire part
								Jointly optimizing process planning and scheduling (IPPS)
Technology Category		Key Approaches						
Geometric Deep Learning	Learning geometric representations from CAD/B-Rep/point-cloud data via CNNs, GNNs, or point-based models	• [87], [88], [89], [90], [91], [92]	• [94], [95]	• [98], [101] ○ [99]	-	-	• [111] ○ [110]	○ [97]
	Learning optimal sequential decisions through Q-learning or actor-critic policies	-	-	• [96], [100]	• [100]	• [107]	• [109]	-

feature recognition (AFR) for rotating parts [93], volumetric feature classification and recognition for free-form surfaces with support for milling machine selection [102] and generating and evaluating assembly plan alternatives by integrating process monitoring plans into assembly planning [112]. However, rule-based approaches strongly depend on predefined rules and the scope of feature representations, and the approach itself can be constrained for geometries or conditions that are not covered by those representations. For example, Al-Wswasi et al. explicitly state that SI-AFR is limited to recognizing planar and circular surface-based features and thus cannot include free-form surface features, and that the internal and external separation procedure is not valid for internal geometries based on spherical or conical surfaces [93]. Consequently, rule-based approaches can exhibit limitations in that it is difficult to generalize and apply the same logic when the assumed geometric representations and rule conditions do not hold.

Optimization-based approaches specify quantifiable objective functions such as time, cost, quality, and energy, and perform decision-making by exploring optimal or near-optimal solutions under constraints. That is, since the objective function provides criteria for comparing and selecting alternatives, it is particularly advantageous when multiple process alternatives coexist or conflicting criteria must be considered simultaneously. Specifically, studies have proposed a decision support system combining fuzzy CBR and fuzzy AHP for cutter planning and control [104], a multi-criteria evaluation that considers time, cost, accuracy, tool life, and surface roughness in machining process planning evaluation [106], and a solution search that applies a learning-induced hybrid genetic algorithm and multi-neighbor search to minimize makespan and total weighted tardiness in integrated process planning–scheduling (IPPS) for reconfigurable manufacturing cell environments [116]. However, optimization-based approaches can yield different solution ranges depending on the settings of the objective function and constraints. Furthermore, if real-world constraints and variability are not sufficiently modeled, the applicability of derived solutions may be limited. Hu et al. explicitly state that because real assembly and testing workshops face diverse disturbances and constraints, it is necessary to consider additional constraints and dynamic scheduling [116]. Therefore, optimization-based approaches can be limited when real-world factors exist that are not captured by the objective function and constraints, which can restrict the applicability of the explored solutions in the field [119].

Knowledge-based reasoning performs semantic inference by leveraging structural knowledge or relational information within the manufacturing domain. It excels at deriving and refining choices based on relational constraints and chained relations, which are difficult to determine solely

from observed inputs. Its strength lies in the ability to reduce candidates or construct decision paths by following relational structures, particularly in situations where processes, resources, and constraints are intertwined, making it difficult to reliably narrow down candidates using simple similarity or rule matching alone. Examples demonstrating this include a framework that formulates process planning as a sequential recommendation problem and recommends subsequent decisions by leveraging explicit and implicit relationships in process knowledge [96], a setup that selects and activates process analysis rules from process elements of design components and then constructs the activated knowledge structure as a search space to explore process paths [97], and a framework that organizes knowledge based on predefined resource schemas and supplements relationships through schema-based inference for industrial resource recommendation under low-resource conditions [105]. These demonstrate how knowledge-based reasoning can structurally refine decision candidates and paths through relational knowledge. However, Zhang et al. explicitly state a limitation that their proposed process inference requires further improvement in generalization performance when the volume of high-quality process data is limited [97]. Thus, even when leveraging relational information, knowledge-based reasoning may face the limitation that the applicability of inference results can be constrained if the scope and quality of the supporting data and knowledge are insufficient [120].

Geometric deep learning processes structural and geometric data such as CAD models, B-Rep structures, and point clouds to learn shape-based semantics and is widely applied in feature-level tasks. Its core strength lies in learning spatial and topological patterns between shape elements, enabling direct judgment in feature identification and classification tasks that are difficult to handle reliably using only rule-based feature definitions or simple similarity matching. Related research has proposed machining feature recognition combining machining feature instance segmentation, feature type identification (semantic segmentation), and bottom face identification using B-Rep-based inputs [87]. Configurations performing machining process label classification (e.g., milling, turning) using STL-based CAD geometry as input are also addressed [94]. Furthermore, a framework has been proposed that formalizes milling process planning as a classification problem to predict planning items, including tool diameter selection [98]. While geometric deep learning is advantageous for feature-level decision-making based on shape patterns, Chung et al. explicitly state a limitation that contour-based features extracted via convolution remain confined to relative shape information and fail to provide scale references for actual dimensions [98]. They also point out limitations of image-based shape semantic learning, in which visually similar shapes with

different scales can be misinterpreted, potentially reducing selection accuracy [98].

Reinforcement learning is used to learn policies and generate executable strategies when sequential decision-making problems are involved, such as operation sequencing or resource allocation. In other words, it models a decision-making process in which state changes and constraints accumulate as a chain of environment–action interactions, and it provides a repeatable decision-making procedure by internalizing step-by-step selections into a policy. Representative application tasks include process planning decisions that combine machining operation sequencing and machining resource (machine/tool) allocation [100], assembly sequence planning that considers dynamic resource conditions [109], and additive manufacturing process parameter optimization (e.g., searching for parameter combinations that satisfy quality metrics) [107]. These cases demonstrated reinforcement learning for generating executable sequences, allocations, or parameter strategies in multi-stage decision-making by linking sequential choices to policy learning [100, 107, 109]. However, Wu et al. explicitly state a limitation that a well-trained policy can handle process planning only within the machining resource environment used for training, and that additional learning is required when new tools or equipment are introduced and the environment changes [100]. This suggests that the scope of policy reuse may be limited under real-world application conditions where resource configurations change [100].

2.3 Limitations of Existing Approaches in Material Selection and Process Planning

The approaches reviewed in the previous section can be systematically categorized into three groups based on what drives decision-making. Knowledge-based approaches are guided by explicitly represented domain knowledge such as rules, constraints, and relational structures, performing decision-making through symbolic manipulation rather than learning from data. Decision-analytic (optimization-based) approaches are driven by predefined mathematical models, such as criterion weights, normalization, objective functions, and constraints, enabling quantitative comparison and selection of alternatives. In contrast, data-driven learning approaches learn decision logic from data or interaction experiences with the environment, deriving variable-performance relationships, representation patterns, or sequential decision policies. They then perform prediction, recommendation, and multi-stage decision-making based on this. However, these approaches share a common constraint in that they require either detailed pre-definition of rules or mathematical models or reliance on sufficient training data and environmental coverage. Consequently, in

practical applications, tasks such as material selection, process planning, and optimization are often handled separately rather than integrated into a unified system. This results in structural limitations, including insufficient system interconnectivity, inflexible user interaction, and an inability to organically combine quantitative data with domain knowledge. Below, these limitations are summarized from three perspectives.

The first major limitation is the fragmentation of systems and the lack of integration. This is because integrating knowledge with different expression units, such as materials and processes, into a single model significantly increases additional formalization costs, including schema alignment, redefinition of criteria, and constraint formulation. As a result, an approach that separates tasks like material selection, process selection, sequencing, and parameter optimization for individual modeling and optimization tends to prevail over integrated design. Indeed, existing decision-support research primarily focuses on specific tasks like material selection and manufacturing process selection [1, 121, 122]. This tendency remains confined to optimizing each stage individually and does not extend to full-process integration. Consequently, material selection systems and process planning systems are often developed and operated independently. This frequently forces users to rely on multiple standalone software tools with different interfaces and data formats to design the entire manufacturing process [123, 124]. This fragmentation creates a structural impediment, since materials and processes are inherently interdependent, but the disconnect among systems makes it difficult to derive a global optimum [1].

The second limitation is the rigidity of system interfaces and the lack of user-friendliness. Most decision support systems are implemented in rule-based or function-based forms that rely on predefined rules or fixed mathematical expressions, requiring users to strictly adhere to the required input formats and procedures to obtain results [125, 187]. This rigid interaction structure cannot flexibly accommodate users with diverse backgrounds or unstructured query intentions. Consequently, it ultimately restricts system flexibility and increases the entry barrier for practical use [125, 126].

The third limitation is the disconnection between heterogeneous layers of information and the lack of data–knowledge fusion. Manufacturing decision-making requires not only quantitative data such as sensor measurements but also qualitative knowledge such as process rules, engineering experience, and information from the literature [127]. However, existing approaches show structural constraints in integrating these heterogeneous elements within a unified reasoning framework. Data-driven approaches focus exclusively on learning statistical patterns from numerical data and therefore often fail to capture the causal relationships

inherent in domain knowledge [128, 129]. At the same time, expert systems rely solely on explicit rules, and limitations have been reported in learning latent patterns embedded in the data and adapting to complex or changing environments [130]. Consequently, the need for an integrated reasoning framework that can organically integrate bottom-up data with top-down knowledge continues to be raised [130].

In conclusion, next-generation manufacturing decision-support systems must move beyond task-specific fragmented structures and rigid data formats to integrate quantitative data with textual knowledge [131]. They must also provide user-friendly interfaces capable of understanding natural-language queries, thereby lowering entry barriers for users with limited expertise [132]. In this context, LLMs can process vast amounts of textual knowledge while interacting with external tools and databases to handle numerical data and perform integrated reasoning [133]. Accordingly, they have emerged as a promising solution to overcome the fragmentation and rigidity of existing systems [134].

3 LLM-based Manufacturing IDSS

3.1 Potentials and Structural Limitations of LLMs

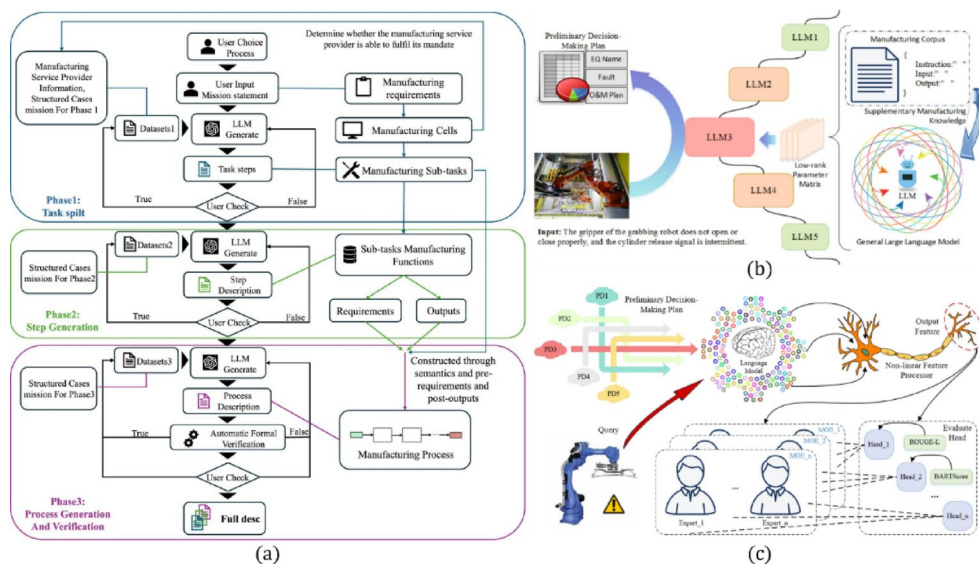
As the structural limitations of existing methodologies discussed in the preceding sections have become evident, the integration of recently advanced LLMs is driving a transformative shift in the paradigm of IDSS. Leveraging their sophisticated natural language processing and reasoning capabilities, LLMs act as cognitive orchestrators that can interpret user's complex query intentions and actively control various external systems. This capability goes beyond simply connecting fragmented individual systems and instead enables the organic integration of heterogeneous

information layers, including both quantitative data and qualitative knowledge. Moreover, LLMs provide a flexible interaction environment that is not constrained by the user's level of expertise, thereby significantly improving system accessibility. Building on these capabilities, a growing body of research in the manufacturing domain is investigating intelligent decision-support systems that utilize LLMs to deliver information to users and assist with complex decision-making tasks.

Ni et al. proposed the LLMAPM framework, which uses LLMs to convert user natural-language input into executable manufacturing task flows, as shown in Fig. 2(a) [135]. The framework employs advanced prompt engineering techniques such as chain-of-thought prompting and in-context learning. It decomposes the user input into specific sub-tasks and defines detailed functional descriptions and parameter specifications required to execute each sub-task. The outputs of these steps are then integrated to define interactions among the sub-tasks and generate the complete workflow.

Du et al. proposed the LLM-MANUF framework, which mitigates manufacturing decision bias through a Mixture of Experts (MoE) architecture that integrates multiple fine-tuned LLMs in parallel, as illustrated in Fig. 2(b) [136]. The input decision requirements are processed by several models fine-tuned on domain-specific datasets, and the decision plans produced by these models are evaluated using the Dynamic Weighted Mixture of Experts Ranking Method (DWMOE), as shown in Fig. 2(c). The model that generates the highest-ranked plan then aggregates the top candidate responses to derive the final decision plan. Using automotive manufacturing O&M datasets, they demonstrated that even combinations of relatively small-scale models can generate concrete and immediately actionable recommendations. These findings show that, through prompt engineering and

Fig. 2 (a) Overview of the LLMAPM framework. Reprinted with permission from ref [135]. Copyright Taylor & Francis, 2025 (b) Generation of Early-Stage Decision-Making Plans in the LLM-MANUF Framework. Reprinted with permission from ref [136]. Copyright Elsevier, 2025 (c) Overview of the Dynamically Weighted Mixture-of-Experts Process in the LLM-MANUF Framework Reprinted with permission from ref [136]



fine-tuning, IDSS based on manufacturing-domain-specific LLMs can effectively propose actions for diverse situations in manufacturing environments.

However, despite these methodological improvements, the structural vulnerabilities inherent in large language models pose a significant challenge for practical application in manufacturing environments. The most prominent issue is the phenomenon of hallucination, where information lacking a factual basis is generated as if it were true [26]. In their survey, Huang et al. analyzed the root causes of LLM hallucination from three perspectives, namely data, training, and inference [26]. From the data perspective, they summarized that (i) misinformation and biases contained in the pretraining corpus can be reproduced and amplified due to the model's memorization tendency, (ii) the knowledge boundary imposed by the coverage of pretraining data makes it easy to fabricate facts for queries outside that boundary, and (iii) low-quality alignment data used during the post-pretraining alignment stage can induce additional hallucinations. From the training perspective, they reported that next-token prediction in pretraining, soft attention dilution in long contexts, and exposure bias caused by the mismatch between training and inference can trigger a snowball effect in which errors accumulate and amplify hallucinations. They also noted that during the SFT stage, when annotated instructions exceed the model's capability or knowledge boundary, the model can be trained to fit out-of-bound responses, and that the combination of overfitting to newly introduced factual knowledge and the inability to express refusal or uncertainty strengthens the tendency to fabricate content rather than refuse to answer out-of-bound questions. Finally, from the inference perspective, they explained that limitations of decoding strategies can directly trigger hallucinations, and that in particular, the risk of hallucination increases as randomness grows in stochastic sampling used to ensure creativity and diversity, while higher temperature exacerbates this by expanding sampling of low-frequency tokens. They further summarized that overconfidence, which reflects excessive reliance on partial outputs during generation, and insufficient contextual attention caused by local attention, including instruction forgetting, can produce context- and instruction-mismatch hallucinations, and that hallucinations may also arise from the representational limitation of output distributions due to the softmax bottleneck and from reasoning failure.

In manufacturing environments, relying on information distorted by hallucinations can lead to critical defects in product integrity and safety. A recent study by Liu et al. [137] quantitatively demonstrated this risk. When asked about the surface treatment process for titanium alloy fasteners used in Boeing 787 fuselage assembly, the LLM recommended electroless nickel plating. However, this process

is explicitly prohibited in HB 8752–2023 due to the risk of hydrogen embrittlement. Furthermore, the model failed to provide even the coating thickness dimensional tolerance ($\pm 5\%$) specified by AS9100D:2016. If such hallucination-induced technical errors were applied to actual processes, the risk associated with specification non-compliance would exceed the FAA acceptable criteria (AC 25.1309) by more than 100 times (two orders of magnitude), which could directly lead to massive economic losses and fatalities.

Additionally, the lack of recency is another issue [138]. Unlike static knowledge domains, the manufacturing industry experiences rapid technological innovation and development, making it highly likely that the information provided by LLMs does not reflect the latest state. This lack of recency reduces the accuracy of manufacturing decisions, significantly limiting the system's practical utility. For example, as the development of high-performance new materials and new processes accelerates, the material property data of newly developed materials and the parameter specifications of newly introduced equipment must be updated promptly. Furthermore, the process–structure–property (PSP) correlation information associated with new material and process development must also be continuously refined. Indeed, in AM processes, rapid thermal cycling forms non-equilibrium microstructures that differ from those in conventional manufacturing methods. Relying solely on historical PSP data can therefore lead to severe errors in material property prediction [139]. Concurrently, the latest requirements from periodically revised industry standards and legal regulations must also be continuously incorporated. For instance, the amendment to the ECHA's PFAS restriction measures announced in October 2025, which applies only a temporary grace period for core materials such as lubricants and seals for construction machinery, underscores the necessity of up-to-date regulatory information during material selection [140]. If an LLM fails to reflect such phase-out schedules or regulatory changes and recommends these materials based on outdated data, companies could face not only premature degradation issues but also significant legal and economic risks due to non-compliance.

The aforementioned advanced techniques, such as prompt engineering, fine-tuning, and MoE architectures, alone cannot fundamentally resolve the issues of hallucinations and the lack of recency. Anh-Hoang et al. analyzed the causes of hallucinations by distinguishing between prompt-induced and model-intrinsic factors. They observed that model-dominant models such as DeepSeek exhibit hallucinations in specific domains, regardless of prompt quality [141]. Furthermore, Chain-of-Thought (CoT) prompting consistently reduced hallucinations across all models, but when a model lacks fundamental knowledge, it can instead produce a “backfire” effect by justifying an incorrect answer

in greater detail. An approach such as the LLM-MANUF framework proposed by Du et al., which builds multiple domain-specialized models through fine-tuning on a manufacturing corpus, integrates them via a Mixture of Experts (MoE) architecture, and then selects the optimal response using a dynamically weighted evaluation scheme, can mitigate hallucinations in specific domain knowledge and improve the reliability of the outputs [135]. However, while fine-tuning is effective for strengthening internal knowledge, continuously retraining the model to synchronize it with the latest data from rapidly changing manufacturing sites is unrealistic due to the substantial computational cost and time constraints. Furthermore, hallucinations can still occur in response to unseen data or dynamic field conditions that arise after fine-tuning, which constitutes a fundamental limitation in maintaining knowledge recency [142].

Addressing this, Handler et al. emphasized that overcoming these limitations in LLM-based IDSS requires research into systematic structural designs that utilize LLMs as a common interface for IDSS [143]. In other words, a technical approach that maintains the LLM's reasoning capabilities while simultaneously referencing the latest external knowledge and data in real-time to strengthen the basis for responses is essential. In this context, RAG is gaining attention as a core technology to overcome the limitations of LLMs and ensure the reliability of manufacturing decisions. Indeed, Lee et al. empirically showed that knowledge graph-based RAG in safety management tasks in the construction domain reduces hallucination risk through accurate and verifiable information retrieval and provides access to up-to-date databases [144]. Their results indicate that this approach is more advantageous than existing AI models in improving the efficiency and reliability of safety management plan formulation. Furthermore, Werner and Arenella showed that an RAG system based on a centralized knowledge base integrating standards, test reports, design records, certification documents, and accident data can reduce the time required for compliance research while improving response consistency and accuracy [145]. This enables rapid, evidence-based regulatory decision-making even in complex global safety standard environments.

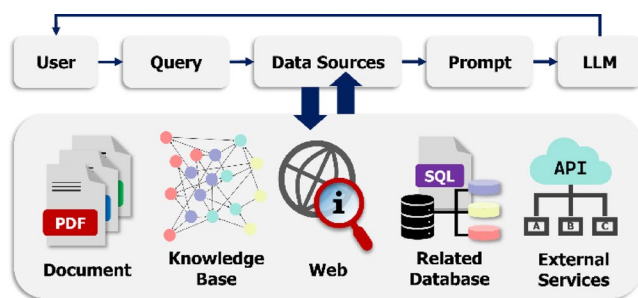


Fig. 3 Overview of the answer generation process in RAG

3.2 Retrieval-Augmented Generation

RAG is a semi-parameterized approach that integrates non-parametric knowledge stored in external knowledge sources with a parameterized LLM to address the hallucination and slow knowledge update issues of LLMs [146]. Figure 3 illustrates the answer generation process through RAG. When a user query is input, an LLM-based search query generation module creates and executes a search query to retrieve documents relevant to the question from external data sources. The retrieved documents, along with the user's question, are injected into the LLM's input prompt for answer generation. Based on this, the LLM generates an answer grounded in the content of the retrieved documents. This approach enables the provision of fact-based, high-credibility answers and efficiently maintains information currency by updating only the external data sources without retraining the model itself [146, 147].

Vector databases are one of the representative external data sources in RAG. They split documents into chunks and store them by converting each chunk into a high-dimensional vector representation using an embedding model [146]. Here, the embedding model projects each chunk into a numerical vector in a high-dimensional semantic space. During this process, texts that share semantic similarity are mapped to occupy nearby locations in that space or to have similar geometric orientations [148, 149]. When a user query is input, it is also embedded, and a similarity search is performed against the vectors stored in the database. Subsequently, the text chunks corresponding to the vectors with the highest similarity scores are included in the prompt, thereby generating the final response [146].

In this process, chunks serve as the fundamental unit of retrieval, and RAG precision and recall are directly influenced by how the original document text is segmented into chunks based on specific criteria and strategies [150, 151]. Bhat et al. systematically analyzed variations in retrieval recall across multiple datasets with diverse characteristics, such as document length and answer type, under different embedding models and chunk size settings, and found that smaller chunks (64–128 tokens) yielded the best performance on datasets with short, fact-based answers, whereas larger chunks (512–1024 tokens) were required for datasets demanding descriptive or technical responses [150]. Furthermore, Stäbler et al. compared the retrieval performance of various chunking strategies and embedding model combinations using token-level Intersection-over-Union (IoU) and reported that sentence-based chunking consistently outperformed semantic chunking and fixed-length token-based chunking across domains, and that mid-range chunk sizes (256–512 tokens) achieved more favorable performance than smaller (128 tokens) or larger (1024 tokens) settings

[151]. These prior findings demonstrate that chunking is not merely a preprocessing step but a core design variable that determines RAG retrieval quality.

Accordingly, various chunking strategies have recently been proposed to simultaneously improve recall and precision (or accuracy) [152]. Zhao et al. proposed a granularity-aware Mixture-of-Chunkers (MoC) framework combining a router that predicts appropriate chunking granularity per document with expert chunkers (meta chunkers) for each granularity level [152]. MoC mitigates the computational cost–accuracy tradeoff while significantly improving chunking quality and RAG performance by having the router select meta-chunkers based on document characteristics, generating and applying chunk boundaries via regular expression (Regex) rules, and reliably extracting the original text chunks through edit-distance-based recovery [152]. Yepes et al. proposed an element-type-based chunking framework that constructs chunks based on structural elements extracted from financial reports by a document understanding model [153]. Their approach demonstrated the effectiveness of structure-aware chunking in the financial domain by improving evidence snippet retrieval quality and RAG Q&A accuracy compared to existing fixed-length and sentence-based chunking in experiments on FinanceBench. Singh et al. proposed a chunking strategy that decomposes documents into sentences during the chunking phase and then sets chunk boundaries where the cosine similarity between sentence embeddings falls below a threshold, thereby generating semantically cohesive variable-length chunks [154]. This chunking method enhances thematic and logical consistency within a single chunk and reduces unnecessary surrounding context, enabling more stable retrieval of chunks containing relevant information in the subsequent retrieval stage.

3.3 Inherent Limitations of RAG and Evidence-Handling Framework

RAG is a prominent approach that provides “factual anchors” to LLM responses by injecting documents retrieved from external knowledge sources into the prompt, thereby mitigating hallucinations. However, RAG does not completely eliminate hallucinations, and failures during retrieval, evidence integration, and answer generation can still lead to hallucinations [155]. Therefore, omissions in retrieval, the inclusion of irrelevant or erroneous evidence, or the injection of ambiguous, conflicting, or incomplete evidence can still result in incorrect answers and hallucinations. Accordingly, Sect. 3.3.1 analyzes these inherent limitations and failure modes of RAG, while Sect. 3.3.2 summarizes mechanisms for mitigating them.

3.3.1 Inherent Limitations and Failure Modes of RAG

The causes of hallucinations in RAG can be broadly categorized into retrieval failure at the retrieval stage and generation deficiency at the answer generation stage [155]. Retrieval failure refers to situations where the system fails to sufficiently retrieve correct evidence relevant to the user’s query from external data sources or retrieves irrelevant or erroneous information alongside it. Conventional RAG obtains evidence through vector similarity search, which embeds user queries and documents and then selects the top-*k* documents with the highest vector similarity scores, such as cosine similarity. However, because vector similarity search prioritizes vectors that are “close” to the query in the embedding space, documents that appear semantically similar but are actually unsuitable as correct evidence for the query may also be retrieved [156]. Furthermore, encoding queries and documents as fixed-length dense vectors can impose limitations in expressive capacity and semantic fidelity, particularly for long documents, because diverse nuances are compressed into a single representation and the distance relationships between queries and documents in the embedding space can be distorted. This can lead to retrieval errors in which irrelevant documents are mapped relatively close to the query [157]. Moreover, because a fixed top-*k* set is returned, if fewer than *k* items containing the correct evidence exist in the vector database, the remaining slots may be filled with relatively high-similarity but irrelevant items, increasing the likelihood that they are retrieved together [158]. Meanwhile, if the embedding model is primarily trained on general-domain corpora, it may fail to capture domain-specific expressions sufficiently, leading to a domain mismatch between the query and the corpus [146]. In such cases, even when correct evidence exists, it may not be mapped sufficiently close in the embedding space and can be excluded from the top-*k* results, resulting in retrieval failure. Consequently, retrieval-stage failures provide the LLM with incomplete context in which correct evidence is missing or irrelevant evidence is mixed in, and this becomes a prerequisite that determines response quality in subsequent stages.

Defects in the answer generation stage refer to failures that occur when an LLM generates responses without sufficiently utilizing or integrating evidence provided through retrieval, or without maintaining factual consistency [155]. When retrieval-stage failures cause irrelevant or redundant documents to be injected into the answer generation prompt, context noise can occur. In such cases, the model may fail to identify key evidence and instead be swayed by peripheral information, plausibly assembling incorrect content [155, 159]. Furthermore, when documents containing conflicting information are retrieved simultaneously, or when

retrieved documents contradict the model's parameterized knowledge, context conflict may occur. In such cases, the model may fail to adequately reflect and aggregate the conflicts, instead relying on a subset of top-ranked documents or using its parameterized knowledge as a tie-breaker, which can yield inconsistent conclusions [155, 160]. Furthermore, according to Zhang et al.'s investigation, in long-context settings, the "middle curse" can occur due to the Transformer's self-attention characteristics and limitations in positional encoding resolution, which reduces the utilization of information in the middle portion of the input and can significantly decrease accuracy when key evidence is located there [155]. Indeed, Liu et al. experimentally demonstrated that even when correct evidence is present in long prompts, it may not be effectively utilized, and performance can exhibit a U-shaped pattern depending on evidence location, being high at the beginning and end of the context but substantially lower in the middle, and in some cases even falling below closed-book performance [161].

3.3.2 Evidence-Handling under Ambiguous, Conflicting, or Incomplete Retrieval Results

As discussed in the previous section, the RAG process can still lead to hallucinations and incorrect answers when evidence is missing, irrelevant or erroneous documents are retrieved, or ambiguous or conflicting documents are injected simultaneously. Particularly in manufacturing, such evidence uncertainty can directly result in specification violations, reduced safety, and cost losses, which creates a need for mechanisms that improve the quality of retrieved documents and systematically evaluate, rank, and integrate evidence. Accordingly, various frameworks have been proposed to mitigate hallucinations in RAG, and these approaches can generally be categorized into improving the retrieval quality of relevant evidence at the retrieval stage and strengthening evidence utilization and consistency in the generated responses.

A representative method for improving retrieval accuracy is to rewrite the user's query [162, 163]. Ma et al. proposed the Rewrite–Retrieve–Read framework, which introduces a query rewriting module prior to the retrieval stage [162]. They achieved consistent performance improvements across various benchmarks by leveraging a large LLM for rewriting and response generation while using a small LLM as a trainable rewriter. In this setting, they employed reinforcement learning and used the reader's response performance as a reward signal. Through this learning scheme, they learned rewritten queries that are aligned with the retrieval results. Chan et al. proposed the RQ-RAG framework, which explicitly learns query rewriting, decomposition, and clarification [163]. They reported an average improvement

of 1.9% points on single-hop QA and improved performance on multi-hop QA on Llama2-7B by incrementally generating refined sub-queries for retrieval and then generating responses grounded in the retrieved evidence.

Research has also been conducted to improve retrieval accuracy by redesigning document representations and index structures at the pre-retrieval stage [164–166]. For example, by introducing dynamic chunking to redesign document representation units and applying high-quality Sentence-Transformers embeddings (all-mpnet-base-v2), the approach improved representation fidelity before retrieval [164]. Furthermore, ARAGOG reported that sentence-window retrieval, which indexes text at the sentence level and then combines surrounding context as needed, achieved the largest improvement in retrieval precision compared to Naive RAG [165]. Moreover, the Document Summary Index, which indexes document summaries rather than indexing the full text directly, also demonstrated competitive performance, suggesting that restructuring the representations to be indexed can contribute to improving retrieval quality. LightRAG constructs a graph-based text index via LLM-based entity and relationship extraction instead of indexing document chunks in a flat manner, and it combines dual-level retrieval that adapts to different levels of query specificity and abstraction with incremental updates to support more comprehensive and consistent evidence retrieval for complex queries [166]. Furthermore, on the UltraDomain benchmark, LightRAG reported an overall win rate of 60.0%–84.8% relative to NaiveRAG under LLM-judge evaluation, providing quantitative evidence of its improvement effect.

Research has also been actively conducted to evaluate and refine the retrieved document set during a post-processing stage, aiming to preemptively remove irrelevant evidence that does not contribute to answer generation [167–169]. Yu et al. proposed the RankRAG framework [167], which integrates the roles of a reranker and a generator into a single model by training one LLM to directly judge and rank document relevance in the post-retrieval stage. They demonstrated consistent performance improvements over strong RAG baselines across various knowledge-intensive QA benchmarks, and they empirically showed that removing irrelevant documents via evidence ranking under constrained context budgets can effectively improve response accuracy and factual consistency. Xu et al. proposed GenRT [168], which jointly performs reranking and truncation within a single model from a list-aware perspective that considers relative importance and interactions among documents in the ranked list. This mitigates context fragmentation and error accumulation that can arise when reranking and truncation are handled as separate pipelines. They conducted benchmarks on web search LTR and

open-domain QA-based RAG, and GenRT consistently outperformed various baselines in reranking while reporting the most competitive performance–efficiency trade-off in truncation. Yan et al. proposed Corrective RAG (CRAG), which pre-evaluates the relevance of retrieved documents using a lightweight retrieval evaluator, refines relevant evidence using a decompose–recompose approach, and supplements insufficient evidence via web search [169].

Even when documents irrelevant to the query are filtered out through relevance evaluation, a knowledge conflict may occur if the retrieved evidence contradicts the model’s internal parameterized knowledge, which can prevent the external evidence from being fully utilized. To mitigate such conflicts, recent studies have introduced mechanisms that compare and verify the LLM’s internal knowledge against retrieved evidence and then integrate or selectively adopt it depending on the context, with the aim of reducing errors during the answer generation stage. Jin et al. analyzed the tug-of-war phenomenon observed in RAG. When internal parameterized knowledge and retrieved evidence conflict, models simultaneously exhibit a tendency to actively use additional evidence to update answers and a tendency to overly rely on a specific knowledge source simply to resolve the conflict. To mitigate this, they proposed Conflict-Disentangle Contrastive Decoding (CD^2) [170]. CD^2 increases the relative probability of tokens supported by reliable evidence by contrasting the logits induced by internal knowledge and those induced by evidence-based generation during inference. They applied CD^2 to Llama2-7B and conducted benchmark tests on the NQ and TriviaQA datasets. Without additional parameter updates, they reported recall improvements over the in-context baseline of 2.05% points (NQ-Inco) and 2.70% points (TriviaQA-Inco), demonstrating improved recall of correct answers via external evidence under conflict settings [170]. Wang et al. proposed Astute RAG, which adaptively elicits query-relevant internal knowledge from the LLM. This knowledge is then refined by comparing and organizing it with retrieved external evidence while preserving source information, integrating consistent content, and distinguishing conflicting content, after which candidate answers are generated from each refined knowledge unit and the final answer is selected through confidence evaluation [171]. They conducted benchmarks on multiple QA datasets and reported that Astute RAG generally outperformed existing robustness-enhanced RAG baselines across all benchmarks. Notably, under conditions where retrieval precision was extremely low and RAG instead degraded performance, other RAG variants underperformed the “No RAG” baseline, whereas Astute RAG maintained strong performance without such degradation.

3.4 RAG Applications in Manufacturing IDSS

Vector databases can store unstructured data without schema constraints, which facilitates the integration of heterogeneous and unstructured data and supports scalability, making them highly versatile for RAG [172]. Consequently, recent research in the manufacturing domain has actively focused on building unstructured knowledge assets based on vector databases and applying them to practical process problem solving [173–175]. Auyeskan et al. developed a material selection system for the L-PBF process based on LLMs. They utilized a vector RAG system based on a vector database to select appropriate materials by considering industry-specific usage frequencies [173]. Álvaro and Barreda proposed a RAG system specialized for quality control in ceramic tile manufacturing processes using a vector database [174]. This system structured key information such as defect types, causes, and solutions based on academic literature and defect data. It maximized the accuracy and efficiency of information processing by combining Bi-encoder-based retrieval with Cross-encoder re-ranking techniques. Li et al. constructed a knowledge base for process routes and operation orders in vector database format. Based on this, they proposed a RAG system that suggests macro-level process routes and micro-level detailed operation instructions to enhance the efficiency of aerospace component machining [175].

RAG can be integrated with data storage systems that require a specific query language during the search phase, such as relational databases and graph databases that store knowledge graphs. In such cases, when a user query is input, an LLM-based query generation module, which understands the database schema, generates a search query language capable of extracting the data needed to generate a response to the user query. It then executes this query to retrieve the information. Based on this, attempts are being made to integrate structured data and complex relational data into the RAG pipeline [176, 177].

Jeon et al. proposed ChatCNC, an interactive monitoring framework combining an LLM-based multi-agent with real-time RAG technology, enabling operators to query and analyze real-time CNC operational data [176]. This research demonstrated the effectiveness of the proposed system by achieving a high accuracy of 93.3% for complex production tracking queries without requiring data engineer assistance, through converting text into SQL queries. Bahr et al. proposed the KG-RAG framework, combining knowledge graphs and RAG, to overcome the limitations of existing Failure Mode and Effects Analysis (FMEA) data [177]. Specifically, during the chunking process, they applied a technique that traverses connected nodes on the graph to integrate information in path units. Unlike simple random

segmentation, this approach preserves structural relationships and context between data, significantly improving the precision and recall of contextual retrieval.

By linking diverse external data sources in this manner, it is possible to build a RAG system optimized for the characteristics of the decision-making problem at hand. Since the RAG model generates answers based on information retrieved from external data sources, the construction method of these data sources becomes a key factor determining system performance. Strategic design is essential—specifically, deciding how to store data and what domain knowledge to include. Therefore, Chap. 4 specifically investigates RAG-based knowledge base construction methodologies and practical application cases, focusing on the fields of material selection and process planning.

4 Review of Knowledge Base for RAG

4.1 Knowledge Base

A Knowledge Base (KB) is traditionally defined as a structured collection of facts and rules that supports logical inference. Beyond this conventional concept, the modern definition of a KB has evolved into that of a dynamic and extensible knowledge network model that not only stores structured information but also integrates diverse types of data and supports reasoning as well as complex question answering [178, 179].

Despite these characteristics, KBs and Databases (DBs) are often conflated. The essential distinction between the two lies in their capabilities for semantic processing and inference. A DB is designed to efficiently store and manage large volumes of structured data and supports basic query processing (e.g., SQL queries), but it lacks the ability to infer new facts that are not explicitly stored [180]. In contrast, a KB is designed to represent relationships and contextual semantics among knowledge elements, enabling the derivation of new knowledge. Due to this fundamental difference, research efforts have continued to explore methods for transforming a DB, which functions as a simple repository, into an inference-capable KB. In the manufacturing domain, KBs have gained attention as a key component for intelligent process design [181]. They serve as an essential bridge connecting product design with actual manufacturing and are expected to significantly contribute to shortening product development cycles and maximizing production efficiency through the integration of extensive experiential data and expert knowledge [182].

In the manufacturing domain, KBs have gained attention as a key component for intelligent process design [181]. They serve as an essential bridge connecting product design

with actual manufacturing and are expected to significantly contribute to shortening product development cycles and maximizing production efficiency through the integration of extensive experiential data and expert knowledge [182].

To realize the potential of KBs and successfully implement intelligent manufacturing, it is essential to thoroughly consider the processes of knowledge acquisition, representation, and storage. In particular, how knowledge is systematically represented and stored so that systems can access it efficiently and utilize it for inference is a crucial challenge that determines the success of KB construction. The following section examines representative approaches for knowledge representation and storage that support the effective development and utilization of KBs within this context.

4.1.1 Frame-based Knowledge Base

One of the representative knowledge representation methods used for constructing a KB is the frame-based approach. This method represents knowledge about a specific concept as a structured chunk that contains slots and supports knowledge inheritance through a hierarchical structure [183]. Lingarkar et al. represented the knowledge of a CNC control system as frames organized in a hierarchical manner [184]. Through this structure, lower-level frames inherited general information from higher-level frames, such as “machine constants” and “sampling frequency,” thereby avoiding information redundancy and supporting efficient knowledge sharing among frames. In addition, Lei et al. employed an approach in which the “green indicator frame,” which includes information such as energy consumption and carbon emissions, inherits the technical properties of the “parts (features) frame,” such as size and tolerance [185]. Through this mechanism, they expanded the existing process planning model by extending it with a knowledge system related to environmental attributes.

Frame-based knowledge representation thus offers clear advantages by enabling systematic structuring of complex domain knowledge through hierarchical modeling and facilitating information sharing and extension through inheritance. However, despite these benefits, the frame-based approach must overcome two fundamental limitations in practical implementation. The first is the knowledge acquisition bottleneck. While frames provide a structural skeleton for organizing knowledge, the detailed rules required for complex decision-making must still be manually defined by human experts. This issue was also noted in the study by Lei et al., in which the authors established the structural foundation of the green indicator frame but emphasized that the “green-process decision rule,” which draws the actual conclusions, still requires further research [185]. The second limitation concerns real-time inference performance.

Although Lingarkar et al. conducted their research on real-time systems such as CNC controllers, they pointed out that the backtracking mechanism used by Prolog, which served as the inference engine, inherently requires exponential time [184]. Consequently, frame-based inference may be unsuitable for decision-making systems that are subject to severe time constraints.

4.1.2 Rule-based Knowledge Base

Rule-based representation is a method of representing knowledge using IF-THEN style rules. Knowledge is expressed as clear condition-consequence relationships, deriving specific actions or conclusions when particular conditions are met [186]. This rule-based representation method has the advantage of being intuitive and easy to understand, as it directly encodes expert knowledge, while also offering high inference efficiency and accuracy. Ipek et al. [66] utilized this approach to formalize material characteristics and physical property requirements based on expert empirical knowledge into IF-THEN rules. As illustrated in Fig. 4(c), their system follows an intuitive structure where suitable values are derived as results when specific condition ranges are satisfied. Based on this structure, the researchers successfully built an expert system for material selection by combining forward chaining and backward chaining reasoning.

However, the fundamental limitation persists in rule-based approaches: experts must manually construct all knowledge and inference rules beforehand. This replicates the knowledge acquisition bottleneck previously noted in the frame approach, leaving a persistent challenge: the system's coverage inevitably becomes significantly lower when faced with new, undefined questions. To mitigate this

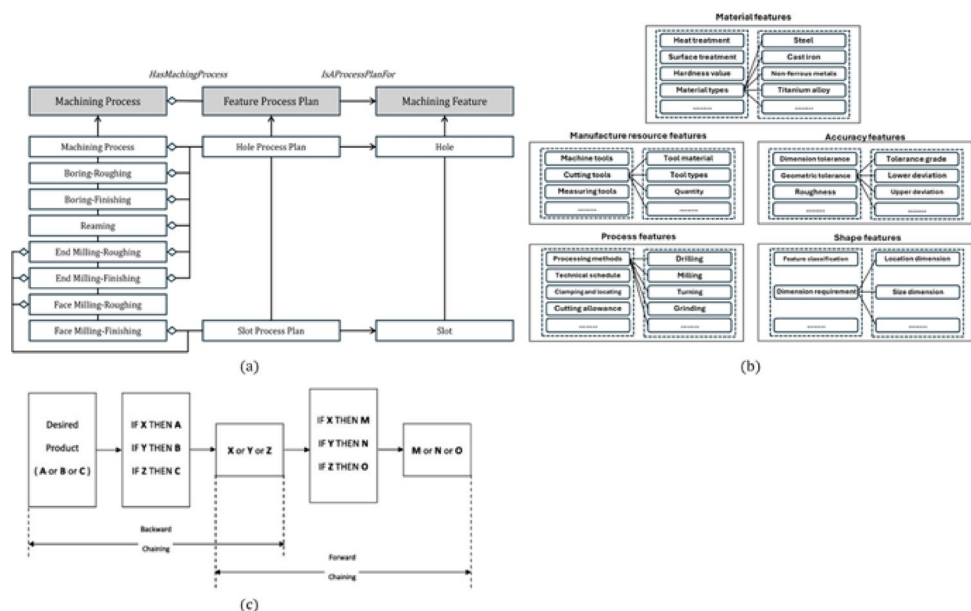
limitation, Thike et al. [187] proposed a rule-case-based hybrid reasoning approach in the field of material failure analysis. This method attempts to solve the low coverage problem by first executing rule-based reasoning (RBR) to find an answer and sequentially executing case-based reasoning (CBR) as a secondary step when the inference fails. However, despite presenting a practical strategy to fill rule gaps through CBR, this RBR-CBR hybrid model has limitations. It remains a complementary device that circumvents the inherent coverage problem of RBR rather than fundamentally solving it.

In conclusion, rule-based representation has the strength of ensuring transparency in the reasoning process and facilitating intuitive understanding by structuring expert knowledge into explicit IF-THEN forms. However, it has the limitation that comprehensively and perfectly capturing all knowledge and exception rules for complex and vast domains manually in advance is extremely difficult in practice, acting as a significant constraint on the system's scalability.

4.1.3 Ontology-based Knowledge Base

Ontology is a knowledge representation approach that explicitly defines the concepts of a specific domain and the relationships among them. The goal of ontology is to classify knowledge into data, relations, and other components and to provide a controlled vocabulary for knowledge representation [188]. As illustrated in Fig. 4(b), ontology clearly defines complex domain knowledge such as material and manufacture resource by visualizing it as a hierarchical structure of classes and properties. This classification scheme has been applied in existing studies. Zhou et al.

Fig. 4 Representative knowledge representation methods., (a) Ontology-based representation Reprinted with permission from ref [190]. Copyright Springer, 2016, (b) Ontology-based structures Adapted With permission ref [189]. Copyright Springer, 2017, (c) Rule-based structures Adapted With permission from ref [66]. 2016, Copyright Elsevier, 2013



used ontology-based representation to formally describe complex domain knowledge and its relationships, including workpieces, cutting tools, and machine tools, and proposed a cutting tool configuration process [189]. They built a system that infers available tools from instances according to pre-specified reasoning rules such as SWRL. Furthermore, as Fig. 4(a) illustrates the organic relationships between machining feature and machining process, Kang et al. proposed an automated approach that selects and sequences machining processes appropriate for a given feature by using a process ontology model together with reasoning rules [190]. In particular, their study highlighted the advantage that, even when process technology advances and process capability changes, it is sufficient to modify only the corresponding instance values in the ontology, without altering the rules themselves, thereby providing flexibility in maintenance and extension. These preceding studies demonstrate that ontology-based knowledge representation can systematically integrate heterogeneous knowledge, thereby enhancing system scalability and reusability and supporting decision-making through explicit reasoning.

However, both studies explicitly pointed out as a limitation that their ontology models are “not yet complete” and that they must be manually extended to include more feature types and process types [189, 190]. This is an inherent limitation arising from the fact that ontology-based reasoning depends on expert knowledge encoded in predefined rules. In other words, the system does not autonomously derive new types of knowledge, and for new reasoning capabilities, experts must necessarily intervene and add new rules, which means that the problem of a “lack of flexibility in knowledge acquisition” still remains.

Table 3 Comparison of key knowledge representation methods: Characteristics, advantages, and limitations

Representation Method	Structure / Characteristics	Advantages	Limitations
Frame-based	• Hierarchical structure with slots • Data inheritance	• Reduces redundancy • Systematic structuring	• Knowledge acquisition bottleneck
Rule-based	• IF-THEN logic • Explicit condition-action pairs	• Intuitive & easy to understand • High inference efficiency	• Manual rule construction required • Low coverage for undefined cases
Ontology-based	• Defined concepts & relationships • Controlled vocabulary	• High interoperability & reusability • Explicit logical reasoning	• Incomplete schema (manual extension) • Rigid pre-defined rules

4.1.4 Implications of Conventional Knowledge Representation Methods

The key characteristics, advantages, and disadvantages of the major knowledge representation methods discussed above are summarized in Table 3. These classical knowledge representation methods, exemplified by frames, rules, and ontologies, are significant in that they explicitly structure experts’ empirical knowledge and thereby provide a robust foundation for systems to perform logical reasoning. However, all three approaches share the problem that the processes of defining and extending knowledge rely excessively on manual expert intervention, which inevitably results in a structural limitation known as the Knowledge Acquisition Bottleneck. Therefore, a new technological paradigm is required that can overcome these constraints by automatically learning knowledge from large-scale unstructured data and enabling its flexible extension. In the following section, we thus discuss in detail Knowledge Graphs as an alternative to these conventional methods and as a next-generation knowledge representation model.

4.2 Knowledge Graph

4.2.1 Definition of Knowledge Graph and Graph RAG

A knowledge graph (KG) is defined as a graph-structured form of knowledge representation in which entities and the relations between them are organized as nodes and edges [191]. Essentially, it inherits the structural advantages of ontology discussed earlier while further maximizing connectivity among data, and this property has recently played a key role in compensating for the limitations of LLMs. In other words, KGs are attracting attention as a core foundation that provides high accuracy and rich context when LLMs access external knowledge.

In related work, Xiong et al. represented Channel Knowledge Map (CKM) data of wireless networks as a KG and conducted experiments to predict channel gain given the positions of transmitters and receivers [192]. Unlike conventional RAG approaches that retrieve text snippets, their study experimentally demonstrated that a KG-RAG approach, which jointly leverages the structural relations encoded in the Knowledge Graph, can generate higher-quality answers. In addition, Bahr et al. built an FMEA KG and combined it with an LLM to perform complex reasoning over FMEA data, thereby developing a question-answering system for failure mode and effects analysis [177]. Their results showed that, compared with the traditional Excel-based approach, usability increased and information retrieval time decreased, and they further confirmed quantitatively that the system achieved high accuracy not only

for analytical and numerical questions but also for semantic questions.

Microsoft introduced the concept of Graph RAG and emphasized that this approach complements the limitations of conventional Vector RAG and enables global sensemaking over large-scale text corpora (the entire collection of original text documents) as well as Query-Focused Summarization (QFS) [193]. They generated global sensemaking questions for which no single clear answer exists and conducted a comparative evaluation against conventional RAG methods. The evaluation results reported that Graph RAG

Table 4 Comparative summary of reported performance metrics in prior LLM-(KG/Graph)-RAG case studies

Task	Baselines	Metrics	Key Results	Refs
Fault Diagnosis QA	LLM-only (GPT-4) vs. Retriever (BM25/embedding/table) vs. KG-LLM	Accuracy	• GPT-4=0.790 • Emb=0.807	[220]
Document QA	Naive RAG vs. KG-RAG	Recall (Top 10 evidence inclusion rate (%))	• Table=0.817 • KG-LLM=0.818 • Naive RAG Recall = 0.500	[28]
Query-based summarization	Vector RAG vs. KG-RAG	Precision, Recall, F1-score Summarization quality	• KG-RAG Recall = 0.650 • QA (multi-hop): KG-RAG↑	[196]
Query-based summarization	Vector RAG vs. KG-RAG	LLM-as-judge win rate: Comprehensiveness, Diversity, Empowerment, Directness	• Summarization: Vector RAG↑ • Comprehensiveness / Diversity (win rate) Graph Rag ↑	[193]
DfAM Q&A	LLM-Only vs. KG-RAG	Answer Recall, Response Accuracy	• Directness (win rate) Vector RAG ↑ • Answer Recall: Non-RAG 62.89 → 71.67 (+8.78%p, KG-RAG, GPT-4o) • Response Accuracy: Non-RAG 62.81 → 70.37 (+7.56%p, KG-RAG, GPT-4o)	[197]

produces superior responses for holistic questions that ask about the overall content of documents Furthermore, this advantage remained consistent when comprehensiveness (the average number of factual claims) and diversity (the average number of factual-claim clusters) were quantitatively assessed based on factual claims extracted from the answers [193]. However, the same study also reported that, in terms of directness included as a comparison criterion, vector-based RAG produces more direct responses [193].

In this literature review, the reported results of the selected case studies are organized and presented in Table 4 using a common framework of Task, Baselines, Metrics, and Key Results. Table 4 is designed to examine, at different evaluation levels, not only final answer performance but also whether supporting evidence is included at the retrieval stage and the performance at the generation stage. However, because datasets and evaluation protocols differ across studies, Table 4 does not claim an absolute ranking across studies rather, it summarizes the directions of improvement and applicability conditions observed within each study's-controlled comparison setting.

In document-based question answering, the proportion of the top 10 retrieved results that contain the answer evidence increases from 0.500 with naive RAG to 0.650 with Graph Rag [28]. This indicates that evidence usable for response generation can be secured more reliably at the retrieval stage, thereby expanding the room for the generative model to construct responses grounded in evidence. In addition, in manufacturing fault-diagnosis question answering, the accuracy of a knowledge-graph-based pipeline (KG-LLM) improves from 0.790 to 0.818 compared with LLM-only (GPT-4) [220], and in additive-manufacturing design-knowledge question answering, KG-based RAG is reported to improve both Answer Recall (62.89→71.67) and Response Accuracy (62.81→70.37) [197]. Accordingly, the cases in Table 4 can serve as quantitative evidence showing that KG/Graph-based integration may yield observable improvements in both (i) evidence acquisition (retrieval stage) and (ii) final response performance (e.g., accuracy and recall).

However, these trends may vary depending on the task and evaluation criteria. For example, in their long-corpus-based query-focused summarization comparison, Microsoft targets global questions for which no single correct answer exists; accordingly, they use an LLM-as-a-judge evaluation protocol, and while Graph Rag is rated relatively better in comprehensiveness and diversity, vector-based RAG is reported to generate more direct responses under the directness criterion [193]. In addition, Han et al. systematically compare RAG and Graph Rag and report that Graph Rag-type methods tend to be relatively advantageous for question answering that requires multi-step reasoning, whereas

for query-based summarization that integrates content across an entire document, vector-based RAG can be superior or comparable under similarity-to-reference metrics (e.g., ROUGE, BERT Score) [196]. In other words, relative strengths may be observed differently depending not only on how summary quality is defined and evaluated but also on the combination of task type and evaluation criteria.

These comparative findings suggest that the effects of KG/Graph-based integration should not be generalized as universal superiority but rather may emerge meaningfully in problem settings where structural relationships among data must be exploited to retrieve evidence and to combine and reason over information. Moreover, the broader body of prior work including this line of discussion—suggests that KGs can contribute to improving system reliability and performance by complementing the structured knowledge and reasoning capabilities that LLMs inherently lack [194, 195]. Building on these technical advantages, the following section examines concrete research cases that illustrate how KGs are structured and applied in the manufacturing domain, where complex data are tightly interwoven.

4.2.2 Knowledge Graph in Manufacturing Domain

In modern manufacturing, data are extremely large in scale and complex in structure, which makes it difficult to achieve efficient management and utilization using only conventional relational databases. As a result, KG technology, which models the knowledge of complex manufacturing systems in terms of entities and relations in order to capture hidden context among data and connect them in a coherent manner, has emerged as a key alternative. Recent studies have applied KGs across various areas such as process planning automation, resource recommendation, and product lifecycle management, providing optimal solutions that satisfy users' complex requirements [110, 198].

First, in decision support systems that select the fundamental resources and materials for manufacturing processes, KGs play an important role [199]. Duan et al. constructed a metal cutting process knowledge graph (MCPKG) by linking cutting tools, materials, features, and other elements based on aero-engine maintenance data [202]. In their work, a personalized PageRank (PPR) algorithm that leverages the connectivity structure of the graph was used to compute the importance of tools according to workpiece features, thereby realizing precise, data-driven recommendations.

Beyond tool selection, there are also studies that aim to support optimal design by clarifying the complex correlations between the microstructure of materials and their properties. Zheng et al. proposed a high-throughput computing framework that combines multiscale simulations (such as CPFEM) with a knowledge graph to analyze the

characteristics of 6XXX series Al-Mg-Si aluminum alloys [237]. In this study, alloy composition, microstructure, and property data were categorized into static and dynamic data, which were then structured into a systematic knowledge network. Based on this, the authors quantitatively analyzed how initial grain orientation and hardening coefficients affect alloy performance, thereby reducing traditional empirical trial-and-error and providing more rational, data-driven support for material design and selection.

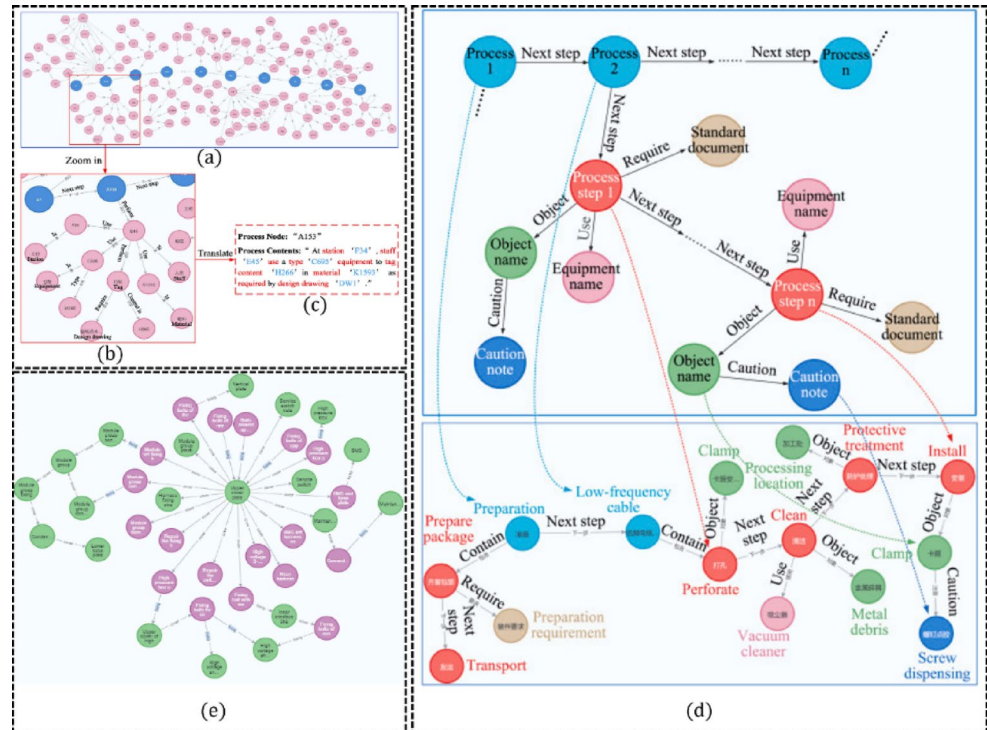
More recently, advanced systems have emerged that go beyond simple recommendation by integrating KGs with LLMs to enable intelligent reasoning and context understanding. Wu et al. successfully implemented a user-oriented hybrid recommendation system for alloy material selection by grouping similar alloys into cluster nodes to enhance structural relatedness and then combining LLM-based data augmentation with embedding analysis [206].

Furthermore, Fan et al. proposed an integrated framework that simultaneously performs material selection and process parameter optimization using LLM agent technology [236]. Based on approximately 300 research articles, they precisely linked core entities such as materials, process parameters, and output characteristics in a triple structure, and defined explicit relations such as “optimal_extrusion_temperature” and “has_property” to form a hierarchical network of relationships. By combining this with LLM agents and RAG techniques, they proposed the AutoMEX framework, which supports multi-hop reasoning. As a result, the system understands the context of user queries and precisely recommends optimal materials and process conditions, thereby enabling even non-experts to achieve high-quality outputs without trial-and-error in an intelligent manufacturing environment.

Once resources and materials have been selected, the next step is process planning, in which the optimal route must be derived under complex constraints. In this field, Xiao et al. used a bidirectional model (GRBE) for extracting knowledge from unstructured historical documents and constructed an assembly process knowledge graph (APKG) that incorporates 5M1E elements such as worker, equipment, and material, as illustrated in Fig. 5(b) [200]. Through the KGMS methodology, they analyzed complex interdependencies and derived the optimal process routes, as depicted in Fig. 5(a). Moreover, they automatically converted the graph structures, which are understood by machines, into natural language work instructions that field operators can immediately comprehend, as shown in Fig. 5(c).

The same research group also carried out work that goes beyond simple retrieval of process knowledge to enable effective reuse by distinguishing the assembly process into an abstract schema layer and a concrete instance layer, as illustrated in Fig. 5(d) [111]. In this study, they combined

Fig. 5 Applications and structures of Knowledge Graphs for manufacturing process planning. (a)–(c) Assembly process planning framework based on GRBE and KGMS Adapted with permission from ref [200] Copyright Springer, 2025, (a) Generated process path, (b) Detailed structure of a process node including 5M1E elements, and (c) Automatic translation to natural language instructions. with permission ref (d) Hierarchical Assembly Process Knowledge Graph (APKG) consisting of schema and instance layers for flexible process reuse Reprinted with permission from ref [111] Copyright Elsevier, 2025, (e) Disassembly information model for battery recycling illustrating physical connection relationships Reprinted with permission from ref [201] Comment. Copyright Springer, 2025



knowledge graph embedding (KGE) techniques with a Siamese Transformer to vectorize deep semantic information across processes, thereby implementing a system capable of accurately recommending similar processes.

There is also a pronounced trend toward enhancing reasoning capabilities by combining KGs with AI algorithms rather than using them solely for knowledge structuring. Zhang et al. proposed the KGPP model, which integrates KGs with reinforcement learning (RL) [96]. By structuring elements such as parts, processes, and devices into sequential and parallel relationships and applying an actor–critic algorithm, they achieved high recommendation accuracy. In a similar context, Zhang et al. developed the KG-MAPP model, which integrates deep learning (DL) [97]. By jointly learning 3D CAD model features and knowledge graph information (TransE), they showed that the model can effectively compensate for process decision-making relationships that cannot be fully captured by geometric information alone.

In addition, Tian et al. proposed case-based process planning through a bottom-up ontology definition [207], and Huang et al. improved the accuracy of process planning for aerospace components by using an mKGMPP multilayer knowledge graph that integrates text and digital models (2D and 3D) and by combining geometric similarity reasoning with knowledge constraints [115].

Beyond individual process steps, the application scope of KGs has extended across the entire product lifecycle, including design, disassembly, and supply chain management

[208–210]. To ensure explainability in the design stage, Jing et al. constructed a manufacturing knowledge graph (MKG) from unstructured design documents and drawings and developed a module that recommends manufacturing knowledge together with designer-trustworthy explanations by using a graph attention network (GAT) [204]. Su et al. achieved advanced knowledge integration in the aerospace domain by modeling heterogeneous data (such as materials and tools) with deep semantic relationships including Usage and Act On [203].

KGs are also utilized in recycling processes at the end of a product's lifecycle. Wu et al. proposed a disassembly information model in which edges represent physical fastening and positional interference relationships among battery components, as depicted in Fig. 5(e), and derived optimal disassembly sequences using a topological sorting algorithm [201]. From a supply chain management perspective, Kosasih et al. structured complex supply chain ecosystems by linking entities such as companies, products, and capabilities, and demonstrated that potential risks in the supply chain that are not explicitly recorded can be handled flexibly by using hidden relations prediction techniques [205].

Taken together, a key technological trend to note here is that KGs are evolving not as standalone technologies but in hybrid forms combined with deep learning and LLMs. The structural explicitness intrinsic to KGs, when combined with the powerful computational capabilities of neural network models, alleviates data sparsity problems and dramatically improves reasoning accuracy [216, 217].

This indicates that KGs serve as a core foundation that provides AI systems with domain-specific knowledge and a reasoning base that they inherently lack. Furthermore, recent studies are expanding their focus beyond model performance alone to also strengthen explainability of results and usability in real industrial environments [218, 219].

Such structured knowledge bases and advanced reasoning mechanisms function as core engines that LLM-based RAG systems must possess in order to move beyond simple text retrieval and generate logical and accurate expert knowledge. In other words, the sophisticated KGs constructed in prior research serve as the source that guarantees the accuracy and reliability of RAG systems.

In conclusion, this body of research suggests that, for the successful introduction and utilization of KGs, a carefully designed graph structure and modeling process that reflects domain-specific characteristics is an essential prerequisite. However, manually performing such complex modeling entails practical limitations. Consequently, new approaches that leverage the superior language capabilities of LLMs to automate and streamline the construction process have recently gained attention. The following section introduces case studies on these LLM-based methodologies for constructing knowledge graphs.

4.3 Knowledge Graph Construction with LLM

Knowledge graphs can function reliably only when domain-specific structures such as schemas and constraints are clearly defined and high-quality data conforming to them are secured [211–213]. However, existing construction methods heavily rely on manual work by domain experts for tasks such as schema design, entity and relationship refinement, and consistency verification, which is costly, time-consuming, and difficult to scale, ultimately causing a knowledge acquisition bottleneck [214, 215]. To mitigate these limitations, recent research focuses on leveraging LLMs' natural language understanding and generation capabilities to automate the construction process, including entity–relationship extraction, normalization, and validation [216–218].

LLMs demonstrate excellent performance in constructing KGs by extracting key entities and relations from fragmented unstructured text. Li et al. proposed a method for extracting structured knowledge triples by using an LLM to analyze the semantic context of unstructured text, such as substation operation data [221]. They collected approximately 1 million words of source text, refined it, and built a knowledge base comprising 2,058 documents. Ultimately, they generated a KG consisting of 4,783 nodes across 4 categories and 9,486 relationships. To evaluate the quality of LLM-based extraction, they randomly selected 400 samples and manually compared the source text with the extracted

results, and evident errors were observed in only 27 samples. These results quantitatively suggest that LLM-based extraction can maintain relatively high consistency even on large-scale unstructured documents. In addition, through research in the medical domain, Yang et al. showed that LLMs can automatically generate KGs by extending existing structures and identifying relationships among entities, and this approach has important implications for the management of complex and dynamic manufacturing data [219].

However, to reliably utilize these extraction results as an actual KG, a quality assurance strategy must be defined. This strategy should specify how extracted entities and relationships map to the domain schema and constraints, and how errors are detected and corrected. In this regard, recent studies generally emphasize the following core components for quality assurance in automatic KG construction: (i) schema and constraint definition, (ii) entity–relationship extraction, (iii) entity normalization and linking (standard naming, synonym resolution, and unit/parameter normalization), and (iv) rule/constraint-based validation and refinement, including LLM-based validation. In practical systems, these components are combined and applied in various sequences depending on the context [222].

Manufacturing KGs often exhibit sparsity and limited observable connections, which can limit the extraction of sufficient structural features from sparse data sources. Accordingly, link prediction and embedding-based methods are used to improve KG completeness by predicting missing associations and supplementing implicit relationships, which is relevant to subsequent retrieval and reasoning tasks [224]. In addition, recent studies have reported graph augmentation approaches that use LLM prompting to generate candidate triples from textual knowledge to enrich KGs and compensate for incomplete or sparse knowledge and missing attributes [225]. However, LLM-based generation may introduce inaccurate triples due to hallucinations, which, once injected, can propagate and compromise KG reliability [225, 226]. In particular, such erroneous triples may later be reused as structural evidence in path-based reasoning and retrieval-based question answering, reinforcing incorrect relations or inducing new errors, thereby allowing the errors to spread across the graph [231]. Zhang et al. illustrated, in the context of automatically constructing KGs with LLMs for an industrial equipment O&M domain, that errors in entity recognition and type classification at the information extraction stage can affect relation judgments and prediction outcomes, ultimately degrading graph quality [227]. For example, they reported cases in which a defect phenomenon entity such as tear was incorrectly identified as the type “cause,” which in turn affected relation judgments and led to inaccurate predictions. They also showed that some entities and relations may not be identified, resulting

in missing knowledge and consequently missing entities/relations in the KG. This suggests that extraction errors in automated construction can manifest in two forms (1) distorted relation judgments due to type misclassification and (2) graph incompleteness due to extraction omissions—thereby reducing the consistency and completeness of the resulting KG [227]. Such hallucination-induced spurious entities or relations can distort the graph structure and lead to performance degradation or bias in downstream analysis and retrieval-/reasoning-based applications [228]. Therefore, LLM-generated candidate triples should be treated as hypotheses and filtered through appropriate validation procedures, such as evidence-grounded checks or additional review, before being incorporated into KGs [26].

To this end, the use of evaluation metrics is required to quantitatively assess the accuracy and reliability of automatically constructed KGs. Mihindukulasooriya et al. compared the set of triples generated by an LLM, G_T , with a reference triple set, E_T , and evaluated how well the generated output aligns with the reference knowledge using exact-match-based metrics such as precision, recall, and the F1-score. Here, precision denotes the proportion of correct triples among the generated triples, recall denotes the proportion of reference triples recovered by the generated output, and the F1-score denotes the harmonic mean of precision and recall. Each metric is defined as follows, with precision, recall, and the F1-score corresponding to Eq. (1), Eq. (2), and Eq. (3), respectively.

$$\text{Precision} = \frac{|G_T \cap E_T|}{|G_T|} \quad (1)$$

$$\text{Recall} = \frac{|G_T \cap E_T|}{|E_T|} \quad (2)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Moreover, this study did not assess performance solely via triple matching. To jointly examine the grounding of generated outputs and compliance with a schema, defined as an allowed set of relation and entity types, it verified at the component level whether the generated subject, relation, and object are supported by the input text and referenced resources. Based on this verification, the study distinguished and measured subject hallucination, relation hallucination, and object hallucination [229].

Such hallucination-based metrics can complement triple-level exact-match metrics such as precision, recall, and F1 by helping identify generation that is not grounded in the input, which is difficult to disentangle using triple matching alone. Concretely, reliability can be further assessed by

reporting the proportion, or hallucination rate, of generated triples that contain relations outside a predefined allowed relation set or contain subjects or objects not grounded in the input text [229, 230].

Beyond simple information extraction, there has also been active research on construction frameworks that ensure data quality while minimizing human intervention. Xu et al. introduced dataset pre-annotation and an LLM-based validator to address the high cost of constructing datasets for intelligent process planning [223]. Using this approach, they established an efficient pipeline for acquiring high-quality datasets while minimizing human involvement, and they reported quantitative benefits, including a 48.58% reduction in construction time, a 46.44% reduction in cost, and a 1.96% improvement in F1 score compared to existing deep learning-based construction methods. These results highlight the practical necessity of validator-based quality control in LLM-based KG construction and provide quantitative evidence that integrating quality assurance elements such as pre-annotation and LLM-based validation can enhance both construction efficiency and output quality in real-world tasks. Furthermore, they suggested that the level of automation could be increased by integrating GAN-based adversarial learning in future work to improve how closely automatically generated annotations approximate real-world data.

In a similar vein, Zhu et al. reported a zero-shot temporal KG completion approach using GANs, in which relation embeddings are generated from textual relation descriptions, validated via a discriminator, and then used to predict missing facts for unseen relations, such as event quadruples in temporal KGs [226]. This approach can help reduce the burden of manual supplementation by supporting KG completion for unobserved relations however, additional safeguards are required to mitigate structural inconsistencies and training stability issues, and thus such GAN-based approaches can increase the level of automation but do not by themselves guarantee fully automated KG construction. Ma et al. proposed the FDRKG-LLM framework for Industry 5.0 environments [220]. This study combined LLMs with normalization and refinement procedures to mitigate issues such as hallucinations, timeliness limitations, and low transparency that can arise when applying LLMs alone to manufacturing-site diagnostics. Accordingly, FDRKG-LLM is designed to ensure graph consistency and reliability by combining LLM-based entity and intent extraction, entity and relationship normalization and linking, KG-based diagnostic subgraph retrieval, and LLM-based path refinement. In an evaluation using 255 industrial equipment failure record cases, the proposed method achieved an accuracy of 0.818, improving over both a standalone LLM (GPT-4: 0.790) and standard RAG (0.807).

These findings indicate that LLMs are expanding beyond simple text-processing tools to become core components integrated with KG-based structuring for leveraging industrial domain knowledge. Specifically, to mitigate issues in manufacturing-site diagnostics such as hallucinations, timeliness limitations, and low transparency, recent approaches integrate LLMs into quality assurance procedures including entity and relationship normalization and linking and path refinement, and perform reasoning based on evidence retrieved through KG search. Furthermore, by executing core steps through prompt-based methods without task-specific parameter updates, these approaches suggest a direction for reducing the manual effort previously required to structure and utilize domain knowledge. Meanwhile, although LLM-enabled automatic KG construction and KG-based question answering are promising in that they can rapidly structure and utilize knowledge from unstructured documents, structural constraints related to reliability must also be considered for industrial deployment [232]. For example, if schema, constraints, and normalization criteria are not sufficiently fixed based on exogenous references, initial biases or errors can accumulate and be amplified through iterative processes in which generated outputs are fed back into the construction criteria, thereby weakening graph consistency [233]. In addition, because manufacturing decision-making is sensitive to physical constraints as well as safety and quality requirements, a configuration that adds verification steps such as validator, annotation, and post-processing, rather than directly using LLM/KG outputs, is widely adopted, particularly in manufacturing applications where high accuracy is critical [234, 235].

In this context, the automatic construction-and-use pipeline should be designed around fixing schema and constraints grounded in exogenous criteria and conducting stepwise quality checks [230]. As needed, it can also incorporate complementary linkage between KG-grounded evidence provision and constraint checking based on physics models, such as analysis and simulation, or on experimental and sensor data [234, 235]. Such measures can help reduce error accumulation and deployment uncertainty in iterative construction processes and, by supporting reliability assurance including checks of physical feasibility and safety, can improve the practicality of industrial adoption.

5 Discussions

As summarized in Table 5, research on knowledge graphs in the manufacturing domain has been expanding the role and scope of their application in multiple directions to meet the diverse requirements of industrial practice. When these studies are analyzed in terms of their technical characteristics

and primary application objectives, several key trends can be identified.

First, many studies use knowledge graphs as a core tool for knowledge structuring and integration to systematically manage fragmented manufacturing data. Manufacturing environments have long faced the chronic problem that heterogeneous data, such as text, 2D/3D drawings, and sensor data, coexist in different formats, making integrated management difficult to achieve using conventional databases. Researchers are therefore focusing on defining domain-specific schemas to extract core entities and relationships from unstructured data and connecting them into coherent networks [202–205]. This approach is particularly important because it transforms tacit knowledge within the manufacturing domain into explicit knowledge, systematizing it as digital assets that systems can understand and utilize.

Furthermore, there is a clear trend toward leveraging accumulated knowledge bases to perform intelligent inference and optimize decision-making. Rather than merely retrieving stored knowledge, these approaches combine AI algorithms such as Deep Learning and Reinforcement Learning with knowledge graphs to address data sparsity and explore optimal solutions. This hybrid approach functions as a core engine that efficiently solves challenging decision-making problems, such as process route optimization under complex constraints, sequence determination, and similar case matching, by learning latent patterns and implicit relationships within knowledge graphs [96, 97, 111, 200, 201, 237].

Recently, a new approach has rapidly emerged that goes a step further: combining KB with LLM to automate the construction process and maximize the utility of KB. While traditional KB construction relied on manual expert labor, resulting in a high-cost, low-efficiency structure, recent research leverages the exceptional language understanding and generation capabilities of LLMs to address this issue. Active efforts are underway to automate the entire construction process, for example by using LLM for data augmentation to compensate for insufficient data and by automatically parsing multimodal information to convert it into structured knowledge [115, 206, 220, 223, 236]. This trend suggests that knowledge graphs are evolving beyond simple analytical tools to become foundational technologies for explainable manufacturing AI, capable of understanding users' natural language queries and generating logically grounded, explainable responses.

In conclusion, manufacturing KB technologies are being utilized both as structuring tools for the physical integration of data and as optimization engines powered by intelligent algorithms. Recently, their role has expanded into the realm of generation and automation through convergence with LLM. In particular, throughout this process, KB has

Table 5 Overview of manufacturing knowledge graph research trends by major technological paradigms

Key Paradigm	Key Characteristics	Core Technologies & Algorithms	Representative Studies (Ref.)
Knowledge Structuring (Structure & Integration)	• Integration of heterogeneous data (documents, drawings)	• Knowledge Graph Construction (Entity/Relation Extraction)	• Jing et al. [204]: Car body knowledge structuring
	• Definition of knowledge centered on Entities and Relations	• Graph Database (e.g., Neo4j, RDF Store)	• Duan et al. [202]: Cutting tool KG construction • Su et al. [203], Tian et al. [207]: Aerospace & Machining Ontology • Kosasih et al. [205]: Supply chain data integration
Intelligent Inference (Hybrid Inference)	• Structured conversion of unstructured data		
	• Integration of KG + AI (Deep Learning/RL)	• Graph Embedding (TransE)	• Zhang et al. [96]: KG + Reinforcement Learning (RL)
	• Addressing data sparsity and learning implicit relations	• GNN / GAT (Graph Attention Networks)	• Zhang et al. [97]: KG + Deep Learning (DL)
	• Optimal path search and importance calculation	• Reinforcement Learning (Actor-Critic)	• Xiao et al. [111]: Embedding similarity & Path generation
Generation & Automation using LLM (LLM Integration)	• Convergence of KG + LLM (Graph RAG)	• LLM (Prompt Engineering)	• Wu et al. [206]: LLM-enhanced data & recommendation
	• Automated extraction and construction from unstructured data	• Generative AI	• Huang et al. [115]: LLM-based text/image parsing
		• Multi-modal Parsing	• Xu et al. [223], Yang et al. [219]: Construction automation
	• Natural language QA and Explainability		• Ma et al. [220], Li et al. [221]: Entity extraction & QA

gained renewed importance as a core knowledge base for RAG (Retrieval-Augmented Generation), which helps suppress LLM hallucinations and supports accurate retrieval of specialized knowledge and logical reasoning. This technological convergence is expected to serve as a powerful driver for building autonomous manufacturing environments in which future manufacturing systems can flexibly handle complex and variable situations while minimizing human intervention.

However, despite these technological advances, studies reported to date have predominantly taken the form of case studies specific to individual companies, lines, or processes. It is difficult to conclude that a representative integrated graph schema encompassing diverse manufacturing domains has become established as a de facto standard. Previous studies have repeatedly pointed out the absence of a generally agreed-upon manufacturing ontology or reference architecture that comprehensively covers materials, products, resources, processes, and the entire manufacturing system [238, 239]. Attempts have been made to design knowledge graphs by referencing core manufacturing domain standards such as DIN 8580 and STEP-NC [240]. However, most of the implementations presented are optimized for specific process groups (particularly machining) or limited application scopes, thereby introducing structural and scope constraints that hinder common reuse across different research and industrial settings.

To address this, two approaches can be considered. First, a top-down approach could emerge: establishing a high-level integrated schema and schema design guidelines that maintain consistency with existing standards while enabling broad applicability, allowing common reference across diverse manufacturing domains. Second, considering the reality that it is difficult to integrate the vast amount of manufacturing data and knowledge already fragmented and accumulated at the enterprise, line, and process levels all at once, a bottom-up approach will also serve as a crucial complementary approach. This involves systematically developing procedural methodologies to build, validate, and operate reliable knowledge graphs for individual manufacturing domains in stages, then progressively aligning and integrating them to ultimately form a knowledge graph usable across the entire manufacturing sector in the long term. If concrete design principles and empirical research on these top-down and bottom-up approaches are systematically established and empirically validated, they could serve as a core foundation for the widespread and sustainable adoption of LLM–KG–RAG technology within the manufacturing sector.

6 Conclusions

This review aimed to clarify the form and content requirements of knowledge bases suitable for use in RAG-based DSS for materials selection and process planning. To this end, prior studies were systematically collected and categorized with a focus on research methodologies for decision support, CAPP studies, and knowledge construction techniques based on knowledge graphs and LLMs, and the findings were reorganized from a RAG perspective.

Through an examination of existing DSS models, it was confirmed that such models are inherently valid only within the distribution of their training data, and that they face fundamental limitations in applicability and reliability when applied beyond the specific application scenarios or process conditions under which the data were obtained. Although improving generalizability requires large-scale, high-quality datasets that encompass diverse variables and conditions, the materials and process domains frequently experience data sparsity due to the high cost of experiments and simulations, which serves as a major bottleneck for dataset acquisition. In addition, the numerical-data-centric nature of conventional approaches reveals structural limitations that make it difficult to directly incorporate or integrate expert heuristics or qualitative design requirements, both of which are hard to quantify yet essential for decision making.

The classification of CAPP-related studies into three task levels (feature-level, operation-level, system-level) and five technical categories (rule-based, optimization, knowledge-driven reasoning, geometric deep learning, reinforcement learning) indicates that process planning problems form a heterogeneous research landscape in which traditional rule-based methods, optimization techniques, knowledge-based reasoning approaches, and recent deep learning and reinforcement learning methods coexist. Table 5 provides an overview of the techniques predominantly used for each task and serves as a useful guide for understanding the overall structure and technological trends of recent CAPP research. This overview is meaningful because it offers foundational insight into how different levels of knowledge should be combined with appropriate algorithmic approaches when designing future RAG-based DSS.

A review of research on knowledge graphs shows that successful adoption and utilization of KGs in manufacturing domains require careful graph design and modeling processes that reflect domain-specific characteristics. However, fully manual modeling of complex schemas and relational structures imposes substantial constraints in terms of cost, time, and required expertise. Consequently, approaches that leverage the strong language understanding and generation capabilities of LLMs to automatically parse unstructured text and multimodal data and extract and refine entities and

relations are being widely adopted as a means to automate and streamline KG construction. This trend suggests that LLMs are emerging not merely as text processing tools but as core technologies capable of alleviating the high-cost and expert-dependent challenges of knowledge graph development.

In summary, the convergence of LLMs and KGs holds the potential to simultaneously mitigate the generalization limits inherent in data-driven models and the bottlenecks of knowledge acquisition. This convergence can fundamentally transform the paradigm of knowledge base construction and management in manufacturing domains and accelerate the realization of RAG-based intelligent decision support systems. From this perspective, this review presents research directions for gradually completing a knowledge base for materials selection and process planning, including an upper-level integrated schema and schema design guidelines that can cover diverse manufacturing domains, a bottom-up methodology for stepwise construction, verification, and operation of trustworthy domain-specific knowledge graphs, and the need for quantitative evaluation, verification, and management frameworks for LLM–KG–RAG pipelines.

Acknowledgements This work was supported by the Technology Innovation Program - Industry Technology Alchemist Project (20025702, Development of smart manufacturing multiverse platform based on multisensory fusion avatar and interactive AI) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). This work was also supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program (No. RS-2024-00423300) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was also supported by the Chung-Ang University Graduate Research Scholarship in 2025.

Funding Open Access funding enabled and organized by Chung-Ang University

Declarations

Conflict of interest 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.

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not

included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Giachetti, R. E. (1998). A decision support system for material and manufacturing process selection. *Journal of Intelligent Manufacturing*, 9(3), 265–276.
- Afteni, C., & Frumușanu, G. (2017). A review on optimization of manufacturing process performance. *International Journal of Modeling and Optimization*, 7(3), 139–144.
- Lukic, D., Milosevic, M., Antic, A., & Ficko, M. (2017). Multi-criteria selection of manufacturing processes in the conceptual process planning. *Advances in Production Engineering & Management*, 12(2).
- Kleban, S. D. (1998, July). Concurrent materials and process selection in conceptual design. In *Proceedings Artificial Intelligence and Manufacturing Research Planning Workshop* (pp. 98–102). AAAI Press.
- Albiñana, J. C., & Vila, C. (2012). A framework for concurrent material and process selection during conceptual product design stages. *Materials & Design*, 41, 433–446.
- Zakeri, S., Chatterjee, P., Konstantas, D., & Ecer, F. (2023). A decision analysis model for material selection using simple ranking process. *Scientific Reports*, 13(1), 8631.
- Xu, H. M., Yuan, M. H., & Li, D. B. (2009). A novel process planning schema based on process knowledge customization. *The International Journal of Advanced Manufacturing Technology*, 44(1), 161–172.
- Sharma, V., & Maan, V. (2025). From cost to sustainability: Evolving trends in MCDM methods for material selection and future perspectives. *Critical Reviews in Solid State and Materials Sciences*, 1–28.
- Jamwal, A., Agrawal, R., Sharma, M., & Kumar, V. (2021). Review on multi-criteria decision analysis in sustainable manufacturing decision making. *International Journal of Sustainable Engineering*, 14(3), 202–225.
- Rahim, A. A., Musa, S. N., Ramesh, S., & Lim, M. K. (2020). A systematic review on material selection methods. *Proceedings of the Institution of Mechanical Engineers Part L: Journal of Materials: Design and Applications*, 234(7), 1032–1059.
- Ghaleb, A. M., Kaid, H., Alsamhan, A., Mian, S. H., & Hidri, L. (2020). Assessment and comparison of various MCDM approaches in the selection of manufacturing process. *Advances in Materials Science and Engineering*, 2020(1), 4039253.
- Zhou, T., Tang, D., Zhu, H., & Wang, L. (2020). Reinforcement learning with composite rewards for production scheduling in a smart factory. *Ieee Access : Practical Innovations, Open Solutions*, 9, 752–766.
- O'donovan, P., Leahy, K., Bruton, K., & O'Sullivan, D. T. (2015). Big data in manufacturing: a systematic mapping study. *Journal of Big Data*, 2(1), 20.
- Keen, P. G. (1980, June). Decision support systems: a research perspective. In *Decision support systems: Issues and challenges: Proceedings of an international task force meeting* (pp. 23–44).
- Fernando, J. G., & Baldeovar, M. (2022). Decision support system: Overview, different types and elements. *Technoarete Trans Intell Data Min Knowl Discov (TTIDMKD)*, 2, 13–18.
- Sikselyte-Butkiene, I., Streimikiene, D., Balezentis, T., & Skulskis, V. (2021). A systematic literature review of multi-criteria decision-making methods for sustainable selection of insulation materials in buildings. *Sustainability*, 13(2), 737.
- Kiritis, D. (1995). A review of knowledge-based expert systems for process planning. Methods and problems. *The International Journal of Advanced Manufacturing Technology*, 10(4), 240–262.
- Leo Kumar, S. P. (2019). Knowledge-based expert system in manufacturing planning: state-of-the-art review. *International Journal of Production Research*, 57(15–16), 4766–4790.
- Xiao, Y., Zheng, S., Shi, J., Du, X., & Hong, J. (2023). Knowledge graph-based manufacturing process planning: A state-of-the-art review. *Journal of Manufacturing Systems*, 70, 417–435.
- Mumali, F., & Kałkowska, J. (2024). Intelligent support in manufacturing process selection based on artificial neural networks, fuzzy logic, and genetic algorithms: Current state and future perspectives. *Computers & Industrial Engineering*, 193, 110272.
- Kumar, S. L. (2017). State of the art-intense review on artificial intelligence systems application in process planning and manufacturing. *Engineering Applications of Artificial Intelligence*, 65, 294–329.
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., & Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5), 1–72.
- Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., & Tang, Y. D. (2024). The application of large language models in medicine: A scoping review. *Iscience*, 27(5).
- Li, J., Gao, Y., Yang, Y., Bai, Y., Zhou, X., Li, Y., & Huang, H. (2025). *Fundamental capabilities and applications of large language models: A survey*. ACM Computing Surveys.
- Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports*, 15(1), 13755.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55.
- Emovon, I., & Ogheniyerovwho, O. S. (2020). Application of MCDM method in material selection for optimal design: A review. *Results in Materials*, 7, 100115.
- Knollmeyer, S., Caymazer, O., & Grossmann, D. (2025). Document GraphRAG: Knowledge Graph Enhanced Retrieval Augmented Generation for Document Question Answering Within the Manufacturing Domain. *Electronics*, 14(11), 2102.
- Hashemi Sohi, F. S., Mansour, S., & Dehghanian, A. (2022). Multi-objective optimization for selecting sustainable materials with simultaneous consideration of several components in a product. *International Journal of Sustainable Engineering*, 15(1), 107–121.
- Ljungberg, L. Y. (2007). Materials selection and design for development of sustainable products. *Materials & Design*, 28(2), 466–479.
- Ferro, P., Bonollo, F., & Cruz, S. A. (2021). Product design from an environmental and critical raw materials perspective. *International Journal of Sustainable Engineering*, 14(1), 1–11.
- Gopalakrishnan, B., & Pandiarajan, V. (1991). Materials and manufacturing processes selection system for product designs in concurrent engineering. *Journal of Materials Processing Technology*, 28(1–2), 93–103.
- Ashby, M. F., Brechet, Y. J. M., Cebon, D., & Salvo, L. (2004). Selection strategies for materials and processes. *Materials & Design*, 25(1), 51–67.
- Van Kesteren, I., De Bruijn, S., & Stappers, P. J. (2008). Evaluation of materials selection activities in user-centred design projects. *Journal of Engineering Design*, 19(5), 417–429.

35. Guisbiers, G., & Wautelet, M. (2007). Materials selection for micro-electromechanical systems. *Materials & design*, 28(1), 246–248.
36. de Oliveira, M. C. L., Ett, G., & Antunes, R. A. (2012). Materials selection for bipolar plates for polymer electrolyte membrane fuel cells using the Ashby approach. *Journal of power sources*, 206, 3–13.
37. Risaliti, E., Del Pero, F., Arcidiacono, G., & Citti, P. (2025). Optimizing lightweight material selection in automotive engineering: a hybrid methodology incorporating Ashby's method and VIKOR analysis. *Machines*, 13(1), 63.
38. Taherdoost, H., & Madanchian, M. (2023). Multi-criteria decision making (MCDM) methods and concepts. *Encyclopedia*, 3(1), 77–87.
39. Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research*, 48(1), 9–26.
40. Khalid, R., Jayamani, E., Soon, K., PVS, H. P., Jeyanthi, S., & Sankar, R. R. (2022). Selection of green composite materials for orthopedic prosthesis using analytical hierarchy process. *Materials Today: Proceedings*, 62, 6857–6863.
41. Amer, A. E., Rahmani, K., & Lebedev, V. A. (2020, August). Using the Analytic Hierarchy Process (AHP) method for selection of phase change materials for solar energy storage applications. In *Journal of Physics: Conference Series* (Vol. 1614, No. 1, p. 012022). IOP Publishing.
42. Daniyan, I., Mpofu, K., & Ramatsetse, B. (2020). The use of Analytical Hierarchy Process (AHP) decision model for materials and assembly method selection during railcar development. *Cogent Engineering*, 7(1), 1833433.
43. Mohamed, N., Mazen, S., & Helmy, W. (2022). E-ahp: An enhanced analytical hierarchy process algorithm for prioritizing large software requirements numbers. *International Journal of Advanced Computer Science and Applications*, 13(7).
44. Madanchian, M., & Taherdoost, H. (2023). A comprehensive guide to the TOPSIS method for multi-criteria decision making. *Madanchian M, Taherdoost H. A comprehensive guide to the TOPSIS method for multi-criteria decision making. Sustainable Social Development*, 1(1), 2220.
45. Mayyas, A., Omar, M. A., & Hayajneh, M. T. (2016). Eco-material selection using fuzzy TOPSIS method. *International Journal of Sustainable Engineering*, 9(5), 292–304.
46. Okokpujie, I. P., Okonkwo, U. C., Bolu, C. A., Ohunakin, O. S., Agboola, M. G., & Atayero, A. A. (2020). Implementation of multi-criteria decision method for selection of suitable material for development of horizontal wind turbine blade for sustainable energy generation. *Heliyon*, 6(1).
47. Yang, W. C., Chon, S. H., Choe, C. M., & Yang, J. Y. (2021). Materials selection method using TOPSIS with some popular normalization methods. *Engineering Research Express*, 3(1), 015020.
48. Taherdoost, H., & Madanchian, M. (2023). VIKOR method—an effective compromising ranking technique for decision making. *Macro Management & Public Policies*, 5(2).
49. Opricovic, S., & Tzeng, G. H. (2007). Extended VIKOR method in comparison with outranking methods. *European journal of operational research*, 178(2), 514–529.
50. Zulkafli, M. N. A., Rasid, M. A. H., & Nafiz, D. M. (2025). Optimizing material selection for brushed DC motor components using the VIKOR method: a comprehensive performance evaluation. *Multiscale and Multidisciplinary Modeling Experiments and Design*, 8(2), 129.
51. Dev, S., Aherwar, A., & Patnaik, A. (2020). Material selection for automotive piston component using entropy-VIKOR method. *Silicon*, 12(1), 155–169.
52. Lin, T. Y., Hung, K. C., Jablonsky, J., & Lin, K. P. (2025). An Enhanced VIKOR and Its Revisit for the Manufacturing Process Application. *Computers Materials & Continua*, 83(2).
53. Figueira, J. R., Greco, S., Roy, B., & Słowiński, R. (2013). An overview of ELECTRE methods and their recent extensions. *Journal of Multi-Criteria Decision Analysis*, 20(1–2), 61–85.
54. Chen, Z. S., Hu, Y. J., Ma, Z., Yang, H. H., Shang, L. L., & Skibniewski, M. J. (2024). Selecting optimal honeycomb structural materials for electronics clean rooms using a Bayesian best-worst method and ELECTRE III. *Journal of Building Engineering*, 85, 108703.
55. Kirişci, M., Demir, I., & Şimşek, N. (2022). Fermatean fuzzy ELECTRE multi-criteria group decision-making and most suitable biomedical material selection. *Artificial Intelligence in Medicine*, 127, 102278.
56. Brans, J. P., & Vincke, P. (1985). Note—A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making). *Management science*, 31(6), 647–656.
57. Mahajan, A., Binaz, V., Singh, I., & Arora, N. (2022). Selection of natural fiber for sustainable composites using hybrid multi criteria decision making techniques. *Composites Part C: Open Access*, 7, 100224.
58. Patnaik, P. K., Mishra, S. K., & Ashish, A. T. (2020, March). Ranking of fiber reinforced composite materials using PSI and PROMETHEE method. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1–5). IEEE.
59. Mousavi-Nasab, S. H., & Sotoudeh-Anvari, A. (2018). A new multi-criteria decision making approach for sustainable material selection problem: A critical study on rank reversal problem. *Journal of Cleaner Production*, 182, 466–484.
60. Wang, Y. M., & Luo, Y. (2009). On rank reversal in decision analysis. *Mathematical and Computer Modelling*, 49(5–6), 1221–1229.
61. Zakeri, S., Konstantas, D., Chatterjee, P., & Zavadskas, E. K. (2025). Soft cluster-rectangle method for eliciting criteria weights in multi-criteria decision-making. *Scientific Reports*, 15(1), 284.
62. Maliene, V., Dixon-Gough, R., & Malys, N. (2018). Dispersion of relative importance values contributes to the ranking uncertainty: Sensitivity analysis of Multiple Criteria Decision-Making methods. *Applied Soft Computing*, 67, 286–298.
63. Liou, J. J., & Tzeng, G. H. (2012). Comments on Multiple criteria decision making (MCDM) methods in economics: an overview. *Technological and Economic Development of Economy*, 18(4), 672–695.
64. Liao, S. H. (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert systems with applications*, 28(1), 93–103.
65. Yang, X., & Zhu, C. (2024). Industrial expert systems review: A comprehensive analysis of typical applications. *Ieee Access : Practical Innovations, Open Solutions*, 12, 88558–88584.
66. İpek, M., Selvi, İ. H., Findik, F., Torkul, O., & Cedimoğlu, I. H. (2013). An expert system based material selection approach to manufacturing. *Materials & Design*, 47, 331–340.
67. Ahmed, S. N., Bhargava, M., & KV, S. S. (2023). Material selection using knowledge-based expert system for racing bicycle forks. *Intelligent Systems with Applications*, 19, 200257.
68. Ouellet, V., Mocq, J., Adlouni, E., S. E., & Krause, S. (2021). Improve performance and robustness of knowledge-based FUZZY LOGIC habitat models. *Environmental Modelling & Software*, 144, 105138.
69. Skrzek, K., Mazgajczyk, E., & Dybała, B. (2025). Application of Fuzzy Logic-Based Expert Advisory Systems in Optimizing the Decision-Making Process for Material Selection in Additive Manufacturing. *Materials*, 18(2), 324.

70. Berman, A. F., Maltugueva, G. S., & Yurin, A. Y. (2018). Application of case-based reasoning and multi-criteria decision-making methods for material selection in petrochemistry. *Proceedings of the Institution of Mechanical Engineers Part L: Journal of Materials: Design and Applications*, 232(3), 204–212.
71. Goel, V., & Chen, J. (1996). Application of expert network for material selection in engineering design. *Computers in industry*, 30(2), 87–101.
72. Pollice, R., dos Passos Gomes, G., Aldeghi, M., Hickman, R. J., Krenn, M., Lavigne, C., & Aspuru-Guzik, A. (2021). Data-driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4), 849–860.
73. Guo, K., Yang, Z., Yu, C. H., & Buehler, M. J. (2021). Artificial intelligence and machine learning in design of mechanical materials. *Materials Horizons*, 8(4), 1153–1172.
74. Jang, S., Goh, C. H., & Choi, H. J. (2015). Multiphase design exploration method for lightweight structural design: Example of vehicle mounted antenna-supporting structure. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 2(3), 281–287.
75. Bauer, J., Ji, Z., Rupp, F., Caydamli, Y., Heudorfer, K., Gompf, B., & Middendorf, P. (2025). Optimizing transparent fiber-reinforced polymer composites: A machine learning approach to material selection. *Composites Part B: Engineering*, 112799.
76. Lee, Y., Han, S., Jang, S., Kim, W., Choi, H. J., & Choi, S. K. (2018). Multidisciplinary materials and geometry optimization of superheater tubes for advanced ultra-supercritical power boilers. *Journal of Mechanical Science and Technology*, 32(7), 3359–3369.
77. Hao, C., Sui, Y., Yuan, Y., Li, P., Jin, H., & Jiang, A. (2025). Composition optimization design and high temperature mechanical properties of cast heat-resistant aluminum alloy via machine learning. *Materials & Design*, 250, 113587.
78. Li, Z., Li, S., & Birbilis, N. J. M. T. C. (2024). A machine learning-driven framework for the property prediction and generative design of multiple principal element alloys. *Materials Today Communications*, 38, 107940.
79. Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., & Xie, T. (2025). A generative model for inorganic materials design. *Nature*, 639(8055), 624–632.
80. Zhou, Z., Shang, Y., Liu, X., & Yang, Y. (2023). A generative deep learning framework for inverse design of compositionally complex bulk metallic glasses. *npj Computational Materials*, 9(1), 15.
81. Kauwe, S. K., Graser, J., Murdock, R., & Sparks, T. D. (2020). Can machine learning find extraordinary materials? *Computational Materials Science*, 174, 109498.
82. Li, K., Rubungo, A. N., Lei, X., Persaud, D., Choudhary, K., DeCost, B., & Hattrick-Simpers, J. (2025). Probing out-of-distribution generalization in machine learning for materials. *Communications Materials*, 6(1), 9.
83. Altling, L., & Zhang, H. (1989). Computer aided process planning: the state-of-the-art survey. *The International Journal of Production Research*, 27(4), 553–585.
84. Tan, W., & Khoshnevis, B. (2000). Integration of process planning and scheduling—a review. *Journal of Intelligent Manufacturing*, 11(1), 51–63.
85. Xu, X., Wang, L., & Newman, S. T. (2011). Computer-aided process planning—A critical review of recent developments and future trends. *International Journal of Computer Integrated Manufacturing*, 24(1), 1–31.
86. Babic, B., Nesic, N., & Miljkovic, Z. (2008). A review of automated feature recognition with rule-based pattern recognition. *Computers in industry*, 59(4), 321–337.
87. Zhang, H., Zhang, S., Zhang, Y., Liang, J., & Wang, Z. (2022). Machining feature recognition based on a novel multi-task deep learning network. *Robotics and Computer-Integrated Manufacturing*, 77, 102369.
88. Wang, P., Yang, W. A., & You, Y. (2023). A hybrid learning framework for manufacturing feature recognition using graph neural networks. *Journal of Manufacturing Processes*, 85, 387–404.
89. Ma, L., & Yang, J. (2024). Adaptive recognition of machining features in sheet metal parts based on a graph class-incremental learning strategy. *Scientific reports*, 14(1), 10656.
90. Lee, J., Lee, H., & Mun, D. (2022). 3D convolutional neural network for machining feature recognition with gradient-based visual explanations from 3D CAD models. *Scientific Reports*, 12(1), 14864.
91. Yeo, C., Kim, B. C., Cheon, S., Lee, J., & Mun, D. (2021). Machining feature recognition based on deep neural networks to support tight integration with 3D CAD systems. *Scientific reports*, 11(1), 22147.
92. Ning, F., Shi, Y., Cai, M., & Xu, W. (2023). Part machining feature recognition based on a deep learning method. *Journal of Intelligent Manufacturing*, 34(2), 809–821.
93. Al-wswasi, M., & Ivanov, A. (2019). A novel and smart interactive feature recognition system for rotational parts using a STEP file. *The International Journal of Advanced Manufacturing Technology*, 104(1), 261–284.
94. Peddireddy, D., Fu, X., Wang, H., Joung, B. G., Aggarwal, V., Sutherland, J. W., & Jun, M. B. G. (2020). Deep learning based approach for identifying conventional machining processes from CAD data. *Procedia Manufacturing*, 48, 915–925.
95. Liu, X., Wang, Z., Melkote, S. N., & Rosen, D. W. (2025). Manufacturing process identification from 3D point cloud models using semantic segmentation. *Journal of Manufacturing Systems*, 82, 858–873.
96. Zhang, L., Wu, H., Chen, Y., Wang, X., & Peng, Y. (2025). A knowledge-guided process planning approach with reinforcement learning. *Journal of Engineering Design*, 36(7–9), 1527–1550.
97. Zhang, Y., Zhang, S., Huang, R., Huang, B., Liang, J., Zhang, H., & Wang, Z. (2022). Combining deep learning with knowledge graph for macro process planning. *Computers in Industry*, 140, 103668.
98. Chung, C., Yang, C. W., & Chang, H. M. (2025). Development of convolutional neural network based autonomous milling process planning system for 2.5 D parts. *The International Journal of Advanced Manufacturing Technology*, 141(7), 4489–4504.
99. Han, Z., Huang, R., Huang, B., Jiang, J., & Li, X. (2023). Data-driven and knowledge-guided approach for NC machining process planning. *Computer-Aided Design*, 162, 103562.
100. Wu, W., Huang, Z., Zeng, J., & Fan, K. (2021). A fast decision-making method for process planning with dynamic machining resources via deep reinforcement learning. *Journal of manufacturing systems*, 58, 392–411.
101. Wang, Z., Zhang, S., Zhang, H., Zhang, Y., Liang, J., Huang, R., & Huang, B. (2024). Machining feature process route planning based on a graph convolutional neural network. *Advanced Engineering Informatics*, 59, 102249.
102. Kataraki, P. S., & Abu Mansor, M. S. (2018). A novel classification of freeform volumetric features and generative CAPP approach for milling machine selection. *The International Journal of Advanced Manufacturing Technology*, 98(1), 985–1009.
103. Xu, T., Chen, Z., Li, J., & Yan, X. (2015). Automatic tool path generation from structuralized machining process integrated with CAD/CAPP/CAM system. *The International Journal of Advanced Manufacturing Technology*, 80(5), 1097–1111.
104. Kasie, F. M., & Bright, G. (2023). Application of Fuzzy Case-Based Reasoning and Fuzzy Analytic Hierarchy Process for Machining Cutter Planning and Control. *Advances in Fuzzy Systems*, 2023(1), 8072930.

105. Liu, Y., Gu, F., Gu, X., Wu, Y., Guo, J., & Zhang, J. (2022). Resource recommendation based on industrial knowledge graph in low-resource conditions. *International Journal of Computational Intelligence Systems*, 15(1), 42.
106. Butdee, S., & Kunhirunbawon, S. (2020). Multi-criteria decision for machining process plan evaluation using fuzzy logic modeling and feature based method. *Materials Today: Proceedings*, 26, 1982–1987.
107. Dharmadhikari, S., Menon, N., & Basak, A. (2023). A reinforcement learning approach for process parameter optimization in additive manufacturing. *Additive Manufacturing*, 71, 103556.
108. Lv, L., Deng, Z., Meng, H., Liu, T., & Wan, L. (2020). A multi-objective decision-making method for machining process plan and an application. *Journal of Cleaner Production*, 260, 121072.
109. Wu, W., Huang, Z., Zeng, J., & Fan, K. (2022). A decision-making method for assembly sequence planning with dynamic resources. *International Journal of Production Research*, 60(15), 4797–4816.
110. Zhou, B., Bao, J., Chen, Z., & Liu, Y. (2022). KGAssembly: Knowledge graph-driven assembly process generation and evaluation for complex components. *International Journal of Computer Integrated Manufacturing*, 35(10–11), 1151–1171.
111. Xiao, Y., Zheng, S., Feng, H., Huang, Y., Leng, J., & Hong, J. (2025). KGESM: A knowledge graph embedding-based similarity matching model for intelligent assembly process generation. *Journal of Manufacturing Systems*, 82, 1110–1124.
112. Gonnermann, C., Hashemi-Petroodi, S. E., Thevenin, S., Dolgui, A., & Daub, R. (2022). A skill-and feature-based approach to planning process monitoring in assembly planning. *The International Journal of Advanced Manufacturing Technology*, 122(5), 2645–2670.
113. Dong, J., Jing, X., Lu, X., Liu, J., Li, H., Cao, X., & Li, L. (2022). Process knowledge graph modeling techniques and application methods for ship heterogeneous models. *Scientific Reports*, 12(1), 2911.
114. Trstenjak, M., Opetuk, T., Cajner, H., & Tosanovic, N. (2020). Process planning in Industry 4.0—Current state, potential and management of transformation. *Sustainability*, 12(15), 5878.
115. Huang, Z., Guo, X., Jiang, C., Yang, M., Xue, H., Zhao, W., & Wang, J. (2025). mKGMPP: A multi-layer knowledge graph integration framework and its inference method for manufacturing process planning. *Advanced Engineering Informatics*, 65, 103266.
116. Hu, Y., Dong, H., Liu, J., Zhuang, C., & Zhang, F. (2025). A learning-guided hybrid genetic algorithm and multi-neighborhood search for the integrated process planning and scheduling problem with reconfigurable manufacturing cells. *Robotics and Computer-Integrated Manufacturing*, 93, 102919.
117. Marzia, S., Azab, A., & Vital-Soto, A. (2025). Integrated Process Planning and Scheduling Framework Using an Optimized Rule-Mining Approach for Smart Manufacturing. *Mathematics*, 13(16), 2605.
118. Lihong, Q., & Shengping, L. (2012). An improved genetic algorithm for integrated process planning and scheduling. *The International Journal of Advanced Manufacturing Technology*, 58(5), 727–740.
119. Li, X., Kambhampati, S., & Shah, J. (2000, September). ASUPPA: A Framework for Interactive and Iterative Synthesis and Improvement of Process Plans. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 35111, pp. 1–15). American Society of Mechanical Engineers.
120. Sosa, D. N., & Altman, R. B. (2022). Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference. *Briefings in bioinformatics*, 23(4), bbac268.
121. Mabkhot, M. M., Al-Samhan, A. M., & Hidri, L. (2019). An Ontology-Enabled Case-Based Reasoning Decision Support System for Manufacturing Process Selection. *Advances in Materials Science and Engineering*, 2019(1), 2505183.
122. Pollini, B., & Rognoli, V. (2021). Early-stage material selection based on life cycle approach: tools, obstacles and opportunities for design. *Sustainable Production and Consumption*, 28, 1130–1139.
123. Kutin, A., Dolgov, V., Sedykh, M., & Ivashin, S. (2018). Integration of different computer-aided systems in product designing and process planning on digital manufacturing. *Procedia Cirp*, 67, 476–481.
124. Villalonga, A., Negri, E., Biscardo, G., Castano, F., Haber, R. E., Fumagalli, L., & Macchi, M. (2021). A decision-making framework for dynamic scheduling of cyber-physical production systems based on digital twins. *Annual Reviews in Control*, 51, 357–373.
125. Silva, S. T., Hak, F., & Machado, J. (2022). Rule-based clinical decision support system using the openehr standard. *Procedia Computer Science*, 201, 726–731.
126. Yang, L. H., Wang, Y. M., & Fu, Y. G. (2018). A consistency analysis-based rule activation method for extended belief-rule-based systems. *Information Sciences*, 445, 50–65.
127. Offermans, T., Szymańska, E., Souza, F. A., & Jansen, J. J. (2024). Process expert knowledge is essential in creating value from data-driven industrial soft sensors. *Computers & Chemical Engineering*, 183, 108602.
128. Zhao, Z., Li, Y., Liu, C., Liu, X., & Gao, J. (2024). Stable Data-Driven Manufacturing Decision-Making by Introducing Causal Relationships for High-Dimensional Data. *IEEE Transactions on Industrial Informatics*.
129. Sun, Y. N., Pan, Y. J., Liu, L. L., Gao, Z. G., & Qin, W. (2024). Reconstructing causal networks from data for the analysis, prediction, and optimization of complex industrial processes. *Engineering Applications of Artificial Intelligence*, 138, 109494.
130. De Gasperi, G., & Facchini, S. D. (2025). A comparative study of rule-based and data-driven approaches in industrial monitoring. *arXiv preprint arXiv:2509.15848*.
131. Ur-Rahman, N. (2015). Textual data mining for next generation intelligent decision making in industrial environment: a survey. *European Scientific Journal*, 11(24).
132. Keskin, Z., Joosten, D., Klasen, N., Huber, M., Liu, C., Drescher, B., & Schmitt, R. H. (2025). Llm-enhanced human-machine interaction for adaptive decision making in dynamic manufacturing process environments. *IEEE access*.
133. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459–9474.
134. Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., & Niforatos, E. (2024). Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. *Frontiers in Artificial intelligence*, 7, 1293084.
135. Ni, M., Wang, T., Leng, J., Chen, C., & Cheng, L. (2025). A large language model-based manufacturing process planning approach under industry 5.0. *International Journal of Production Research*, 1–20.
136. Du, K., Yang, B., Xie, K., Dong, N., Zhang, Z., Wang, S., & Mo, F. (2025). LLM-MANUF: An integrated framework of Fine-Tuning large language models for intelligent Decision-Making in manufacturing. *Advanced Engineering Informatics*, 65, 103263.
137. Liu, B., Cui, Z., Hu, S., Li, X., Lin, H., & Zhang, Z. (2025). Llm evaluation based on aerospace manufacturing expertise: Automated generation and multi-model question answering. *arXiv preprint arXiv:2501.17183*.

138. Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., ... Luong, M. T. (2024, August). Freshllms: Refreshing large language models with search engine augmentation. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 13697–13720).
139. Hu, Z., & Yan, W. (2024). Data-driven modeling of process-structure-property relationships in metal additive manufacturing. *Npj Advanced Manufacturing*, 1(1), 3.
140. CECE. (2025, October 29). *ECHA updated PFAS restriction proposal: What it means for construction equipment—from lubricants to machinery applications*. Committee for European Construction Equipment. <https://www.cece.eu/news/echa-update-d-pfas-restriction-proposal-what-it-means-for-construction-equipment-from-lubricants-to-machinery-applications>
141. Anh-Hoang, D., Tran, V., & Nguyen, L. M. (2025). Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8, 1622292.
142. Meng, C., Ling, H., Wang, J., Liu, Y., Zhang, S., Hong, D., ... Han, N. (2025, September). Balancing Fine-tuning and RAG: A Hybrid Strategy for Dynamic LLM Recommendation Updates. In Proceedings of the Nineteenth ACM Conference on Recommender Systems (pp. 919–922).
143. Handler, A., Larsen, K. R., & Hackathorn, R. (2024). Large language models present new questions for decision support. *International Journal of Information Management*, 79, 102811.
144. Lee, J., Ahn, S., Kim, D., & Kim, D. (2024). Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. *Automation in Construction*, 168, 105846.
145. Werner, J., & Arenella, K. (2025, May). AI-Powered Compliance: A RAG-Based System for Product Safety Design Engineering. In *2025 IEEE International Symposium on Product Compliance Engineering (ISPC)* (pp. 1–6). IEEE.
146. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).
147. Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., ... Xue, C. J. (2024). Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.
148. Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14(6), 1006–1033.
149. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
150. Bhat, S. R., Rudat, M., Spiekermann, J., & Flores-Herr, N. (2025). Rethinking Chunk Size For Long-Document Retrieval: A Multi-Dataset Analysis. *arXiv preprint arXiv:2505.21700*.
151. Stäbler, M., Turnbull, S., Müller, T., Langdon, C., Marx-Gómez, J., & Köster, F. (2025, August). The impact of chunking strategies on domain-specific information retrieval in RAG systems. In *2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS)* (pp. 1–6). IEEE.
152. Zhao, J., Ji, Z., Fan, Z., Wang, H., Niu, S., Tang, B., ... Li, Z. (2025). MoC: Mixtures of Text Chunking Learners for Retrieval-Augmented Generation System. *arXiv preprint arXiv:2503.09600*.
153. Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024). Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131*.
154. Singh, I. S., Aggarwal, R., Allahverdiyev, I., Taha, M., Akalin, A., Zhu, K., & O'Brien, S. (2024). Chunkrag: Novel llm-chunk filtering method for rag systems. *arXiv preprint arXiv:2410.19572*.
155. Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: a review. *Mathematics*, 13(5), 856.
156. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., ... Silvestri, F. (2024, July). The power of noise: Redefining retrieval for rag systems. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 719–729).
157. Luan, Y., Eisenstein, J., Toutanova, K., & Collins, M. (2021). Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9, 329–345.
158. Rossi, N., Lin, J., Liu, F., Yang, Z., Lee, T., Magnani, A., & Liao, C. (2024, October). Relevance filtering for embedding-based retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 4828–4835).
159. Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024, April). Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI* (pp. 194–199).
160. Chen, H. T., Zhang, M., & Choi, E. (2022, December). Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2292–2307).
161. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12, 157–173.
162. Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023, December). Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5303–5315).
163. Chan, C. M., Xu, C., Yuan, R., Luo, H., Xue, W., Guo, Y., & Fu, J. (2024). Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.
164. Tanyildiz, D., Ayvaz, S., & Amasyali, M. F. (2024). Enhancing Retrieval-Augmented Generation Accuracy with Dynamic Chunking and Optimized Vector Search. *Orclever Proceedings of Research and Development*, 5(1), 215–225.
165. Eibich, M., Nagpal, S., & Fred-Ojala, A. (2024). Aragog: Advanced rag output grading. *arXiv preprint arXiv:2404.01037*.
166. Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2024). Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.
167. Yu, Y., Ping, W., Liu, Z., Wang, B., You, J., Zhang, C., ... Catanzaro, B. (2024). Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37, 121156–121184.
168. Xu, S., Pang, L., Xu, J., Shen, H., & Cheng, X. (2024, May). List-aware reranking-truncation joint model for search and retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2024* (pp. 1330–1340).
169. Yan, S. Q., Gu, J. C., Zhu, Y., & Ling, Z. H. (2024). Corrective retrieval augmented generation.
170. Jin, Z., Cao, P., Chen, Y., Liu, K., Jiang, X., Xu, J., ... Zhao, J. (2024, May). Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024) (pp. 16867–16878).
171. Wang, F., Wan, X., Sun, R., Chen, J., & Arik, S. O. (2025, July). Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings*

- of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 30553–30571).
172. Ma, L., Zhang, R., Han, Y., Yu, S., Wang, Z., Ning, Z., ... Lu, C. T. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv preprint arXiv:2310.11703*.
 173. Auyeskan, U., Turyzbekov, G., Roshaven, S. P., Perveen, A., & Talamona, D. (2025). Decision-making framework supported by techno-economic analysis of laser powder bed fusion: a novel approach using Retrieved Augmentation Generation (RAG). *Progress in Additive Manufacturing*, 1–17.
 174. Álvaro, J. A. H., & Barreda, J. G. (2025). An advanced retrieval-augmented generation system for manufacturing quality control. *Advanced Engineering Informatics*, 64, 103007.
 175. Li, G., Xu, W., & Du, T. (2025, May). An Automatic Method for Machining Process Route Generation Based on Large Language Models and Retrieval-Augmented Generation. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Smart Manufacturing* (pp. 638–642).
 176. Jeon, J., Sim, Y., Lee, H., Han, C., Yun, D., Kim, E., ... Lee, J. (2025). ChatCNC: Conversational machine monitoring via large language model and real-time data retrieval augmented generation. *Journal of Manufacturing Systems*, 79, 504–514.
 177. Bahr, L., Wehner, C., Wewerka, J., Bittencourt, J., Schmid, U., & Daub, R. (2025). Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis. *Journal of Industrial Information Integration*, 45, 100807.
 178. Gilmore, J. F. (1984, December). Knowledge base systems in computer aided technology. In *The 23rd IEEE Conference on Decision and Control* (pp. 586–590). IEEE.
 179. Patel, A., & Jain, S. (2018). Formalisms of representing knowledge. *Procedia Computer Science*, 125, 542–549.
 180. Bhuiyan, M. H., Bhattacharjee, A., & Nath, R. P. D. (2017, December). DB2KB: A framework to publish a database as a knowledge base. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1–7). IEEE.
 181. Szejka, A. L., Junior, O. C., & Mas, F. (2024). Knowledge-based expert system to drive an informationally interoperable manufacturing system: An experimental application in the Aerospace Industry. *Journal of Industrial Information Integration*, 41, 100661.
 182. Guo, L., Yan, F., Li, T., Yang, T., & Lu, Y. (2022). An automatic method for constructing machining process knowledge base from knowledge graph. *Robotics and Computer-Integrated Manufacturing*, 73, 102222.
 183. Assal, H., & Myers, L. (1990, August). Implementation of a frame-based representation in CLIPS. In *NASA Johnson Space Center, First CLIPS Conference Proceedings, Volume 2*.
 184. Lingarkar, R., Liu, L., Elbestawi, M. A., & Sinha, N. K. (2002). Knowledge-based adaptive computer control in manufacturing systems: a case study. *IEEE Transactions on systems man and cybernetics*, 20(3), 606–618.
 185. Lei, Q., Wang, H., & Song, Y. (2016). Hybrid knowledge model of process planning and its green extension. *Journal of Intelligent Manufacturing*, 27(5), 975–990.
 186. Reddy, B., & Fields, R. (2022, April). From past to present: a comprehensive technical review of rule-based expert systems from 1980–2021. In *Proceedings of the 2022 ACM Southeast Conference* (pp. 167–172).
 187. Thike, P. H., Xu, Z., Cheng, Y., Jin, Y., & Shi, P. (2019). Materials failure analysis utilizing rule-case based hybrid reasoning method. *Engineering Failure Analysis*, 95, 300–311.
 188. Schmetz, A., Lee, T. H., Hoeren, M., Berger, M., Ehret, S., Zontar, D., ... Brecher, C. (2020). Evaluation of industry 4.0 data formats for digital twin of optical components. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 7(3), 573–584.
 189. Zhou, G., Lu, Q., Xiao, Z., Zhou, C., Yuan, S., & Zhang, C. (2017). Ontology-based cutting tool configuration considering carbon emissions. *International Journal of Precision Engineering and Manufacturing*, 18(11), 1641–1657.
 190. Kang, M., Kim, G., Lee, T., Jung, C. H., Eum, K., Park, M. W., & Kim, J. K. (2016). Selection and sequencing of machining processes for prismatic parts using process ontology model. *International Journal of Precision Engineering and Manufacturing*, 17(3), 387–394.
 191. Heist, N., Hertling, S., Ringler, D., & Paulheim, H. (2020). Knowledge graphs on the web—an overview. *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, 3–22.
 192. Xiong, Y., Zhang, R., Liu, Y., Niyato, D., Xiong, Z., Liang, Y. C., & Mao, S. (2024). When graph meets retrieval augmented generation for wireless networks: A tutorial and case study. *arXiv preprint arXiv:2412.07189*.
 193. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
 194. Tan, X., Wang, X., Liu, Q., Xu, X., Yuan, X., & Zhang, W. (2025, April). Paths-over-graph: Knowledge graph empowered large language model reasoning. In *Proceedings of the ACM on Web Conference 2025* (pp. 3505–3522).
 195. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 3580–3599.
 196. Han, H., Ma, L., Shomer, H., Wang, Y., Lei, Y., Guo, K., ... Tang, J. (2025). Rag vs.graphrag: A systematic evaluation and key insights. *arXiv preprint arXiv:2502.11371*.
 197. Wan, Y., Liu, Y., Zammit, J. P., Chen, Z., Li, L., & Francalanza, E. (2025, June). Facilitating design for additive manufacturing with KG-based retrieval-augmented generation. In *2025 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)* (pp. 1–8). IEEE.
 198. Hua, Y., Wang, R., Wang, Z., Wang, G., & Yan, Y. (2025). Knowledge graph with deep reinforcement learning for intelligent generation of machining process design. *Journal of Engineering Design*, 36(11), 2072–2106.
 199. Hussong, M., Ruediger-Flore, P., Klar, M., Kloft, M., & Aurich, J. C. (2025). Selection of manufacturing processes using graph neural networks. *Journal of Manufacturing Systems*, 80, 176–193.
 200. Xiao, Y., Zheng, S., Leng, J., Gao, R., Fu, Z., & Hong, J. (2025). An assembly process planning pipeline for industrial electronic equipment based on knowledge graph with bidirectional extracted knowledge from historical process documents. *Journal of Intelligent Manufacturing*, 36(5), 3647–3667.
 201. Wu, H., Jiang, Z., Zhu, S., & Zhang, H. (2024). A knowledge graph based disassembly sequence planning for end-of-life power battery. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 11(3), 849–861.
 202. Duan, Y., Hou, L., & Leng, S. (2021). A novel cutting tool selection approach based on a metal cutting process knowledge graph. *The International Journal of Advanced Manufacturing Technology*, 112(11), 3201–3214.
 203. Su, C., Jiang, Q., Han, Y., Wang, T., & He, Q. (2025). Knowledge graph-driven decision support for manufacturing process: A graph neural network-based knowledge reasoning approach. *Advanced Engineering Informatics*, 64, 103098.
 204. Jing, Y., Zhou, G., Zhang, C., Chang, F., Yan, H., & Xiao, Z. (2024). XMKR: Explainable manufacturing knowledge recommendation for collaborative design with graph embedding learning. *Advanced Engineering Informatics*, 59, 102339.
 205. Kosasih, E. E., Margaroli, F., Gelli, S., Aziz, A., Wildgoose, N., & Brintrup, A. (2024). Towards knowledge graph reasoning for

- supply chain risk management using graph neural networks. *International Journal of Production Research*, 62(15), 5596–5612.
206. Wu, T., Du, S., Zhang, Y., & Li, H. (2025). Knowledge-Enriched Recommendations: Bridging the Gap in Alloy Material Selection With Large Language Models. *IEEE Access*.
 207. Tian, Y., Xu, S., Fu, J., Chen, Y., Wang, H., & Wang, T. (2025). Process reuse-based machining process knowledge graph construction and process planning approach. *Flexible Services and Manufacturing Journal*, 1–33.
 208. Rajabi, E., & Etmiani, K. (2024). Knowledge-graph-based explainable AI: A systematic review. *Journal of information science*, 50(4), 1019–1029.
 209. Liu, X., Mao, T., Shi, Y., & Ren, Y. (2024). Overview of knowledge reasoning for knowledge graph. *Neurocomputing*, 585, 127571.
 210. Ottersen, S. G., Pinheiro, F., & Bação, F. (2024). Triplet extraction leveraging sentence transformers and dependency parsing. *Array*, 21, 100334.
 211. Yu, H., Li, H., Mao, D., & Cai, Q. (2020). A relationship extraction method for domain knowledge graph construction. *World Wide Web*, 23(2), 735–753.
 212. Rong, Z., Yuan, L., & Yang, L. (2024). Enhanced knowledge graph recommendation algorithm based on multi-level contrastive learning. *Scientific Reports*, 14(1), 23051.
 213. Wang, W., Shen, X., Yi, B., Zhang, H., Liu, J., & Dai, C. (2024). Knowledge-aware fine-grained attention networks with refined knowledge graph embedding for personalized recommendation. *Expert Systems with Applications*, 249, 123710.
 214. Ma, T., Huang, L., Lu, Q., & Hu, S. (2023). Kr-gcn: Knowledge-aware reasoning with graph convolution network for explainable recommendation. *ACM Transactions on Information Systems*, 41(1), 1–27.
 215. Zhong, L., Wu, J., Li, Q., Peng, H., & Wu, X. (2023). A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4), 1–62.
 216. Deng, J., He, C., Chen, J., Qin, B., Wu, J., Huang, Q., & Li, Y. (2025). Constructing a knowledge graph-driven intelligent data-enabled design system for mold using deep semantic understanding and intelligent decision support. *Scientific Reports*, 15(1), 7322.
 217. Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature communications*, 15(1), 1418.
 218. Bachhofner, S., Kiesling, E., Revoredo, K., Waibel, P., & Polleres, A. (2022, July). Automated process knowledge graph construction from BPMN models. In *International Conference on Database and Expert Systems Applications* (pp. 32–47). Cham: Springer International Publishing.
 219. Yang, H., Xiao, L., Zhu, R., Liu, Z., & Chen, J. (2024, December). An LLM supported approach to ontology and knowledge graph construction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 5240–5246). IEEE.
 220. Ma, Y., Zheng, S., Yang, Z., Pan, H., & Hong, J. (2025). A knowledge-graph enhanced large language model-based fault diagnostic reasoning and maintenance decision support pipeline towards industry 5.0. *International Journal of Production Research*, 1–22.
 221. Li, X., Zheng, J., Su, Z., Chen, Y., Wang, Y., Chen, C., ... Wang, K. (2024, November). Construction of Knowledge Graph of Substation Main Equipment Based on LLM. In *2024 3rd Asia Power and Electrical Technology Conference (APET)* (pp. 746–750). IEEE.
 222. Zhang, B., & Soh, H. (2024). Extract, define, canonicalize: An llm-based framework for knowledge graph construction. *arXiv preprint arXiv:2404.03868*.
 223. Xu, Q., Qiu, F., Zhou, G., Zhang, C., Ding, K., Chang, F., ... Liu, J. (2025). A large language model-enabled machining process knowledge graph construction method for intelligent process planning. *Advanced Engineering Informatics*, 65, 103244.
 224. Wan, Y., Liu, Y., Chen, Z., Chen, C., Li, X., Hu, F., & Packianather, M. (2024). Making knowledge graphs work for smart manufacturing: Research topics, applications and prospects. *Journal of manufacturing systems*, 76, 103–132.
 225. Spillo, G., Musto, C., Mannavola, M., de Gemmis, M., Lops, P., & Semeraro, G. (2025, June). GAL-KARS: Exploiting LLMs for Graph Augmentation in Knowledge-Aware Recommender Systems. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 73–82).
 226. Zhang, W., & Serban, O. (2025). LLM-based Reranking and Validation of Knowledge Graph Completion. In *Sixth International Workshop on Knowledge Graph Construction@ ESWC2025*.
 227. Zhang, Z., Yu, J., Yang, B., Du, K., Wang, S., & Qi, X. (2025). A knowledge graphs construction method enhanced by multimodal large language model for industrial equipment operation and maintenance. *Advanced Engineering Informatics*, 68, 103705.
 228. Cai, E., & O'Connor, B. (2025). Understanding the effect of knowledge graph extraction error on downstream graph analyses: a case study on affiliation graphs. *Applied Network Science*, 10(1), 64.
 229. Mihindukulasooriya, N., Tiwari, S., Enguix, C. F., & Lata, K. (2023, October). Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In *International semantic web conference* (pp. 247–265). Cham: Springer Nature Switzerland.
 230. van Cauter, Z., & Yakovets, N. (2024, August). Ontology-guided knowledge graph construction from maintenance short texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)* (pp. 75–84).
 231. Zhu, L., Gong, Y., & Bai, L. (2026). Zero-shot temporal knowledge graph completion based on generative adversarial network. *World Wide Web*, 29(1), 1.
 232. Jarnac, L., Chabot, Y., & Couceiro, M. (2024). Uncertainty Management in the Construction of Knowledge Graphs: a Survey. *arXiv preprint arXiv:2405.16929*.
 233. Wiharja, K., Pan, J. Z., Kollingbaum, M. J., & Deng, Y. (2020). Schema aware iterative knowledge graph completion. *Journal of Web Semantics*, 65, 100616.
 234. Zhang, H., Schmidt, W. J., Shen, X., Cao, Q., Monka, S., & Paschke, A. (2025). Knowledge Graph Construction towards a Graph RAG-Enhanced Intelligent Maintenance Chatbot. In *International Workshop on Scaling Knowledge Graphs for Industry 2025*.
 235. De Santis, A., Balduini, M., De Santis, F., Proia, A., Leo, A., Brambilla, M., & Della Valle, E. (2024, November). Integrating large language models and knowledge graphs for extraction and validation of textual test data. In *International Semantic Web Conference* (pp. 304–323). Cham: Springer Nature Switzerland.
 236. Fan, H., Huang, J., Xu, J., Zhou, Y., Fuh, J. Y. H., Lu, W. F., & Li, B. (2025). AutoMEX: Streamlining material extrusion with AI agents powered by large language models and knowledge graphs. *Materials & Design*, 251, 113644.
 237. Zheng, X., Kong, Y., Chang, T., Liao, X., Ma, Y., & Du, Y. (2022). High-throughput computing assisted by knowledge graph to study the correlation between microstructure and mechanical properties of 6XXX aluminum alloy. *Materials*, 15(15), 5296.
 238. Ramos, L. (2015). Semantic Web for manufacturing, trends and open issues: Toward a state of the art. *Computers & Industrial Engineering*, 90, 444–460.
 239. Kim, S. W., Kong, J. H., Lee, S. W., & Lee, S. (2022). Recent advances of artificial intelligence in manufacturing industrial

sectors: A review. *International Journal of Precision Engineering and Manufacturing*, 23(1), 111–129.

240. Xiao, W., Qiu, T., Guo, J., & Zhao, G. (2025). MetaFactory: A cloud-based framework to configure and generate dynamic data structures from the STEP-NC knowledge graph. *Journal of Manufacturing Systems*, 80, 89–107.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Geonhwi Lee is a Ph.D student in the School of Mechanical Engineering, Chung-Ang University (CAU) in Seoul, Korea. He received his M.S degree in the Department of Mechanical Engineering in 2025 from Chung-Ang University. His academic interests include generative AI, machine learning, modeling and simulation, optimization, and smart manufacturing.



Solchan Kim is a M.S. student in the School of Mechanical Engineering, Chung-Ang University (CAU) in Seoul, Korea. He received his B.S. degree in the Department of Automotive Engineering in 2025 from Dong-Eui University. His research interests include generative AI, large vision-language models (LVLMs), and video-driven knowledge bases for manufacturing.



Yulseok Byun is a graduate student of the School of Mechanical Engineering, Chung-Ang University, Seoul, Korea. He received his B.S. degree in mechanical engineering from Chung-Ang University, Seoul, Korea, in 2025. His research interests include smart manufacturing, additive manufacturing, design for robotic assembly, and unmanned aerial system (UAS).



Jiho Lee is currently an Assistant Professor of Mechanical Engineering at the University of North Texas (UNT), Denton, TX, USA. Prior to joining UNT, he served as a Research Scientist and Postdoctoral Researcher at Purdue University, West Lafayette, IN, USA. He holds both a Ph.D. and B.S. in Mechanical Engineering from Sungkyunkwan University, Suwon, Korea. His research focuses on AI-driven smart manufacturing, with an emphasis on Physics-informed AI, Industrial IoT-based monitoring, autonomous robotics, and digital twins.



Hae-Jin Choi received M.S. and Ph.D. degrees in mechanical engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 2001 and 2005, respectively. He was an Assistant Professor at Nanyang Technological University, Singapore. He was also a Postdoctoral Fellow at the GWW School of Mechanical Engineering, Georgia Tech. He is currently a Professor at the School of Mechanical Engineering, Chung-Ang University, Seoul, Korea. His research interests include fault diagnosis, prognostics and health management, machine learning, simulation-based design optimization, management of uncertainty, and integrated materials and products.