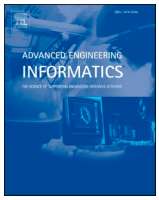
Advanced Engineering Informatics 65 (2025) 103263

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



LLM-MANUF: An integrated framework of Fine-Tuning large language models for intelligent Decision-Making in manufacturing

Kaze Du [a], Bo Yang [a,*], Keqiang Xie [b,c], Nan Dong [c], Zhengping Zhang [a,d], Shilong Wang [a], Fan Mo [e,f]

[a] State Key Laboratory of Mechanical Transmission for Advanced Equipment, Chongqing University, Chongqing 400044, China
[b] School of Mechanical Science & Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[c] The Fifth Electronics Research Institute of the Ministry Industry and Information Technology, Guangzhou 510000, China
[d] Seres Group Co. Ltd., Chongqing 401335, China
[e] Centre for Human-inspired Artificial Intelligence, University of Cambridge, CB2 1SB Britain, United Kingdom
[f] Department of Computer Science, University of Oxford, OX1 3QG Britain, United Kingdom

ARTICLE INFO

Keywords:
Large Language Model
Integrated Framework
Fine-Tuning
Manufacturing
Intelligent Decision-Making

ABSTRACT

Intelligent decision-making is pivotal for unlocking the value of industrial knowledge and enhancing the manufacturing sector across diverse scenarios. However, traditional decision-making methods in manufacturing fail to fully capture the complex interrelationships among various components, often resulting in biased decisions. As a novel productivity tools, large language models (LLMs) have strong contextual semantic parsing capabilities. Therefore, this paper proposes a fine-tuned LLMs integration framework for intelligent decision-making in manufacturing. The framework enables the extraction of decision-making information from diverse feature subspaces through multiple parallel fine-tuned LLMs, which generate several preliminary decision-making plans. Subsequently, the framework models the probabilities of these plans to derive a ranked list of candidates. It then employs RoBERTa and a Dynamic Weighted Mixture of Experts Ranking Method (DWMOE) to perform multi-dimensional feature extraction and candidate ranking, guided by a multi-metric head. Finally, the best fine-tuned LLM is used to fuse the top-ranked candidates, minimizing bias in the final decision-making process. To evaluate the efficacy of LLM-MANUF, we construct a dataset of manufacturing product equipment operation and maintenance texts based on a specific automotive enterprise. The results indicate that the LLM-MANUF not only outperforms individual fine-tuned LLMs but also matches the performance of LLMs with 30B parameters, achieving a BLEU-4 score of 83.37 points, which demonstrates exceptional reliability and effectiveness. LLM-MANUF provides a powerful intelligent decision-making support tool for manufacturing decision-making models.

1. Introduction

Decision-making is a crucial activity in manufacturing systems, encompassing production, operations management, and continuous improvement [1]. In this context, decision-makers often rely on their expertise and practical experience to develop multiple alternatives tailored to the specific needs of a given site, ensuring well-informed choices [2]. These decisions permeate the entire manufacturing lifecycle, including the design and scheduling of production lines, the selection of appropriate processing methods, and the operation and maintenance of equipment [3]. Efficient and accurate decision-making plays a critical role in optimizing the entire production process, helping manufacturing companies streamline decision-making process, reduce uncertainty, improve operational efficiency, and enhance overall performance. This is especially true when unforeseen events, such as quality defects or equipment failures, arise during production [4]. In such cases, rapid and precise decision-making can prevent issues like product quality deterioration, production delays, and even safety incidents. However, achieving fully continuous optimal decision-making is nearly impossible, as the manufacturing decision-making process is

* Corresponding author at: intelligent algorithm technology, State Key Laboratory of Mechanical Transmission for Advanced Equipment, Chongqing University, 174 Shazheng Street, Shapingba District, Chongqing 400044, China.
E-mail addresses: dkz98@cqu.edu.cn (K. Du), yangbo61@cqu.edu.cn (B. Yang), xiekq@hust.edu.cn (K. Xie), dongnan@ceprei.com (N. Dong), Aa517307130@163.com (Z. Zhang), wang.sl@cqu.edu.cn (S. Wang), fm651@cam.ac.uk (F. Mo).

https://doi.org/10.1016/j.aei.2025.103263
Received 24 October 2024; Received in revised form 7 February 2025; Accepted 10 March 2025
Available online 17 March 2025
1474-0346/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

inevitably influenced by various internal and external factors, leading to decision-making bias [5]. Therefore, developing an intelligent manufacturing decision-making framework to effectively mitigate decision-making bias is essential for organizations to address complex production challenges, ensure consistent product quality, and maintain efficient production capacity.

The causes of decision-making bias in manufacturing can be categorized into objective and subjective factors. On the objective side, the increasing complexity of modern manufacturing systems presents new challenges [6]. Manufacturing systems involve intricate interactions among equipment, information flow, process flow, and human–computer interfaces, creating new demands for decision-making methods. With the rapid advancement of AI technologies, such as machine learning and deep learning, these techniques are being gradually integrated into manufacturing decision-making support systems [7]. Unlike traditional rule-based decision trees and expert systems, which rely on fixed rule sets or templates, AI-driven approaches automatically identify patterns and make predictions by training models on manufacturing data [8], as illustrated in Fig. 1. AI applications in manufacturing decision-making are primarily divided into two paradigms: data-driven and knowledge-driven. Data-driven approaches focus on extracting patterns from large datasets [9] and are particularly suited for applications like condition monitoring and predictive maintenance. While data-driven approaches assist decision-making, they typically do not directly generate actionable recommendations. In contrast, knowledge-driven approaches leverage knowledge graphs [10] or pre-trained models [11] to derive decision-making plans by parsing in-depth manufacturing knowledge, thus compensating for the limitations of data-driven methods. Furthermore, the emergence of LLMs has significantly enhanced the understanding of manufacturing domain expertise. With their advanced semantic parsing capabilities, LLMs can effectively identify and extract implicit semantic information from the manufacturing process, thereby enhancing the intelligence of the manufacturing decision-making framework.

On the subjective level, as manufacturing systems become increasingly complex, the frequency and diversity of potential problems also grow [12]. This heightened complexity introduces more uncertainties into the manufacturing process. While existing decision-making frameworks can support engineers, most are based on closed-domain knowledge systems [13], and their development and optimization heavily rely on the experience and expertise of domain experts. Due to variations in the knowledge, experience, and preferences of different experts, their interpretations and solutions for the same manufacturing problem may differ [14], directly impacting the consistency and reliability of decision-making. This issue is particularly significant in complex manufacturing environments with highly interconnected equipment and processes, where subjective differences can accumulate, compromising product quality and production stability. Additionally, engineers often face substantial information processing burdens when interacting with manufacturing decision-making frameworks [15]. Limited cognitive capacity and time constraints hinder their ability to make optimal decisions under information overload, further amplifying the risk of decision-making bias.

To address the issue of manufacturing decision-making bias, this paper proposes a fine-tuned LLMs integration framework, LLM-MANUF, for intelligent decision-making in manufacturing. Specifically, the LLM-MANUF framework incorporates multiple small-scale LLMs, each fine-tuned on a manufacturing corpus. LLMs enhance the contextual parsing capability of the decision-making framework, enabling it to better understand the complex relationships within manufacturing systems. Moreover, it reduces the likelihood of decision-making bias through a ranking strategy and a fusion strategy that integrate the strengths of each model. The scientific novelty of this work can be summarized as follows:

(1) An LLM integration framework, LLM-MANUF, is designed to mitigate the manufacturing decision bias problem, which uses multiple fine-tuned LLMs to jointly analyze the same decision-making task, and subsequently applies a ranking and fusion strategy that integrates the strengths of LLMs to generate the optimal decision-making plan.
(2) Fine-tuning LLMs for intelligent decision-making in manufacturing are investigated. A manufacturing corpus is obtained by organizing manufacturing-related documents, several LLMs with low computational resource requirements are selected, and LoRA fine-tuning method is applied to obtain a fine-tuning LLM based on the aforementioned manufacturing-specific corpus with targeted fine-tuning of LLMs.
(3) A Dynamically Weighted Mixture of Experts Ranking Method is introduced. This method models the probability of the best candidate and generates a candidate ranking. Semantic feature extraction is performed using the RoBERTa semantic parser, while the probability distribution model is constructed using the DWMOE. Candidate ranking is guided by a multi-metric header.
(4) Adaptive Fusion Decision Strategy (AFDS): This strategy dynamically selects the most suitable LLM to fuse the best candidate solutions based on the specific requirements of the manufacturing decision scenario. The fusion process synthesizes the strengths of each candidate solution, improving the accuracy and reliability of the final decision-making plan and significantly reducing decision-making bias.

The structure of the remainder of this paper is as follows: Section 2 reviews relevant research advances in manufacturing decision bias and

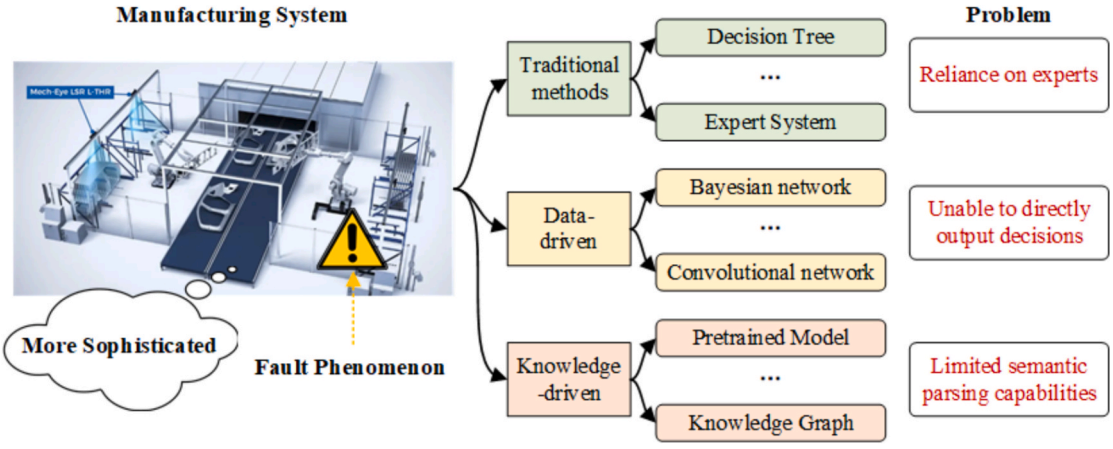


Fig. 1. Issues with Existing Manufacturing Decision-Making Methods.

integration frameworks. Section 3 outlines the key components of the LLM-MANUF framework. Section 4 presents a case study on manufacturing decision-making, conducted for validation purposes. Finally, Section 5 provides a summary of the research and proposes directions for future work.

2. Related Works

2.1. Research on decision-making in manufacturing domain

Manufacturing decision-making bias refers to the difficulty in fully understanding the complex interrelationships within a manufacturing system due to the cognitive limitations of the decision maker or the decision-making system [16], leading to deficiencies in the processes of acquiring, processing, and analyzing information [6]. Current research focuses on alleviating bias by enhancing the decision-making capabilities of manufacturing methods [17] or improving the decision-making abilities of individuals involved [18].

In terms of enhancing decision-making methods, research has primarily concentrated on improving the understanding of the complex associations within manufacturing systems, which can be achieved through both data-driven and knowledge-driven approaches. Data-driven approaches primarily focus on monitoring and diagnosing production states but cannot directly support decision-making. In contrast, knowledge-driven approaches emphasize guiding the decision-making process by leveraging accumulated expertise and experience [19]. This approach not only provides direct decision-making advice but also promotes a deeper understanding and optimization of manufacturing systems through systematic management and generalization of domain expertise. Su et al. [20] facilitated the collection and management of manufacturing process knowledge using knowledge graphs, enabling dynamic modeling and iteration through knowledge matching and reasoning. Jing et al. [21] proposed a novel, interpretable manufacturing knowledge recommendation method based on graph embedding, which enables designer-oriented knowledge reuse, proactively avoiding design errors and minimizing iterative cycles. Xu et al. [22] introduced a two-phase semantic embedding method for manufacturing processes, based on the Sentence-BERT pre-trained model, achieving accurate defect case predictions.

Although deep learning models such as graph neural networks and BERT have demonstrated substantial semantic parsing capabilities in natural language processing, their understanding of manufacturing-specific semantics remains limited when applied to manufacturing systems. In contrast, large-scale language models, as exemplars of the new generation of artificial intelligence, offer promising solutions to these challenges through their advanced contextual understanding and reasoning capabilities. José et al. [23] constructed an advanced retrieval-enhanced generative system for manufacturing quality control based on the GPT-3.5-Turbo model, which diagnoses quality defects and provides solutions. Fu et al. [24] developed a steelmaking process management system integrating a visual language model and a language model, enabling textual description of slab defects and production data analysis. Liu et al. [25] proposed a knowledge-enhanced federated model that integrated an aerospace assembly knowledge graph into a large language model, allowing for in-depth analysis of complex knowledge, accurate fault localization, and troubleshooting solution recommendations. The large language model offers a viable and effective approach for parsing the complex associations in manufacturing systems, with its advanced contextual understanding, reasoning capabilities, and deep semantic learning specific to the manufacturing domain.

It is also essential to focus on enhancing the decision-making abilities of decision makers to reduce the possibility of decision bias. Melanie et al. [18] argued that human-centered planning and control process design can mitigate biased decision-making by systematically reducing subjective bias, thereby improving the overall performance of smart manufacturing systems. They developed a human-centered production planning and control decision-making framework that takes into account key influencing factors, challenges, and the performance of the decision-making process. Chen et al. [26], aiming to minimize the impact of a decision maker's unknown information in the manufacturing decision-making process, framed the decision maker's uncertainty as a regression problem, which can be estimated using machine learning and model predictive control techniques, thus aiding the decision maker in making the best decision. However, optimizing the manufacturing decision framework is generally more effective than merely enhancing the personal capabilities of decision makers. Several studies have demonstrated that the reliability and accuracy of decision-making can be significantly improved by integrating multiple decision-making models into a comprehensive framework. For instance, Jiang et al. [27] achieved LLM integration by distinguishing differences between candidate outputs through candidate pairing and improving candidate outputs using GENFUSER. Zhu et al. [28] established a multi-task-oriented LLM integration framework that enables task assignment and LLM invocation through a hybrid expert model, enhancing the LLM's ability to understand multimodal and multiformat input data and generate output.

Although the studies mentioned above have yielded valuable results in mitigating manufacturing decision bias, they have some limitations. Therefore, this paper proposes an approach based on fine-tuned large-scale LLMs to capture the multidimensional semantics of manufacturing knowledge, providing robust support for a deeper understanding of the complex interrelationships within manufacturing systems. Building on this, we propose an integrated framework for intelligent manufacturing decision-making that combines multiple fine-tuned LLMs to synthesize decision-making information from various feature subspaces. This framework enables a comprehensive analysis of key factors in the manufacturing process from multiple perspectives, thus ensuring more accurate and reliable decision-making.

2.2. Research on integration methods

Combining decision-making information from multiple models represents an effective strategy for enhancing model performance and mitigating decision-making bias issues. An integrated decision framework typically comprises three components: a base model, a ranking method, and a fusion method [29]. The base model serves as the fundamental building block of the framework, tasked with generating candidate information directly. The ranking method evaluates and compares the quality of candidate information produced by various base models. Meanwhile, the fusion method integrates the ranked information to produce a final solution. The effectiveness of the base model sets the initial performance level of the integrated framework, whereas the integration rules, formed by the ranking and fusion methods, dictate the overall efficacy of the framework.

The objective of a ranking method is to filter high-quality candidates and provide prioritization information for the fusion process. It is common practice to utilize pairwise or listwise loss functions in ranking methods [30]. Bhattacharyya et al. [31] introduced an energy-based ranking approach, guiding the model to prioritize candidate samples with high BLEU scores through marginal and joint energy models. Liu et al. [32] employed a sequence-to-sequence model for generating candidate options, followed by training a parametric evaluation model using comparative learning to rank these options. However, directly training the model on the loss function between predicted and true values may reduce the generalization ability, especially when handling complex associations in manufacturing processes. Conversely, modeling the probability of the best candidate can better preserve model generalization [33]. Mathieu et al. [33] implemented a hybrid expert system to model each abstract summary directly, leveraging deeper features and correlation information for optimal ranking of summaries. Notably, relying solely on a single indicator for evaluating and ranking candidate information has limitations and fails to reflect comprehensive quality

accurately [34]. To address this, Sun et al. [35] proposed a low-rank multi-metric learning model for supervised classification, reducing misclassification probability and enhancing stability through multiple local class indicators. Direct probabilistic modeling of candidate plans and guidance from multiple metrics can improve the accuracy of ranking manufacturing decision-making plans.

After ranking the candidate information, although a decision-making plan can be directly outputted, relying solely on ranking results may not fully meet the complex requirements of manufacturing decisions. Therefore, to further enhance the accuracy and reliability of manufacturing decisions, it is necessary to conduct fusion processing on the top-ranked manufacturing decision-making plans. The core objective of the fusion approach is to overcome potential limitations of single solutions by integrating the strengths of multiple high-quality candidates, thereby providing decision-makers with more precise support [27]. Yu et al. [36] introduced a ranking information fusion model, Fusion-in-T5, which achieves information fusion by integrating features like text matching, ranking features, and global document information into a unified template while utilizing an attention mechanism. Jiang et al. [27] proposed a seq2seq-based fusion module based on a Flan-T5-XL model as a fuser, generating enhanced outputs by feeding a set of candidate commands conditioned on input commands into the fuser. Bianco et al. [37] utilized GPT-3 as a fuser for merging the two best image captions after models such as BLIP had completed their descriptions. However, these methods often introduce new models for the fusing process, potentially requiring the fuser to relearn context-specific semantics. In contrast, selecting the base model that performs best in the task as the fuser offers significant advantages, given its thorough understanding of specific contexts and direct applicability to the fusion task.

Despite existing research advancing the effectiveness of ranking and fusion methods, several limitations remain that hinder their performance in complex manufacturing scenarios. To address these challenges, this paper introduces a dynamically weighted mixture of experts ranking method alongside an adaptive fusion decision strategy, tailored to the specific needs of intelligent manufacturing decision-making. This approach aims to enhance the accuracy of manufacturing decision-making plans by effectively ranking and fusing candidate decision information.

3. Methods

In this section, we first present the general framework of LLM-MANUF and then provide a detailed description of its various components.

3.1. LLM-MANUF:A new integration framework for decision-making

The LLM-AMNUF framework consists of several fine-tuned LLMs, a DWMOE ranking method and a AFDS fusion method, as illustrated in Fig. 2. it is inputted into multiple fine-tuned LLMs $\{LLM_1, LLM_2, \cdots, LLM_n\}$, and each LLM_i generates a preliminary decision-making plan PD_i. These preliminary decision-making plans are then ranked by the DWMOE, and the final ranking of all preliminary decision-making plans is performed by defining a comprehensive ranking function $Rank(PD_i)$ and weighting the ranking based on multiple ranking metrics $\{\boldsymbol{f_1}(\boldsymbol{i}), \boldsymbol{f_2}(\boldsymbol{i}), \cdots, \boldsymbol{f_m}(\boldsymbol{i})\}$. The top-ranked decision candidates are then selected and spliced together through AFDS to form a fusion instruction, and the fine-tuned LLM that is most suitable for the task at hand is selected to serve as a fuser to generate the final manufacturing decision-making plan.

In this section, Section 3.2 outlines the fine-tuning method and the selection of LLMs employed within the framework. Section 3.3 describes the dynamically weighted mixture of experts ranking method, while Section 3.4 discusses the adaptive fusion decision-making strategy.

3.2. Preliminary Decision-Making plans generation

The main role of the fine-tuned LLMs is to explore different feature spaces to generate preliminary decision-making plans based on the decision-making demand. This process emphasizes LLMs fine-tuning for intelligent decision-making in manufacturing. Furthermore, it involves selecting multiple LLMs and generating preliminary decision-making plans. Details are illustrated in Fig. 3.

3.2.1. LORA fine-tuning method

The LLMs are pre-trained on generalized domain corpora, which equip them with strong performance in multi-task processing, such as question answering and reasoning. However, the non-specific nature of the pre-training corpus results in a deficiency in manufacturing domain knowledge. To address this gap, fine-tuning the LLM is necessary. Fine-tuning allows the model to effectively integrate and utilize expertise from the manufacturing domain, thereby enhancing its efficiency and accuracy in applications and better meeting the needs of manufacturing users.

LORA is an efficient, lightweight fine-tuning method that primarily adjusts weights through the introduction of a low-rank matrix, as illustrated in Fig. 4. Specifically, it involves freezing the weight matrix of the LLM while incorporating a trainable low-rank matrix. This low-rank matrix $\triangle W$ is represented as the product of two smaller matrices $\boldsymbol{A} \in \mathbb{R}^{d\times r}, \boldsymbol{r \ll d}, \boldsymbol{B} \in \mathbb{R}^{k\times r}, \boldsymbol{r \ll k}$:

$$\Delta W = AB^T \tag{1}$$

where $r(A) = r(B) = r$.

Given a decision-making requirement Q, it can be represented as an input feature vector $\boldsymbol{x_Q} \in \mathbb{R}^k$. This vector $\boldsymbol{x_Q}$ is then subjected to a feature mapping through a linear transformation. The fine-tuning weight matrix used in this transformation is denoted as:

$$\boldsymbol{W}^{'} = \boldsymbol{W} + \boldsymbol{\Delta W} = \boldsymbol{W} + \boldsymbol{AB^T} \tag{2}$$

The result of the linear transformation of the decision-making

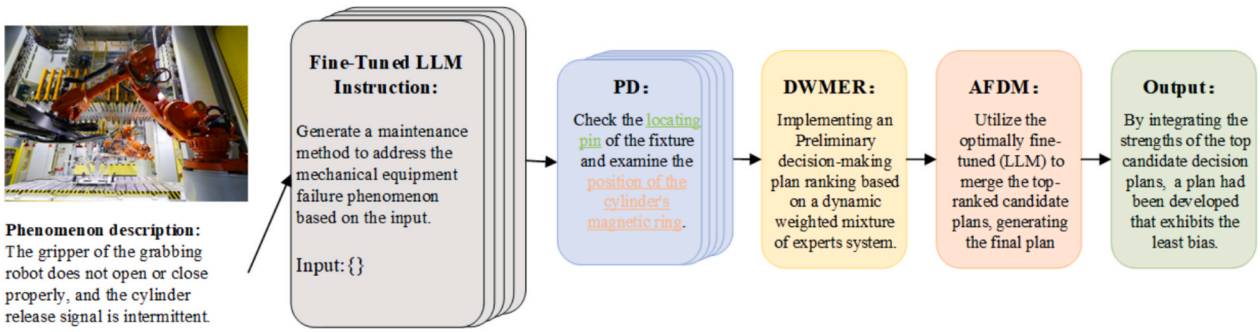


Fig. 2. Overview Of The Proposed LLM-MANUF Framework.

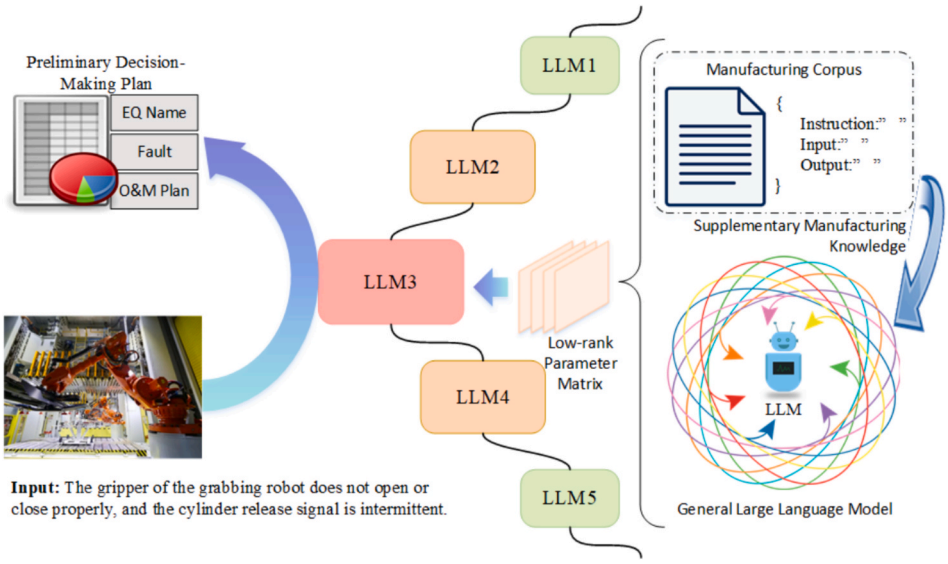


Fig. 3. Preliminary Decision-Making Plans Generation.

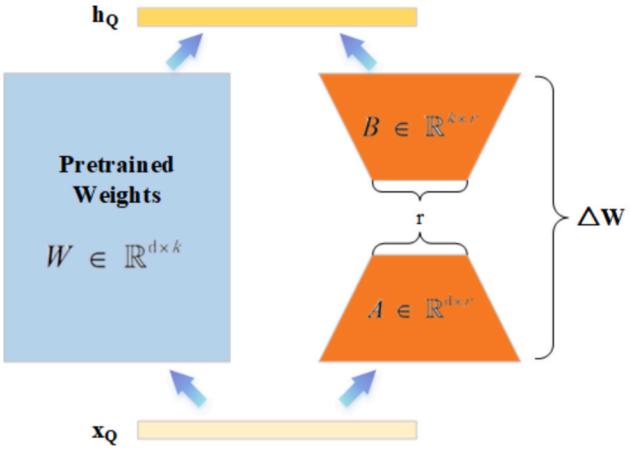


Fig. 4. Low-Rank Adaptation.

requirement Q becomes:

$$h_Q = W'x_Q = (W + AB^T)x_Q = Wx_Q + A(B^Tx_Q) \tag{3}$$

where the matrix $A(B^Tx_Q)$ can be seen as a feature adjustment to the input requirement Q, containing both the model's pre-trained knowledge and new low-rank information of manufacturing.

3.2.2. Preliminary decision-making feature acquisition

We carefully select five LLMs recognized for their performance, repeatability, scalability, and popularity as preliminary decision-making models. Due to constraints on computational resources, this integration framework adopts a smaller set of LLMs, each with a parameter scale of approximately 7B. Additionally, the number of models and their parameter scales can be flexibly adjusted based on available computational resources. This allows for accommodating integration needs for more models and diverse application scenarios, such as those with 2B, 13B, or 70B parameters. In this article, the selected models include Llama3, Qwen2.5, Gemma2, ChatGLM3, and Baichuan2. For LLMs with multiple variants, the instructed version is prioritized, followed by the chat version, and finally the base model. Regarding language preference, if an LLM is trained on an English corpus, versions fine-tuned on a Chinese corpus are preferred. The primary features of fine-tuned LLMs are shown in Table 1.

Table 1
the detail information of fine-tuned LLMs.

Model	Version	Params	Original Language
LLAMA3	Llama3-Chinese-8B-Instruct	8B	English
QWEN2.5	qwen2.5-7b-instruct	7B	Chinese
Gemma2	Gemma2-9B-it	9B	Chinese/English
CHATGLM3	chatglm3-6b	6B	Chinese
BAICHUAN2	Baichuan2-7B-Chat	7B	Chinese

The LLM-MANUF framework integrates multiple fine-tuned LLMs, each of which combines its own pre-trained general knowledge and manufacturing knowledge to deeply parse the manufacturing decision-making requirements, realize the multi-dimensional semantic capture of manufacturing knowledge, and deeply understand the complex interconnections of manufacturing systems. At the same time, the parallel fine-tuned LLMs perform multi-dimensional feature transformation and extraction in multiple feature subspaces, obtain relevant manufacturing decision-making information in different feature subspaces, and generate preliminary decision plans $\{PD_1, PD_2, \cdots, PD_n\}$ that are interrelated but have different decision-making information.

3.3. Dynamic weighted mixture of experts ranking method

The main objective of the DWMOE is to rank the preliminary decision-making plans $\{PD_1, PD_2, \cdots, PD_n\}$ generated by LLMs that have been fine-tuned in the manufacturing domain. The DWMOE ranking method is based on modeling the probability of the best candidate to obtain the candidate ranking. Semantic feature extraction is realized by using RoBERTa semantic parser, probability distribution modeling is realized based on dynamic weighted mixture of expert, and candidate ranking is realized under the guidance of multi-metric header. The details are shown in Fig. 5.

The semantic parsing component primarily includes a semantic parser S and a nonlinear feature processor M. Pre-trained models, such as BERT, RoBERTa, and DeBERTa, have proven effective in enhancing the quality of semantic features. In this study, RoBERTa is chosen as the semantic parser S. The features captured by the semantic parser undergo nonlinear feature extraction, facilitating the capture of complex feature relationships to produce more representative feature representations.

To enhance the semantic parser S ability to derive superior semantic features and fully utilize contextual information. Decision requirements Q and preliminary decision-making plans are concatenated to form input

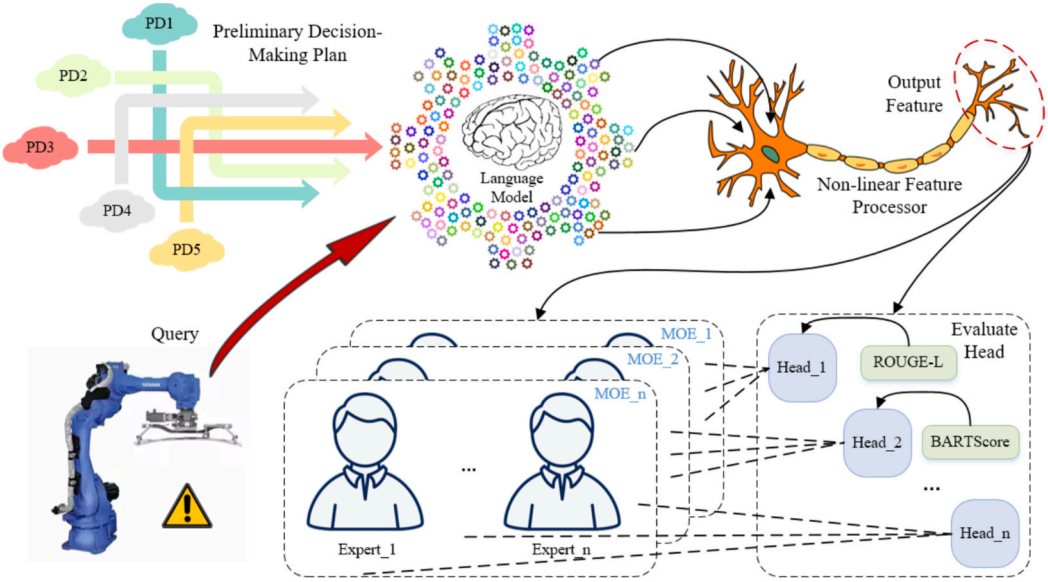


Fig. 5. Dynamically Weighted Mixture of Experts.

representations $x_{input} = [Q; PD_i]$. These representations, structured as "[CLS] Requirement [SEP] Preliminary Decision-Making", are then fed into the semantic parser S. The feature representations $h_{RoBERTa}(x_{input})$ of the final layer is extracted and processed by a multilayer perceptron (MLP) for nonlinear feature extraction. The nonlinear feature can be represented $x_{hidden} = MLP(h_{RoBERTa}(x_{input}))$. By integrating the decision requirement with preliminary decision-making plans, the pre-trained model can fully explore the associative semantics, resulting in richer semantic representations compared to scenarios lacking this integration.

The mixture of experts system component primarily comprises a sparse mixture of experts and multi-metric heads. Evaluating the quality of a preliminary decision-making plan is multifaceted. Relying solely on a single evaluation metric, such as ROUGE-L, is insufficient for comprehensive assessment. Employing multi-metric heads for evaluation effectively mitigates the bias associated with using a single index. Additionally, to accommodate diverse decision-making needs, and to enhance the robustness and generalization of the entire ranking method, the sparse mixture of experts is introduced in this ranking method.

Two metrics, ROUGE-L and BARTScore, are selected as metric heads H_R, H_B for the ranking method. ROUGE-L evaluates similarity through the longest common subsequence between texts, considering both word frequency and order. BARTScore is based on the generative model BART, utilizing a distinct word-matching mechanism to directly assess text quality. Two indicators employ different evaluation mechanisms. Thus, when applied to the same text, they offer complementary insights, enhancing the comprehensiveness and accuracy of the evaluation.

For a metric head H, a set of sparse mixture of experts model $E_H = \{e_{H,1}, e_{H,2}, \cdots, e_{H,n}\}$ is introduced, each comprising a total of n experts. To dynamically select the most suitable experts during ranking, we employ an attention mechanism that assigns contribution weights to each expert. For each metric head, the attention weights are computed based on the hidden feature x_{hidden}. Specifically, for a metric head H, each expert $e_{H,i}$ in its corresponding expert set receives a weight $e_{H,i}$, expressed as follows:

$$\alpha_{H,i} = \frac{\exp(score(x_{hidden}, e_{H,i}))}{\sum_{j=1}^{n} \exp(score(x_{hidden}, e_{H,i}))} \tag{4}$$

where $score(\cdot,\cdot)$ is the evaluation function in the attention mechanism, used to measure the relevance of the hidden feature x_{hidden} to each expert $e_{H,i}$, and the dot product is used here for evaluation and scoring:

$$score(x_{hidden}, e_{H,i}) = x_{hidden}^{T} W_{e_{H,i}} \tag{5}$$

where $W_{e_{H,i}}$ is a weight matrix.

Immediately after calculating the weights of each expert, the outputs of the experts are weighted and summed to obtain the score s_H of the metric head H:

$$s_H = \sum_{i=1}^{n} \alpha_{H,i} e_{H,i}(x_{hidden}) \tag{6}$$

Based on the above equation and the selected metrics, the scores of the two metric heads $s_{ROUGE-L}, s_{BARTScore}$ can be derived and the corresponding loss function can be expressed as:

$$\mathscr{L} = \lambda_{ROUGE-L}\mathscr{L}_{ROUGE-L}(s_{ROUGE-L}, y) + \lambda_{BARTScore}\mathscr{L}_{BARTScore}(s_{BARTScore}, y) \tag{7}$$

where $\mathscr{L}\ominus$ is a loss function, $\lambda_{ROUGE-L}, \lambda_{BARTScore}$ are the weights of the two metric heads, $\mathscr{L}_{ROUGE-L}, \mathscr{L}_{BARTScore}$ are the metric loss functions.

Finally, based on the weights $\lambda_{ROUGE-L}, \lambda_{BARTScore}$ of the above metric heads, the scores $s_{ROUGE-L}, s_{BARTScore}$ of the two metric heads are weighted:

$$s = \frac{\lambda_{ROUGE-L} s_{ROUGE-L} + \lambda_{BARTScore} s_{BARTScore}}{2} \tag{8}$$

where s is the metric score for each preliminary decision-making plan based on which the preliminary decision-making plan are ranked.

3.4. Adaptive fusion decision strategy

The main purpose of AFDS is to adaptively select the appropriate model as a fuser to realize the advantageous fusion of the best solutions based on the best candidate solutions to arrive at the final decision-making plan. The AFDS aims to dynamically select the most suitable model for fusing candidate solutions, thereby facilitating an effective decision-making process. By identifying and integrating the best candidate solutions, AFDS enhances the quality of the final decision-making plan and mitigates manufacturing decision-making bias.

Initially, the top K candidate decision-making plans from the decision ranking module are fused. These superior candidate plans often possess complementary strengths or exhibit similar trends in certain aspects. By merging these schemes, we can retain their respective

advantages while effectively mitigating the weaknesses of individual plans, thereby enhancing the overall decision plan's accuracy and robustness.

Unlike the approach of redesigning or selecting new models proposed in some literature [27,31], this paper employs an adaptive fusion decision-making strategy. It directly uses the first-ranked fine-tuned LLM in the ranking method as a fuser, aiming to fully leverage the effectiveness of existing superior models. Specifically, the inputs for the fuser comprise the decision requirements Q and the set of top K candidate decision-making plans $\{CD_1, CD_2, \cdots, CD_n\}$ (e.g., K = 2). The final decision-making plan is generated by sequentially integrating inputs and candidate solutions, delineated by identifiers (e.g., '[candidate_id]'), followed by the rational optimization of instructions for the LLM. This process guides the fuser to effectively extract and utilize pertinent information from the optimal candidate solution.

4. Case study

4.1. Dataset

The operation and maintenance (O&M) of production equipment serves as a case study to evaluate the effectiveness of the LLM-MANUF integration framework in intelligent decision-making within the manufacturing field. The O&M dataset constructed in this paper is derived from the O&M records of workshop equipment at an automobile manufacturing workshop in Chongqing, China. The records span three years and consist of 23 k entries, encompassing 113 types of equipment and 303 types of components. High-frequency fault components include robots, inductors, and cylinders, with typical equipment faults summarized in Table 2. Specifically, robot faults, which account for 9.25 % of entries, include issues such as no robot action and axis deviation errors, totaling 2,191 entries. Inductor faults, primarily characterized by the inability to detect workpieces, total 1,237 entries, representing 5.22 %. Cylinder faults, including instances where the cylinder does not open correctly and the cylinder signal flashes off, also total 1,237 entries, accounting for 5.22 %. Additionally, there are 563 entries related to other cylinder faults, comprising 2.38 %.

Since the workshop equipment operation and maintenance records are semi-structured data, they cannot be directly utilized by the framework and require pre-processing. Each column in the records corresponds to a specific type of information, encompassing, but not limited to, the equipment name, various functional groups (VFGs), fault phenomena, and operations and O&M measures, as depicted in Fig. 6. The VFG refers to the primary function module of the equipment; for example, the VFG of a gripper robot is the gripper unit, the VFG of a spot welding robot is the welding clamp, and the VFG of an arc welding robot is the gun head.

Given that workshop equipment operation and maintenance records consist of semi-structured data, their formats are incompatible with existing frameworks, necessitating data preprocessing. In this study, a generalized LLM processes four types of key information, including the construction of equipment-component hierarchies and the establishment of relationships between equipment and failure phenomena, as illustrated in Fig. 7. Utilizing a gripper robot as an example, its VFG is identified as "gripper, " with a fault phenomenon described as "not opening or closing; cylinder release signal intermittent. " This information is integrated by the LLM to generate a semantically coherent sentence: "The gripper of the robot does not open or close properly, and the cylinder release signal is intermittent. " Subsequently, O&M measures corresponding to the fault phenomenon are extracted, and "fault-O&M " pairs are constructed in Alpaca format, culminating in the formation of a production equipment O&M dataset. Related details are presented in Table 3.

To standardize the feature variability within the dataset, several measures have been implemented, including the removal of samples with missing information and the elimination of excessively long texts in O&M measures or fault descriptions. These steps ensure that the length of all fault-O&M pairs remains relatively consistent and semantically coherent.

Table 2
Typical equipment fault in the vehicle manufacturing process.

Order	Fault Phenomenon	Number	Percentage
1	cylinder failure	563	2.38 %
2	sensor failure	1237	5.22 %
3	robot failure	2191	9.25 %
4	component interference	695	2.93 %
5	mold and fixture switching failure	827	3.49 %

4.2. Evaluate metrics

Indicators BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L and F1 are used to evaluate the effectiveness of the LLM-MANUF framework. BLEU-4 is a metric that evaluates text generation quality by calculating the n-gram similarity, specifically for n-grams ranging from 1 to 4 words, between the generated text and the reference text. The formula for BLEU is as follows:

$$BLEU = BP \times \exp(\sum_{n=1}^{N} \omega_n \log p_n) \tag{9}$$

where the n-gram accuracy of the generated decision-making plan relative to the reference decision-making plan is denoted as $\boldsymbol{p_n}$. Here, $\boldsymbol{\omega_n}$ represents the weight of the n-gram, and N signifies the maximum n-gram length, which is 4 in this paper. BP stands for the brevity penalty, calculated as the ratio of the lengths of the generated decision scheme to the reference decision-making plan.

ROUGE-1, ROUGE-2, and ROUGE-L are metrics from the same series used to assess text similarity. While ROUGE-1 and ROUGE-2 are similar in their calculation methods, the key difference lies in their focus: ROUGE-1 evaluates the overlap of 1-grams (individual words), whereas ROUGE-2 examines the overlap of 2-grams (pairs of consecutive words). The expression for ROUGE-N is provided below:

$$Rouge - N = \frac{\sum_{S \in \{label\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{label\}} \sum_{gram_n \in S} Count(gram_n)} \tag{10}$$

where $\{label\}$ represents all labels, which refer to the reference decision-making plan. $Count_{match}(gram_n)$ denotes the n-grams that appear in both the generated decision-making plan and the reference decision-making plan. $Count_{match}(gram_n)$ indicates all possible n-grams within the reference decision-making plan.

ROUGE-L evaluates the similarity between the generated text and the reference text by examining the Longest Common Subsequence (LCS). The calculation uses the following expression:

$$Rouge - L = \frac{LCS(RD, label)}{m} \tag{11}$$

where RD represents the generation decision-making plan, and m denotes the length of the reference decision-making plan.

The F1 score is the reconciled average of precision and recall and is commonly used to measure the accuracy of a classification model. Its expression is:

$$\boldsymbol{F1} = \frac{\boldsymbol{2 \cdot TP}}{\boldsymbol{2 \cdot TP + FP + FN}} \tag{12}$$

where TP denotes a positive example that was correctly predicted to be a positive example. FP denotes a positive example that was predicted to be a negative example in the prediction. TN denotes a negative example that was correctly predicted to be a negative example. FN denotes a negative example that was predicted to be a positive example in the

Equipment Name	VFG	Fault Phenomenon	O&M measures
gripper robot	gripper	not open or close, cylinder release signal intermittent	Adjust the magnetic ring to the optimal position.
arc welding robot	wire feeder	clogged contact tip	Clear the clogged nozzle and resume wire feeding.
process equipment	jig	the front beam pops out	Reopen the clamp and install the plate component.

Fig. 6. Workshop Equipment Operation and Maintenance Records.

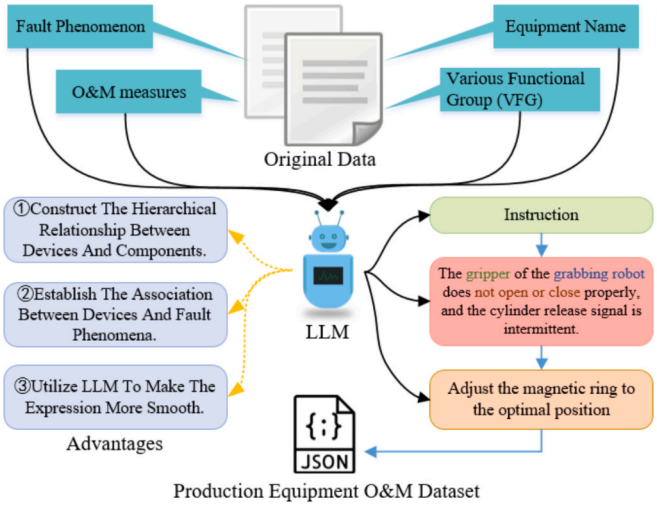


Fig. 7. Data pre-processing.

Table 3
Instruction fine-tuning dataset format.

Order	Data
Insturction	Generate an operation and maintenance method to address the mechanical equipment failure phenomenon based on the input.
1	Input: The gripper of the grabbing robot does not open or close properly, and the cylinder release signal is intermittent.
2	"Input": "The arc welding robot experiences wire tangling in the wire feeder due to a clogged contact tip."
3	"Input": "The fixture switches to the post-cut storage in advance, and when the fixture reaches the post-cut storage and waits for the workstation to complete, the front beam pops out."

prediction.

4.3. Parameter settings and hardware

All the fine-tuning of the LLM is implemented through the llama_-factory [38] tool, and the decision ranking module is based on the pytorch framework. LORA is used for fine-tuning, the maximum length cutoff_len of each input sample is 1024 tokens, the batch size of training on each device batch_size is 1, the number of gradient accumulation steps is 8, and the learning rate learning_rate is 0.0001. The learning rate is gradually increased using a cosine annealing learning rate scheduler with a warm-up phase of 10 %.

All the experiments in this paper are performed on a server with Ubuntu 20.04.6 LTS as the server operating system, AMD 5995WX as the CPU, 8 × 32 GB of RAM, and 4 × NVIDIA A6000 48 GB as the GPU.

4.4. Experiment results

Relevant experiments are conducted to verify the effectiveness of the LLM-MANUF framework proposed in this paper, which integrates the advantages of multiple LLMs. The specific results are detailed in Table 4.

4.4.1. LLM-MANUF framework integrating the strengths of all parties

As shown in Table 4, after LoRA fine-tuning, the selected 7B-scale models—Llama3, Qwen2.5, Gemma2, Baichuan2, and Chatglm3—demonstrated strong performance on the production equipment O&M dataset's test set. Regarding text quality metrics, the Qwen2.5 model achieved the highest BLEU-4 score of 73.83, whereas the Chatglm3 scored the lowest with a BLEU-4 of 66.87, showing a difference of 6.96. For decision accuracy measured by F1 score, Qwen2.5 again led with an F1 value of 83.98, while Baichuan2 had the lowest F1 score at 77.81. The LLM-MANUF framework outperformed all individual models, achieving BLEU-4 and F1 scores of 82.55 and 92.03, respectively. Compared to Qwen2.5, the best-performing single fine-tuned model, the LLM-MANUF framework improved the BLEU-4 score by 8.72 points and the F1 score by 8.05 points. This indicates that our proposed LLM-MANUF framework not only generates higher-quality manufacturing decision schemes but also ensures high decision accuracy.

To further assess the framework's performance, we incorporated full-parameter fine-tuning in our experiments. Generally, full-parameter fine-tuning exhibits greater effectiveness compared to LoRA fine-tuning, as detailed in Table 4. Specifically, all models showed improved performance under full-parameter fine-tuning relative to LoRA fine-tuning, with Qwen2.5 achieving top rankings across all metrics—ROUGE-1, ROUGE-L, and F1 scores. Conversely, Chatglm3 maintained the lowest performance levels. By employing fully fine-tuned LLMs instead of those fine-tuned via LoRA, we constructed the LLM-MANUF(full) framework. Although the score improvements in LLM-MANUF(full) were not as pronounced as in the LLM-MANUF (LoRA) framework, its final scores notably surpassed those of the LoRA-based framework, achieving a BLEU-4 score of 83.37 and an F1 value of 93.73.

4.4.2. LLM-MANUF framework challenging the limits of high parameters

In general, LLMs with a large number of parameters tend to outperform those with fewer parameters. However, the limits imposed by parameter magnitude are not insurmountable [39]. To illustrate the efficacy of the LLM-MANUF framework in overcoming high parameter limits, we incorporate LLMs with 13B and 30B parameters into our experiments. Specifically, we consider the Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, and Llama2-Chinese-13B-Chat-ms models. Notably, since the LLM3 model lacks a 13B version, we use the preceding Llama2 model from the same family. To minimize the effects of version differences, we also include the 7B-level Llama2 in our experiments.

As shown in Table 4, the Llama2-7B model is less effective than Llama3 under both LORA fine-tuned and full-parameter fine-tuned conditions. In contrast, the Llama2-13B model outperforms both Llama3-7B models in terms of effectiveness. This finding indicates that while Llama2 may not match Llama3 at the 7B level, increasing the number of parameters can mitigate performance differences across model versions. The impact of parameter levels on LLM performance has been highlighted.

Table 4
Experiment Results.

Model	Fine-tuning	Param	Version	BLEU-4	ROUGE-1	ROUGE-2	ROUG E-L	F1
Llama3	LORA	8B	Chinese-Instruct	69.29	73.32	65.59	73.22	81.84
Llama3	Full	8B	Chinese-Instruct	71.94	74.87	69.25	74.83	85.89
Llama2	LORA	7B	Chinese-Chat-ms	64.99	70.11	58.78	70.05	79.13
Llama2	Full	7B	Chinese-Chat-ms	69.01	70.83	64.23	70.82	83.38
Llama2	LORA	13B	Chinese-Chat-ms	73.03	77.28	67.88	77.26	86.62
QWEN2.5	LORA	7B	Instruct	73.83	77.52	70.08	77.46	83.98
QWEN2.5	Full	7B	Instruct	74.53	78.23	71.86	78.00	87.41
QWEN2.5	LORA	14B	Instruct	78.05	79.75	73.21	79.67	88.40
QWEN2.5	LORA	32B	Instruct	**83.46**	84.92	**81.07**	84.90	92.44
Gemma2	LORA	6B	Instruct	68.36	70.95	63.46	70.89	80.01
Gemma2	Full	6B	Instruct	71.72	73.87	69.67	73.76	82.93
Chatglm3	LORA	6B	Base	66.87	70.72	64.40	70.46	78.29
Chatglm3	Full	6B	Base	69.57	72.58	67.26	72.48	80.45
BAICHUAN2	LORA	7B	Chat	63.88	68.24	59.23	68.18	77.81
BAICHUAN2	Full	7B	Chat	75.62	77.81	71.92	77.79	81.50
LLM-MANUF + Llama3				79.25	82.54	75.70	82.51	89.69
LLM-MANUF + Llama3*				77.45	80.73	73.85	80.66	87.09
LLM-MANUF + MLM				77.21	79.58	73.25	79.44	86.79
LLM-MANUF + SimCLS				79.59	82.00	75.90	81.89	89.07
LLM-MANUF (ours-lora)				82.55	85.14	77.67	84.94	92.03
LLM-MANUF (ours-full)				83.37	**85.35**	80.58	**85.33**	**93.73**

Note: * indicates that the fuser has not been fine-tuned.

Given that the Qwen2.5 series spans parameter sizes from 7B to 30B, their performance comparison is particularly illustrative. As indicated in Table 4, the Qwen2.5-14B and Qwen2.5-32B models exhibit strong performance, with the Qwen2.5-14B achieving a BLEU-4 score of 78.05 and an F1 value of 88.40, while the Qwen2.5-32B scores 83.46 in BLEU-4 and 92.44 in F1. Notably, our LLM-MANUF framework surpasses the 14B models (be it Llama2 or Qwen2.5) in performance metrics. Regarding the 32B models, the LLM-MANUF framework shows advantages in ROUGE-1, ROUGE-L, and F1 metrics, while its BLEU-4 and ROUGE-2 scores are closely comparable to those of the Qwen2.5-32B. Overall, the LLM-MANUF framework demonstrates superior results in overcoming high parametric limitations.

Overall, the LLM-MANUF framework effectively surpasses the performance of 14B LLMs and demonstrates superior results in overcoming high parameter limits.

4.4.3. *LLM-MANUF framework as A Whole*

The LLM-MANUF framework integrates multiple fine-tuned LLMs, ranking methods, and adaptive fusion decision strategies, each of which is essential for its functionality. Beyond evaluating the fine-tuned LLMs, assessing the sequencing methods and fusion strategy is equally important. Table 4 illustrates that substituting the ranking method with MLM [40] and SimCLS [32] allows the LLM-MANUF framework to maintain satisfactory accuracy; however, its performance improvement does not match that achieved with the DWMOE method, showing a difference in BLEU-4 and F1 scores of nearly 3 points. The adaptive fusion decision strategy, despite being a minor modification, eliminates the need for designing a new model while simultaneously enhancing the overall performance of the framework. Fixing the fuser, such as using llama3 without adjustments, inevitably decreases the framework's effectiveness. Moreover, whether the llama3 model serving as the fuser is fine-tuned impacts the framework's performance. Specifically, the BLEU-4 score using the original llama3 was 1.8 points lower than its fine-tuned counterpart, and the F1 score showed a decrease of 4.94 points.

4.5. *Discussion*

4.5.1. *Detailed Decision-Making plan requires A professional model*

To thoroughly analyze the differences between models in intelligent decision-making for manufacturing, Table 5 presents the responses of each LLM model in specific cases, providing a visual representation of the details. The left side of the table displays the Chinese responses, while the right side contains the corresponding English translations. This approach reflects the reality that engineers typically do not use English for inquiries in practical applications. Thus, the English part of the experiment does not represent the model's direct output.

In the specific case examined, the fault phenomenon indicates that the cylinder signal flashes off. All major models can generate corresponding O&M decision-making schemes for this issue. However, there are notable differences in detail when compared to the schemes produced by the LLM-MANUF framework. Whether it is ChatGPT-4o, Claude3.5, or Qwen2, the O&M decision-making schemes generally encompass aspects such as the cylinder air source, solenoid valve, signal line, control system, and wear interference. But these schemes only offer broad operational suggestions, which provide limited assistance in actual equipment operation and maintenance.

The individual decision-making models within the LLM-MANUF framework have been fine-tuned using a corpus of manufacturing domain knowledge. Thus, more targeted expertise have been acquired by LLMs. As illustrated in the Table 5, the LLM-MANUF framework offers more specific O&M measures for the cylinder signal flashing issue, including checking the cylinder positioning pin for interference, examining the position of the magnetic ring, and adjusting the inductor. While these solutions overlap with those proposed by generic large models, the LLM-MANUF framework presented in this paper is more practical for real-world operational processes.

4.5.2. *The Extent of A Decision-Maker's contribution*

To effectively demonstrate the contribution of each preliminary decision-making model to the LLM-MANUF framework, we analyze the decision participation of each model in the experiment. Specifically, we measure the proportion of preliminary decision-making plans generated by these models that ranked in the top two positions, as shown in Table 6. The results indicate that both Llama3 and Qwen2.5 have a high percentage of top-two votes across the entire experiment. Their shares exceed 25 % in both Rank 1 and Rank 2. Notably, Qwen2.5 accounts for 34.03 % in Rank 1. By combining these results with those of the two fine-tuned models presented in Table 4, we observe that better-performing individual fine-tuned LLMs contribute more significantly to the LLM-MANUF framework. Although the Chatglm3 model underperforms both individually and within the framework, its vote share indicates that it still contributes to the overall performance enhancement of the LLM-MANUF framework.

Table 5
O&M Case Study.

Question	Instance
	Input: The gripper of the grabbing robot does not open or close properly, and the cylinder release signal is intermittent.
ChatGPT 4o	1. Check the air supply: Ensure that the air pressure is normal and that there are no leaks or blockages in the air pipes to guarantee that the cylinder receives adequate power. 2. Inspect the solenoid valve: Verify the working condition of the solenoid valve to ensure it can switch normally and control the cylinder's movement. Clean or replace the solenoid valve if necessary. 3. Adjust the cylinder: Check for wear or internal leakage in the cylinder and perform lubrication, cleaning, or replacement as needed. 4. Signal line maintenance: Check the stability of signal line connections to eliminate issues caused by poor contact that may lead to signal interruptions. 5. Control system check: Diagnose the control system to ensure correct program settings and absence of erroneous commands that could cause the gripper to malfunction. 6. Mechanical inspection: Examine the mechanical components of the gripper robot for any jamming or obstructions, and maintain cleanliness and lubrication of the gripper parts.
Claude3.5	1. Check the pneumatic system to ensure normal air pressure and no leaks in the lines. 2. Inspect the cylinder and solenoid valve for proper operation, cleaning any dirt or foreign objects if present. 3. Examine the gripper mechanism for any binding, and lubricate if necessary. 4. Check related sensors and wiring connections to rule out signal transmission issues. 5. Adjust the position of cylinder limit switches to ensure stable signals. 6. Review the control program to eliminate any logical errors. 7. If the problem persists, consider replacing relevant components.
QWEN2	1. Inspect Cylinder Connections: Firstly, inspect the connection between the cylinder and the gripper to ensure the cylinder can properly drive the gripper. Check if the mounting screws of the cylinder are loose and if the connection components between the cylinder and gripper are worn out or damaged. 2. Pressure Check: Verify the air circuit for normalcy, including whether the air supply pressure meets the requirements, and if the air tubes are blocked or leaking. Use a pressure gauge to ensure the air pressure is stable and meets the equipment's requirements. 3. Sensor Verification: Check if the sensors controlling the cylinder's action are functioning properly. Sensors may fail to send accurate signals due to dust, dirt, or damage. Clean the sensor surfaces, ensure their mounting positions are correct, and calibrate or replace them if necessary. 4. Inspect Control System: Verify the programming settings of the PLC (Programmable Logic Controller) to ensure control signals are correctly sent. Check if the PLC's input and output points are functioning properly and if the signal lines are reliably connected. 5. Maintenance and Lubrication: Check the lubrication status of the gripper and cylinder to ensure moving parts are adequately lubricated to reduce friction and wear. 6. Professional Maintenance: If the above steps do not solve the problem, it is recommended to contact professional technicians for a more thorough inspection and maintenance. It may be necessary to disassemble the cylinder and gripper to check if internal parts are damaged.
LLM-MANUF	1. Check the Positioning Pins and Fixtures: Carefully inspect the position of the fixture positioning pins to ensure they are fully in place. Use appropriate tapping tools to gently adjust the positioning pins, avoiding excessive force that could cause damage. 2. Adjust the Cylinder Magnet Ring: Ensure the cylinder magnet ring is in the correct position and adjust it according to the equipment manual. Check for wear on the magnet ring and replace it if necessary. 3. Equipment Reset and Startup: After adjustments are complete, reset the equipment according to operating procedures. Confirm all safety measures are in place before starting the equipment for testing. 4. Manual Recovery Operations: In cases where the automatic recovery function fails, adopt manual recovery methods. When performing manual operations, strictly adhere to safety operating procedures. 5. Adjust the Sensor Position: Ensure the sensor position is accurate and calibrate if necessary. Regularly check the sensitivity and response time of the sensor to ensure its normal function. 6. Documentation and Preventive Maintenance: Record the details of each maintenance operation, including identified issues, actions taken, and outcomes. Develop a preventive maintenance plan to regularly inspect and service key components.

Table 6
Preliminary Decision-Making Model Integration Involvement.

Ranking	Preliminary Decision-Making Model				
	Llama3	Qwen2.5	Gemma2	Chatglm3	Baichuan2
Rank1	25.62 %	34.03 %	15.64 %	8.87 %	15.84 %
Rank2	32.80 %	28.15 %	19.54 %	9.11 %	10.40 %

4.5.3. Better prompts enhance performance

During the fine-tuning of LLMs, tailored instructions can optimize the models for specific tasks. In specialized domains like manufacturing, effective instructions enable the model to concentrate on industry-specific knowledge and task requirements. We adjust the instructions in the constructed manufacturing product equipment O&M dataset, as illustrated in Table 7. Following this adjustment, individual LLMs are fine-tuned using the modified dataset. The preliminary decision-making models, adjusted with new instructions, demonstrate notable improvements across various decision-making task metrics. For instance, the LLAMA3 model demonstrates an increase of 0.65 points in the BLEU-4 score, achieving a total score of 69.94. Additionally, it records a Rouge-1 score of 74.02, a Rouge-2 score of 67.80, and a Rouge-L score of 73.94.

To more intuitively demonstrate the impact of prompt optimization on framework performance, this paper analyzes several examples, with detailed case information provided in Table 8. The analysis reveals that an optimized prompt not only facilitates more effective learning from the fine-tuned corpus but also better leverages the pre-existing knowledge of the large language model (LLM), thereby enhancing the robustness of the LLM-MANUF framework.

Regarding the LLM-MANUF framework, its performance also improved in tandem with the enhanced fine-tuned preliminary decision-making model, consistent with patterns observed in previous experiments. These experimental results indicate that for different manufacturing decisions-making tasks, customized instructions can be designed to meet the specific needs of decision-making tasks and data presentation characteristics. This approach enables the LLMs to thoroughly comprehend task requirements and focus on the pertinent knowledge embedded in the fine-tuning data.

4.5.4. Time required to generate a response

The time required for decision-making plans generation is a critical factor in the manufacturing decision-making process. We evaluated the average decision-making plan generation times for both the LLM-MANUF framework and fine-tuned LLMs, with the results presented in Fig. 8. The decision generation time for a single LLM increases as the number of parameters grows. Specifically, within the parameter range of 7B to 14B, the increase in decision generation time is relatively modest, rising from approximately 18 to 21 s. However, when the number of parameters reaches 30B, the decision generation time escalates significantly, reaching nearly nine times that of the 7B parameter-level LLMs. Notably, while the LLM-MANUF framework requires about 6 times

Table 7
Customized Instruction.

Order	Improved Prompt
Role	Mechanical Engineering Specialist
Background	Users need a system that can provide decision support for equipment operation and maintenance to ensure efficient and stable operation of equipment, as well as quick response and resolution when problems are encountered.
Profile	You are a senior mechanical engineering expert with extensive equipment management experience and deep technical knowledge to provide specialized solutions to a wide range of equipment problems.
Skills	You have key competencies in troubleshooting, preventive maintenance, emergency response and decision-making, and are familiar with the linkages between complex machinery and equipment, enabling you to provide comprehensive O&M support to users.
Goals	Provide a complete equipment O&M decision-making program, including failure analysis as well as equipment O&M decision-making program.
Constrains	Decision making solutions must be based on the latest industry standards and best practices.
OutputFormat	Provide a structured manufacturing decision solution that includes a fault description, probable cause, recommended solution, and repair steps.
Workflow	1. infer possible causes of the failure based on the failure description. 2. provide solutions and repair steps for each possible cause. 3. advises the user on how to proceed with failure verification and subsequent maintenance measures.
Examples	Example 1: Fault description: The robot does not move while BS09R4 is resharpening. Possible cause: Inconsistent signal feedback to PLC when robot is resharpening (PLC is signal 1, robot feedback 56). Solution: Manually operate the robot to trim and optimize the robot trimming program. Example 2: Fault description: MT12L7 gripper cylinder sensor shifted, did not sense the part. Possible cause: The sensor is not sensing the incoming part. Solution: Manually adjust the sensor and optimize the sensor position. Example 3: Fault description: BC02R7 robot does not release parts and waits for PLC cycle completion signal. Possible cause: Interaction signal with plc flashes off. Solution: Initialize robot.

longer than the 7B parameter-level LLM (111.28 s), it still exhibits a timeliness advantage over the 30B parameter-level LLM.

5. Conclusion

In the modern manufacturing industry, decision-making processes span the entire product life cycle, from design to production, and can significantly enhance both efficiency and productivity. However, as manufacturing systems grow in complexity, traditional decision-making methods often struggle to address these intricate, interrelated challenges, leading to decision-making biases. Moreover, the variability in expertise among decision-makers exacerbates this issue.

To mitigate manufacturing decision-making bias, this paper introduces LLM-MANUF, an intelligent decision-making framework that integrates fine-tuned LLMs specialized in the manufacturing domain. LLM-MANUF aims to enhance the comprehension and parsing of manufacturing decision-making requirements by leveraging the advanced contextual semantic reasoning capabilities and pre-trained knowledge base of LLMs. This approach not only minimizes reliance on expert-driven traditional corpora but also promotes more objective and consistent decision-making. The framework incorporates a DWMOE method, which effectively prioritizes candidate options through probabilistic modeling guided by multi-metric evaluation indices. This process ensures that less relevant options are excluded, thereby aligning

Table 8
Impact of Prompt Changes on Decision-Making Plan.

Input	Pre-Prompt	Post-Prompt
FB05R6 robot welding alarm, no current alarm.	1. Check and adjust welding parameters 2. Check the welding gun and wire feeding mechanism, clean or replace damaged parts 3. Calibrate or replace sensors 4. Check the connection line 5. Review PLC program 6. Evaluate and improve working environment	1. Check welding torch and cables 2. Check gas lines for leaks or blockages 3. Check welding parameter settings 4. Check that the wire passes smoothly through the wire feed mechanism and is not tangled or blocked 5. Check sensor and feedback system 6. Check software and control systems 7. Check all moving parts, including arms, joints, etc., to make sure they are running smoothly with no jams or unusual noises 8. Check and lubricate the necessary mechanical parts 9. Check whether the working environment meets the requirements, such as temperature, humidity, etc.
No signal from the sensor when the robot is grasping the gun, no movement of the robot.	1. Check the connection cable between the sensor and the robot to make sure it is not loose, broken or damaged. 2. Check the power supply voltage of the sensor 3. Test the sensor function 4. Ensure that the sensor is installed in the correct position, and the detection range of the sensor can cover the change of robotic disk surface 5. Check whether the sensitivity of the sensor is set appropriately. 6. Check the control system 7. Check whether the mechanical movement of the changeover disk is normal.	1. Check the working voltage and output signal of the sensor, and replace the sensor in time if there is any problem. 2. Manually adjust the position of the sensor to ensure that it can accurately detect the target on the surface of the robbing disk, the use of appropriate fixtures to fix the sensor 3. Check the connection cable from the sensor to the control system. 4. Review the PLC program 5. Check the condition of the changeover disk and other related mechanical parts. 6. Add appropriate amount of lubricant to the parts that need lubrication to ensure smooth movement of mechanical parts; 7. Tighten the loose parts to ensure their normal operation.

decisions more closely with actual needs. Furthermore, LLM-MANUF employs an AFDS method, dynamically selecting the most appropriate LLM for specific manufacturing scenarios. By integrating the strengths of multiple top-ranked candidates, LLM-MANUF generates robust manufacturing decision-making solutions, improving the accuracy and reliability of the decision-making process. Ultimately, LLM-MANUF mitigates decision-making bias arising from system complexity and varying levels of expert proficiency. We select production equipment O&M as a case study and developed a corresponding dataset for experiment. The results demonstrate that the proposed LLM-MANUF framework can produce more accurate decision-making schemes than

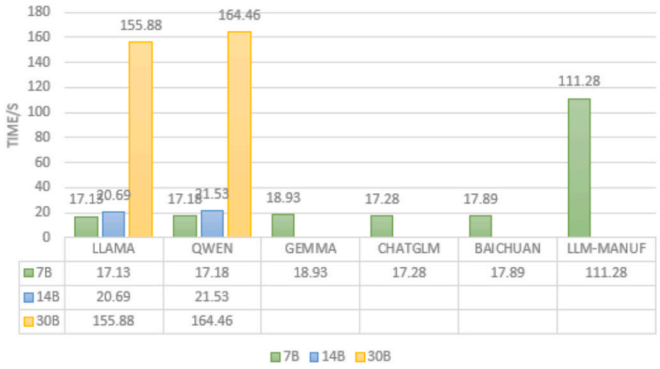


Fig. 8. The time spent on decision-making.

individual fine-tuned models, achieving the performance of higher parameter-level LLMs with reduced computational demands.

Despite the framework's superior performance in managing production equipment O&M cases, it exhibits certain limitations. Firstly, incorporating more advanced LLM architectures could enhance its efficiency, effectiveness, and modal coverage. Secondly, this study focused solely on textual information, neglecting other modalities that could enrich the decision-making process. Future improvements might include:

(1) Integrating more advanced LLM models, such as Deepseek R1, Qwen2.5-Max, and Llama3.3, aims to reduce computational requirements, enhance performance, and support broader modalities.
(2) Implementing intelligent decision-making processes based on multimodal information, including images, graphics, and text. Leveraging LLMs as core feature extraction tools enables deep feature extraction and characterization across various data modalities through their robust semantic comprehension and cross-modal feature extraction capabilities. This forms modality-specific manufacturing decision-making information. Subsequently, the integration framework fuses multimodal decision-making information, capitalizing on the complementarity and correlations between different modalities to generate a comprehensive decision-making plan, thereby achieving knowledge-driven, multimodal intelligent decision-making for manufacturing.

CRediT authorship contribution statement

Kaze Du: Writing – review & editing, Writing – original draft, Methodology. **Bo Yang:** Supervision, Funding acquisition. **Keqiang Xie:** Investigation. **Nan Dong:** Validation, Formal analysis. **Zhengping Zhang:** Data curation. **Shilong Wang:** Resources. **Fan Mo:** Resources.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Key Research and Development Program of China (no. 2023YFB3306800), the New Chongqing Young Innovative Talent Program (CSTB2024NSCQ-QCXMX0028), the Key Laboratory of Industrial Software Engineering and Application Technology of Ministry of Industry and Information Technology (HK202303540), the Fundamental Research Funds for the Central Universities (2023CDJKYJH033).

Data availability

The data that has been used is confidential.

References

[1] C. Yu, A. Matta, Q. Semeraro, Group decision making in manufacturing systems: an approach using spatial preference information and indifference zone, J. Manuf. Syst. 55 (2020) 109–125.
[2] F. Ahmed, K.Y. Kim, Recursive approach to combine expert knowledge and data-driven RSW weldability certification decision making process, Rob. Comput. Integr. Manuf. 79 (2023) 102428.
[3] Q. Li, Y. Yang, M. Yang, et al., A graphical model for formalizing health maintenance activities in the context of the whole equipment lifecycle, Adv. Eng. Inf. 58 (2023) 102226.
[4] C.L. Hu, L. Wang, M.L. Chen, et al., A real-time interactive decision-making and control framework for complex cyber-physical-human systems, Annu. Rev. Control. 57 (2024) 100938.
[5] P.M. Babu, J. Seadon, D. Moore, Cognitive biases that influence Lean implementation and practices in a multicultural environment, Int. J. Lean Six Sigma (2023) (ahead-of-print).
[6] F. El Kalach, R. Wickramarachchi, R. Harik, et al., A semantic web approach to fault tolerant autonomous manufacturing, IEEE Intell. Syst. 38 (1) (2023) 69–75.
[7] B. Yang, H. Shen, F. Bi, et al., Rapid heat transfer simulation of composites curing process based on cGANs and MPGNNs, Int. J. Heat Mass Transf. 241 (2025) 126752.
[8] F. Franke, S. Franke, R. Riedel, AI-based improvement of decision-makers' knowledge in production planning and control, IFAC-PapersOnLine 55 (10) (2022) 2240–2245.
[9] A. Bousdekis, K. Lepenioti, D. Apostolou, et al., A review of data-driven decision-making methods for industry 4.0 maintenance applications, Electronics 10 (7) (2021) 828.
[10] S. Wang, J. Yang, B. Yang, et al., An intelligent quality control method for manufacturing processes based on a Human–Cyber–Physical knowledge graph, Engineering 41 (2024) 242–260.
[11] H. Fan, X. Liu, J.Y.H. Fuh, et al., Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics, J. Intell. Manuf. (2024) 1–17.
[12] B. Alkan, D.A. Vera, M. Ahmad, et al., Complexity in manufacturing systems and its measures: a literature review, European J. Industrial Eng. 12 (1) (2018) 116–150.
[13] J. Dong, J. Wang, S. Chen, Knowledge graph construction based on knowledge enhanced word embedding model in manufacturing domain, J. Intell. Fuzzy Syst. 41 (2) (2021) 3603–3613.
[14] U.K. Uz Zaman, M. Rivette, A. Siadat, et al., Integrated product-process design: material and manufacturing process selection for additive manufacturing using multi-criteria decision making, Rob. Comput. Integr. Manuf. 51 (2018) 169–180.
[15] T. Yang, X. Yi, S. Lu, et al., Intelligent manufacturing for the process industry driven by industrial artificial intelligence, Eng. 7 (9) (2021) 1224–1230.
[16] J. Ehrlinger, W.O. Readinger, B. Kim, Decision-making and cognitive biases, Encyclopedia of Mental Health 12 (3) (2016) 83–87.
[17] A. Jamwal, R. Agrawal, M. Sharma, et al., Review on multi-criteria decision analysis in sustainable manufacturing decision making, Int. J. Sustain. Eng. 14 (3) (2021) 202–225.
[18] M. Kessler, J.C. Arlinghaus, A framework for human-centered production planning and control in smart manufacturing, J. Manuf. Syst. 65 (2022) 220–232.
[19] L. He, W. Guo, P. Jiang, A decision-making model for knowledge collaboration and reuse through scientific workflow, Adv. Eng. Inf. 49 (2021) 101345.
[20] C. Su, Y. Han, X. Tang, et al., Knowledge-based digital twin system: using a knowlege-driven approach for manufacturing process modeling, Comput. Ind. 159 (2024) 104101.
[21] Y. Jing, G. Zhou, C. Zhang, et al., XMKR: explainable manufacturing knowledge recommendation for collaborative design with graph embedding learning, Adv. Eng. Inf. 59 (2024) 102339.
[22] Y. Xu, T. Peng, J. Tao, et al., A representation learning-based approach to enhancing manufacturing quality for low-voltage electrical products, Adv. Eng. Inf. 62 (2024) 102636.
[23] J.A.H. Álvaro, J.G. Barreda, An advanced retrieval-augmented generation system for manufacturing quality control, Adv. Eng. Inf. 64 (2025) 103007.
[24] T. Fu, S. Liu, P. Li, Intelligent smelting process, management system: efficient and intelligent management strategy by incorporating large language model, Frontiers of Eng. Manage. 11 (3) (2024) 396–412.
[25] L.I.U. Peifeng, L. Qian, X. Zhao, et al., Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly, IEEE Trans. Ind. Inf. (2024).
[26] Y. Chen, Y. Zhou, Y. Zhang, Machine learning-based model predictive control for collaborative production planning problem with unknown information, Electronics 10 (15) (2021) 1818.
[27] D. Jiang, X. Ren, B.Y. Lin, Llm-blender: ensembling large language models with pairwise ranking and generative fusion, Proceedings of the 61st Annual Meeting of the the Association-for-Computational-Linguistics (ACL) (2023) 14165–14178.
[28] Zhu B, Ning M, Jin P, et al. LLMBind: A unified modality-task integration framework. arXiv preprint, 2024:2402.14891.
[29] A. Chandler, D. Surve, H. Su, Detecting Errors through Ensembling Prompts (DEEP): an End-to-End LLM framework for detecting factual errors, in: Proceedings

of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 13120–13133.
[30] M.S. Zahedi, M. Rahgozar, R.A. Zoroofi, MATER: Bi-level matching-aggregation model for time-aware expert recommendation, Expert Syst. Appl. 237 (2024) 121576.
[31] S. Bhattacharyya, A. Rooshenas, S. Naskar, et al., Energy-based reranking: improving neural machine translation using energy-based models, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) (2021) 4528–4537.
[32] Liu Y, Liu P. SimCLS: A simple framework for contrastive learning of abstractive summarization. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021(2):1065-1072.
[33] Ravaut M, Joty S, Chen N F. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022(1):4504-4524.
[34] P. Li, H. Xie, Y. Jiang, et al., Neighborhood-adaptive multi-cluster ranking for deep metric learning, IEEE Trans. Circuits Syst. Video Technol. 33 (4) (2022) 1952–1965.
[35] P. Sun, L. Yang, Low-rank supervised and semi-supervised multi-metric learning for classification, Knowl.-Based Syst. 236 (2022) 107787.
[36] S. Yu, C. Fan, C. Xiong, et al., Fusion-in-T5: unifying variant signals for simple and effective document ranking with attention fusion, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) (2024) 7556–7561.
[37] Bianco S, Celona L, Donzella M, et al. Improving image captioning descriptiveness by ranking and llm-based fusion. arXiv preprint, 2023:2306.11593.
[38] Y. Zheng, R. Zhang, J. Zhang, et al., Llamafactory: unified efficient fine-tuning of 100+ language models, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (2024) 400–410.
[39] A.Q. Jiang, A. Sablayrolles, A. Roux, et al., Mixtral of experts, Arxiv Preprint 2401 (2024) 04088.
[40] Salazar J, Liang D, Nguyen T Q, et al. Masked language model scoring. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020:2699-2712.