10.37965/jdmd.2026.1561

A Method for Generating Safety Measures of Hydropower Plant Work Tickets Integrating Knowledge Graph and Large Language Models

Xinyang Yi[1, 2], Ran Duan[3, 4, 5], Yitao Fei[1], Wanru Liu[1], Chaoying Yang[6, *], Jie Liu[1, 3], Zebin Cheng[2]

(1. School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China;

2. Hubei Qingjiang Hydropower Development Co., Ltd, Yichang 443000, China;

3. Hubei Key Laboratory of Digital Valley Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;

4. Changjiang Survey Planning Design Research Co., Ltd, Wuhan 430014, China;

5. Key Laboratory of Internet Plus Smart Water, CWRC, Wuhan 430014, China;

6. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China)

* Chaoying Yang: yangchaoying211@gmail.com

Received 25 June 2026; Revised 21 July 2026; Accepted 04 August 2026; Published online 11 August 2026

Abstract:

To address the low efficiency and error-prone nature of manually formulating safety measures in hydropower plant work tickets, as well as the difficulty of existing methods in balancing accuracy and compliance, this paper proposes a safety measure generation method integrating a knowledge graph (KG) and large language models (LLMs). A work ticket KG is constructed to capture historical associations between work contents and safety measures, while a safety regulation vector database is established as a rigid compliance constraint, forming a dual-modal safety knowledge base. Named entity recognition and a weighted similarity retrieval algorithm are employed to obtain relevant subgraphs from the KG. Based on the retrieved context, a locally deployed LLM generates candidate safety measures. Furthermore, a compliance review mechanism, combining opposite-action veto and KG co-occurrence statistics, is introduced to evaluate both compliance and necessity. Experiments on 1,600 test work tickets show that the proposed method achieves a precision of 91.5%, a recall of 87.8%, and a compliance score of 86.7, outperforming all baseline configurations and improving the reliability and automation of work-ticket processing.

Keywords: hydropower plant work tickets; knowledge graph; large language models; safety measure generation; compliance review

I Introduction

With the advancement of new power systems, hydropower plays an increasingly important role in peak shaving, frequency regulation, and emergency reserves [1]. Ensuring its safe and stable operation is therefore essential for power system reliability [2]. The work ticket system is a core component of power production safety management and provides the procedural basis for field operations [3]. However, hydropower operation and maintenance (O&M) tasks often involve multiple interconnected devices and complex electrical and mechanical isolation measures, making safety measure formulation highly dependent on expert knowledge [4,5]. Although existing generation management systems support workflow management, they provide limited intelligent assistance for safety measure formulation. In practice, operators still rely on personal experience or historical work tickets, which can lead to omitted safety measures and inconsistent implementation under complex operating conditions [6]. Therefore, developing an automated and reliable safety measure generation method has become an urgent requirement.

Owing to the extensive variety and intricate interdependencies of equipment in hydropower stations, conventional work ticket generation methods based on rule bases or expert systems suffer from high maintenance costs and limited generalizability [7]. To improve automation, researchers have introduced Natural Language Processing (NLP) techniques into power O&M applications[8]. For example, Zhang et al. proposed a text mining and natural language processing method for construction site accident analysis to improve safety knowledge extraction efficiency [9]. Lawal et al. developed machine learning models for optimising tunnel support design, demonstrating the potential of data-driven methods in underground engineering [10]. Although these methods improve information extraction and text processing capabilities, they mainly rely on shallow semantic analysis and struggle to capture complex operational logic, limiting the quality of generated safety measures.

The emergence of LLMs has created new opportunities for intelligent power O&M [11,12]. Li et al. proposed SafetyGPT, a multimodal safety agent for detecting unsafe worker behaviors through human–machine collaboration [13]. Wang et al. combined KGs with LLMs to develop a prompt-enhanced question-answering framework for electric power applications [14]. Li et al. developed an LLM-based question-answering system for power safety regulations using text classification and vector retrieval [15]. Cheng et al. introduced GAIA for power dispatch tasks, including operation adjustment, operation monitoring, and black-start support [16]. Xiang et al. integrated domain-specific prompting with fuzzy evaluation to assess power-system operating states and generate dispatch recommendations [17]. Kaya et al. evaluated LLMs in STPA- and FRAM-based safety assessment [18]. The models generated substantial information rapidly. However, they showed weaknesses in systems thinking and output consistency. Despite these advances, LLMs remain vulnerable to hallucinations in domain-specific tasks [19]. They may generate incorrect equipment references, use non-standard terminology, or omit critical operational steps. These risks limit their direct application to safety-critical work-ticket generation.

To mitigate the reliability degradation caused by hallucinations, integrating KGs with retrieval-augmented generation (RAG) has become a major research direction, where KGs provide precise and reliable knowledge retrieval to guide LLM generation [20]. Owing to their structured relational representation and strong interpretability, KGs have been widely applied in engineering decision support [21]. For example, Zhang et al. constructed causally embedded multi-knowledge graphs for geological fault identification, significantly improving feature representation under complex geological conditions [22]. Feng et al. proposed a KG-based framework for generating calculation schemes in water project scheduling to improve generation reliability through structured knowledge representation [23]. Wang et al. constructed a domain-specific industrial safety knowledge graph to support hazard analysis and decision-making in complex industrial processes [24]. Mao et al. developed a multimodal RAG framework that jointly exploits textual, visual, and topological information to enhance knowledge retrieval for power grid work tickets [25]. Nevertheless, existing studies generally rely on a single knowledge source and lack effective coordination between historical operational experience and safety regulation constraints[26]. Consequently, methods based solely on knowledge retrieval or LLM generation cannot simultaneously guarantee the accuracy and compliance of generated safety measures.

To address the above limitations, this paper proposes a safety measure generation method for hydropower plant work tickets that integrates KG and LLMs. The main innovations are as follows:

(1) A dual-modal knowledge base integrating a historical work ticket KG with a safety regulation vector database is constructed to simultaneously support experience reuse and compliance verification.

(2) A graph-enhanced retrieval framework based on LLMs is developed to jointly exploit KG information and semantic reasoning capability, thereby improving the accuracy and contextual relevance of generated safety measures.

(3) A review mechanism integrating feature decoupling, opposite-action veto mechanism, and graph co-occurrence support is proposed to enforce safety regulation constraints, thereby enhancing the compliance and reliability of the generated safety measures.

II Methodology

The overall framework of the proposed safety measure generation method is illustrated in Fig. 1. The framework comprises three main stages: dual-modal knowledge base construction, graph-enhanced safety measure generation, and safety measure review. A historical work-ticket KG and a safety regulation vector database are first constructed to provide structured operational experience and compliance knowledge. Based on the retrieved graph context, a large language model generates candidate safety measures through graph-enhanced retrieval. The generated results are then verified using a review mechanism that integrates feature decoupling, opposite-action veto mechanism, and graph co-occurrence support to ensure both semantic correctness and regulatory compliance. The details of each stage are described in the following subsections.

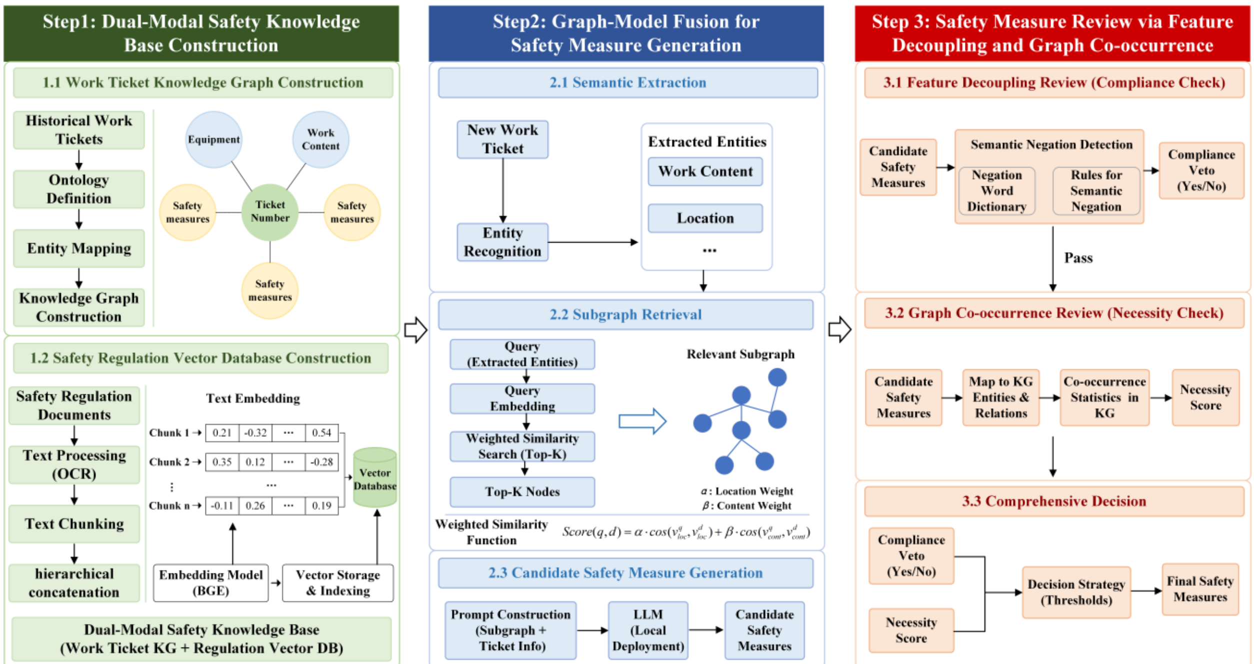

Fig. 1 Flowchart of the Proposed Framework

2.1 Construction of a Dual-Modal Knowledge Base for Safety Measures

2.1.1 Multi-Source Data Preprocessing

Multi-source power data require different preprocessing strategies. Historical work tickets are processed through field mapping to preserve factual information, whereas safety regulations require specialized parsing due to their complex hierarchical structures.

To retain document semantics, a structure restoration method is adopted. PDF files are first parsed using OCR, and a document logic tree is reconstructed based on layout features such as font size, indentation, and line spacing. The parsed content is then converted into Markdown format and manually verified. This process preserves heading hierarchies and contextual relationships, providing support for subsequent text chunking and knowledge modeling. The workflow is shown in Fig. 2.

10.37965/jdmd.2026.1561

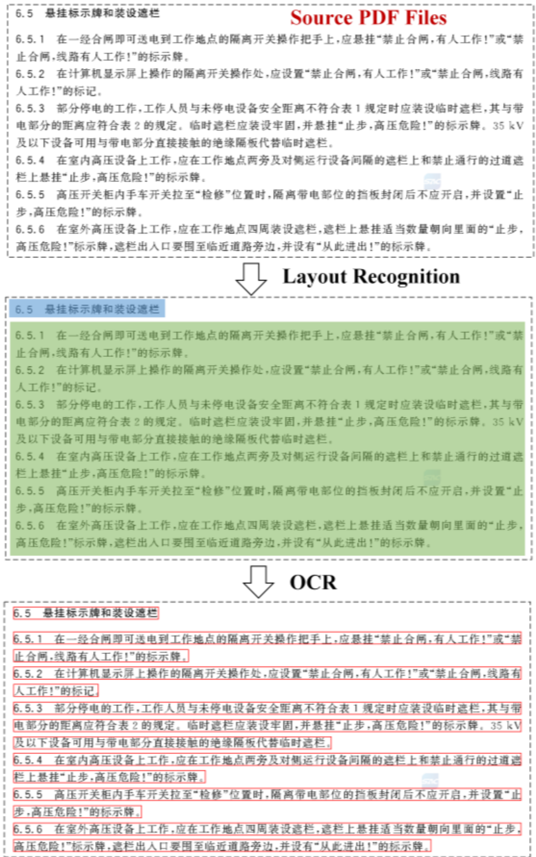

Fig. 2 OCR Recognition Process

2.1.2 Construction of the Work Ticket KG

To address the entity associations and business logic in work ticket data, a Neo4j graph database is used to construct a historical work ticket KG. This achieves the structured storage and association mining of historical operation information. To describe the graph structure, the work ticket KG is formally defined as a directed multigraph $G = (E, R, T)$. Here, E represents the set of all entities, R represents the set of relationships between entities, and T represents the set of triplets.

Automated scripts are used to read the structured Excel data. Using the "work ticket number" entity as the core, fields such as "work ticket number," "equipment name," "work content," and "safety measures" are mapped and instantiated as graph nodes. Based on this, the semantic relationships between entities are strictly defined. The relationship set is defined as $R = \{r_1, r_2, r_3, r_4\}$. These correspond to "operation object," "task," "power outage status," and "required measures," respectively. The triplet set T can be expressed as follows:

$$T = \{(h,r,t) \mid h \in E_{ticket}, t \in E_{attribute}, r \in R\} \quad (1)$$

where h represents the head entity (i.e., the work ticket number), t represents the tail entity, and r denotes the semantic edge connecting them. The structure of the work ticket KG is shown in Fig. 3.

Through the batch import and construction of these relationships, flat data tables are transformed into a networked knowledge structure. This provides knowledge support for subsequent graph-based complex retrieval and reasoning.

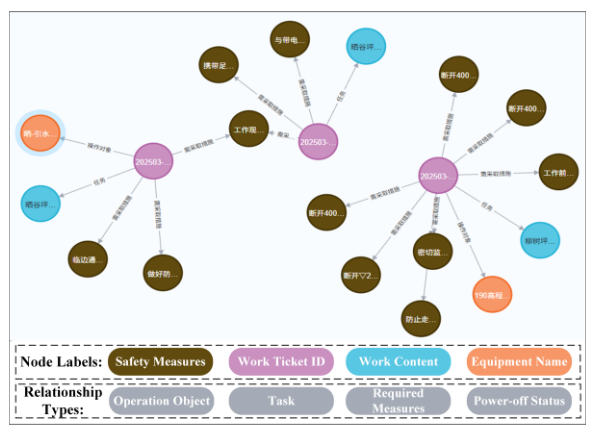

Fig. 3 Work Ticket KG Structure

2.1.3 Vectorized Representation of Safety Regulation Knowledge

To solve the problems of semantic loss and context fragmentation when processing long safety regulation documents in RAG tasks, a semantic chunking strategy based on the document heading hierarchy is adopted.

This strategy replaces the traditional fixed-character chunking method. By following the hierarchical structure of safety regulations, this approach preserves the semantic completeness of each provision or operational rule within a chunk. It also avoids semantic fragmentation caused by mechanical truncation. The knowledge vectorization process is shown in Fig. 4.

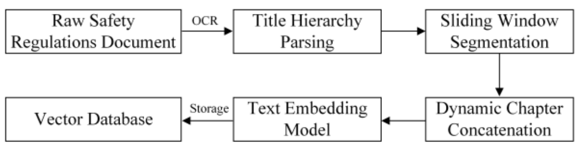

Fig. 4 Knowledge Vectorization Process

The original safety regulation document sequence is defined as P. It is divided into a set of L chunks using the heading-level chunking strategy, $S=\{s_1,s_2,...,s_L\}$. Meanwhile, to enhance the semantic coherence between chunks, a sliding window mechanism is introduced. For the i-th chunk s_i, its text range is constrained by the context overlap rate λ. Furthermore, to address the lack of global background after the chunks are isolated, a hierarchical concatenation strategy is adopted, as shown in Fig. 5. Specifically, the multi-level heading path of each chunk is dynamically concatenated to its head. The heading path H_i of the current chunk is concatenated before the main content C_i to form the input sequence X_i. This explicitly preserves the hierarchical context information of the regulations. The construction process is defined as follows:

$$X_i = Concat(H_i, SEP, C_i) \tag{2}$$

where *Concat* represents the concatenation function, and SEP denotes the separator.

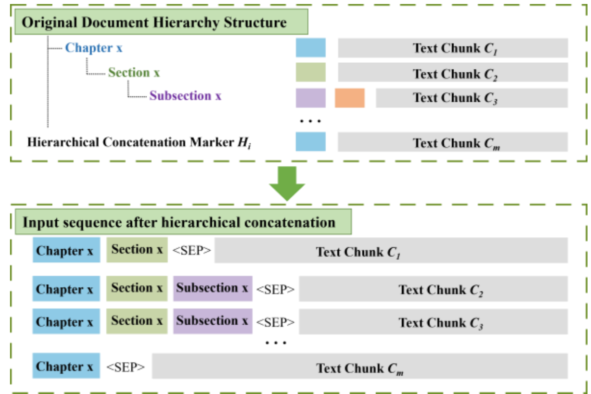

Fig. 5 Hierarchical Concatenation Strategy

These strategies have different but complementary functions. Heading-level chunking preserves the semantic integrity of each provision. The sliding-window mechanism reduces information loss at chunk boundaries. Hierarchical concatenation adds the regulatory context of each chunk. Together, these strategies improve retrieval completeness and accuracy. They also reduce irrelevant matches caused by isolated text fragments.

The processed sequences are then encoded into high-dimensional vectors using a pretrained text-embedding model. The vectors are stored in a vector database for subsequent indexing and semantic retrieval.

2.2 Generation of Safety Measures Based on Graph-Model Fusion

2.2.1 Semantic Extraction Based on Context Decoupling

In power operation scenarios, work ticket descriptions are usually written in unstructured natural language. They often couple the operation location with the work content, resulting in high semantic coupling and inaccurate knowledge retrieval. To solve this problem, an LLM-based context semantic decoupling method is proposed. It converts unstructured text into structured semantic slots through prompt-guided entity extraction. Fig. 6 illustrates the process. For example, given the input *"Inspection of the diversion penstock and high slope at Xisiping Hydropower Station"*, the LLM outputs `{"Location": "Xisiping Hydropower Station", "Content": "Inspection of the diversion penstock and high slope"}`.

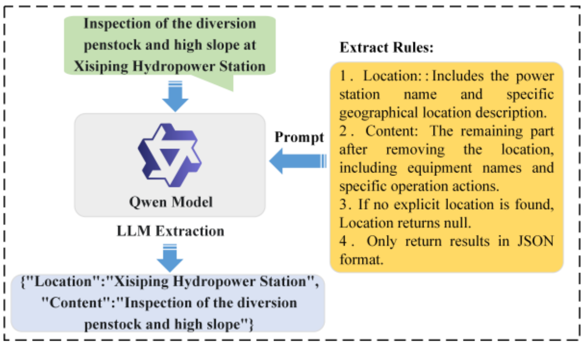

Fig. 6 Semantic Decoupling Extraction

2.2.2 KG-Based Subgraph Retrieval of Operational Experience

To effectively reuse historical safety measure experience, a KG retrieval algorithm based on weighted semantic similarity is proposed. Since work content is more critical than station location for safety measure formulation, an asymmetric weighting mechanism is introduced. Specifically, a text embedding model maps the decoupled query vector v_q to the feature vector v_d of historical graph nodes, and a weighted similarity function is defined to reduce retrieval noise caused by location differences among similar operations:

$$Score(q,d) = \alpha \cdot cos(v_{loc}^{q}, v_{loc}^{d}) + \beta \cdot cos(v_{cont}^{q}, v_{cont}^{d}) \tag{3}$$

where $Score(q,d)$ denotes the scoring function, and $cos(\cdot)$ represents the cosine similarity calculation. α and β are the weight coefficients for the location and the work content, respectively, satisfying $\alpha+\beta=1$.

Based on this scoring function, a set of Top-K nodes with the highest similarities, denoted as N_{top}, is retrieved from all historical nodes. Subsequently, each node in N_{top} is used as an anchor to perform a 1-hop traversal in the KG. The associated triplets are extracted to construct a subgraph G_{sub} containing historical operation experience. The triplet set T_{sub} of this subgraph can be expressed as:

$$T_{sub} = \{(h,r,t) \mid h \in N_{top} \vee t \in N_{top}, (h,r,t) \in G_{kg}\} \tag{4}$$

2.2.3 LLM-Driven Safety Measure Generation

After retrieving the operation experience subgraph, the KG triplets are converted into natural language safety measures using a prompt-based RAG framework. The retrieved triplets serve as contextual knowledge and are combined with task instructions and generation constraints to construct the prompt. A locally deployed Qwen model then generates the safety measures. As shown in Fig. 7, the model consists of stacked Transformer decoder layers, each including a multi-head self-attention mechanism and a feed-forward neural network[27].

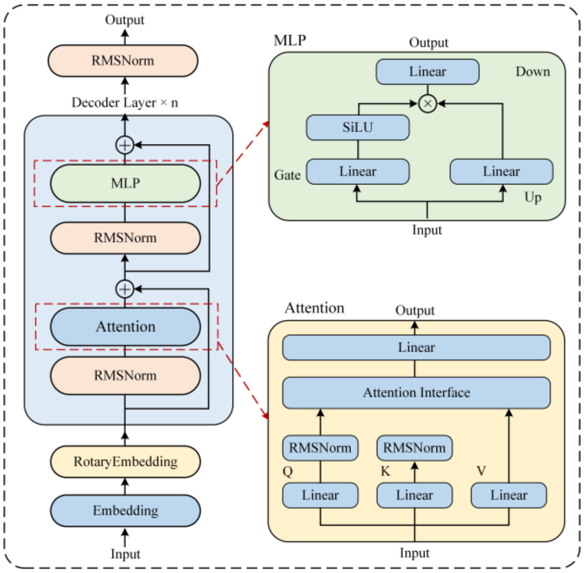

Fig. 7 Qwen Model Architecture Diagram

To ensure the reliability and traceability of the generated content, the prompt requires the model to generate safety measures based on the retrieved knowledge and explicitly indicate the reference sources. The final results are output in a structured JSON format. The safety measure generation prompt is constructed as follows.

Safety Measure Generation Prompt

System: *You are an expert in the formulation of operational safety measures for hydropower plants. Your task is to generate safety measures based on the provided information.*

User:

Requirements:

1. Prioritize the facts provided in the knowledge base and do not fabricate information.

2. The plant/station location referenced in the safety measures must be consistent with the query text.

3. Output must be strictly in JSON format with the following fields:

{

"measures": ["Measure 1", "Measure 2", "..."],

"citations": ["Reference 1", "Reference 2", "..."]

}

- `measures` must be an array, and each item must be a plain-text safety measure without numbering or line breaks; at least one measure must be included.

- `citations` may be an empty array; if provided, they should briefly describe the source content.

4. Do not output any text other than the JSON content.

5. Some retrieved content may be irrelevant to the task; ensure that the generated safety measures are relevant to the task.

6. Safety measures should be practical and feasible, avoiding vague or generic statements.

2.3 Safety Measure Review Based on Feature Decoupling and Graph Co-occurrence

In power operations, work ticket safety measures must be logically correct and operationally reliable. Although LLMs can generate candidate measures, hallucinations may cause equipment misidentification or incorrect operations, posing safety risks. Moreover, text vector similarity alone is insufficient for review. It cannot reliably distinguish opposite operations (e.g., “disconnect” vs. “close” or “fully close” vs. “fully open”) and fails to assess whether a safety measure is necessary in a specific operation scenario.

To address these issues, a review mechanism integrating feature decoupling, opposite-action veto, and graph co-occurrence statistics is proposed. By constructing a compliance and necessity review framework, the generated candidate safety measures are intelligently reviewed.

2.3.1 Rule Template-Based Feature Decoupling of Safety Measures

Given the standardized nature of hydropower safety instructions, a rule template-based feature decoupling algorithm is designed to convert LLM-generated safety measures into structured representations. A standard operation action dictionary, A_{dict} is first constructed from power safety regulations and historical work tickets. It includes operation verbs such as "fully close," "disconnect," "put into service," and "take out of service." Candidate safety measures are then classified into operational isolation and general specification categories according to whether they contain actions in A_{dict}. Different feature decoupling strategies are subsequently applied. The classification criteria and decoupling rules are listed in Table 1. For general specifications, the complete sentence is retained as the retrieval query, and compliance is evaluated through semantic matching without applying the opposite-action penalty.

Table 1 Safety Measure Classification and Decoupling Rules

Measure type	Trigger condition	Extracted elements	Rule	Example	Structured result
Operational isolation	Contains action verbs in a_{dict}	Action, equipment	Regex + a_{dict} Matching	"fully close unit 1f inlet butterfly valve 1df, gao self-provided power station"	Action: fully close; Equipment: inlet butterfly valve 1df of unit 1f, gao self-owned power station
General specification	No action verb matched	Complete sentence	No processing	"enter the work site and complete the "two wears & one wear" requirements."	Complete sentence retained

2.3.2 Compliance Review Based on an Opposite-Action Veto Mechanism

The compliance review enforces safety regulations and prevents operation errors. For operational isolation measures, it incorporates an opposite-action veto mechanism based on regulation vector retrieval. Let m_i be a candidate safety measure generated by the LLM, and r_j be the most semantically similar clause retrieved from the regulation vector database. A power operation opposite-action dictionary, denoted as D_{ant}, is constructed. This dictionary contains mutually exclusive action pairs. It is used to modify the traditional cosine similarity calculation. The modified compliance matching function S_{comp} is defined as follows:

$$S_{comp}(m_i, r_j) = sim(m_i, r_j) \cdot [1 - P_{conflict}(m_i, r_j)] \tag{5}$$

where $sim(m_i, r_j)$ is the cosine similarity based on the text embedding model, and $P_{conflict} \in \{0,1\}$ is the penalty function. If the action intents of m_i and r_j belong to an opposite-action pair in D_{ant}, $P_{conflict}$ equals 1; otherwise, it equals 0. If the following condition is met:

$$S_{comp} \geq \delta \tag{6}$$

where δ is a preset compliance threshold, the safety measure m_i is considered compliant at the regulation level.

2.3.3 Necessity Review Based on Historical Graph Co-occurrence

The necessity review verifies whether candidate safety measures satisfy the minimum safety requirements of the current operation. It uses the relational structure of the work ticket KG to perform probabilistic reasoning based on co-occurrence support, ensuring consistency with the operation scenario and avoiding redundant or missing measures.

For a given work task w, the set of historical work ticket nodes containing w, denoted as $N_{ticket}(w)$, is retrieved from the KG. For any candidate safety measure m_k, its historical co-occurrence support in this task is defined as $S_{sup}(m_k \mid w)$:

$$S_{sup}(m_k \mid w) = \frac{|N_{ticket}(w) \cap N_{ticket}(m_k)|}{|N_{ticket}(w)|} \tag{7}$$

where $N_{ticket}(m_k)$ is the set of historical work tickets in the KG containing the candidate safety measure m_k, and $|\cdot|$ represents the cardinality of this set. $S_{sup}(m_k \mid w)$ reflects the probability of adopting this safety measure in similar historical operations. If the candidate safety measure m_k satisfies the following condition:

$$S_{sup}(m_k \mid w) \geq \gamma \tag{8}$$

where γ is a preset necessity threshold. If this condition is met, the safety measure is considered to be strongly correlated with the current work content. Conversely, if a safety measure meets the regulatory requirements but is not supported by similar historical operation scenarios, it is classified as a redundant measure.

III Experimental Validation

3.1 Experimental Setup

To evaluate the proposed method, an experimental dataset was constructed from real-world work tickets and industry safety regulations. The operational data consist of historical work tickets collected from a cascade hydropower plant between January 2015 and October 2025. The dataset covers primary electrical, secondary electrical, and hydro-mechanical operations and includes task descriptions, equipment information, safety measures, and related records. The regulation corpus was built from national, industry, and enterprise safety standards in the power sector.

(1) Work ticket KG: After deduplication, cleaning, and desensitization, 8,000 valid work tickets were retained. Among them, 6,400 (80%) were used to construct the knowledge base, and the remaining 1,600 served as the test set. The resulting KG follows a four-level hierarchy of *station–equipment–task–safety measure* and contains 32,695 entities and 53,621 triplets.

(2) Regulation knowledge base: Five core safety regulations were selected as the

compliance basis, including the *Safety Work Regulations for Power Grids (Electrical Part of Power Plants and Substations)* and the *Two-Ticket Management Standard* of a hydropower company. Using a sliding-window strategy (window size: 512 tokens, overlap: $\lambda = 10\%$), 2,600 regulation vector indices were constructed.

To ensure data security, all experiments were conducted in a local offline environment. The framework was implemented with Python 3.10 and PyTorch 2.6.0. Neo4j and Chroma were used to store the work ticket KG and regulation vector database, respectively, while vLLM provided the LLM inference service. All experiments were performed on an Ubuntu 22.04 server. The hardware configurations are shown in Table 2.

Table 2 Experimental Hardware Parameter Configuration

Configuration Item	Value
Operating System	Ubuntu 22.04
CPU	Intel(R) Xeon(R) Gold 6330
GPU	NVIDIA GeForce RTX 4090D 24GB

Qwen3-8B was selected as the base model for local deployment, considering model performance, size, and GPU memory constraints. During inference, the temperature was set to 0.1 and Top-P to 0.9 to ensure deterministic outputs. The BGE-Large-zh-v1.5 model was used for semantic embedding with a dimension of 1024, supporting accurate KG retrieval and regulation matching.

For the review module, the compliance threshold δ and necessity threshold γ were determined using 100 safety measure–regulation pairs and 100 safety measure–task pairs sampled from the test set. These samples were used to compute S_{comp} and S_{sup}, and their distributions are shown in Fig. 8.

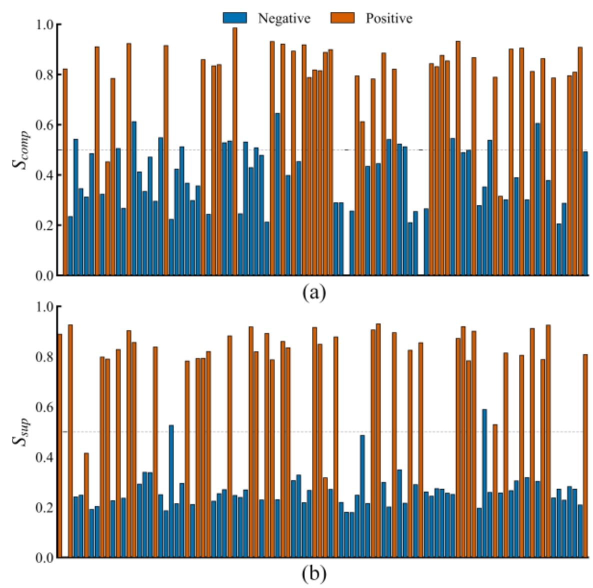

Fig. 8 Sample Similarity Distribution Characteristics: (a) S_{comp} Score; (b) S_{sup} Score

As shown in Fig. 8(a), compliance scores of positive samples are mainly distributed in the range of 0.78–0.95, whereas negative samples are concentrated below 0.60. Therefore, the distribution valley at δ=0.70 was selected as the compliance threshold.

As shown in Fig. 8(b), the co-occurrence support of positive samples is generally above 0.75, while that of negative samples is mostly below 0.35. Accordingly, the necessity threshold was set to γ=0.75.

3.2 Parameter Sensitivity Analysis

The station location weight α and work content weight β control the retrieval strategy of the proposed weighted similarity algorithm. To validate the importance of work content and determine the optimal parameter settings, a sensitivity analysis was conducted on the test set.

In the absence of large-scale annotated retrieval pairs, weight parameters were evaluated using safety measure consistency. Specifically, the Average Safety Measure Semantic Similarity (AMSS) measures the semantic similarity between the ground-truth safety measures and the Top-1 retrieved historical safety measures. The mathematical definition is:

$$AMSS(\alpha)=\frac{1}{|D_{test}|}\sum_{i\in D_{test}} sim(M_i^T, M_i^R(\alpha)) \quad (9)$$

where D_{test} is the test set, M_i^T is the ground-truth safety measure set of the i-th test sample, $M_i^R(\alpha)$ is the safety measure set of the Top-1 historical node retrieved from the knowledge base under the parameter α configuration, and $sim(\cdot)$ is calculated using cosine similarity based on the BGE embedding model.

Because $\alpha+\beta=1$, a grid search was performed for α in the range [0,1] with a step size of 0.1. The AMSS values under different settings are shown in Fig. 9.

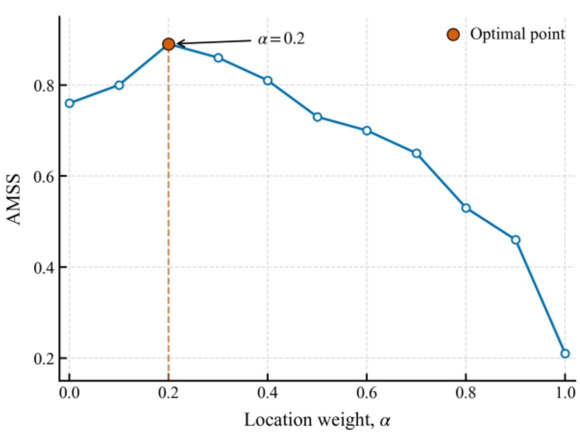

Fig. 9 AMSS Value Variation Curve

As α increases, the AMSS first increases and then decreases. When α=0 and α=1, the AMSS is 0.76 and 0.21, respectively, indicating that neither work content nor location alone achieves satisfactory retrieval performance. The best result is obtained at α=0.2, where the AMSS reaches 0.89, suggesting that an 80% work content weight combined with a 20% location weight provides the most effective retrieval.

3.3 Experimental Analysis

3.3.1 Evaluation Metrics

To evaluate the proposed method, experiments were conducted on a work ticket test set covering typical maintenance, defect elimination, and equipment commissioning scenarios. Performance was assessed using Precision (P), Recall (R), and Compliance Score (C_{score}).

(1) Precision (P): Measures the proportion of correct and executable safety measures in the generated results.

$$P=\frac{N_{correct}}{N_{total}}\times 100\% \quad (10)$$

where $N_{correct}$ is the number of safety measure items judged by experts as correct, unambiguous, and specific to the current work content, and N_{total} is the total number of safety measure items generated by the model.

(2) Recall (R): Measures the coverage of standard safety measures by the generated results.

$$R=\frac{N_{correct}}{N_{standard}}\times 100\% \quad (11)$$

where $N_{standard}$ is the total number of safety measure items that should be included in the

10.37965/jdmd.2026.1561

standard work ticket corresponding to the work content.

(3) Compliance Score (C_{score}): This metric evaluates the regulatory compliance of the generated safety measures on a 100-point scale. The outputs of different methods were anonymized and randomly ordered before evaluation. Three hydropower experts with more than five years of experience in work-ticket preparation independently scored each sample according to the deduction rules listed in Table 3. The arithmetic mean of the three expert ratings was taken as the compliance score for each sample, thereby reducing the influence of individual scoring differences. The final C_{score} was calculated by averaging the sample-level scores over the entire test set. The detailed scoring criteria are listed in Table 3.

Table 3 Compliance Scoring Rules

Evaluation Item	Penalty	Description
Compliance Score	100 points	Full score
Critical Defect	−100	Violates mandatory safety rules (e.g., no power isolation or grounding)
General Defect	−10 / item	Incorrect description or terminology
Redundant Item	−2 / item	Irrelevant or repeated measures

3.3.2 Baseline Methods

To evaluate the effectiveness of different knowledge enhancement and review strategies, four baseline methods were compared with the proposed approach (M5). The configurations of all experimental groups are summarized in Table 4.

Table 4 Configurations of Experimental Groups

Group	Configuration	Incremental Component
M1	Foundation LLM	None
M2	M1 + Regulation Vector database	Regulation knowledge
M3	M1 + Work Ticket KG	Historical knowledge
M4	M1 + KG + Regulation Vector database	Dual-modal knowledge
M5	M4 + Review Mechanism	Compliance & necessity review

3.3.3 Experimental Results

Fig. 10 compares the performance of the five methods in terms of P, R, C_{score}.

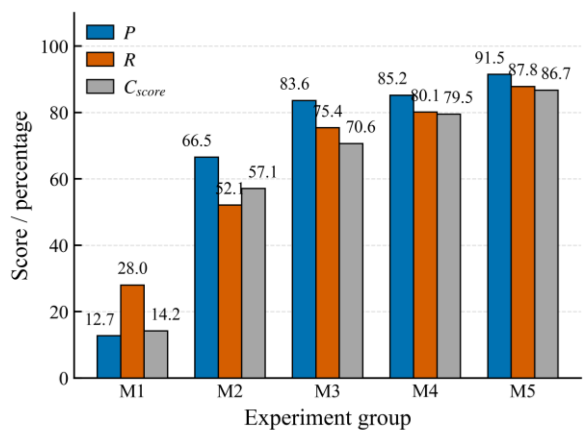

Fig. 10 Performance Comparison of the Five Experimental Configurations.

As shown in Fig. 10, the proposed method (M5) achieves the best performance across all evaluation metrics, with a precision of 91.5%, a recall of 87.8%, and a compliance score of 86.7. Compared with the baseline LLM (M1), introducing either regulation knowledge (M2) or historical work ticket knowledge (M3) significantly improves generation quality. Specifically, M2 substantially increases the compliance score, whereas M3 achieves higher precision and recall. Integrating both knowledge sources (M4) further improves overall performance, demonstrating their complementary benefits.

The additional gains of M5 result from the proposed review mechanism. Without post-review, M4 still suffers from hallucinations, including confusion between opposite operational actions and redundant safety measures. In contrast, the opposite-action veto mechanism effectively eliminates conflicting operations, while graph co-occurrence reasoning removes redundant measures and supplements missing ones. Consequently, M5 achieves the highest accuracy, coverage, and regulatory compliance among all methods.

To further evaluate result stability, 200 samples were randomly selected from the test set. The distribution of compliance scores for the five methods is shown in the box plot in Fig. 11.

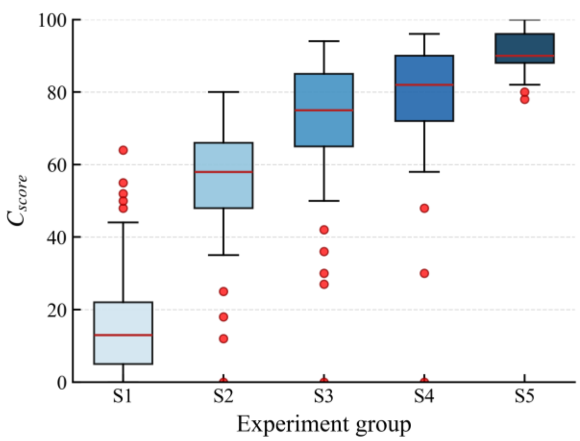

Fig. 11 Box plot of Safety Measure Compliance Scores

As shown in Fig. 11, M1 exhibits the lowest median score (13), indicating that the foundation LLM cannot generate engineering-feasible safety measures without domain knowledge. Compared with M1, both M2 and M3 substantially improve compliance scores but still produce numerous low-score outliers, reflecting the limitations of relying solely on regulation knowledge or historical experience. Although M4 further shifts the score distribution upward, low-score outliers remain because of LLM hallucinations.

In contrast, M5 achieves the highest median score and the smallest interquartile range while effectively eliminating most extreme low-score outliers. These results demonstrate that the proposed review mechanism enhances both the compliance and robustness of safety measure generation by resolving conflicting operations and reducing redundant or missing measures.

IV Conclusions

Traditional methods struggle to balance historical operational experience with safety regulation constraints in hydropower plant O&M. To address this issue, this paper proposes an intelligent safety measure generation method that integrates a historical work ticket KG with safety regulation constraints. Experimental results demonstrate its effectiveness in improving both generation quality and regulatory compliance. The main conclusions are summarized as follows:

(1) A dual-modal knowledge base combining a historical work ticket KG and a regulation vector database was constructed, enabling effective reuse of operational experience while providing reliable compliance constraints.

(2) A graph-enhanced retrieval framework with weighted semantic similarity was developed to jointly exploit work content and location information. Experiments show

that dual-modal knowledge significantly improves the precision and recall of safety measure generation compared with single-source methods.

(3) A review mechanism integrating opposite-action veto and graph co-occurrence reasoning was proposed to perform compliance and necessity verification, further improving the reliability and robustness of the generated safety measures.

The proposed method still depends on the quality of historical work tickets and the structured processing of safety regulations, and its generalizability to other hydropower plants requires further validation. Future work will integrate multimodal equipment information, plant topology, and real-time operational data to support more adaptive and reliable safety-measure generation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 52575115), the Natural Science Foundation of Hubei Province (No. 2024AFB428), and the Independent Innovation Project of CISPDR (No. CX2022Z16-1).

References

[1] S. J. W. Klein and E. L. B. Fox, "A review of small hydropower performance and cost," *Renew. Sustain. Energy Rev.*, vol. 169, Art. no. 112898, 2022.

[2] Z. Kang *et al.*, "Design and application of multi basin hydropower and new energy centralized control center system," *J. Phys.: Conf. Ser.*, vol. 3150, no. 1, pp. 012029, 2025.

[3] C. Huang *et al.*, "Intelligent management model for two-ticket system of power grid enterprise safety production by big data," *in Proc. 3rd Int. Conf. Electron. Inf. Technol. (EIT)*, pp. 302–305, 2024.

[4] J. Xiang, "Research on construction safety evaluation and management system of water conservancy and hydropower projects," *IOP Conf. Ser.: Earth Environ. Sci*., vol. 638, Art. no. 012087, 2021.

[5] E. Özcan *et al.*, "Risk based maintenance in the hydroelectric power plants," *Energies*, vol. 12, no. 8, pp. 1502, 2019.

[6] J. Bai *et al.*, "Research on an intelligent judgment method for grounding protection range of power work points based on large models and multi-source data fusion," *in Proc. 2025 IEEE 7th Int. Conf. Civil Aviat. Saf. Inf. Technol. (ICCASIT)*, pp. 1711–1716, 2025.

[7] F. Hao *et al.*, "An intelligent work ticket generation method for distribution networks based on domain knowledge augmentation and contextualised feature encoding," *IET Smart Grid*, vol. 9, no. 1, Art. no. e70080, 2026.

[8] M. Zhou *et al.*, "Progress in neural NLP: Modeling, learning, and reasoning," *Engineering*, vol. 6, no. 3, pp. 275–290, 2020.

[9] F. Zhang *et al.*, "Construction site accident analysis using text mining and natural language processing techniques," *Autom. Constr.*, vol. 99, pp. 238–248, 2019.

[10] A. I. Lawal *et al.*, "Optimising tunnel support design with machine learning

models," *Environ. Earth Sci.*, vol. 84, Art. no. 599, 2025.

[11] H. Mirshekali *et al*., "A Review of Large Language Models for Energy Systems: Applications, Challenges, and Future Prospects," *IEEE Access*, vol. 13, pp. 163162–163188, 2025.

[12] S. Majumder *et al*., "Exploring the capabilities and limitations of large language models in the electric energy sector," *Joule*, vol. 8, no. 6, pp. 1544–1549, 2024.

[13] W. Li *et al*., "SafetyGPT: An autonomous agent of electrical safety risks for monitoring workers' unsafe behaviors," *Int. J. Electr. Power Energy Syst*., vol. 168, Art. no. 110672, 2025.

[14] C. Wang *et al*., "Knowledge graphs enhanced large language model prompt for electric power question answering," *in Proc. 2023 7th Int. Conf. Electron. Inf. Technol. Comput. Eng. (EITCE)*, Xiamen, China, 2023.

[15] B. Li *et al*., "The design and implementation of an intelligent Q&A system for electric power safety regulations based on large language model technology," *in Proc. 2024 6th Int. Conf. Energy Syst. Electr. Power (ICESEP)*, pp. 604–607, 2024.

[16] Y. Cheng *et al*., "A large language model for advanced power dispatch," *Sci. Rep.*, vol. 15, Art. no. 8925, 2025.

[17] Y. Xiang *et al.*, "Large language model for secure operation of power systems," *Smart Energy Syst. Res.*, vol. 1, no. 1, Art. no. 10005, 2025.

[18] G. K. Kaya *et al.*, "Large language models powered system safety assessment: Applying STPA and FRAM," *Safety Sci.*, vol. 191, Art. no. 106960, 2025.

[19] D. Janiak *et al*., "The illusion of progress: Re-evaluating hallucination detection in LLMs," *in Proc. 2025 Conf. Empirical Methods in Natural Language Processing*, pp. 34716–34733, 2025.

[20] D. Sanmartin, "KG-RAG: Bridging the gap between knowledge and creativity," *arXiv preprint* arXiv:2405.12035, 2024.

[21] A. Hogan *et al*., "Knowledge graphs," *ACM Comput. Surv*., vol. 54, no. 4, pp. 1–37, 2021.

[22] F.-Y. Zhang *et al*., "Geological fault identification via multi-knowledge causally embedding graph representation applied in underground gas storage," *Pet. Sci*., vol. 23, no. 8, pp. 5234–5251, Aug. 2026.

[23] J. Feng *et al*., "Construction and application of water engineering dispatching knowledge graph based on large language model," *J. Front. Comput. Sci. Technol.*, vol. 18, no. 6, pp. 1637–1647, 2024.

[24] Z. Wang, B. Zhang, and D. Gao, "A novel knowledge graph development for industry design: A case study on indirect coal liquefaction process," *Comput. Ind*., vol. 139, Art. no. 103647, 2022.

[25] J. Mao *et al*., "MGRAG: MultiModal grid-aware retrieval augmentation generation framework for power grid work tickets," *Pattern Recognit*., vol. 169, Art. no. 111845, 2026.

10.37965/jdmd.2026.1561

[26] F.-Y. Zhang *et al.*, “Transferable generation of complex fracture morphologies: Fine-tuned knowledge from graph models learned in inter-well shale stimulation,” *Adv. Eng. Inform.*, vol. 76, Art. no. 105090, 2026.

[27] A. Yang *et al.*, “Qwen3 technical report,” *arXiv preprint* arXiv:2505.09388, 2025.