Renewable Energy and Power Quality Journal
https://repqj.com/index.php/repqj/index
RE&PQJ, Vol. 23, No. 6, 2025

Improving the Precision of Hidden Danger Recognition in Power Dispatch Duty Logs through RLHF Multi-round Human Feedback Mechanism

Siwu Yu*, Yumin He, Guobang Ban, Jintong Ma, Guanghui Xi, Lingwen Meng, Shasha Luo, Siqi Guo

Electric Power Research Institute of Guizhou Power Grid Co, Guiyang, 550002, China
*Corresponding author's email: swyu2012@163.com

Abstract. The problem of hidden danger identification in power dispatch duty logs is that professional text semantics are complex and expert annotations are scarce, resulting in insufficient recognition accuracy. This paper proposes an optimization method based on multi-round RLHF (Reinforcement Learning from Human Feedback). The reward model is trained through interactive expert feedback to drive the fine-tuning of the BERT model, and active learning is combined to screen high-value samples to achieve continuous improvement in the accuracy of hidden danger identification. A multi-dimensional reward function based on semantic similarity and hidden danger severity is designed. The reward model is trained using real-time expert scoring of the model output to quantify the recognition accuracy. With the reward model as the optimization target, the PPO (Proximal Policy Optimization) algorithm is used to fine-tune the pre-trained BERT model for multiple rounds. Active learning combines uncertainty sampling and diversity sampling strategies to give priority to log texts with low model prediction confidence and large semantic differences. Expert annotation data, reward model output, and active learning samples are jointly included in the training cycle to gradually improve model performance. Experiments show that the multi-round RLHF optimization framework significantly improves the precision and recall of hidden danger identification, can effectively deal with the scarcity of expert annotations, and shows a high coverage rate in long-tail hidden danger identification, demonstrating strong professional text semantic understanding capabilities and practical value.

Key words. Power dispatch, Duty log, Hidden danger recognition, Multiple rounds of human feedback, Reinforcement learning from human feedback

Table 1. Abbreviations.

Abbreviation	Full Name
RLHF	Reinforcement Learning from Human Feedback
BERT	Bidirectional Encoder Representations from Transformers
PPO	Proximal Policy Optimization
BiLSTM-CRF	Bidirectional Long Short-Term Memory - Conditional Random Field
CDF	Cumulative Distribution Function
KL	Kullback-Leibler Divergence
MSE	Mean Squared Error
SGD	Stochastic Gradient Descent

Table 1 summarizes the key technology abbreviations and full English names involved in this paper, covering core terms in natural language processing, optimization algorithm and statistical analysis.

1. Introduction

The power dispatch duty log is the core record of power grid operation monitoring, carrying key information such as equipment status, operating instructions, and abnormal events [1,2]. Efficient and accurate recognition of potential safety hazards from the log is of great significance for preventing power grid failures and ensuring power supply safety [3,4]. However, the current automated hidden danger recognition of power dispatch logs still faces significant challenges. Log texts are highly professional and contain a large number of industry terms, abbreviations, and unstructured

descriptions, making it difficult for general natural language processing models to accurately understand their semantics [5,6]. The safety standards of the power system are strict. Hidden danger recognition requires not only the discovery of explicit anomalies but also the inference of potential risks based on industry knowledge [7,8], which places higher demands on the domain adaptability of the model. More importantly, the expert annotation resources in the power field are extremely limited [9,10]. Since log analysis requires deep industry experience, the training cycle of qualified annotators is long, and the cost is high, making it difficult to obtain large-scale high-quality annotated data; Specifically, although expert annotation resources have professional capabilities, they still face practical challenges: log analysis needs to be combined with real-time operation scenarios and historical experience, and the annotation process needs to check the equipment status, operation associations and potential chain risks one by one, which is extremely time-consuming. For example, a complete evaluation of a single complex log may take 20-30 minutes, and the average daily log volume exceeds 10,000, making it difficult for experts to cover large-scale data needs even if they are full-time annotation. In addition, under high-load operation and maintenance tasks, the deployment of experts to participate in annotation will directly affect the efficiency of real-time monitoring of the power grid, further increasing labor costs [11,12]. Existing supervised learning methods rely heavily on the scale of labeled data and often perform poorly when samples are insufficient, especially for the recognition of "long-tail hidden dangers" that occur infrequently but are highly harmful [13,14]. Although such hidden dangers do not account for a large proportion of historical data, once missed, they may trigger a chain reaction and cause major safety accidents. Traditional methods usually use rule matching or static machine learning models, which lack a continuous optimization mechanism and cannot adapt to the dynamic changes of log data [15,16]. It is also difficult to make full use of limited expert knowledge for iterative improvement. Existing automated recognition systems often have a balance problem between false positives and false negatives [17,18]. Overly strict screening may cause a large number of normal logs to be misjudged as hidden dangers, increasing the burden of manual review; while overly loose filtering misses real risks and weakens the warning effect [19,20]. This contradiction is particularly prominent in power logs with strong professionalism and complex semantics. Therefore, how to build a hidden danger recognition framework that can continuously learn and adaptively optimize with limited expert participation has become a key issue in improving the level of intelligent power dispatching. This study is aimed at this demand, exploring how to gradually improve the model's recognition accuracy for complex hidden dangers through a human-machine collaborative reinforcement learning mechanism, guided by a small amount of expert feedback, and especially enhance the detection capability for long-tail and rare risk patterns.

The core goal of this paper is to build an intelligent hidden danger recognition framework for power dispatch logs. Through the deep collaboration of reinforcement learning and human expert feedback, the bottlenecks of natural language processing in professional fields are broken through: the professionalism of semantic understanding, the scarcity of labeled data, and the recognition effectiveness of long-tail distribution. Different from the static supervised learning paradigm, this study applies the multi-round RLHF mechanism into the field of power text analysis. Its innovation is reflected in multiple dimensions: at the modeling level, a composite reward function that integrates domain knowledge is designed to quantify the accuracy of hidden danger recognition as a weighted score of semantic matching and risk severity, so that the reinforcement learning process can simultaneously optimize text representation and risk rating. Compared with the traditional single-dimensional reward design, this scheme realizes the joint modeling of power professional terminology understanding and safety level evaluation. In terms of algorithm architecture, based on the dynamic fine-tuning strategy of PPO, an incremental optimization target is constructed through the real-time scoring of model output by experts. Although the existing BERT-based power text analysis method has made breakthroughs in semantic understanding, it is still limited by static annotated data and a single optimization goal. The framework in this paper transforms expert feedback into dynamic reward signals through the RLHF mechanism, achieving continuous alignment of domain knowledge and model capabilities. This method breaks through the limitations of static fine-tuning of the BERT model and enables the pre-trained language model to continuously adapt to the language evolution and new hidden danger patterns of power logs. In terms of system efficiency, a hybrid active learning strategy was developed to enable the model to actively mark emerging operational risks; actual accident reports were introduced as feedback signals during the training process, and a reward function was constructed in combination with expert scores to achieve accurate response to real accident scenarios. This solution focuses expert feedback on log segments where the model's cognition is ambiguous and the representation differences are significant, thereby improving the performance gain brought by unit annotation investment. This mechanism is particularly suitable for scenarios in the power sector where the annotation cost is high, and provides a feasible path for continuous optimization of models in small sample environments. The industrial value of the research lies in the establishment of a scalable hidden danger recognition enhancement system. Its core innovation is to transform the domain cognition of human experts into quantifiable reinforcement signals, and to achieve a gradual improvement in model capabilities through multiple rounds of interaction.

2. Related Work

For the recognition of hidden dangers in power logs, existing studies mainly use three types of methods: rule-based methods, traditional machine learning methods, and deep learning methods. The rule engine [21,22] relies on expert experience to construct regular expressions and logical rules. Although it is highly interpretable, it is difficult to adapt to the semantically variable log representation, resulting in a low recall rate. Shi X proposed a structured representation method for assembly process planning based on knowledge graphs, using a Bidirectional Long Short-Term Memory Conditional Random Field (BiLSTM-CRF) model for named entity recognition, which improved the accuracy of recognition and extraction in power safety and verified the effectiveness of the method [23]. Sequence annotation models such as BiLSTM-CRF [24,25] improve recognition effects by capturing contextual dependencies, but their generalization ability for professional terms is insufficient, and their performance drops sharply when there is insufficient annotated data. In recent years, the BERT pre-trained language model [26,27] has achieved certain breakthroughs through fine-tuning. Jiamiao Y proposed a method for automatic risk rating of power grid field operation based on the BERT model, combining text enhancement and error rating correction strategies to effectively improve the accuracy of risk rating. This method performed well in processing risk classification tables and actual operation texts, and had more semantic understanding advantages than traditional models [28]; however, its optimization relied on statically labeled data and could not be dynamically adjusted using real-time feedback from experts. More importantly, rule-based methods cannot proactively identify emerging operational risks and lack dynamic adaptability. Machine learning improves generalization capabilities through data-driven methods, but is limited by static labeled data and lacks the ability to proactively discover new risks. None of the above methods effectively solves the long-tail distribution problem. Existing studies have attempted to alleviate data imbalance through oversampling or cost-sensitive learning, but they cannot fundamentally improve the model's ability to recognize rare patterns.

To deal with the problems of scarce annotations and long-tail distribution, some studies have begun to explore interactive learning methods. Active learning [29,30] selects samples with large amounts of information through uncertainty sampling to reduce annotation costs, but only optimizes the data selection strategy and does not change the model training mechanism. Reinforcement learning [31,32] optimizes the model through reward signals in text analysis, but relies on preset rules to design reward functions, which makes it difficult to adapt to the complex semantics of power logs. RLHF [33,34] has been demonstrated in dialogue systems that human feedback can significantly improve model alignment capabilities. Shi H reviewed data-enabled smart grids and discovered and proposed a large-scale language model using RLHF to accelerate the large-scale application of smart grids [35]. However, when it is directly applied to professional fields, it faces problems such as single reward function design and low feedback efficiency. In contrast, the multi-round RLHF framework constructed in this paper innovatively combines three aspects: 1) a multi-dimensional reward function that integrates semantic similarity and hidden danger severity; 2) a multi-round reinforcement fine-tuning mechanism based on the BERT model; 3) an expert feedback optimization guided by active learning. This comprehensive solution can overcome the limitations of existing methods in professional terminology understanding, long-tail hidden danger recognition, and annotation efficiency.

3. RLHF-Based Hidden Danger Recognition Optimization Framework

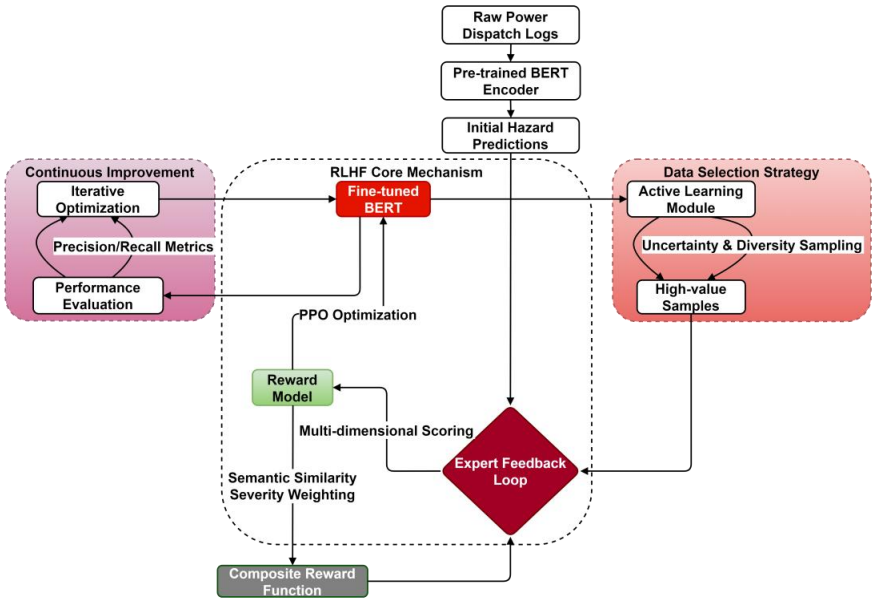

Figure 1. RLHF-based hidden danger recognition optimization framework.

Figure 1 shows the multi-round RLHF power dispatch log hidden danger recognition optimization framework, which contains three coupled subsystems at its core: the pre-trained BERT encoder processes the original log to generate initial predictions, forming the basis for semantic understanding; experts drive reward model training through multi-dimensional scoring and use the PPO algorithm to implement BERT reinforcement fine-tuning to form the RLHF core optimization loop; the active learning module selects high-value samples through uncertainty sampling and diversity screening strategies to maximize the utility of expert feedback. The framework triggers iterative optimization through the precision/recall rate indicators of the performance evaluation module to form a closed-loop learning mechanism. The entire framework realizes a continuous enhancement cycle from initial prediction → expert feedback → model optimization → sample selection, providing a scalable and intelligent solution for power dispatch log analysis.

A. Expert Feedback-driven Reward Model Construction

A multi-dimensional reward function based on semantic similarity and hidden danger severity is designed, and the reward model is trained through real-time scoring of the model output by experts to quantify the accuracy of hidden danger recognition.

1) Multi-dimensional Scoring Mechanism for Semantic Similarity and Severity Evaluation

The expert feedback output and the model prediction results are first encoded into vector representations. The vector cosine similarity is used to measure the closeness of the two in the semantic space:

$$r_{\text{sem}} = \frac{\boldsymbol{e}_{\text{exp}} \cdot \boldsymbol{e}_{\text{pred}}}{\left\|\boldsymbol{e}_{\text{exp}}\right\| \left\|\boldsymbol{e}_{\text{pred}}\right\|} \quad (1)$$

In the formula, $\boldsymbol{e}_{\text{exp}}$ represents the encoding output of the expert's manually annotated text, and $\boldsymbol{e}_{\text{pred}}$ represents the encoding output of the model for the same text; the numerator is the dot product of the two vectors, and the denominator is the product of their norms. This dimensional score reflects the degree of fit between the model judgment and the expert judgment at the contextual semantic level.

For different types of hidden dangers in the log, a severity weight S is attached. This value is determined by the industry manual and the accident level standard and mapped to the interval [0,1]. The final reward score is obtained by combining the following formula:

$$r = \lambda_1 r_{\text{sem}} + \lambda_2 S \quad (2)$$

In the formula, λ_1 and λ_2 are the semantic similarity and severity weight coefficients, respectively, satisfying $\lambda_1 + \lambda_2 = 1$, which are fixed after expert calibration to ensure that the composite score reflects both the model's language understanding effect and the importance of high-risk events.

2) Training Process of Reward Model Driven by Real-time Expert Scoring

Experts record the log hidden danger judgment given by the model in a discrete grade evaluation (0-5 points) on the interface, and these annotated data are entered into the training set together with the corresponding text features. The reward model adopts a multi-layer feedforward neural network architecture. The network input is the expanded and spliced text feature vector r_{sem} and the severity weight S, and the output is the normalized composite score. The training goal is to minimize the mean square error between the model output and the manual score. The update rule adopts stochastic gradient descent optimization; the batch size is set to 32; the learning rate is adjusted according to the exponential decay method.

After each round of feedback is completed, the system automatically selects several samples with the largest and smallest errors in this round, and summarizes them into the expert review queue to ensure that extreme cases are confirmed twice. The confirmed sample labels are used together with the original training set for the next round of model update, and the number of iterations is dynamically determined according to the convergence of the verification set indicators. In the online environment, the reward model weight is deployed to the fine-tuning process after each round of update, so that the subsequent hidden danger judgment scores are more in line with the expert experience, so as to continuously improve the overall recognition precision.

B. Multiple Rounds of Reinforcement Fine-tuning of the BERT Model

Taking the reward model as the optimization target, the PPO algorithm is used to perform multiple rounds of fine-tuning on the pre-trained BERT model to dynamically adjust the model's ability to capture professional terms and long-tail hidden dangers.

Figure 2 shows the process of reinforcement fine-tuning. Experts score the output hidden danger recognition results in real-time and quantify them around multiple dimensions such as term accuracy and risk judgment, and the scoring results are directly involved in the construction of the reward function after aggregation. The reward function generates a reward signal based on semantic similarity and hidden danger severity, driving the strategy optimization module to iteratively adjust the model parameters. The updated BERT re-enters the next round of recognition process to form a closed-loop

reinforcement mechanism. Expert scores are not only used for result evaluation but also embedded in the model update path as a key feedback signal to ensure that the fine-tuning process continues to optimize the ability to capture professional terms and perceive long-tail hidden dangers, thereby achieving a dual improvement in semantic precision and risk identification capabilities.

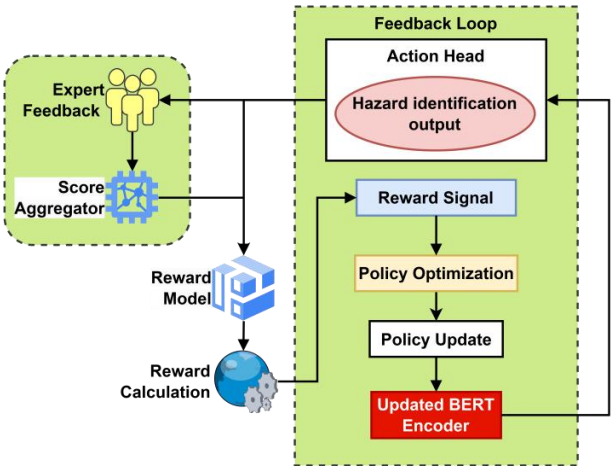

Figure 2. Reinforcement fine-tuning process.

1) Strategy Update Mechanism Guided by Reinforcement Learning Signals

The initial state of the model comes from the BERT encoder after the general language modeling task has been completed. The structure is not modified, and only the strategy probability distribution generator is applied in the output layer. After each result is generated, the aforementioned reward model score is used as environmental feedback to form a reinforcement signal. The policy optimization process adopts the clipping probability ratio policy adjustment method to constrain the change range between the current policy and the old policy to prevent the policy collapse problem caused by too large a step length. The core update target is defined as follows:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_t\left[\min\left(\rho_t A_t, \text{clip}\left(\rho_t, 1-\ , 1+\ \right)\cdot A_t\right)\right] \quad (3)$$

In the formula, $\rho_t = \dfrac{\pi_\theta\left(a_t \middle| s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \middle| s_t\right)}$ represents the probability ratio of the current policy and the old policy to select action a_t under state s_t; ϵ (Epsilon) is the clipping threshold; A_t is the advantage function, which is used to measure the superiority of the current action compared with the average policy, and is specifically composed of the reward value and the state value function estimate. The core intention of this formula is to balance the policy improvement range and behavior stability, and maintain the continuity and effectiveness of the policy behavior during fine-tuning.

The policy iteration adopts batch sampling to dynamically select a subset of samples containing professional terms and rare hidden danger descriptions in the training set to improve the responsiveness to the long-tail characteristics of the domain. After each sampling, the current policy is used to generate an output sequence and is linked with the expert scoring system to obtain an instant scoring result; then, the scoring feedback is mapped to a reward scalar to guide the policy gradient calculation. The optimization direction is completely determined by the reward function to ensure that the update path is close to the domain goal. The strategy distribution, reward score, and historical strategy of all sampled samples are recorded synchronously for the next iteration to prevent training from falling into local optimality.

2) Parameter Fine-tuning Path Control under Multi-round Interactive Feedback

To avoid bias on high-frequency words or template structures during training, a normalization control module is applied after each round of parameter update to recalibrate the word vector distribution. This module uses the drift degree of the semantic distribution center in each round as the adjustment factor to reverse the direction of over-convergence and ensure that the model still has sufficient responsiveness to rare terms in the input text. In terms of specific implementation, a multi-head attention adjustment channel is embedded between the word-level embedding layer and the output layer, so that the attention weight obtained by low-frequency words in high-reward samples can be explicitly retained, thereby increasing their participation intensity in gradient updates.

At the end of each round of training iteration, the parameter state is uniformly sent to the sliding window evaluation area, and combined with the fluctuation trend of the historical five-round fine-tuning indicators, it is automatically determined whether to enter the strategy freezing stage. When the cumulative change rate is lower than the set threshold, the model stops actively updating and only retains the passive response weight to samples with significant rating deviations, forming a local self-stabilization mechanism to slow down the overfitting trend. The loss function form during training is adjusted as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO}} + \beta_{\text{x}}\,\text{KL}\left[\pi_{\theta_{\text{old}}} \middle\| \pi_\theta\right] \quad (4)$$

Among them, the second term is the KL (Kullback-Leibler) divergence between the old strategy and the current strategy, and the coefficient β_{x} controls the strategy convergence speed to avoid deviating too far from the original behavior strategy range and causing a decrease in generalization ability.

During the training process, all professional terminology samples, abnormal structure sentence samples, and historical feedback misjudgment samples are set as priority weight items, and their gradients participate in parameter updates at twice the rate during back propagation, strengthening the model's ability to recognize a small number of high-risk categories. The final output strategy significantly improves the recognition accuracy of long-tail categories, and the semantic generation is more in line with the expert language habits and hidden danger expression paradigm. By continuously strengthening the training path to converge to expert logic, BERT is guided to build a more discriminating language recognition model in the field of safety hidden dangers.

C. High-value Sample Screening under Active Learning

Combining uncertainty sampling and diversity sampling strategies, log texts with low model prediction confidence and large semantic differences are prioritized to maximize the information gain of expert feedback.

1) Uncertainty-driven Candidate Text Screening

Using the output probability distribution of the model in the log text classification task, the distribution entropy index is calculated for each log to measure the model's cognitive ambiguity of the sample. Assuming that for the j -th log, the model outputs the category probability vector $\boldsymbol{p}^{(j)} = \left[p_1^{(j)}, p_2^{(j)}, \cdots, p_C^{(j)} \right]$, where C represents the total number of hidden danger categories. The entropy value is defined as follows:

$$H^{(j)} = -\sum_{i=1}^{C} p_i^{(j)} \ln p_i^{(j)} \quad (5)$$

In the formula, $p_i^{(j)}$ represents the probability that the model judges the j -th log as category i ; the natural logarithm function ln calculates the amount of information; the higher the entropy $H^{(j)}$ value, the stronger the uncertainty of the log inside the model. For the entire unlabeled log pool, each iteration selects the log with the highest entropy value to form an uncertainty candidate set. This step focuses on a few samples that are easily confused by the model, ensuring that each feedback from the expert can focus on the most valuable fuzzy area, thereby accelerating the location and correction of the long-tail hidden danger pattern.

After the candidate set is generated, the system automatically records the entropy value and predicted label confidence interval corresponding to each log, and pushes it to the expert together with the original log text and the intercepted key fields. Experts score and correct these high entropy samples in the manual interface, and the feedback obtained is marked as high-information samples for subsequent training. This screening strategy effectively avoids repeated labeling of low-value, high-confidence samples, saves expert resources, and enables the model to obtain more targeted training examples when facing rare or complex expressions.

2) Sample Selection with Enhanced Semantic Diversity

Based on the uncertainty candidate set, the text embedding space is further used to measure the diversity distance to ensure that the selected samples are evenly distributed at the semantic level. For each selected log, a vector representation $\boldsymbol{v}^{(j)}$ is calculated, which comes from the output of the model's intermediate layer or a specially trained domain embedder. The semantic distance is defined as the complement of the cosine similarity:

$$d_{jk} = 1 - \frac{\boldsymbol{v}^{(j)} \cdot \boldsymbol{v}^{(k)}}{\left\| \boldsymbol{v}^{(j)} \right\| \left\| \boldsymbol{v}^{(k)} \right\|} \quad (6)$$

In the formula, $\boldsymbol{v}^{(j)} \cdot \boldsymbol{v}^{(k)}$ is the inner product operation; $\| \ \|$ represents the vector norm; the larger d_{jk} is, the more significant the difference between the j -th and k -th logs in the semantic space. The initial diversity seed set selects several logs with the highest dispersion as the first batch of annotation objects. Subsequently, a greedy algorithm is used to iterate: at each step, the sample with the largest minimum distance value among the logs that have not yet been selected is included in the set until the preset number is reached.

Finally, the uncertainty ranking and diversity score are linearly fused to construct a composite priority index:

$$\text{Score}^{(j)} = \alpha \frac{H^{(j)}}{\max\limits_{l} H^{(l)}} + (1-\alpha) \frac{\min\limits_{s \in S} d_{ls}}{\max\limits_{l} \min\limits_{s \in S} d_{ls}} \quad (7)$$

Among them, α controls the uncertainty and diversity weights; the denominators $\max\limits_{l} H^{(l)}$ and $\max\limits_{l} \min\limits_{s \in S} d_{ls}$ are used for normalization. The system finally selects samples in descending order of $\text{Score}^{(j)}$ to ensure that the logs submitted to the experts in each round reflect both the model's doubt areas and different clusters in the semantic space, maximizing the feedback value. This strategy not only ensures the labeling efficiency but also enhances the representativeness of the training set in the expression of professional terms and long-tail hidden dangers, and promotes the model to continuously optimize the recognition performance in a scarce data environment.

In order to improve the real-time performance and situational awareness of log analysis, this paper explores the integration mechanism of dispatch logs and SCADA (Supervisory Control and Data Acquisition) systems. By constructing a cross-modal association framework, the log text description (such as equipment status changes, abnormal operation records) and the power grid operation indicators collected by SCADA (such as load imbalance, voltage sag and other numerical data) are timestamp aligned and feature fused to achieve dynamic mapping of text insights and power grid parameters. For example, when the "main transformer overload" is mentioned in the recognition log, the system will synchronously retrieve the load data of the corresponding period to verify the accuracy of the hidden danger description and supplement the quantitative evaluation. In addition, the study also designed a multi-source information fusion module based on the attention mechanism, which enables the model to automatically capture the potential correlation between text clues and power grid indicators, thereby improving the interpretability and response speed of risk judgment.

D. Iterative Optimization Mechanism of Feedback Loop

The high-value samples corrected by experts and the original unlabeled logs form the initial training pool. The reward model gives a composite score r_i to each sample as the basis for calculating the sample weight. Assuming that the expert annotation indicator is e_i, when the i-th sample carries a manual score, $e_i = 1$, otherwise $e_i = 0$; the annotation quality function q_i represents the consistency of the expert score, which is calculated by the difference in scores of the same log in adjacent rounds. The final weight of the sample is defined as:

$$w_i = \alpha e_i + \beta r_i - \gamma q_i \quad (8)$$

Among them, α and β are the weight coefficients of manual annotation and reward score, respectively, satisfying $\alpha + \beta = 1$; γ is the consistency penalty coefficient, which is used to reduce the impact of repeated information on model updates. The weight w_i acts on the loss function, allowing the model to give priority to the gradient contribution of high-value samples during back propagation, and organically couple expert knowledge with reward signals.

After each round of fine-tuning, the system automatically merges the latest expert annotations and the active learning screening samples of this round into the training set, and updates the weight configuration of all samples at the same time. The merging strategy follows the principle of “small batch-high frequency”: new samples are injected in small batches, not exceeding 5% of the total number of samples each time, to ensure that the model can perceive new information each time it is updated, rather than swallowing a large amount of untested data in one go. In addition, the weight of historical samples is adjusted in an exponential decay manner as the round increases, and the weight decay formula is as follows:

$$w_i^{(t+1)} = w_i^{(t)} \times \delta^{\Delta t} \quad (9)$$

In the formula, $w_i^{(t)}$ represents the weight of the i-th sample in the t-th round; Δt represents the interval from the last time the sample is reviewed by experts to the current round; Δ is the weight decay factor, which ranges from 0 to 1. This mechanism allows newly labeled and high-reward samples to be learned first, while avoiding expired samples from occupying an excessive proportion in long-term training.

The training process incorporates the reward model output, active learning sample feedback, and validation set indicators into the monitoring system. Assuming that the validation set accuracy is P_t, and the missed detection rate is F_t, the comprehensive performance index is calculated for each round update:

$$M_t = \omega P_t - (1-\omega) F_t \quad (10)$$

Among them, ω is the performance trade-off coefficient, ranging from [0,1]; an increase in M_t value represents an increase in comprehensive ability. If the M_t increases by more than the preset threshold compared with the previous round, the cycle continues; if the improvement threshold is not reached, the retrospective sample re-evaluation phase is triggered, and several samples with the worst performance in the threshold round are fed back to the expert for review.

After the retrospective sample review is completed, its weight is recalculated and added to the next round of training set, and the training process is restarted. If the M_t increase for three consecutive rounds is lower than the preset threshold, the model enters the “strategy freeze” state, and the update sub-process is activated only when a new abnormal log is detected. This sub-process uses the reward model accumulated by the accumulated expert feedback to prioritize the screening of new samples to avoid the model being trapped in the “micro-oscillation” range for a long time.

This closed-loop architecture adaptively adjusts the sample weight and injection quantity in each training iteration. The expert feedback, reward signal, and active sampling work together to continuously improve the performance of the model in the task of recognizing hidden dangers in power dispatch logs, effectively improving the ability to capture professional terms and long-tail hidden dangers.

4. Method Effect Evaluation

In view of the unstructured characteristics of the dispatch log, the domain dictionary-enhanced word segmentation technology (such as Jieba word segmentation combined with the power terminology library) is used, and irrelevant noise is filtered through regular expressions. Subsequently, the domain adaptive pre-training based on BERT (continued to be fine-tuned on the power corpus) strengthens the semantic representation of professional terms. Key anomaly detection is achieved in two steps: 1) the rule template preliminarily screens high-risk keywords; 2) the semantic similarity model compares the log content with the predefined hidden danger description library to distinguish between routine operations and safety incidents. Table 2 lists the key experimental parameters used in the multi-round RLHF optimization framework of this paper and their value ranges and applicable scopes.

Table 2. Enhanced training strategy parameter configuration table.

Parameter Name	Value/Range	Application Scope
Learning Rate	Initial: 5e-5	BERT fine-tuning
	Exponential	-
Batch Size	32	Reward model training, BERT fine-tuning
Policy Clipping Threshold (ε)	[0.1, 0.2]	Reinforcement learning optimization
Sliding Window Length	2 rounds	Training stability assessment
KL Divergence Coefficient	[0.1, 0.3]	Preventing policy deviation
Advantage Estimation Window Width	32	Policy gradient calculation
Gradient Amplification Factor	2	Increasing update weight for key samples
Weight Decay Factor	0.95	Dynamic sample weight adjustment
Performance Trade-off Coefficient	0.6	Model iteration decision-making
Consistency Penalty Coefficient	0.8	Sample weight calculation
Uncertainty Entropy Threshold	≥ 1.5	Active learning candidate selection
Diversity Distance Threshold	≥ 0.7	Diversity sampling

A. Impact of Semantic Similarity and Hidden Danger Severity Weight on Scoring and Rewards

150 hidden danger inspection records from a large energy company from 2020 to 2023 are collected, covering three types of typical accident hazards: electrical, mechanical, and environmental. Each hidden danger is submitted by front-line operators and then reviewed and annotated with severity levels by safety supervision experts to ensure the accuracy and actual representativeness of data annotation. The output results of the hidden danger recognition model on multiple real cases are combined with the real-time scoring records of each hidden danger description by the expert group. The experimental simulation platform uses Python 3.8 (PyTorch framework) to implement BERT model training and RLHF optimization, and the reward model training and data analysis are assisted by Matlab R2022a. Each data includes the semantic similarity value calculated by the model and the subjective evaluation score of the expert for the severity of the hidden danger. The semantic similarity is obtained by calculating the cosine similarity after vectorizing the hidden danger text, reflecting the accuracy of the model's semantic matching of the hidden danger. The expert score is assigned according to the preset severity grading standard. Different severity weight settings are achieved by adjusting the weighting coefficient in the reward function to simulate the model's response to hidden danger recognition under different focus points. The semantic similarity output by the model is combined with the expert score to generate a multi-dimensional reward score for statistical analysis and visualization.

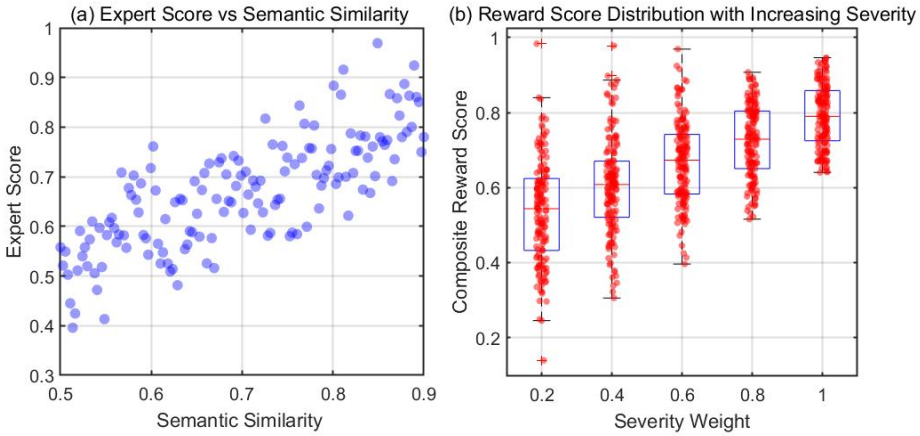

Figure 3. Impact of semantic similarity and hidden danger severity weight on scoring and rewards.

Figure 3 shows the impact of semantic similarity and hidden danger severity weight on scoring and rewards:

Figure 3(a) shows the relationship between semantic similarity and expert scores. The overall distribution shows a positive correlation trend, that is, the higher the semantic similarity, the larger the corresponding score, indicating that the model is easier to be recognized by experts on the basis of reasonable similarity judgment. However, there is still a certain degree of discreteness in the scatter points. This score fluctuation may be due to the fact that although individual samples are semantically close, there are slight differences in the context, which leads to subjective deviations in the perception of risk representation by experts during evaluation. This phenomenon suggests that although semantic similarity is an important influencing factor in scoring, it is difficult to fully represent the actual severity of the hidden danger description, so it is necessary to integrate more dimensions to regulate the reward function.

Figure 3(b) presents the reward distribution characteristics under different severity weights from a global perspective. It can be observed that with the increase of severity weight, the concentration of reward scores is significantly enhanced; the box height gradually shrinks; the number of extreme values is relatively reduced. This change shows that after giving a higher weight to severity in model judgment, expert scores become more consistent, and the score distribution becomes more stable. This centralization trend stems from the fact that high-severity samples usually have clearer feature representations, and experts are more consistent in their recognition judgments, which effectively reduces the subjective fluctuations in the scoring system. This result verifies that the reasonable application of severity factors in the reward function of the model can enhance the controllability and discrimination efficiency of the score, and has practical significance for improving the stability of model training.

B. Uncertain Entropy Value Distribution

Based on the initial original log text collected from large energy companies, after text cleaning, word segmentation, and vectorization processing, it is input into the current fine-tuned BERT model. The model outputs a multi-category probability distribution for each log and then calculates its predicted entropy value. The higher the entropy value, the more uncertain the model is. 150 samples with low prediction confidence are selected and sorted, and finally, the uncertain entropy value distribution is generated, reflecting the uncertainty area of the model in the current recognition of heterogeneous hidden dangers.

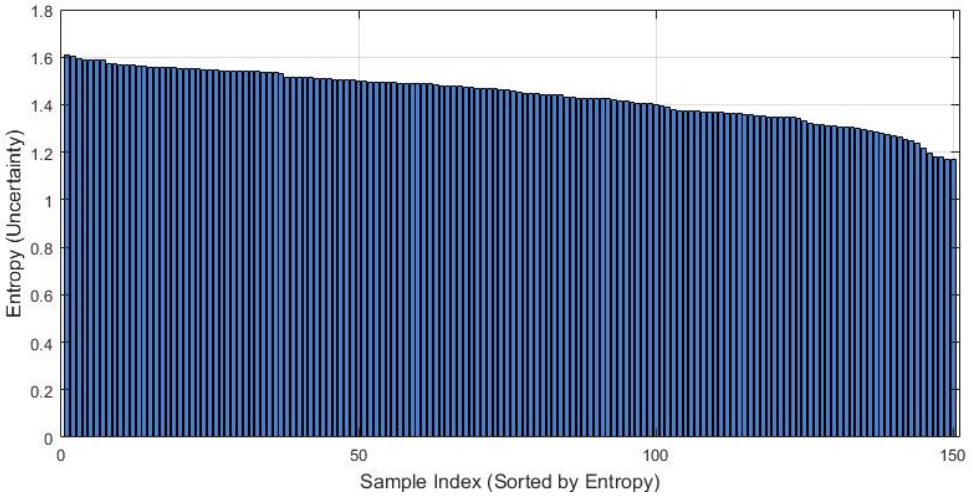

Figure 4. Uncertainty entropy distribution of log samples.

Figure 4 shows the distribution of uncertainty entropy values of 150 log samples under multi-category prediction, reflecting the difference in the confidence of the model in judging each sample. The horizontal axis is the sample number sorted from high to low according to the entropy value, and the vertical axis is the corresponding entropy value, which is used to quantify the information uncertainty of each log under the current classification system. From the overall distribution trend, it can be seen that the entropy value of the first section samples is higher, indicating that the model has a greater judgment ambiguity on these samples; the output category probabilities are closer; a clear tendency prediction cannot be formed. Such high entropy samples often contain long-tail events, complex semantic structures, or implicit professional terms, and are the key objects that should be screened first in active learning. As the entropy value gradually decreases, the model shows a clearer judgment tendency, indicating that it has good generalization ability on most regular logs or samples with stable structures. The difference in entropy values comes from the diversity of the semantics of the samples themselves and the learning bias of the model for atypical patterns, highlighting the key role of uncertainty evaluation in screening representative samples. The overall results provide a quantitative basis for triggering the expert feedback mechanism, allowing high-value samples to be quickly located, effectively

improving the information density of the supervision signal, and accelerating the model's adaptation process to rare hidden danger types.

C. Cumulative Probability Distribution and Prediction Residuals

Based on the collected hidden danger inspection records, the training sample set is constructed in combination with the annotation results of industry experts. Each sample data includes a text description of the hidden danger, an expert's score (1–5 discrete values), and a corresponding severity level label (low, high). After repeated training through the reinforcement learning framework, the model outputs the reward result, compares it with the expert score to calculate the residual, and forms a residual distribution; the reward values of different severity levels are calculated using the cumulative distribution function (CDF) to calculate their cumulative probability distribution. All processing processes are completed in matlab and visualized by drawing.

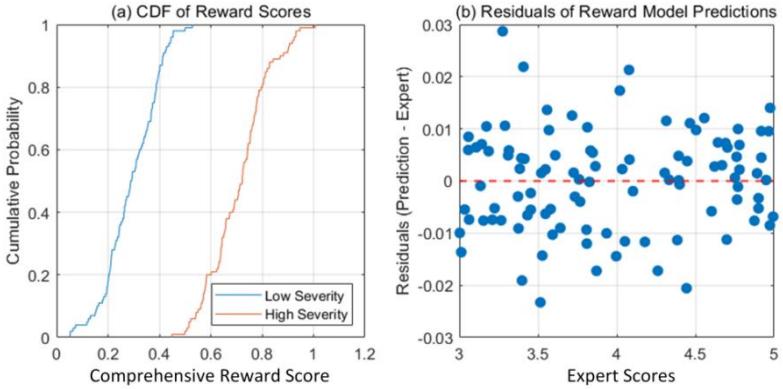

Figure 5. CDF comparison and prediction residual analysis.

Figure 5 shows the cumulative probability distribution and prediction residual:

Figure 5(a) is a comparison of the cumulative probability distribution of the reward score. The horizontal axis is the comprehensive reward score, and the vertical axis is the cumulative probability, which reveals the response trend of the model in distinguishing different hidden danger severity levels. The distribution of reward scores corresponding to low severity is overall biased to the left, while the distribution of reward scores for high severity is shifted to the right, indicating that the model can output higher incentive values when facing more critical hidden dangers. This distribution structure reflects the synergy between the semantic similarity term and the severity weighted term in the reward function, effectively widening the score range of low- and high-risk scenarios, and helping the model to achieve risk prioritization in subsequent learning.

Figure 5(b) shows the residual distribution between expert scores and model prediction rewards. The horizontal axis is the hidden danger score directly evaluated by experts, and the vertical axis is the difference between model prediction and expert judgment. The residuals are concentrated around zero, between -0.03 and 0.03, indicating that the reward model can stably approach manual evaluation in most cases. The scatter points do not show a trend of drastic fluctuations with the increase in scores, which indirectly verifies that the multi-dimensional reward function also maintains strong prediction consistency in high-scoring areas. The small deviation of the scatter points comes from the misjudgment of fuzzy expressions by semantic matching items, suggesting that the context processing ability of semantic embedding needs to be further optimized. Overall, the cumulative probability distribution and prediction residuals jointly verify the design rationality of the reward model structure in terms of robustness and discriminability, and provide theoretical and empirical support for subsequent strategy optimization.

D. Strategy Stability and Term Recognition Accuracy Evolution in Multiple Rounds of Fine-tuning

Table 3. Training sample priority and category distribution design.

Sample Type	Weight Adjustment	Training Frequency Ratio	Typical Feature Description	Reinforcement Focus
Domain-Specific Terms	2.0×	15%	Industry-specific terms and abbreviations	Precision in recognition and semantic understanding
Long-Tail Hidden Danger Descriptions	2.0×	10%	Low-frequency but high-risk hidden danger expressions	Anomaly detection and risk coverage
Common Samples	1.0×	75%	Standard hidden danger reports and conventional patterns	Overall model stability and generalization

Table 3 presents the refined design of weight adjustment, frequency configuration, and reinforcement direction of different types of samples during the training process. To improve the model's ability to recognize domain-specific terms and low-frequency high-risk hidden dangers, Domain-Specific Terms and Long-Tail Hidden Danger Descriptions samples are given a weight of 2.0 times respectively to enhance their participation in the same training cycle. Among them, Domain-Specific Terms samples account for 15% of the total training data, mainly covering industry abbreviations and technical expressions, emphasizing precise semantic capture; Long-Tail Hidden Danger Descriptions samples account for 10%, focusing on rare but potentially high-threatening sentences, aiming to improve the model's perception of boundary risks. The remaining regular samples maintain the default weight, accounting for 75% as the training subject, to ensure that the model maintains stable generalization ability when dealing with common hidden dangers. This sample allocation mechanism forms a high-value sample guidance strategy in enhanced training, laying the foundation for model semantic transfer and structural robustness.

Based on the constructed reward model and the BERT output log after multiple rounds of fine-tuning, after each round of fine-tuning, the values of the three stability indicators, KL divergence, policy entropy, and ε clipping threshold, during the strategy update process are recorded to generate a strategy stability trend. The term recognition accuracy is obtained by statistically analyzing the accuracy changes of five types of high-frequency professional terms in the test set. All indicators take the average of multiple verifications to ensure that the trends reflected by the data are representative and consistent.

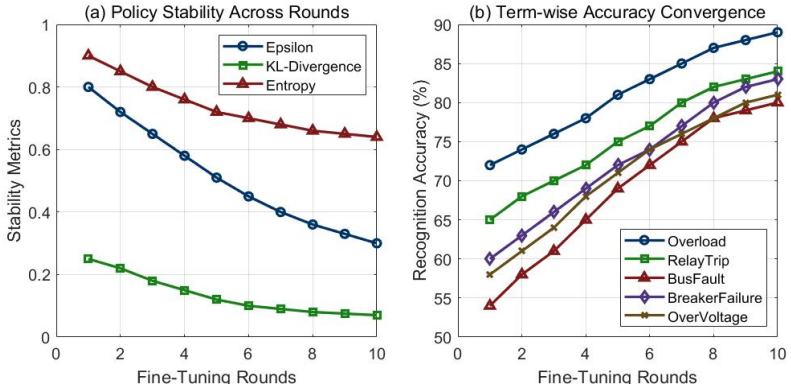

Figure 6. Evolution of policy stability and term recognition accuracy in multiple rounds of fine-tuning.

Figure 6 shows the evolution of policy stability and term recognition accuracy in multiple rounds of fine-tuning:

Figure 6(a) shows the trend of policy stability changing from round to round during the fine-tuning process, including three indicators, ε clipping threshold, KL divergence, and policy entropy. As the optimization rounds progress, the three indicators all show a consistent convergence trend, indicating that the model strategy gradually stabilizes. Among them, the gradual contraction of the ε value reflects the gradual reduction of exploration behavior, that is, the model's dependence on high-confidence strategies continues to increase; the decrease in KL divergence indicates that the distribution difference between the new and old strategies continues to shrink, which means that the policy update is more robust; the decrease in policy entropy indicates that the uncertainty of the policy output is reduced, and the model gradually establishes a preference for high-value behaviors in multiple rounds of optimization. The above phenomenon comprehensively reflects that under the reward-driven mechanism, the strategy tends to be deterministic and convergent, reflecting the effectiveness of PPO optimization in guiding the improvement of model stability.

Figure 6(b) reflects the changes in the recognition accuracy of five key professional terms in multiple rounds of model fine-tuning. The overall trend is a steady improvement and tends to be stable in the later period. In the 10th round of fine-tuning, the recognition accuracy of Overload, RelayTrip, BusFault, BreakerFailure, and OverVoltage is 89%, 84%, 80%, 83%, and 81%, respectively, showing that the model has gradually mastered the semantic distribution characteristics and contextual dependencies. The terms are improved rapidly in the early stage, reflecting that their contextual semantic features are more significant and easy for the model to capture; the terms with low initial recognition are mainly limited by semantic ambiguity or sample sparseness, which can be effectively compensated by strategy tuning after multiple rounds of fine-tuning. The data trend reveals the significant role of the dynamic adjustment mechanism in improving the terminology learning ability, indicating that the multi-round optimization strategy guided by rewards can effectively enhance the model's semantic sensitivity and generalization ability in professional contexts.

E. Comparison of Precision and Recall

In multiple rounds of strategy iteration, the BERT model continuously adjusts the strategy parameters through the PPO mechanism to make its output closer to the expert evaluation criteria. In the process of active sample selection, the system prioritizes the selection of log texts

with low confidence and significant semantic differences to increase the proportion of high-value samples in the training set. The labels and reward feedback corrected by experts are integrated into the training set to build a closed-loop update mechanism to achieve progressive optimization of model accuracy and recall. In the evaluation stage, the accuracy and recall of each hidden danger category are calculated by comparing the unified predicted output with the true label, and the standardized bar chart method is used for visualization analysis to ensure that the results are interpretable and reproducible. At the same time, the differences between this paper's multi-round RLHF optimization framework and BiLSTM-CRF, traditional BERT fine-tuning, and rule engine are compared.

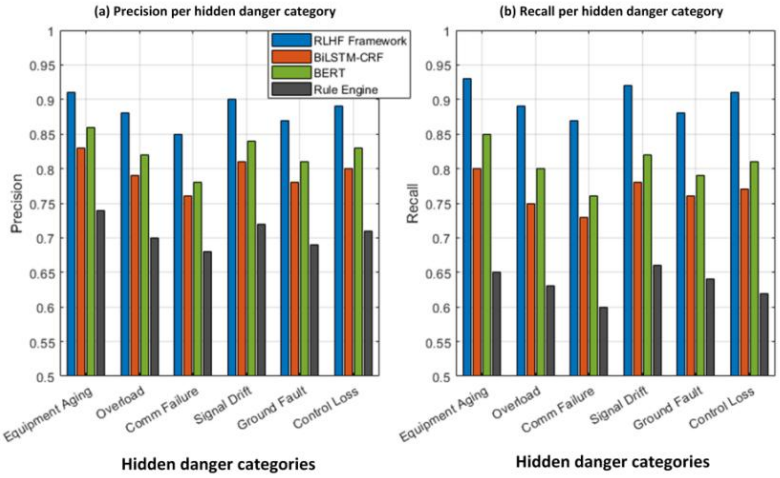

Figure 7. Comparison of precision and recall on different hidden danger categories.

Figure 7 shows the precision and recall performance of various models on different hidden danger categories. Overall, the constructed multi-round RLHF optimization framework achieves high indicators in most types of hidden dangers, with an average precision of about 88.3% and an average recall of 90.0%, showing strong hidden danger recognition capabilities. This advantage mainly comes from the use of expert feedback to guide the model to continuously adjust during the training process, which improves the capture effect of complex professional terms and hidden dangers. In contrast, the traditional BiLSTM-CRF and baseline BERT fine-tuning models show certain performance bottlenecks. Especially in the hidden danger categories with high semantic complexity, both the precision and recall rates have decreased, reflecting the model's lack of sensitivity to fine-grained information. In addition, the performance of the rule engine is significantly behind in all categories because the rule method has limited adaptability to abnormal changes and is difficult to cover diverse and dynamic hidden danger manifestations.

Further analysis shows that the balance between precision and recall reflects the model's ability to control false positives. The high recall rate of the multi-round RLHF framework shows its comprehensive capture ability of hidden dangers, while the improvement in precision reflects the effective suppression of misjudgments. This phenomenon shows that the model successfully uses the reward mechanism to adjust the strategy during the training process and strengthens the ability to distinguish hidden danger characteristics. The performance differences of different hidden danger categories also reveal the limitations of the model's adaptability. Some hidden dangers still require more detailed feature mining due to their rarity and complex expressions. The overall trend shows that the training process that integrates expert feedback and dynamic optimization is of great significance to improving the practical value of the hidden danger recognition system.

F. Quantitative Indicators of Expert Feedback Efficiency

Multiple rounds of iterative training are performed under the same expert feedback frequency and feedback sample size. After each round of training, the model calculates the accuracy on the validation set and records the difference between the accuracy of the current round and the previous round as the accuracy increment of the round. This process continues until the 60th round to capture the dynamic trend of model performance improvement. A fixed time window is used to ensure that the accuracy increment of feedback per unit time is comparable. The multi-round RLHF optimization framework, BiLSTM-CRF, traditional BERT fine-tuning, and rule engine are executed separately under the same data and training strategy to eliminate external interference. The results reflect the response speed and effect of each model to expert feedback information, especially the significant improvement of the multi-round RLHF optimization framework in the early rounds and the stable convergence state in the later stage, verifying its adaptability and iterative efficiency in an environment with scarce annotations.

Table 4. Comparison of accuracy increment driven by expert feedback.

Model	10th Round	20th Round	30th Round	40th Round	50th Round	60th Round
Multi-round RLHF Framework	0.041	0.038	0.035	0.03	0.018	0.009
BiLSTM-CRF	0.021	0.018	0.015	0.012	0.008	0.004
Traditional BERT Fine-tuning	0.027	0.024	0.02	0.016	0.01	0.005
Rule Engine	0.009	0.007	0.005	0.004	0.002	0.001

Table 4 reflects the change trend of the accuracy increment of different models in multiple iteration stages after receiving expert feedback. The multi-round RLHF optimization framework shows significant performance gains in the first 30 rounds, and the accuracy increment remains at 0.035 and above, indicating that it is more responsive to expert information in the early stages. In contrast, the traditional BERT fine-tuning and BiLSTM-CRF models show a smaller increase under the same feedback conditions, and the increment slows down significantly after 40 rounds. Since the rule engine does not have the ability to learn, its increment always remains at a low level, and it is close to the plateau in the early stage. The accuracy increments of all models gradually converge in the 50th and 60th rounds, indicating that under the continuous effect of expert feedback, the model enters a learning saturation state. Overall, this comparison process effectively reveals the potential of the multi-round reinforcement feedback mechanism in improving data utilization efficiency and adapting to environments with limited annotation resources.

G. Recognition Coverage of Long-tail Hidden Dangers

To evaluate the coverage of each model in the recognition of long-tail hidden dangers, a low-frequency category set is constructed, and the hidden danger types that appear less than 5 times in the log are selected as evaluation objects. Relying on the expert-annotated dataset, the results generated by each model are compared with the real labels one by one. The number of samples correctly recognized in each type of low-frequency hidden danger is counted, and the ratio is calculated with the total number of samples that actually exist to obtain the coverage index. The above process is performed on a unified sample to ensure input consistency and reduce interference factors beyond the model's capabilities. The coverage data is all derived from the model's inference results on a unified low-frequency sample set, which reflects its ability to capture long-tail patterns and the difference in generalization level.

Figure 8 shows the comparison of the recognition coverage of the four models on the low-frequency hidden danger category, covering the recognition coverage of a variety of long-tail hidden dangers including rare leakage, minor corrosion, signal noise, intermittent faults, and voltage fluctuations. These hidden dangers have extremely low occurrence frequencies, which poses a great challenge to the model's generalization ability and recognition stability. The multi-round RLHF optimization framework shows a relatively superior coverage effect in all categories, demonstrating its strong ability to capture the characteristics of rare hidden dangers. The recognition coverage of rare leakage, minor corrosion, signal noise, intermittent faults, and voltage fluctuation long-tail hidden dangers reaches 0.95, 0.94, 0.93, 0.96, and 0.95, respectively. The model uses continuous reward signals to drive the fine-tuning process, which effectively enhances the sensitivity to professional terms and complex hidden danger patterns, thereby improving the recognition accuracy of low-frequency samples. In contrast, BiLSTM-CRF and traditional BERT fine-tuning perform slightly worse in some hidden danger categories, mainly because they learn the frequently occurring hidden danger samples more fully during training, but the features of low-frequency samples are insufficient, resulting in a decrease in recognition coverage. The rule engine has the most limited performance because it relies on manually set rules and lacks the ability to adapt to complex hidden danger changes. This also reflects the limitations of pure rule methods when facing long-tail risks.

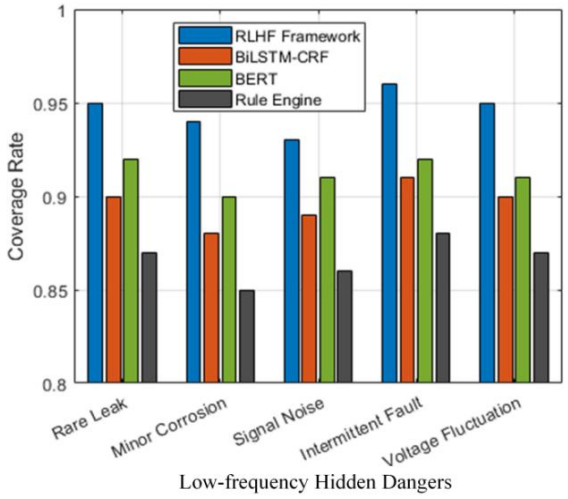

Figure 8. Coverage of low-frequency hidden dangers.

The performance fluctuations of the model in different hidden danger categories are partly due to the scarcity of data in the low-frequency categories themselves and the significant differences in features, which increases the difficulty of learning. Overall, the multi-round RLHF optimization framework significantly improves the model's adaptability and coverage of long-tail hidden dangers through the combination of multi-stage reinforcement learning and expert feedback, providing

more robust technical support for hidden danger management in complex industrial environments.

H. Comparison of False Positive Rates and Consistency of Scores by Different Experts

Each model is inferred separately, and the number of samples in its output that are inconsistent with the actual labels but are predicted to be high-severity hidden dangers is recorded, and the false positive rate is calculated accordingly. All models are trained and evaluated under the same distribution. The RLHF model applies expert feedback for semantic adjustment after each round of reinforcement fine-tuning, and finally, a unified test comparison is performed after the 60th round of model stabilization. False positive samples are reviewed by multiple experts to eliminate human label errors and ensure the consistency and objectivity of the evaluation process.

Table 5. Comparison of false positive rates of different models in high-severity hidden danger recognition.

Model Type	High-Severity Samples	False Positives	False Positive Rate (%)	Common Causes of False Positives
Rule Engine	120	9	7.5	Inflexible rule scope; struggles with linguistic variation
Static BERT Fine-tuning	120	7	5.83	Misclassification under vague wording
BiLSTM-CRF	120	8	6.67	Limited context extraction; poor handling of inter-sentence cues
Multi-round RLHF Framework	120	2	1.67	Domain-specific phrasing underrepresented

Table 5 shows the false positive rate performance of different models in the high-severity hidden danger recognition task, reflecting the reliability differences of the models in handling key risk events. The false positive rate of the rule engine is 7.50%, which is mainly limited by the rigidity of the rule template. When faced with semantic changes or undefined expressions, it cannot be effectively adapted, resulting in false triggering of risk labels. The false positive rates of the BiLSTM-CRF and static BERT models are 6.67% and 5.83%, respectively. Although they are better than the rule method in semantic modeling, they still have a tendency to misjudge when dealing with samples with drastic context changes or fuzzy semantic boundaries, especially the lack of deep understanding of term ambiguity and event context. In contrast, the false positive rate of the multi-round RLHF optimization model is only 1.67%. Its significant advantage comes from the continuous correction and strengthening of the model's semantic representation ability under multiple rounds of expert feedback, and it has a higher sensitivity to professional terms, fuzzy expressions, and contextual signals. Even in the face of rare expressions, the model can combine the context to achieve more precise judgment and avoid mislabeling, but there is still room for improvement in the recognition of domain-specific wording. Most false positives are concentrated on new terms or unstructured expressions that are not covered by the training samples, which further proves the core role of active learning and feedback mechanisms in controlling false positives in serious scenarios. Building an expert feedback closed-loop can significantly improve the accuracy and credibility of high-risk text recognition.

Table 6. Consistency analysis of different experts' scores on log samples.

Sample Type	Number of Experts	Mean Rating Variance	Krippendorff's α	Disagreement Level	Model Handling Strategy
Highly Consistent	15	0.02	0.89	Very Low	Direct inclusion in training
Generally Consistent	15	0.07	0.78	Acceptable	Weighted training instance
Mild Disagreement	15	0.13	0.61	Moderate	Label determined by majority vote
Clear Disagreement	15	0.25	0.42	High	Expert consensus re-evaluation
Hard-to-correct Sample	15	0.31	0.33	Severe	Excluded from training set

Table 6 systematically quantifies the consistency of expert ratings in different types of log samples, using Krippendorff's α coefficient as the main indicator, combined with the score variance to reflect the degree of disagreement in the subjective judgment of the sample. Under the premise that the number of experts is fixed at 15, the typical consistent sample reaches 0.89, indicating that the label reliability is extremely high and can be directly used as high-confidence training data; although there are slight differences in general consistent samples, they are still in an acceptable range, so they are weighted when included to control the impact of noise. For samples with mild and obvious disagreements, the model uses majority voting and consensus review mechanisms to ensure label accuracy. For samples with extremely low consistency and difficult to reach consensus, the model actively removes them to avoid training deviation. The overall strategy reflects the robustness and fine-grained

response ability of the model design to label quality control under the heterogeneity of expert opinions.

5. Conclusions

The hidden danger recognition system for power dispatch logs constructed in this paper, relying on multi-dimensional reward function design and multi-round RLHF fine-tuning process, effectively enhances the model's ability to capture industry terms and rare hidden danger patterns. The closed-loop of high-value samples selected by the active learning strategy and expert scoring enables continuous tracking and correction of long-tail risks in an environment with limited annotation resources. The reward and punishment signals and strategy updates run in parallel, which enables the model's judgment accuracy and coverage depth to be improved simultaneously, significantly optimizing the limitations of traditional static fine-tuning and rule-driven methods in dynamic log analysis scenarios. After 10 rounds of fine-tuning, the recognition accuracy of five categories of terms such as "overload" reached more than 80%, the average accuracy of six types of hidden dangers was 88.3%, and the recall rate was 90.0%; the accuracy increment in the first 30 rounds of iterations exceeded 0.035, and the coverage rate of rare hidden dangers exceeded 0.93, demonstrating excellent long-tail recognition capabilities and expert response capabilities. This achievement provides a replicable optimization path for small sample semantic analysis in the power industry, and lays a theoretical and practical foundation for the subsequent collaborative application of cross-domain expert feedback mechanisms and deep language models. Future research can focus on the adaptive adjustment of reward and punishment design and the deployment of larger-scale real-time feedback systems to further improve model robustness and operational efficiency.

Funding

This work is financially supported by the Science and technology project of China Southern Power Grid Co.Ltd. (No.GZKJXM20232503).

References

[1] Miller, M. Arbabzadeh, E. Gençer. Hourly power grid variations, electric vehicle charging patterns, and operating emissions. Environmental Science & Technology, 2020, 54(24), 16071-16085. DOI: 10.1021/acs.est.0c02312

[2] P. Mahmoudi-Nasr. Toward modeling alarm handling in SCADA system: A colored Petri nets approach. IEEE Transactions on Power Systems, 2019, 34(6), 4525-4532. DOI: 10.1109/TPWRS.2019.2916025

[3] J.M. Liu, Z.Y. Zhao, J. Ji, M.L. Hu. Research and application of wireless sensor network technology in power transmission and distribution system. Intelligent and Converged Networks, 2020, 1(2), 199-220. DOI: 10.23919/ICN.2020.0016

[4] L. Han, R.C. Zhang, X.S. Wang, A. Bao, H.T. Jing. Multi-step wind power forecast based on VMD-LSTM. IET Renewable Power Generation, 2019, 13(10), 1690-1700. DOI: 10.1049/iet-rpg.2018.5781

[5] C. Paul, P.K. Roy, V. Mukherjee. Optimal solution of combined heat and power dispatch problem using whale optimization algorithm. International Journal of Applied Metaheuristic Computing (IJAMC), 2022, 13(1), 1-26. DOI: 10.4018/IJAMC.290532

[6] R. Ucheniya, A. Saraswat, S.A. Siddiqui. Decision making under wind power generation and load demand uncertainties: a two-stage stochastic optimal reactive power dispatch problem. International Journal of Modelling and Simulation, 2022, 42(1), 47-62. DOI: 10.1080/02286203.2020.1829443

[7] A. Mousaei, M. Gheisarnejad, M.H. Khooban. Challenges and opportunities of FACTS devices interacting with electric vehicles in distribution networks: A technological review. Journal of Energy Storage, 2023, 73, 108860. DOI: 10.1016/j.est.2023.108860

[8] M.H. Ali, A.M.A Soliman, A.H. Adel. Optimization of reactive power dispatch considering DG units uncertainty by dandelion optimizer algorithm. International Journal of Renewable Energy Research (IJRER), 2022, 12(4), 1805-1818. DOI: 10.20508/ijrer.v12i4.13573.g8606

[9] I.L. Carreño, A. Scaglione, S.S Saha, D. Arnold, N. Sy-Toan, C. Roberts. Log (v) 3LPF: A linear power flow formulation for unbalanced three-phase distribution systems. IEEE Transactions on Power Systems, 2022, 38(1), 100-113. DOI: 10.1109/TPWRS.2022.3166725

[10] S. Sengupta, T. Spencer, N. Rodrigues, R. Pachouri, S. Thakare, P.J. Adams, et al. Current and future estimates of marginal emission factors for Indian power generation. Environmental Science & Technology, 2022, 56(13), 9237-9250. DOI: 10.1021/acs.est.1c07500

[11] M.M. Hosseini, L. Rodriguez-Garcia, M. Parvania. Hierarchical combination of deep reinforcement learning and quadratic programming for distribution system restoration. IEEE Transactions on Sustainable Energy, 2023, 14(2), 1088-1098. DOI: 10.1109/TSTE.2023.3245090

[12] Y. Chen, W. Wei. Robust generation dispatch with strategic renewable power curtailment and decision-dependent uncertainty. IEEE Transactions on Power Systems, 2022, 38(5), 4640-4654. DOI: 10.1109/TPWRS.2022.3214856

[13] S. Acharya, S. Ganesan, D.V. Kumar, S. Subramanian. Optimization of cost and emission for dynamic load dispatch problem with hybrid renewable energy sources. Soft Computing, 2023, 27(20), 14969-15001. DOI: 10.1007/s00500-023-08584-0

[14] L.P. Huang, C.S. Lai, Z.L. Zhao, G.Y. Yang, B. Zhong, L.L. Lai. Robust $ Nk $ Security-constrained Optimal Power Flow Incorporating Preventive and Corrective Generation Dispatch to Improve Power System Reliability. CSEE Journal of Power and Energy Systems, 2022, 9(1), 351-364. DOI: 10.17775/CSEEJPES.2021.06560

[15] S.Y. Wang, J. Liu, H.T. Chen, R. Bo, Y.H. Chen. Modeling state transition and head-dependent efficiency curve for pumped storage hydro in look-ahead dispatch. IEEE Transactions on Power Systems, 2021, 36(6), 5396-5407. DOI: 10.1109/TPWRS.2021.3084909

[16] J.K. Pattanaik, M. Basu, D.P. Dash. Improved real-coded genetic algorithm for reactive power dispatch. IETE Journal of Research, 2022, 68(2), 1462-1474. DOI: 10.1080/03772063.2019.1654933

[17] W.B. Chen, M. Tanneau, P. Van Hentenryck. End-to-end feasible optimization proxies for large-scale economic dispatch. IEEE Transactions on Power Systems, 2023,

39(2), 4723-4734. DOI: 10.1080/03772063.2019.1654933
[18] M. Kamruzzaman, J.D. Duan, D. Shi, M. Benidris. A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources. IEEE Transactions on Power Systems, 2021, 36(6), 5525-5536. DOI: 10.1109/TPWRS.2021.3078446
[19] A. Bin Thaneya, A. Horvath. Exploring regional fine particulate matter (PM2. 5) exposure reduction pathways using an optimal power flow model: the case of the Illinois power grid. Environmental Science & Technology, 2023, 57(21), 7989-8001. DOI: 10.1021/acs.est.2c08698
[20] S. Pandya, H.R. Jariwala. Single-and multiobjective optimal power flow with stochastic wind and solar power plants using moth flame optimization algorithm. Smart Science, 2022, 10(2), 77-117. DOI: 10.1080/23080477.2021.1964692
[21] H.Y. Lai, M. Nissim. A survey on automatic generation of figurative language: From rule-based systems to large language models. ACM Computing Surveys, 2024, 56(10), 1-34. DOI: 10.1145/3654795
[22] F.E. Ayo, J.B. Awotunde, L.A. Ogundele, O.O. Solanke, B. Brahma, R. Panigrahi, et al. Ontology-based layered rule-based network intrusion detection system for cybercrimes detection. Knowledge and Information Systems, 2024, 66(6), 3355-3392. DOI: 10.1007/s10115-024-02068-9
[23] X.L. Shi, X.T. Tian, L.P. Ma, X. Wu, J.G. Gu. A knowledge graph–based structured representation of assembly process planning combined with deep learning. The International Journal of Advanced Manufacturing Technology, 2024, 133(3), 1807-1821. DOI: 10.1007/s00170-024-13785-4
[24] P. Yang, Q.J. Li, L. Zhu, Y.J. Zhang. Research of lighting system fault diagnosis method based on knowledge graph. Journal of Computational Methods in Science and Engineering, 2024, 24(4-5), 2135-2151. DOI: 10.3233/JCM-24723
[25] H.G. Wang, X. Ji, X.L. Zhao, Y.D. He, T. Yu. Power data quality assessment and verification governance based on knowledge graph. Intelligent Decision Technologies, 2024, 18(2), 1271-1286. DOI: 10.3233/IDT-240054
[26] Z.W.Y. Gong, Z.G. Cao, S. Zhou, F. Yang, C.Y. Shuai, X. Ouyang, et al. Thermal Fault Detection of High-Voltage Isolating Switches based on Hybrid Data and BERT. Arabian Journal for Science and Engineering, 2024, 49(5), 6429-6443. DOI: 10.1007/s13369-023-08272-z
[27] K.P. Yu, L. Tan, S. Mumtaz, S. Al-Rubaye, A. Al-Dulaimi, A.K. Bashir, et al. Securing critical infrastructures: Deep-learning-based threat detection in IIoT. IEEE Communications Magazine, 2021, 59(10), 76-82. DOI: 10.1109/MCOM.101.2001126
[28] J.M. Yu, H.F. Wang, Y.X. Zhang, Z.M. Fei, H. Zhou, L.W. Wang. Automatic Risk Rating Method for Power Grid Field Operation Based on BERT. Power System Technology, 2023, 47(11), 4746-4754. DOI: 10.13335/j.1000-3673.pst.2022.2512
[29] A.B. Jeddi, A. Shafieezadeh, J. Hur, J.G. Ha, D. Hahm, M.K. Kim. Multi-hazard typhoon and earthquake collapse fragility models for transmission towers: An active learning reliability approach using gradient boosting classifiers. Earthquake Engineering & Structural Dynamics, 2022, 51(15), 3552-3573. DOI: 10.1002/eqe.3735
[30] D. Lombardi, T.F. Shipley, J.M. Bailey, P.S. Bretones, E.E. Prather, C.J. Ballen, et al. The curious construct of active learning. Psychological Science in the Public Interest, 2021, 22(1), 8-43. DOI: 10.1177/1529100620973974
[31] M.S. Li, H.M. Zhang, T.Y. Ji, Q.H. Wu. Fault identification in power network based on deep reinforcement learning. CSEE Journal of Power and Energy Systems, 2021, 8(3), 721-731. DOI: 10.17775/CSEEJPES.2020.04520
[32] J.X. Hu, Q. Wang, Y.J. Ye, T. Yi. Toward online power system model identification: A deep reinforcement learning approach. IEEE Transactions on Power Systems, 2022, 38(3), 2580-2593. DOI: 10.1109/TPWRS.2022.3180415
[33] Q. Yu, D.D. Liang, M. Qin, J.C. Chen, H.B. Zhou, J. Ren, et al. Cybertwin based cloud native networks. Journal of Communications and Information Networks, 2023, 8(3), 187-202. DOI: 10.23919/JCIN.2023.10272347
[34] N. Rane, S. Choudhary, J. Rane. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. Journal of Applied Artificial Intelligence, 2024, 5(1), 69-93. DOI: 10.48185/jaai.v5i1.1052
[35] H. Shi, L.R. Fang, X.Y. Chen, C.H. Gu, K. Ma, X.S. Zhang, et al. Review of the opportunities and challenges to accelerate mass-scale application of smart grids with large-language models. IET Smart Grid, 2024, 7(6), 737-759. DOI: 10.1049/stg2.12191