Engineering 53 (2025) 311–322

Contents lists available at ScienceDirect

Engineering

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

Research
Engineering Management—Article

Intelligent Fault Diagnosis for CNC Through the Integration of Large Language Models and Domain Knowledge Graphs

Yuhan Liu [a], Yuan Zhou [b], Yufei Liu [c,*], Zhen Xu [a], Yixin He [a]

[a] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[b] School of Public Policy and Management, Tsinghua University, Beijing 100084, China
[c] Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

ARTICLE INFO

Article history:
Received 26 September 2024
Revised 7 April 2025
Accepted 7 April 2025
Available online 10 April 2025

Keywords:
Large language model
Domain knowledge graph
Knowledge graph-based retrieval augmented generation
Learning mechanism
Decision support system

ABSTRACT

As large language models (LLMs) continue to demonstrate their potential in handling complex tasks, their value in knowledge-intensive industrial scenarios is becoming increasingly evident. Fault diagnosis, a critical domain in the industrial sector, has long faced the dual challenges of managing vast amounts of experiential knowledge and improving human–machine collaboration efficiency. Traditional fault diagnosis systems, which are primarily based on expert systems, suffer from three major limitations: ① ineffective organization of fault diagnosis knowledge, ② lack of adaptability between static knowledge frameworks and dynamic engineering environments, and ③ difficulties in integrating expert knowledge with real-time data streams. These systemic shortcomings restrict the ability of conventional approaches to handle uncertainty. In this study, we proposed an intelligent computer numerical control (CNC) fault diagnosis system, integrating LLMs with knowledge graph (KG). First, we constructed a comprehensive KG that consolidated multi-source data for structured representation. Second, we designed a retrieval-augmented generation (RAG) framework leveraging the KG to support multi-turn interactive fault diagnosis while incorporating real-time engineering data into the decision-making process. Finally, we introduced a learning mechanism to facilitate dynamic knowledge updates. The experimental results demonstrated that our system significantly improved fault diagnosis accuracy, outperforming engineers with two years of professional experience on our constructed benchmark datasets. By integrating LLMs and KG, our framework surpassed the limitations of traditional expert systems rooted in symbolic reasoning, offering a novel approach to addressing the cognitive paradox of unstructured knowledge modeling and dynamic environment adaptation in industrial settings.

© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

New-generation intelligent manufacturing represents a paradigm shift that is characterized by the synergistic integration of next-generation artificial intelligence (enabling technologies) with advanced manufacturing systems (foundational technologies). This paradigm shift has positioned this technology as the cornerstone of an ongoing industrial revolution [1]. The accelerated maturation of large language models (LLMs) is currently augmenting this transformation and establishing LLMs as dual-function catalysts that simultaneously reinforce core technological capabilities and unlock emergent innovation pathways within human–cyber–physical systems (HCPS). LLMs have demonstrated strong capability and performance in understanding and logical reasoning, content generation, human–computer interaction, and other areas, and they have shown great application potential in industrial scenarios. There are several challenges associated with the application of LLMs in engineering scenarios. Industrial applications demand a high level of reliability for which errors and hallucinations are not acceptable. In addition, the integration of domain-specific knowledge such as technical documentation, operational data, forward design principles, and expert experience remains a complex task [2]. Some problems are more prominent in fault diagnosis. Faults are usually highly dynamic with diverse and complex causes, and new fault cases are constantly appearing [3,4]. Therefore, the effective integration of domain-specific knowledge, complex fault scenarios, continuously generated new fault cases, and dynamic feedback information in LLMs poses a major challenge to the applications of these models to fault diagnosis [5].

* Corresponding author.
E-mail address: liuyf@cae.cn (Y. Liu).

https://doi.org/10.1016/j.eng.2025.04.003
2095-8099/© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Fault diagnosis is critical to improving the operational reliability and efficiency of computer numerical control (CNC) machine tools, implementing predictive maintenance to extend the life of CNC systems, and ensuring the smooth operation of production lines [5–7]. The traditional fault diagnosis of CNC systems mainly relies on experienced engineers. Although engineer-based diagnosis is direct, it is also time-consuming and dependent on personal experience, with diagnostic effectiveness limited by an engineer's expertise and experience [8,9]. When a system fault occurs, a long diagnostic cycle is often required for maintenance, which can potentially lead to production halts and significant economic losses. The advent of expert systems has enhanced fault diagnosis capabilities, providing near-human expert-level problem-solving abilities in specific domains and the effective handling of knowledge reasoning issues [10,11]. However, expert systems are highly dependent on expert knowledge input, and a lack of updates to knowledge or an insufficient knowledge base can affect the accuracy of reasoning [12,13]. In addition, expert systems may not be user-friendly in terms of interaction design, making them difficult to understand and difficult to operate for non-experts. The ongoing development of CNC systems and artificial intelligence lays a solid foundation for intelligent fault diagnosis in CNC systems. Fault diagnosis should be based on real-time monitoring, and advancements in modern sensor technology have enabled CNC systems to collect critical performance indicators such as vibration signals and current signals in real-time [14–16]. Furthermore, when a fault occurs in a CNC system, the corresponding alarm information is automatically recorded and transmitted to the backend, thereby providing a solid premise for fault information collection.

The rapid development of LLMs offers new methods and perspectives for fault diagnosis in CNC systems. LLMs are powerful tools for processing vast amounts of data, and they possess advanced language understanding and generation capabilities, which enable them to perform language-related tasks with high efficiency and accuracy [17]. We believe that retrieval-augmented generation (RAG) is an effective approach for applying LLMs to vertical domains. In particular, the use of RAG with a domain-specific knowledge base is a mainstream method because it can address the hallucination problem in LLMs to some extent and improve the accuracy of question-answering in vertical domains [18]. Microsoft's release of GraphRAG validates our design, though we also recognize that general-purpose graphs cannot fully address the specific challenges in our domain. These capabilities are especially important for understanding and responding to the complex queries posed during fault diagnosis.

In this study, first, we designed a domain-specific LLM benchmark to address the issue of model adaptation to the domain. Second, to overcome the challenges of low accuracy in typical RAG responses as well as the difficulty in handling complex tasks, we developed a knowledge graph-based RAG framework. This framework not only resolved the inherent issues in traditional RAG frameworks but also better supported multi-turn dialogues by mapping these dialogues to traversals within the knowledge graph (KG). To implement the knowledge graph-based RAG, we constructed a KG tailored to fault diagnosis scenarios that integrated multi-source data to support fault diagnosis. Then, to meet the demands of different tasks, we designed customized prompt engineering to enhance the model's ability to accurately respond to diverse fault scenarios based on its understanding of user intent. Based on this foundation, to continuously improve the system's applicability in vertical domains, we designed a dynamic learning mechanism based on LLMs and expert input. This mechanism allowed the system to continuously learn and conduct optimization through user interactions during actual use, enabling it to update its KG and enhance its diagnostic capabilities. Finally, we integrated the entire system into a remote operation and maintenance system to achieve intelligent fault diagnosis and maintenance support.

2. Literature review

2.1. Applications of LLMs in vertical domains

In vertical domains, the primary task of LLMs is language understanding, which supports tasks such as entity extraction, relationship extraction, and document classification [19]. These models have had widespread applications in fields such as healthcare and industry [18,20]. Currently, prompt-based methods are the dominant approach. For example, in the healthcare field, models such as ERQA [21], MentaLLaMA [22], ArgMed-Agents [23], and diagnostic reasoning prompts [24] are prominent. In addition, studies have combined LLMs with KGs. These KGs help organize complex information, thereby enhancing the applicability of LLMs [25–27].

Building on prompt-based methods, fine-tuning has also become a common way to apply LLMs to vertical domains. Fine-tuning can deepen the integration of domain knowledge with LLMs, as has been seen in biomedical domain studies [28]. Notably, in a study by Li et al. [29], the key challenge was determining how to effectively leverage domain-specific datasets for fine-tuning. Liu et al. [20] proposed a typical application in manufacturing by combining KGs with LLM fine-tuning, and the KG provided high-quality fine-tuning data to support this process.

In addition, some studies have explored "lower-level" modifications of LLMs through fine-tuning, which were more common in early research and have been proven to improve model adaptability in specific domains. However, as model sizes have increased, pre-training costs have risen significantly. Consequently, researchers working on vertical applications currently tend to focus their efforts on established open-source benchmark models, employing fine-tuning or RAG methods.

Some studies have explored the enhancement of domain-specific capabilities through pre-training. Pre-training has been more commonly seen in earlier research [30,31] and has been shown to improve the adaptability of models in specific domains. However, with the increasing sales of models, the cost of pre-training has also risen significantly. As a result, researchers focusing on vertical applications currently tend to concentrate on adapting large models to domain-specific tasks through fine-tuning or RAG methods based on well-established, high-performance open-source benchmark models [32]. This approach helps to improve the performance of large models in solving domain-specific problems while mitigating the high costs of pre-training.

2.2. Existing intelligent fault diagnosis systems

Early fault diagnosis systems were primarily based on expert systems [33,34]. However, expert systems have inherent limitations, particularly in terms of knowledge updating, which is often challenging. Consequently, data-driven and signal analysis-driven methods have become essential tools for fault diagnosis [5,35]. With the popularity of machine learning and deep learning methods, these approaches have also been applied to fault diagnoses across various industrial scenarios [36,37].

As KG technology has matured, numerous studies have begun to explore knowledge graph-based approaches to fault diagnosis, with typical applications including knowledge graph-based fault diagnosis question-answering systems [38–40]. With the advent

of LLMs, researchers have also attempted to combine KGs with LLMs to enhance the intelligence of fault diagnosis systems. For example, in Liu et al.'s study [41], filtered information from KGs was fed into LLMs to mitigate the "hallucination" issue. Guo et al.'s study [42] further considered multi-hop path-finding within KGs to support optimal retrieval paths, thereby improving the accuracy of fault diagnosis. These studies have provided valuable inspiration and support for the implementation of an intelligent fault diagnosis system in the CNC domain based on LLMs and KGs.

3. Methodology

3.1. Framework

The intelligent fault diagnosis decision support system proposed in this study is depicted in Fig. 1. This system is divided into three main parts. Initially, the data foundation stage involved the construction of a comprehensive KG for CNC systems to collect and integrate knowledge relevant to fault diagnosis. Concurrently, advanced data processing techniques were employed to extract features from engineering data, thus enhancing data usability and diagnostic accuracy. Subsequently, the diagnosis based on the KG included the benchmark testing of LLMs and fine-tuning to ensure optimal performance, as well as the local deployment of the model to guarantee real-time responses and data security. Building on this, the system employed knowledge graph-based retrieval enhancement to generate multi-turn question-answering capabilities, which enabled the system to understand and respond accurately to complex diagnostic queries and thereby improve the efficiency and quality of interactions. Finally, the evaluation system and continuous evolution were addressed. We provided a CNC system evaluation dataset to assess the fault diagnosis system and utilized two learning mechanisms for ongoing evolution: interaction content review based on LLMs and knowledge updates based on expert input.

3.2. Fault diagnosis KG

3.2.1. Ontology design for fault diagnosis KG

The construction of the ontology for a KG relies on the integration of domain expert knowledge with data analysis, employing both top-down and bottom-up approaches [43,44]. The top-down method utilizes expert systems and existing data schemas to guide the construction of a KG, whereas the bottom-up approach identifies and integrates relevant knowledge through information extraction techniques from semi-structured or unstructured data. These strategies not only fully leverage in-depth domain knowledge but also reflect the importance of data-driven methods in the discovery and verification of new knowledge. In our CNC system fault diagnosis KG project, the design of the ontology also stemmed from the general standards of fault diagnosis and the experience of experts [45]. We defined seven key entity categories: equipment (machine tool), equipment modules, parameters, alarm numbers/information, phenomena (symptoms), causes, and solutions. The relationship between these entities is shown in Fig. 2. The structure of the ontology maintained a moderate level of complexity, ensuring basic fault diagnosis capabilities while also being sufficiently generalizable for broader applications. Some

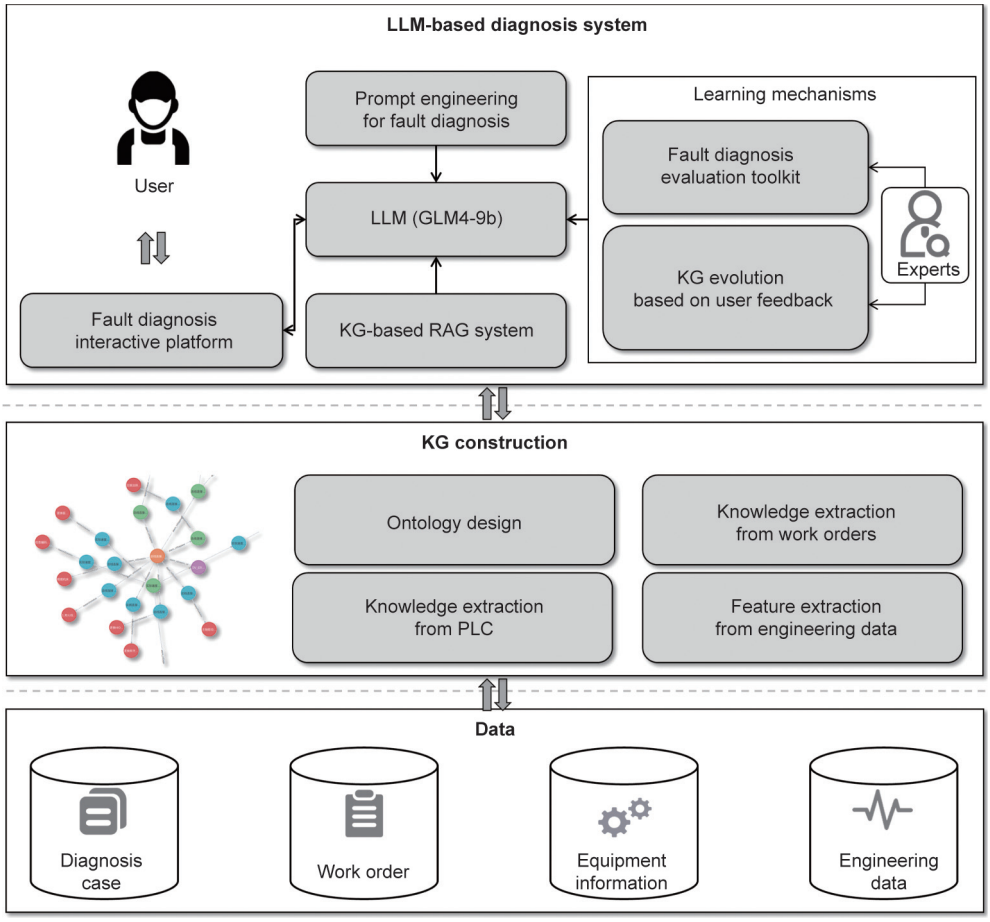

Fig. 1. Framework of fault diagnosis decision support system based on LLM. PLC: programmable logic controller.

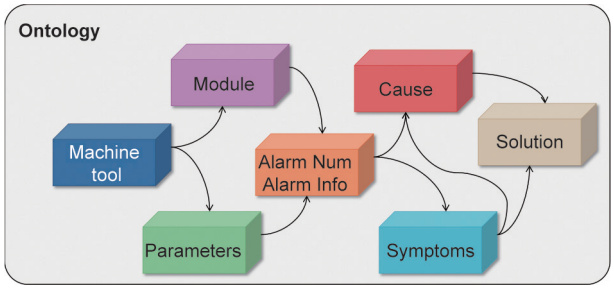

Fig. 2. Ontology design of CNC system fault diagnosis knowledge graph. Num: number; Info: information.

equipment modules incorporated engineering data as features, providing robust support for in-depth fault analysis.

3.2.2. Knowledge extraction from PLC ladder diagrams

In a CNC system controlled by programmable logic controller (PLC), faults can be divided into hardware and software categories [46]. By leveraging the PLC alarms specific to CNC systems, one can rapidly pinpoint and address these faults. The ladder diagram, with its intuitive and debug-friendly design, serves as an effective tool to represent discrete points [47]. This diagram not only simplifies fault diagnosis by providing structured and visual information but also significantly enhances the management of diagnostic processes and the capability of decision support.

However, the specialized nature of PLC ladder diagrams can present challenges for engineers during fault diagnosis. To address this, this paper proposes a novel method of converting ladder diagrams into textual information and extracting knowledge to construct a KG [48]. By analyzing the switch states and their corresponding serial and parallel logic within the PLC ladder diagrams, faults were categorized into operational actions and conditions under which faults occurred, which were then extracted as nodes in a KG. By integrating the fault alarm data with the corresponding solutions, a comprehensive KG pathway was constructed. This approach not only facilitated the rapid and accurate localization of faults but also enhanced the effectiveness of subsequent interventions.

3.2.3. Knowledge extraction from historical work orders

In the operation of a CNC system, faults must be handled on-site by engineers, who then fill out detailed fault work orders. These work orders document important fault cases, describing fault phenomena and the resolution strategies of the engineers, serving as an important resource for enriching the fault diagnosis knowledge base. However, these work orders are often unstructured, voluminous, and vary in content quality, posing significant challenges for data processing.

Leveraging the powerful deep semantic understanding capabilities of LLMs, this study employed LLMs and prompting engineering techniques to guide the behavior of the models, effectively extracting key information such as "fault phenomena," "fault causes," and "solutions" from historical work orders. By designing templates to guide the model's output, we ensured that the extracted information aligned with the needs of the KG. Subsequently, through a series of data cleaning and expert review steps, the extracted data were transformed into high-quality structured information to construct and expand the existing fault diagnosis KG.

3.2.4. Feature extraction from engineering data

In the research on CNC system fault diagnosis, feature extraction from engineering data is a crucial step [49]. Deep learning techniques have become the mainstream method for vibration signal processing and fault analysis, improving the accuracy and efficiency of diagnoses. With the introduction of new technologies such as graph neural networks and physics-based machine learning, this field continues to show immense development potential.

A three-stage engineering data feature extraction strategy was proposed to provide more refined features for CNC system fault diagnosis. The feature extraction framework is shown in Fig. 3. First, the single-channel feature extraction focused on extracting features from a single sensor signal, which served as the basis for identifying key fault indicators [50,51]. Second, multi-channel feature extraction integrated data from different sensors, enabling a broader analysis of the equipment's status and improving fault prediction accuracy [52]. Finally, instruction-based multi-channel feature extraction leveraged machine learning algorithms to deeply analyze multi-source data. This not only captured time dependencies but also predicted potential anomalies based on changes in operational instructions, thereby providing more sophisticated fault diagnosis capabilities [53].

Based on this, we built the associations between engineering data features and device module faults. A training set was created based on historical data for several typical faults, and a fault classification model was trained. In practical fault diagnosis, engineering data features can help more accurately and efficiently pinpoint faulty device modules or alarm information, leading to more precise fault diagnoses.

3.3. LLM-based diagnosis system

3.3.1. KG-based RAG system

In practical industrial applications, conventional RAG models combine external knowledge retrieval with LLMs to improve the accuracy of LLM responses, which partially mitigates the hallucination problem. However, RAG still faces limitations in specific fault diagnosis tasks such as those in CNC systems [54–56]. In CNC system fault diagnosis, while the responses from large models need to be accurate, they also need to be concise so that engineers can quickly use the information for troubleshooting. This necessitates effective organization of the knowledge base. However, due to the complexity of CNC systems, users often require multi-turn interactions to accurately diagnose faults, which demands strong management capabilities for related information in a database to support prolonged and high-quality interactions.

To address these limitations, in this study, an innovative knowledge graph-based RAG method was developed. In this approach, fault diagnosis elements such as "fault symptoms," "fault causes," and "solutions" were treated as entities within a KG. When the system received a user alert or descriptive information, it quickly identified the relevant entities and associated subgraphs within the KG through entity and semantic recognition. During multi-turn interactions with the user, the system continuously recognized user intent and feedback, optimizing the traversal direction of paths within the KG to provide more accurate diagnostic support [57,58].

Utilizing this method, the system narrowed the scope of the knowledge it retrieved by identifying entities and semantic information from the user input, effectively reducing hallucinations and enhancing the focus and accuracy of the responses. In addition, the system supported user interactions to guide the traversal direction within the KG, better meeting user needs. This improvement significantly enhanced the efficiency and reliability of the model in the domain of fault diagnosis.

3.3.2. Prompt engineering for KG-based RAG

A prompt is used to provide text or instructions to guide a model to generate a specific output. The more specific the prompt instructions are, the more closely the LLMs' response aligns with user needs [59]. In this research, we categorized prompts into role prompts and task prompts. The role prompts guided the LLMs to

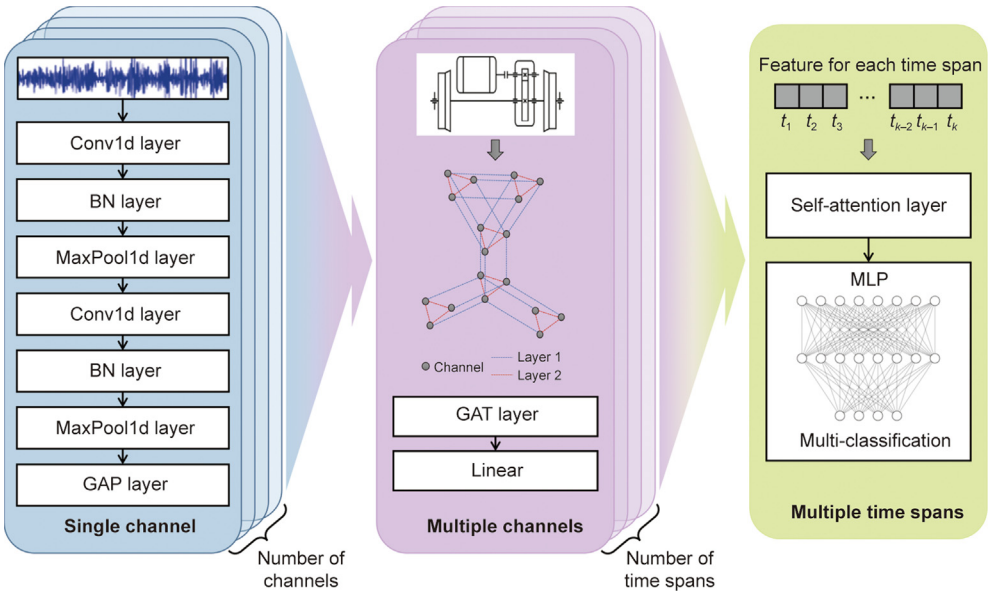

Fig. 3. Feature extraction framework for CNC system engineering data. Conv1d: convolution for 1 dimension; BN: batch normalization; MaxPool1d: max polling for 1 dimension; GAP: global average pooling; GAT: graph attention networks; MLP: multilayer perceptron; *t*: time span; *k*: index of time spans.

communicate in a specific role by constructing a virtual character with a unique perspective, professional knowledge, and behavioral patterns. For instance, in this project, the artificial intelligence (AI) played the role of a CNC fault diagnosis assistant. We continually added instructions to deepen the LLM's understanding of its responsibilities. The task prompts, however, were text inputs used to guide or instruct the model to perform a specific task. When generating diagnostic results, we particularly emphasized the importance of precise task prompt settings, especially for KG-directed walks [60]. These specially crafted prompts were crucial for ensuring the accuracy of diagnostic outputs. We constructed task prompts to constrain the paths of the KG walks and produce answers that were highly relevant to the field of CNC system fault diagnosis.

3.4. Learning mechanisms

A learning mechanism is a crucial component of AI development that enables systems to continuously learn and improve [61]. In this project, we enhanced the KG by expanding edge attributes and introducing the concept of path weights, thereby allowing the system to identify potential errors or deficiencies during interactions. By analyzing user satisfaction with current answers, the system adjusted the weights of the KG paths accordingly.

The framework of the learning mechanism is shown in Fig. 4. Specifically, for answers without feedback, we increased the weight of the corresponding path to ensure that paths with higher weights were prioritized in future retrievals. For answers that included feedback, we extracted the paths using the LLM and added these paths to the feedback database [57,58]. Engineers regularly reviewed the paths in the feedback database, and those that passed the review process were incorporated into the KG. This approach enabled the continuous growth of the knowledge base, further enhancing the fault diagnosis capabilities.

4. Results and discussion

4.1. LLM benchmark for CNC system

Due to differences in the pre-training processes of LLMs, the adaptability of LLMs to various domains is variable. To assess the application effectiveness of LLMs in the CNC system domain, we constructed a multiple-choice question dataset based on the massive multitask language understanding (MMLU) [62] method. This dataset, named CNC language understanding (CNCLU), contained 200 questions that covered essential knowledge points, including basic terminology definitions, functional introductions, and operational processes of CNC systems. Each question provided a correct answer along with three options that simulated common misunderstandings and errors, allowing for a more accurate evaluation and comparison of different LLMs in terms of their understanding and the application of professional knowledge in the CNC system. We also introduced the AI2 Reasoning Challenge (AI2-ARC) dataset, which was used to evaluate the basic reasoning capabilities of large models, thus serving as an additional reference for model selection.

In selecting the LLMs for comparison, we focused on evaluating ChatGLM3-6b, GLM4-9b-Chat [63], Qwen1.5-7b-Chat, Qwen2-7b-Instruct, and Qwen2.5-7b-Instruct [64] (b: billion), all of which had previously shown strong performance in Chinese language tasks. Considering the balance between accuracy, concurrent resources, and response speed, we chose models with sizes ranging from 6b to 9b for testing. This size selection ensured that the system could meet model concurrency and response time requirements while maximizing diagnostic accuracy. Quantization becomes an important consideration when considering the eventual practical deployment of LLMs. Quantization techniques can significantly reduce the storage and computational requirements of models, making them more suitable for operations in resource-constrained environments. In this study, we evaluated two quantization methods, accurate post-training quantization for generative pre-trained transformers (GPTQ) [65] and GPT-Generated Unified Format (GGUF), to analyze the impact of quantization on a model's accuracy and efficiency.

By comparing the performance of these methods in MMLU tasks, AI2-ARC, and CNC system fault diagnosis scenarios, we obtained the performance results shown in Table 1 and Fig. 5. First, regarding quantization methods, we observed that although GPTQ reduced the model size, the most significant issue was that it did not compress the inference time. By contrast, GGUF achieved higher compression rates but significantly reduced the inference speed. From the perspective of accuracy, the models using GGUF showed better performance (out of 12 comparable datasets, 11 sets

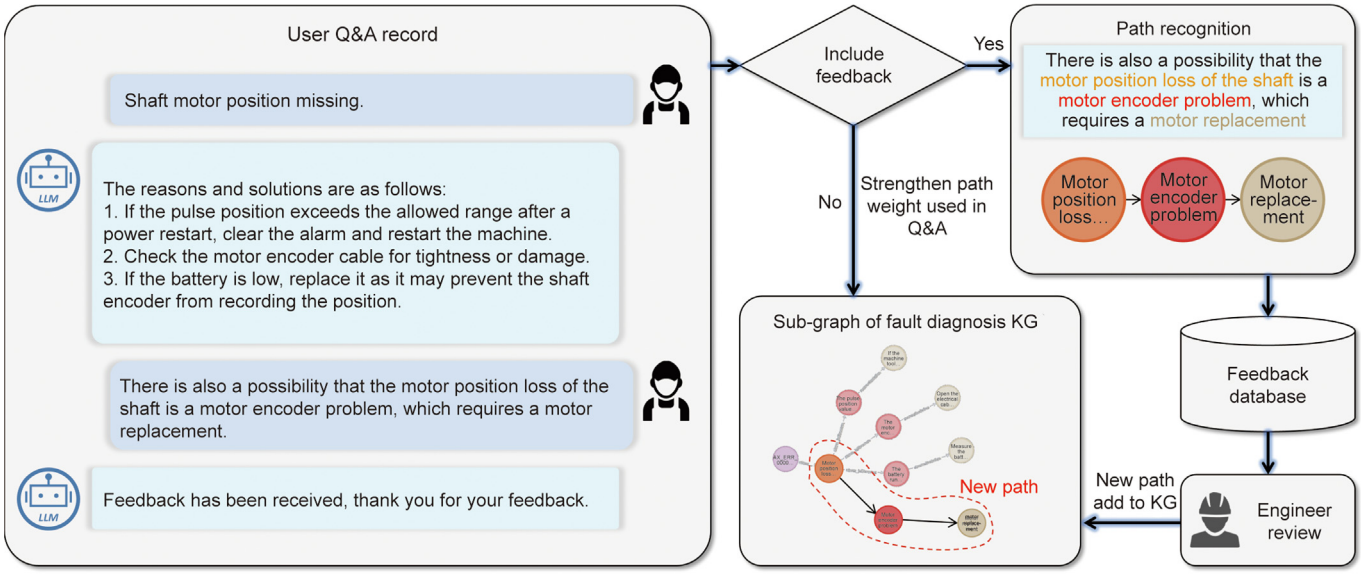

Fig. 4. Feature extraction framework for CNC system engineering data. Q&A: question and answer.

Table 1
Comparison of locally deployed LLM and quantization methods.

Model	Quantization method	Model size (gigabyte, GB)	MMLU		AI2-ARC		CNCLU	
			Score	Time (s)	Score	Time (s)	Score	Time (s)
ChatGLM3-6b	Original	11.63	0.5000	2464.00	0.728	401.26	0.43	21.41
	GPTQ	—	—	—	—	—	—	—
	GGUF	3.34	0.4882	873.80	0.7144	157.75	0.385	9.62
GLM4-9b-Chat	Original	17.51	0.6840	3235.75	0.9198	521.20	0.640	25.96
	GPTQ	6.26	0.6661	2908.14	0.9011	458.22	0.638	22.62
	GGUF	5.08	0.6687	1192.30	0.9074	222.26	0.600	14.52
Qwen1.5-7b-Chat	Original	14.38	0.5834	2484.33	0.8224	402.06	0.428	23.37
	GPTQ	5.46	0.5758	2735.93	0.8122	533.07	0.412	30.79
	GGUF	4.20	0.5770	1118.32	0.8187	187.56	0.435	11.56
Qwen2-7b-Instruct	Original	14.19	0.6909	2469.22	0.8892	400.71	0.541	22.43
	GPTQ	5.19	0.6772	2790.94	0.8805	555.66	0.543	31.46
	GGUF	4.13	0.6809	956.40	0.8873	182.98	0.570	11.20
Qwen2.5-7b-Instruct	Original	14.19	0.7373	2472.20	0.9099	403.83	0.548	22.51
	GPTQ	5.19	0.7198	2769.06	0.9031	549.83	0.548	31.11
	GGUF	4.13	0.7319	961.19	0.9062	182.95	0.605	11.18

In January 2025, for ChatGLM3-6b GPTQ, the quantization information is unavailable; the rest of the quantization information is from llama.cpp, AutoGPTQ, and GPTOModel.

showed better performance with GGUF). For model selection, we observed that for the unquantized models, GLM4-9b-Chat performed the best in both AI2-ARC and CNCLU. However, the newly released Qwen2.5-7b-Instruct performed better in MMLU. For the GGUF-quantized models, we found that Qwen2.5-7b-Instruct and GLM4-9b-Chat performed similarly in AI2-ARC and CNCLU. Since Qwen2.5-7b-Instruct was smaller and more efficient for inferences, we selected the GGUF-quantized Qwen2.5-7b-Instruct as the baseline model.

4.2. Prompt engineering

The role prompts ensured that the system provided generative conversational services as an AI fault diagnosis assistant, thereby enhancing the user experience and improving system security by rejecting questions unrelated to fault diagnosis. The task prompts further refined the user's queries, enabling the LLM to generate more accurate responses. Upon receiving a query, intent recognition was performed, and the corresponding task prompts were designed based on the identified intent, which might have included fault code and information inquiries, phenomenon inquiries, and feedback. Fault code and phenomena inquiries proceeded to the RAG stage based on the KG, while feedback activated the system's learning mechanism. This approach allowed the system to more effectively understand and address user needs while continuously learning and performing optimization in real-world applications, thereby enhancing the overall accuracy and efficiency of fault diagnosis.

4.3. KG for RAG

The fault diagnosis decision support system utilized four types of data: fault diagnosis cases, historical work orders, equipment information, and engineering data. Each type of data played a crucial role in constructing the KG and supporting fault diagnosis.

Fault diagnosis cases: The fault diagnosis cases were provided by the Huazhong CNC Division and encompassed a variety of typical fault modes. This dataset included approximately 500 fault diagnosis cases, covering mechanical, electrical, and software-related failures. Each case contained key information such as fault descriptions, diagnosis results, fault types, affected components, and the time of occurrence.

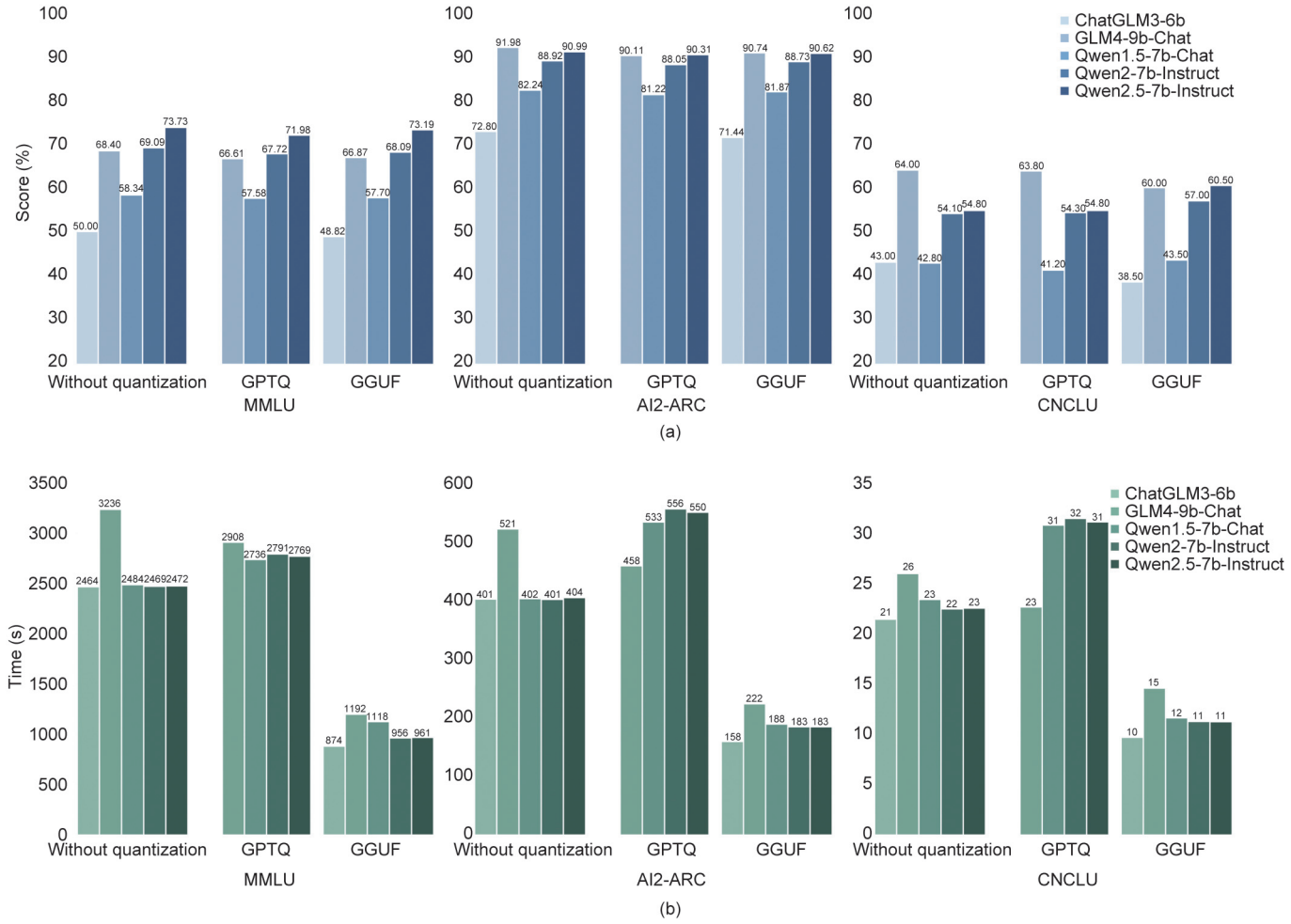

Fig. 5. LLM benchmarks in MMLU, AI2-ARC, and CNCLU, (a) Score (%) of LLMs in different tasks and different quantization methods; (b) time (s) spent by LLMs is in different tasks and different quantization methods.

Work orders: The work orders came from the internal records of the company, including equipment maintenance and fault troubleshooting reports. This dataset comprised approximately 5000 work orders, detailing work order numbers, equipment types, maintenance dates, maintenance contents, fault causes, handling procedures, and the personnel involved.

Equipment information: The equipment information was sourced from machine tool manufacturers and their subsequent updates in maintenance logs. The data included equipment models, production dates, operation logs, equipment configurations, maintenance records, and operational status.

Engineering data: The engineering data came from the CNC system's drive recorder, specifically from the "black box" data recorded in the 10 s preceding an alarm. These data included sensor measurements such as temperature, vibration frequency, current-voltage fluctuations, temperature change rates, and motion speed. These data were critical for fault diagnosis because they provided real-time operational insights that could help detect potential issues and prevent failures.

Based on the fault data provided by the Huazhong CNC Division, we constructed a total of 1549 entities, with 1334 relationships among them. The KG is shown in Fig. 6. The KG contains both complex graphs and simple paths. These data covered a wide range of fault scenarios, providing a solid foundation for subsequent fault diagnosis and knowledge inference. We used Neo4j as the database for storing and querying the KG, ensuring efficient retrieval and updates. The following diagram illustrates the structure of the KG we constructed and its performance in practical applications.

In this study, to address the limitations exhibited by traditional RAG models in handling user contexts and complex requirements, we proposed an improved RAG model. Traditional RAG models often struggle with accurately diagnosing faults in complex dialogues, particularly in multi-turn interactions and intricate fault scenarios, due to their inability to effectively leverage contextual information. To overcome these challenges, we integrated deep learning techniques with KG technology, proposing an enhanced RAG model based on dynamic subgraph partitioning and a multi-turn dialogue mechanism. Specifically, we used fault phenomenon descriptions and alarm codes as the partitioning criteria to divide the KG into multiple subgraphs, each corresponding to a set of solution paths associated with specific fault causes. This subgraph partitioning not only helped narrow the retrieval scope but also improved the system's ability to handle complex fault scenarios.

During operation, the system first identified key entities and semantic information in the user's input to quickly locate the relevant subgraph and provide the most likely fault causes and corresponding solutions, as shown in Fig. 7(a). The multi-turn fault diagnosis process based on KG is shown in Fig. 7(b). As the dialogue progressed, the system dynamically adjusted and optimized the solution paths within the subgraph based on user feedback. This approach not only enhanced the coherence of the dialogue but also significantly improved the accuracy of fault diagnosis and repair

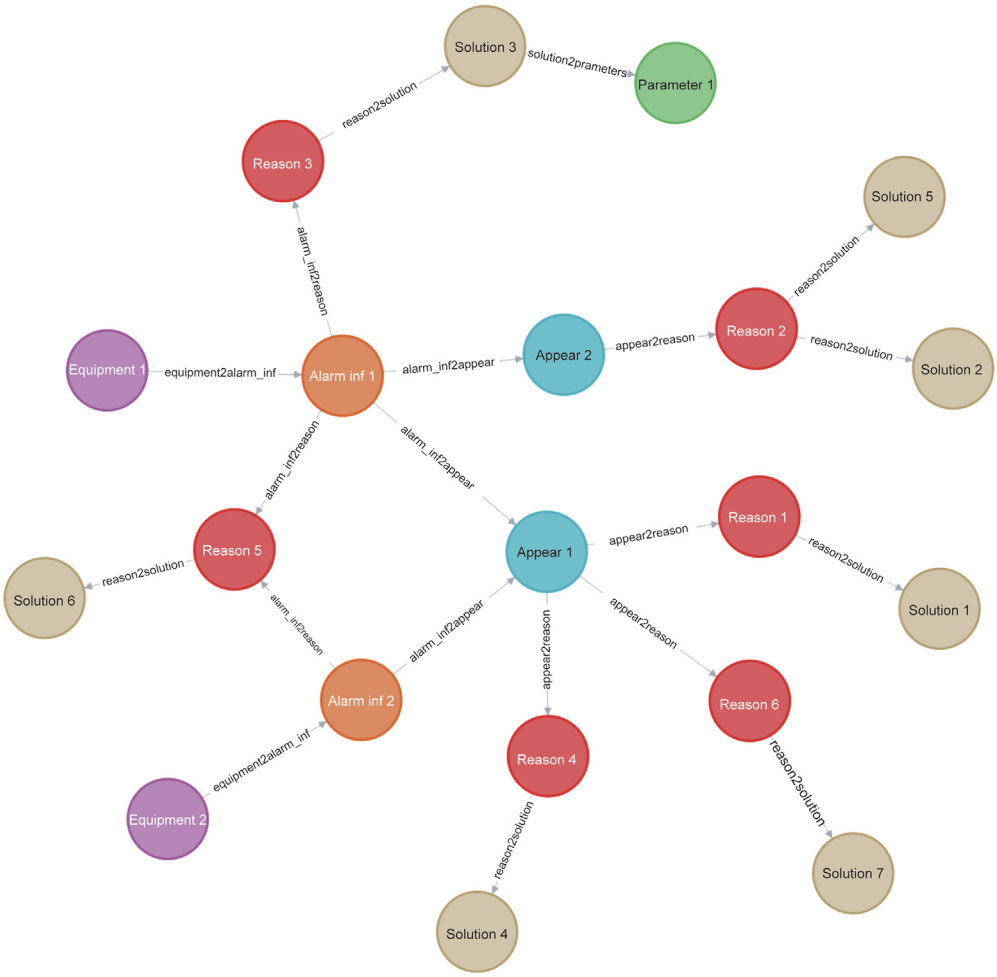

Fig. 6. CNC system fault diagnosis KG.

recommendations through iterative optimization, ensuring that the system better met user needs. In addition, this multi-turn interaction strategy made the system's performance more stable and reliable in complex scenarios, enabling it to handle a broader range of real-world applications.

Building on this process, engineering data could be used to more accurately and efficiently pinpoint faulty device modules or alarm information, leading to more precise fault diagnosis. Specifically, during system operations, engineering data such as sensor readings and environmental conditions were continuously monitored. Key features were extracted from these data and classified through a fault classification model. The results might point to specific device modules or alarm information, which could then be used to filter the subgraph, improving the efficiency and accuracy of the diagnostic process.

Moreover, the KG continuously integrated new data and feedback from the diagnostic process to refine the association between data features and fault categories. This enabled the system to adapt to changing operational conditions and improve its diagnostic accuracy over time. As a result, the system not only enhanced real-time fault detection capabilities but also gained a deeper understanding of fault patterns, supporting more proactive and preventive long-term maintenance strategies. The fusion of engineering data with KG-based fault diagnosis provided a powerful framework for tackling complex real-world challenges in industrial environments.

4.4. Learning mechanisms

To endow the generative fault diagnosis system based on LLMs with self-learning capabilities and growth potential, this study introduced a learning mechanism designed to continuously optimize and enhance the system's performance. The system was equipped with a processing workflow that extracted knowledge from the continuously generated fault diagnosis work orders and dynamically integrated new knowledge into the KG, thereby enriching the fault diagnosis case library. This process significantly improved the system's knowledge base in terms of growth and scalability. Moreover, we optimized the structure of the KG by expanding its connection attributes and introducing the concept of path weighting. Specifically, based on feedback from engineers and users, the weights of the knowledge paths that effectively resolved issues were increased, while the paths identified as problematic by engineers underwent manual review and weight adjustment. This strategy ensured that the knowledge base more accurately reflected real-world conditions and enabled continuous learning and optimization based on user feedback.

We constructed a test dataset consisting of 41 common fault scenarios, with answers generated by the fault diagnosis system. These answers were then rated by experienced engineers and converted into a percentage score. The performance of the system over time is illustrated in Fig. 8. With the introduction of this learning mechanism, the iterative supplementation and optimization of

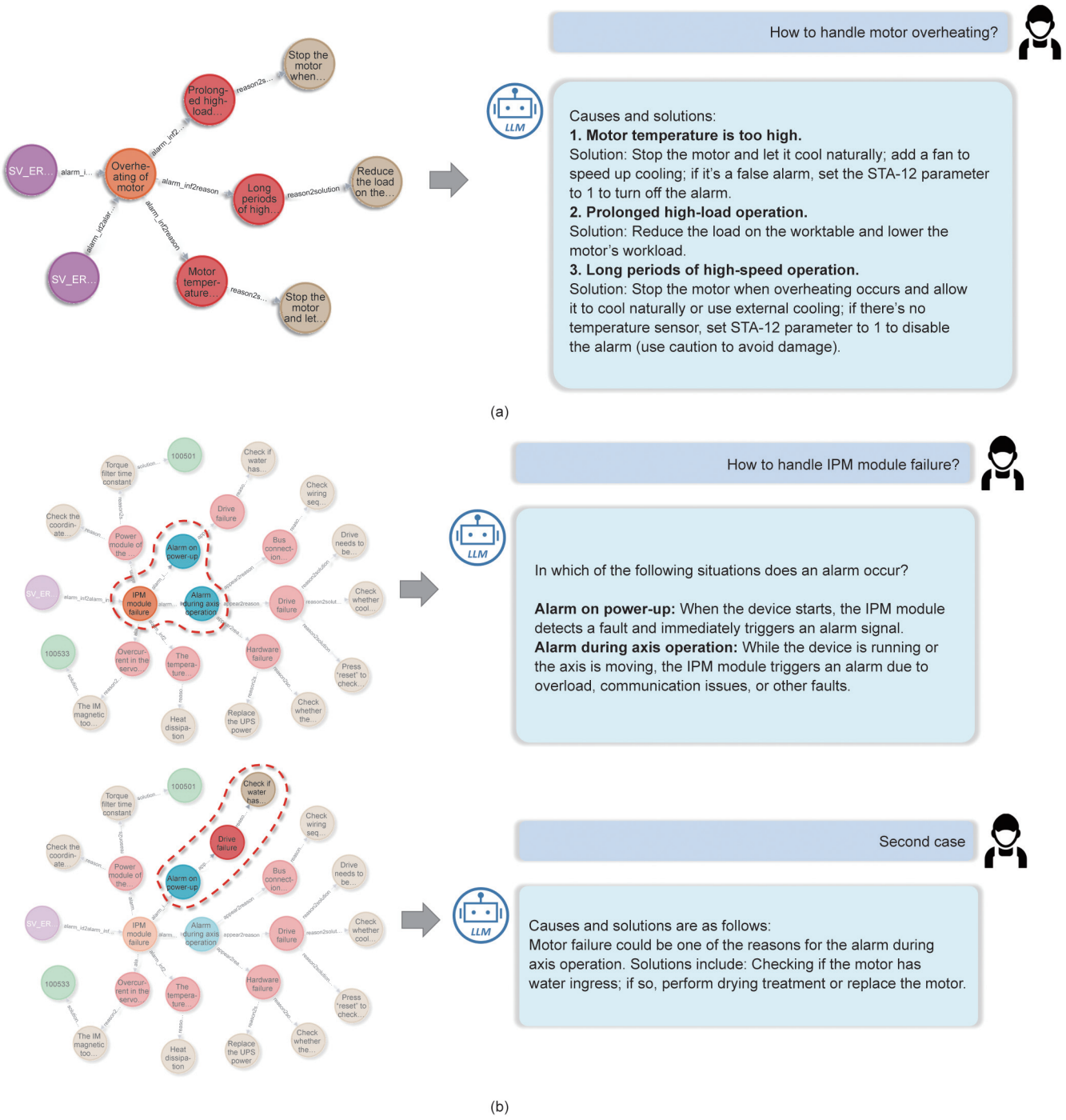

Fig. 7. Multi-turn fault diagnosis interaction based on sub-graph walk. (a) Identifying entities to locate the relevant subgraph and provide the fault causes and solutions; (b) multi-turn fault diagnosis process based on KG. STA: Status; IPM: intelligent power module.

knowledge allowed the system's diagnostic capability to surpass that of an engineer with two years of experience. This demonstrated the system's potential for continuous self-optimization in real-world applications.

In real industrial environments, the fault types and diagnostic requirements may change rapidly, and conflicting information may arise from different user feedback. To address this issue, a panel of experts was assembled in the feedback review section to regularly evaluate user-generated feedback. Once sufficient data were accumulated, a large model-based review filter could be trained to assist experts in the initial screening process.

4.5. Fault diagnosis system

In this research, we developed a generative fault diagnosis decision support system based on LLMs, with the goal of providing users with causes and solutions for CNC system faults. This system was integrated into the CNC Cloud Manager APP by Huazhong CNC, and a connection was established with the CNC system through quick response (QR) code scanning to obtain basic information and the current fault alarm code. When interacting with the system, users could pose specific questions based on actual situations. The system performed an analysis and generated responses using

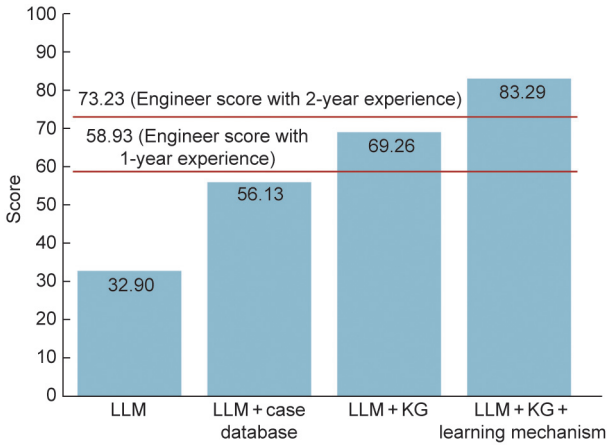

Fig. 8. Comparison of learning mechanisms in CNC system fault diagnosis and engineers.

prompt engineering and the RAG method based on the KG. In addition, the system incorporated a learning mechanism that captured and processed effective user feedback, which resulted in the continuous optimization and updating of its KG to improve diagnostic accuracy and efficiency.

The system supported the real-time acquisition of sensor signals and used these signals to assess the operating status of the CNC system, enabling real-time monitoring and maintenance. The system also combined single-channel feature extraction with multi-channel feature extraction, first extracting features from individual sensors through single-channel analysis, and then integrating data from multiple sensors through multi-channel analysis. Finally, the system conducted a comprehensive analysis of the data across multiple time segments using instruction-domain-based multi-channel feature extraction. These features were used as attributes in the machine tool module of the KG, providing critical support for fault diagnosis analysis.

The generative fault diagnosis system based on LLMs also incorporated faults that were caused by abnormal CNC system parameters. The constructed KG included 37 cases of CNC system faults caused by system parameters, with a total of 63 parameters. After providing the cause and solution for these faults, the system allowed users to establish a connection with the CNC system via mobile QR code scanning and download the potentially faulty parameters to the CNC system. The CNC system could then check and modify these parameters, thereby assisting users in locating and resolving faults online.

4.6. Discussion

This study explored the adaptability of different LLMs in the domain of a CNC system and proposed a simple and effective evaluation method. We constructed a multiple-choice question dataset based on the MMLU method to test the abilities of the models to understand and apply domain-specific knowledge. The evaluation results showed that different models had varying levels of adaptability to the domain, with the GLM4-9b model exhibiting the highest suitability for a CNC system. In addition, we conducted quantization tests to enhance model efficiency and reduce computational resource consumption while maintaining model performance, which is crucial for applications in production processes. Effective quantization meant that with the same computational resources, the system could connect to more devices and provide faster fault diagnosis support.

In this study, the use of KGs as a structured form of knowledge representation significantly enhanced the system's fault diagnosis capabilities, particularly in accurately identifying and responding to complex fault patterns. By integrating role prompts and task prompts into the design strategy, we further improved the overall performance of the system. Building on this foundation, the RAG model based on the KG provided distinct advantages in diagnosing CNC system faults. First, the system was able to deliver concise and accurate answers to specific alarm codes, alarm information, or fault phenomena, thereby eliminating the need for users to sift through lengthy texts to find the "right answer." Second, the system's ability to support multi-turn dialogues enabled it to more deeply understand and resolve user issues, significantly enhancing the system's usability.

The learning mechanism continuously optimized the diagnostic process by analyzing user feedback in real-time, demonstrating exceptional adaptability and improvement. This mechanism not only enabled the system to gradually reduce errors during routine operations but also enhanced its ability to diagnose new types of faults. However, the feedback-based learning process could sometimes be influenced by erroneous feedback, presenting a challenge for the system. To address this, we implemented a two-stage feedback review mechanism. In the first stage, the LLM filtered out a significant amount of non-valuable feedback, aiding engineers in the review process. In the second stage, domain experts further refined this selection, incorporating valuable insights into the KG. Compared with traditional review mechanisms, the introduction of the LLM significantly improved review efficiency, which played a crucial role in the continuous optimization of the KG and the overall improvement of the fault diagnosis system.

5. Conclusions

In this study, an intelligent fault diagnosis system based on LLMs and KGs was successfully developed, validating its effectiveness in CNC systems and demonstrating the feasibility of this approach in practical industrial environments. First, the system integrated multi-source data through KGs, forming a robust data foundation that not only encompassed a wide range of fault cases and related information but also effectively organized multi-dimensional data from CNC systems, providing strong support for efficient fault diagnosis. Second, the system was designed with targeted prompt engineering and employed a knowledge graph-based RAG framework, enabling it to respond to user fault diagnosis requests with high efficiency and accuracy. Third, the introduction of multi-turn dialogue and interactive query capabilities further enhanced the system's usability and user experience. Building on this foundation, the system incorporated a learning mechanism that continuously optimized its performance by analyzing user feedback, thus demonstrating outstanding learning capabilities and ensuring reliability and effectiveness over long-term use.

This research provides a template and standardized framework for the application of LLMs in the industrial domain, offering significant practical value and laying the groundwork for future similar applications. As discussed earlier, addressing the "hallucination" problem of large models, particularly through the use of the RAG framework, is crucial for the application of large models to specialized domains. Similar applications have been observed in the biomedical field. Building on this foundation, KGs offer distinct advantages. First, many industries already have well-established KGs, and the processed data from these graphs can provide high-density information to enhance the performance of large language models. Second, the linking capability of KGs allows for the retrieval of both broader and more detailed information. Finally, user feedback can be effectively recorded within a KG, leading to continuous improvements in system accuracy and performance throughout a user's interaction with the system.

Although this study showcased the great potential of LLMs and KGs in industrial applications, there remains room for improvement in addressing more complex industrial scenarios and a broader range of fault types. Future research will focus on further exploring prompt design, fine-tuning strategies, and even the pre-training process of LLMs to achieve broader industrial applications and higher system performance.

CRediT authorship contribution statement

Yuhan Liu: Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Yuan Zhou:** Supervision, Resources, Methodology, Conceptualization. **Yufei Liu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization. **Zhen Xu:** Writing – original draft, Validation, Software, Methodology, Data curation. **Yixin He:** Writing – original draft, Validation, Software, Methodology, Data curation.

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 research was funded by the National Natural Science Foundation of China (72104224, L2424237, 71974107, L2224059, L2124002, and 91646102), the Beijing Natural Science Foundation (9232015), the Beijing Social Science Foundation (24GLC058); the Construction Project of China Knowledge Center for Engineering Sciences and Technology (CKCEST-2023-1-7), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (16JDGC011), the Tsinghua University Initiative Scientific Research Program (2019Z02CAU), and the Tsinghua University Project of Volvo-Supported Green Economy and Sustainable Development (20183910020). We thank LetPub for its linguistic assistance during the preparation of this manuscript.

References

[1] Zhou J, Zhou Y, Wang B, Zang J. Human–Cyber–Physical Systems (HCPSs) in the context of new-generation intelligent manufacturing. Engineering 2019;5(4):624.
[2] Wu F, Shen T, Bäck T, Chen J, Huang G, Jin Y, et al. Knowledge-empowered, collaborative, and co-evolving ai models: the post-LLM roadmap. Engineering 2025;44:87–100.
[3] Guo L, Li R, Jiang B. Fault detection and diagnosis using statistic feature and improved broad learning for traction systems in high-speed trains. IEEE Trans Artif Intell 2023;4(4):679–88.
[4] Skliros C, Esperon Miguez M, Fakhre A, Jennions I. A review of model based and data driven methods targeting hardware systems diagnostics. Diagnostyka 2018;20(1):3–21.
[5] Aldrini J, Chihi I, Sidhom L. Fault diagnosis and self-healing for smart manufacturing: a review. J Intell Manuf 2024;35(6):2441–73.
[6] Gao Z, Cecati C, Ding SX. A survey of fault diagnosis and fault-tolerant techniques—part II: fault diagnosis with knowledge-based and hybrid/active approaches. IEEE Trans Ind Electron 2015;62(6):3768–74.
[7] Gao Z, Cecati C, Ding SX. A Survey of fault diagnosis and fault-tolerant techniques—part I: fault diagnosis with model-based and signal-based approaches. IEEE Trans Ind Electron 2015;62(6):3757–67.
[8] Link P, Poursanidis M, Schmid J, Zache R, von Kurnatowski M, Teicher U, et al. Capturing and incorporating expert knowledge into machine learning models for quality prediction in manufacturing. J Intell Manuf 2022;33:2129–42.
[9] Wei Y, Li Y, Xu M, Huang W. A review of early fault diagnosis approaches and their applications in rotating machinery. Entropy 2019;21(4):409.
[10] Deng XW, Gao QS, Zhang C, Hu D, Yang T. Rule-based fault diagnosis expert system for wind turbine. In: Proceedings of the 2017 International Conference on Information Science and Technology (IST 2017); 2017 Mar 24–26; Wuhan, China. Les Ulis: EDP Sciences;2017. p. 07005.
[11] Xu X, Yan X, Sheng C, Yuan C, Xu D, Yang J. A belief rule-based expert system for fault diagnosis of marine diesel engines. IEEE Trans Syst Man Cybern Syst 2020;50(2):656–72.
[12] Jimenez-Roa LA, Heskes T, Tinga T, Stoelinga M. Automatic inference of fault tree models via multi-objective evolutionary algorithms. IEEE Trans Depend Secure Comput 2023;20(4):3317–27.
[13] Wang X. Research on fault diagnosis method of CNC machine tools based on integrated MPA optimised random forests. EAI Endorsed Trans Scalable Inf Syst 2024;11(5):5785.
[14] Pignati M, Zanni L, Romano P, Cherkaoui R, Paolone M. Fault detection and faulted line identification in active distribution networks using synchrophasors-based real-time state estimation. IEEE Trans Power Deliv 2017;32(1):381–92.
[15] Canizo M, Onieva E, Conde A, Charramendieta S, Trujillo S. Real-time predictive maintenance for wind turbines using Big Data frameworks. In: Proceedings of the 2017 IEEE International Conference on Prognostics and Health Management (ICPHM); 2017 Jun 19–21; Dallas, TX, USA. New York City: IEEE; 2017. p. 70–7.
[16] Khorsheed RM, Beyca OF. An integrated machine learning: utility theory framework for real-time predictive maintenance in pumping systems. Proc Inst Mech Eng, B J Eng Manuf 2021;235(5):887–901.
[17] Chen J, Qian J, Zhang X, Song Z. Root-KGD: a novel framework for root cause diagnosis based on knowledge graph and industrial data. 2024. SSRN.4933706.
[18] Liu P, Qian L, Zhao X, Tao B. Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly. IEEE Trans Industr Inform 2024;20(6):8160–9.
[19] Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. Engineering 2023;25:51–65.
[20] Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics 2024;11(3):57.
[21] Liu Y, Li X, Luo Y, Du J, Zhang Y, Lv T, et al. Toward a large language model-driven medical knowledge retrieval and QA system: framework design and evaluation. Engineering 2025;50:270–82.
[22] Yang K, Zhang T, Kuang Z, Xie Q, Huang J, Ananiadou S. MentaLLaMA: interpretable mental health analysis on social media with large language models. In: Proceedings of the ACM Web Conference 2024; 2024 May 13–17; Held virtually. New York City: Association for Computing Machinery; 2024. p. 4489–500.
[23] Hong S, Xiao L, Zhang X, Chen J. ArgMed-agents: explainable clinical decision reasoning with LLM disscusion via argumentation schemes. In: Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2024 Dec 3–6; Lisbon, Portugal. New York City: IEEE; 2024.
[24] Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. npj Digit Med 2024;7(20):1–7.
[25] Gao Y, Li R, Croxford E, Caskey J, Patterson BW, Churpek M, et al. Leveraging a medical knowledge graph into large language models for diagnosis prediction: design and application study. JMIR AI 2025;4:e58670.
[26] Hu Z, Xu Y, Yu W, Wang S, Yang Z, Zhu C, et al. Empowering language models with knowledge graph reasoning for open-domain question answering. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language; 2022 Dec 7–11; Abu Dhabi, UAE. Melbourne: ACL Anthology; 2022. p. 9562–81.
[27] Yasunaga M, Ren H, Bosselut A, Liang P, Leskovec J. QA-GNN: reasoning with language models and knowledge graphs for question answering. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2011 Jun 6–11; Held virtually. Melbourne: ACL Anthology; 2021. p. 535–46.
[28] Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-tuning large neural language models for biomedical natural language processing. Patterns 2023;4(4):100729.
[29] Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-ai (LLaMA) using medical domain knowledge. Cureus 2023;15(6):e40895.
[30] Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: retrieval-augmented language model pre-training. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Held virtually. New York City: Association for Computing Machinery; 2020 .p. 3929–38.
[31] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Held virtually. Melbourne: ACL Anthology; 2020. p. 8342–60.
[32] Ovadia O, Brief M, Mishaeli M, Elisha O. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024 Nov 12– 16; Miami, FL, USA. Melbourne: Association for Computational Linguistics; 2024. p. 237–50.
[33] Nan C, Khan F, Iqbal MT. Real-time fault diagnosis using knowledge-based expert system. Process Saf Environ Prot 2008;86(1):55–71.
[34] Angeli C. Online expert systems for fault diagnosis in technical processes. Expert Syst 2008;25(2):115–32.
[35] Ávila Okada KF, Silva de Morais A, Oliveira-Lopes LC, Ribeiro L. A survey on fault detection and diagnosis methods. In: Proceedings of the 2021 14th IEEE International Conference on Industry Applications (INDUSCON); 2021 Aug 15–18; Sao Paulo, Brazil. New York City: IEEE; 2021. p. 1422–9.

[36] Kafeel A, Aziz S, Awais M, Khan MA, Afaq K, Idris SA, et al. An expert system for rotating machine fault detection using vibration signal analysis. Sensors 2021;21(22):7587.
[37] Richardson WB, Meyer J, Von Solms S. Towards machine learning and low data rate IoT for fault detection in data driven predictive maintenance. In: Proceedings of the 2021 IEEE World AI IoT Congress (AIIoT); 2021 May 10–13; Held virtually. New York City: IEEE;2021. p. 202–8.
[38] Su L, Wang Z, Ji Y, Guo X. A survey based on knowledge graph in fault diagnosis, analysis and prediction: key technologies and challenges. In: Proceedings of the 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE); 2020 Oct 23–25; Beijing, China. New York City: IEEE; 2020. p. 458–62.
[39] Chen Z, Mo S, Xu J, Xu Z. Remote fault diagnosis method for traction drive systems based on domain-adapted knowledge graphs. In: Proceedings of the 2024 6th International Conference on Industrial Artificial Intelligence (IAI); 2024 Aug 21–24; Liaoning, China. New York City: IEEE; 2024. p. 1–2.
[40] Han H, Wang J, Wang X. Leveraging knowledge graph reasoning in a multihop question answering system for hot rolling line fault diagnosis. IEEE Trans Instrum Meas 2024;73:1–14.
[41] Liu B, Wu X, Pan D, Chen Y, Huang J, Liao F, et al. Enhancing large language models with graph-based node sampling for fault attribution in power distribution networks. In: Proceedings of the Advanced Intelligent Computing Technology and Applications: 20th International Conference, ICIC 2024; 2024 Aug 5–8; Tianjin, China. Berlin: Springer; 2024. p. 444–55.
[42] Guo T, Yang Q, Wang C, Liu Y, Li P, Tang J, et al. Knowledge navigator: leveraging large language models for enhanced reasoning over knowledge graph. Complex Intell Syst 2024;10(5):7063–76.
[43] Chen Q, Li Q, Wu J, Mao C, Peng G, Wang D. Application of knowledge graph in power system fault diagnosis and disposal: a critical review and perspectives. Front Energy Res 2022;10:988280.
[44] Li Q, Jiang P, Wang J, Yang M, Yang Y. A kind of intelligent dynamic industrial event knowledge graph and its application in process stability evaluation. J Intell Manuf 2025;36(3):1801–18.
[45] Chen S, Cheng L, Deng J, Wang T. Multi-feature fusion event argument entity recognition method for industrial robot fault diagnosis. Appl Sci 2022;12(23):12359.
[46] Tao F, Qi Q, Wang L, Nee AYC. Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: correlation and comparison. Engineering 2019;5(4):653–61.
[47] Wang J, Yin W, Gao J. The construction and application of knowledge graph-based fault diagnostic system of CNC machine tool. In: Proceedings of the 3rd International Conference on Mechanical, Electronics, and Electrical and Automation Control (METMS 2023); 2023 Jul 17–19; Hangzhou, China. Bellingham: SPIE; 2023. p. 3.
[48] An Y, Qin F, Sun D, Wu H. A multi-facets ontology matching approach for generating PLC domain knowledge graphs. IFAC-PapersOnLine 2020;53(2):10929–34.
[49] Xue R, Zhang P, Huang Z, Wang J. Digital twin-driven fault diagnosis for CNC machine tool. Int J Adv Manuf Technol 2024;131(11):5457–70.
[50] Tama BA, Vania M, Lee S, Lim S. Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals. Artif Intell Rev 2023;56(5):4667–709.
[51] Wang X, Liu M, Liu C, Ling L, Zhang X. Data-driven and knowledge-based predictive maintenance method for industrial robots for the production stability of intelligent manufacturing. Expert Syst Appl 2023;234:121136.
[52] Ding A, Qin Y, Wang B, Guo L, Jia L, Cheng X. Evolvable graph neural network for system-level incremental fault diagnosis of train transmission systems. Mech Syst Signal Process 2024;210:111175.
[53] Miao Z, Xia Y, Zhou F, Yuan X. Fault diagnosis of wheeled robot based on prior knowledge and spatial-temporal difference graph convolutional network. IEEE Trans Industr Inform 2023;19(5):7055–65.
[54] Xu S, Chen M, Chen S. Enhancing retrieval-augmented generation models with knowledge graphs: innovative practices through a dual-pathway approach. In: Proceedings of the 20th International Conference, ICIC 2024; 2023 Aug 5–8; Tianjin, China. Singapore: Springer Singapore; 2024. p. 398–409.
[55] Bahr L, Wehner C, Wewerka J, Bittencourt J, Schmid U, Daub R. Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis. J Ind Inf Integr 2024;45:100807.
[56] Sarmah B, Hall B, Rao R, Patel S, Pasquali S, Mehta D. HybridRAG: integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In: Proceedings of the 5th ACM International Conference on AI in Finance; 2024 Nov 14–17; New York City: Association for Computing Machinery (ACM); 2024. p. 608–16.
[57] Zhang Y, Hao Y. Traditional Chinese medicine knowledge graph construction based on large language models. Electronics (Basel) 2024;13(7):1395.
[58] Bi Z, Chen J, Jiang Y, Xiong F, Guo W, Zhang N. CodeKGC: code language model for generative knowledge graph construction. ACM Trans Asian Low-Resour Lang Inf Process 2024;23(3):1–16.
[59] Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A systematic survey of prompt engineering in large language models: techniques and applications. 2024. arXiv:2402.07927.
[60] Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in large language models. 2023. arXiv:2210.03493.
[61] Allen BP, Stork L, Groth P. Knowledge engineering using large language models. 2023. arXiv:2310.00637.
[62] Hendrycks D, Burns C, Basart S, Zou A, Mazsika M, Sonh D, et al. Measuring massive multitask language understanding. 2020. arXiv:2009.03300.
[63] GLM T, Zeng A, Xu B, Wang B, Zhang C, Yin D, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. 2024. arXiv.2406.12793.
[64] Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen technical report. 2023. arXiv.2309.16609.
[65] Frantar E, Ashkboos S, Hoefler T, Alistarh D. GPTQ: accurate post-training quantization for generative pre-trained transformers. 2023. arXiv.2210.17323.