Risk Management and Healthcare Policy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

AI-Driven Medical Device Risk Management: A New Paradigm Integrating Large Language Models and Prompt Engineering for Standard-Risk Knowledge Graph Construction and Application

Wanting Zhu[1], Peiming Zhang[1], Wenke Xia[1], Ziming Gao[2], Weiqi Li[1], Ruixue Tian[3], Li Wang[4]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Educational Institution, Shanghai, People's Republic of China; [2]Oriental Pan-Vascular Devices Innovation College, University of Shanghai for Science and Technology, Educational Institution, Shanghai, People's Republic of China; [3]Lin-Gang Medical Device Innovation Center, Other Institution, Shanghai, People's Republic of China; [4]Henan Drug Evaluation Center, Regulatory Institution, Zhengzhou, People's Republic of China

Correspondence: Peiming Zhang, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516, Jungong Road, Yangpu District, Shanghai, People's Republic of China, Email zpmking@163.com

Purpose: To address the problems in medical electrical equipment risk management caused by the disconnection between unstructured medical electrical equipment standard documents and adverse event data, the lack of high-quality annotated data, and the reliance on manual combing for risk analysis.

Methods: This paper proposes a novel method for constructing a risk knowledge graph that integrates large language models and prompting engineering standards. Using adverse event data from early childhood incubators as a case study, it integrates multi-source standards to construct a three-layer risk knowledge system. It designs multi-angle prompting strategies involving entity relationships and employs a dual strategy of entity disambiguation and aggregation to achieve knowledge integration and standardization.

Results: The thought chain reasoning suggestion has the best performance (mean F1 score of 0.871). The constructed knowledge graph contains 24,106 nodes and 18,053 relationships, achieving a complete "fault-standard-measure" link. Based on this, a question-answering system for intelligent risk retrieval was developed.

Conclusion: This provides a low-cost, reusable knowledge graph construction path for the resource-constrained medical device field, promoting the transformation of risk management towards AI empowerment and assisting in intelligent supervision of adverse events related to medical devices.

Keywords: knowledge graph, large language model, prompt engineering, medical electrical equipment standards documents, intelligent risk supervision

Introduction

Medical Electrical (ME) equipment is the core pillar of modern healthcare. Its operational safety and clinical effectiveness are directly related to the life and health of patients. The risk management of this equipment relies on the standard document system established by the state and the industry for regulation and constraints.[1] However, the ME equipment standard documentation system is extensive, highly specialized, and structurally complex.[2] It is mostly stored in dispersed PDF files, with multiple layers of cross-references between standards,[3] leading to a disrupted linkage between standard clauses and clinical risks. For instance, the "temperature fluctuation limit" standard for infant incubators cannot be directly correlated with "adverse events of neonatal burns"; Simultaneously, unstructured standard documents and fragmented adverse event data rely on manual processing, resulting in delays in the identification of risk information.[4] For example, in the FDA database on adverse events involving infant incubators, more than 80% of malfunctions are caused by device failures, and of these, over 60% can be traced back to non-compliance with standard requirements; In

https://doi.org/10.2147/RMHP.S571156
Received: 30 September 2025
Accepted: 23 December 2025
Published: 9 January 2026

addition, risk analysis relies on expert experience, and the scarcity of regulatory experts in this field greatly limits the feasibility of risk screening.[5] These issues have caused ME equipment risk management to remain at the "passive post-event handling" stage, making it difficult to provide advance warnings. With the breakthrough of artificial intelligence technology in the field of medical risk management, how to use AI tools to open up the entire chain of "standard interpretation-risk identification-regulatory response" has become a key issue in promoting the intelligent transformation of healthcare risk management and ensuring patient safety.

As a powerful knowledge management and application tool,[6] knowledge graphs can provide technical path support for ME device risk management. By constructing a standardized knowledge network, disparate standard requirements are linked with adverse event data, enabling risk retrieval and in-depth analysis, thus providing strong support for regulatory decision-making. One of the core tasks in building a knowledge graph is the extraction of entities and relationships,[7] but traditional methods struggle to meet the practical requirements of risk management in this domain: Rule-driven methods rely on expert-defined "risk-standard" pattern matching mechanisms, which are costly to construct; machine learning-based methods require a large amount of annotated relational data.[8] However, the constant revisions to standards and the emergence of new faults make it difficult to implement; Even when based on deep learning methods such as BioBERT, PubMedBERT and the Att-BiLSTM-CRF medical pre-training model developed by Luo et al performs excellently in medical entity relationship tasks, yet it is still limited by the stringent requirements for data quality and scale inherent to such approaches.[9–11] Furthermore, these models have inadequate generalization ability for out-of-domain data, and their training processes consume substantial computational resources, making it difficult to achieve a balance between efficiency and adaptability in real-world settings with limited regulatory resources.

Generative large language models (LLMs) represented by ChatGPT have become a key artificial intelligence technology for breaking through the bottleneck of extracting risk information from medical devices due to their strong generalization ability in zero-shot/few-shot scenarios.[12,13] Through lightweight optimization methods such as prompt engineering, it precisely focuses on risk-related semantics and achieves efficient information extraction in low-resource scenarios.[14] Research indicates that LLMs possess significant potential in information extraction tasks.[15] Wang et al utilized LLMs to extract relevant entities from medical literature.[16] Chen et al proposed a problem decomposition approach to guide LLMs in the annotation and selection of medical named entity data.[17] Carl et al demonstrated that LLMs can provide patients with key medical information, thereby alleviating the time constraints of healthcare services.[18] This thesis proposes the development of a standard- and risk-oriented knowledge graph that integrates LLMs with prompt engineering, leveraging the powerful semantic understanding and relational reasoning capabilities of LLMs.[19,20] Realizing the semantic link of "standard clauses - failure risk - control measures" helps regulatory authorities connect the standard system with risk prevention and control needs, provides intelligent support for medical device risk management, and lays the foundation for strengthening post-market monitoring of medical products.[21] Furthermore, this study aligns with the regulatory requirements of core medical device risk management standards such as the IEC 60601 series and ISO 14971, as well as relevant FDA guidelines. The proposed AI-driven approach places greater emphasis on the traceability and verification of the risk chain. Compared to traditional biomedical informatics risk assessment tools mentioned above, this method offers significant advantages in reducing reliance on large-scale labeled data, better adapting to multi-level cross-referencing between standards and adverse events, improving risk handling efficiency, and enhancing scalability and interpretability. Specifically, the contributions of this thesis are as follows:

1. Taking the medical device infant incubator as the research object, a multi-source data fusion strategy was adopted to integrate adverse event data and relevant standard documents to form a "standard-risk" oriented knowledge graph, which solved the problem of standard-risk decoupling.
2. In the entity-relationship extraction phase, six targeted prompt-thinking optimization templates were designed to explore LLMs' capabilities in extracting entities and relationships related to risk connections across fault manifestations, standard clauses, cause descriptions, and control measures. It was verified that this approach saves a substantial amount of time compared to manual construction. Through a dual strategy of entity disambiguation and relationship aggregation for knowledge integration, the interference of model hallucinations on the accuracy of risk information is effectively mitigated.

3. On the basis of constructing a good knowledge graph, we will further build a question-answering system to achieve rapid retrieval of risk knowledge and intelligent analysis of corresponding standards, thereby improving the efficiency of regulatory decision-making.

Related Research

Standard Methods for Constructing Knowledge Graphs

Studies have shown that knowledge graphs have great research and application value in standard systems and risk analysis in multiple fields. Li et al developed a standardized knowledge graph for construction project plans in the building industry, enabling precise matching between construction risks and regulatory requirements.[22] Chen et al addressed the issue of risk identification bias caused by imbalanced entity categories in the food safety standards knowledge graph.[23] Liu et al proposed a knowledge graph-based digital modeling method for standard emergency response, enhancing the efficiency of risk response to sudden incidents.[24] Chen et al introduced the ALBERT-BiLSTM-CRF model to construct a knowledge graph for safety management standards in hydraulic engineering projects, effectively reducing the complexity of applying safety risk management standards in this domain.[25] Most of the aforementioned methods use deep learning models to extract textual information and construct knowledge graphs. However, in the field of medical electrical equipment where high-quality annotated data is lacking, on the one hand, the time and labor costs are unbearable, and on the other hand, there is a lack of integration of multi-dimensional data. This limitation leads to a serious disconnect between ME equipment standard knowledge and actual risk management needs, making it difficult to truly achieve end-to-end risk management. This thesis aims to break through this technical bottleneck and limitation, explore a new paradigm for building knowledge graphs that adapt to the risk characteristics of ME devices, and promote the digitalization of standards and the efficiency of risk management capabilities.

Prompting Method

Prompt engineering guides LLMs to produce more accurate and requirement-compliant outputs by optimizing input instructions. It is a lightweight, parameter-efficient fine-tuning method (PEFT).[26] Compared to full-parameter fine-tuning, prompt engineering only requires designing customized templates for risk management tasks.[27] It improves performance while keeping model parameters unchanged, reduces computational costs, and mitigates the issue of overfitting in small-sample scenarios. In knowledge extraction tasks, it demonstrates outstanding few-shot learning capability, making it more suitable for practical situations where ME device risk data is scarce and regulatory resources are limited. Li et al evaluated the effects of different prompting strategies on LLMs and confirmed that the characteristics of clinical questions influence the performance of LLMs.[28] Chen et al enhanced the knowledge reasoning capabilities of pre-trained language models by encoding structured prompts through knowledge graphs, effectively addressing the loss of structured information during knowledge injection.[29] Soman et al proposed the KG-RAG framework in the biomedical field to address challenges related to medical texts.[30] Xu et al designed a three-stage prompting approach combined with the TwoStepChat method to construct a heart failure knowledge graph, significantly reducing manual annotation costs.[31] Kan et al developed a strategy for composable prompts to improve generalization in low-resource scenarios.[32] Inspired by the above, this study fills the gap in the generalization ability of existing AI risk management tools in the medical device field. Six different suggestion optimization methods are designed to extract entity relationships from ME device standard documents and adverse event texts, laying the first step in building a risk-oriented knowledge graph.

Methods

This study uses LLMs as an extraction tool to construct a risk-oriented knowledge graph for medical electrical equipment series standards and implements a full-process framework for building a question-answering system, as shown in Figure 1. This framework uses customized prompting methods to guide LLMs to automatically extract entity relationship quintuples of risk and standard texts in zero-sample/few-sample scenarios, solving the real pain point of high cost of labeling ME equipment standard documents and adverse events, significantly reducing the workload of manual sorting, and improving risk management efficiency.

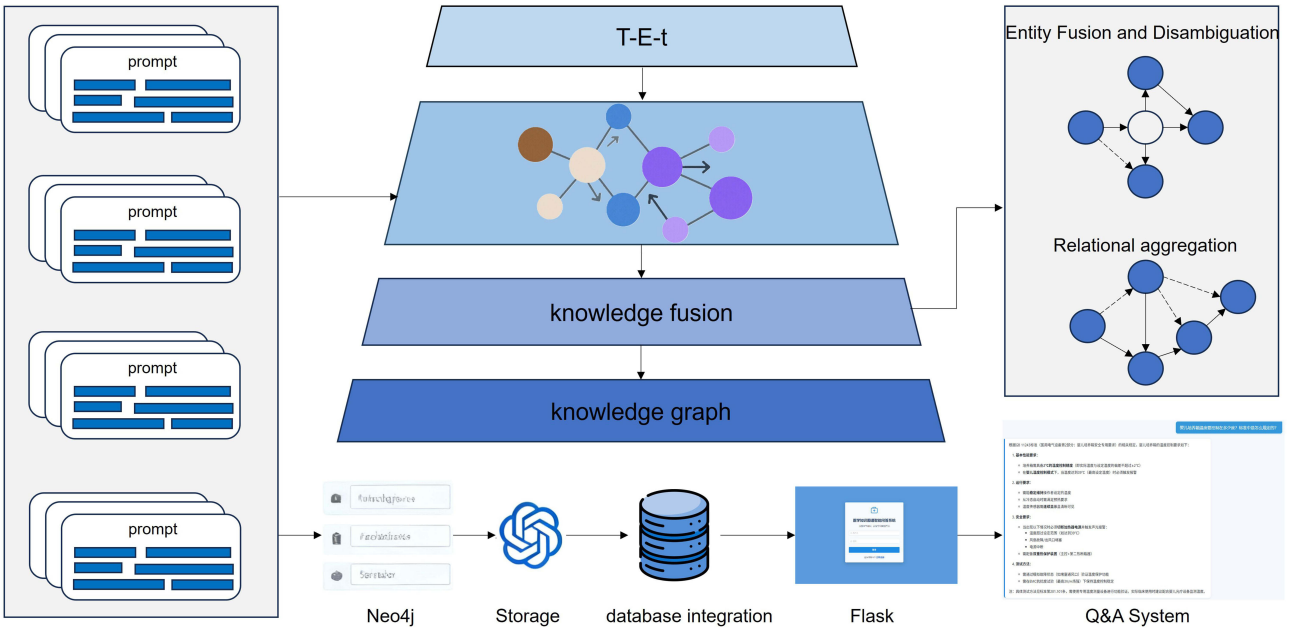

Figure 1 Medical Electrical Equipment Series Standards - Risk-Oriented Knowledge Graph and Q&A System Overall Architecture.

Data Collection and Processing

We selected Chinese standard documents for ME equipment converted from IEC international standards, including 10 core PDF standard documents covering general requirements such as general principles, environment, reliability, and availability, as well as four specific standards for infant incubators, focusing on extracting clauses directly related to risk management. Adverse event report data for Class III medical devices, such as infant incubators, was collected and, through structured processing, risk-related information such as "device failure manifestations," "usage process," "time cause analysis," and "specific control measures" was extracted to form a real-world failure case library. This complementary design of standards and clinical practice data breaks down the barriers between standards and risks, providing multi-dimensional data support for the construction of a risk knowledge map in the ME device field.

Given that medical electrical equipment standard PDF documents are mostly in image format, this study adopts the GOT-OCR2.0 framework. The standard PDF files are subjected to OCR recognition through prompt encoding and tensor conversion, overcoming issues of encoding errors, structural disorder, and content loss found in traditional methods. Adverse event data for infant incubators is extracted in a structured manner from Excel to TXT using a custom Python script.

Construction of the Standard and Risk System Framework

Based on the actual needs of medical electrical equipment risk management scenarios, we use a top-down modeling approach to build an ontology, propose a standard-risk-oriented T-E-t (Type-Element-text) three-layer knowledge system architecture, and progressively deconstruct standard documents to adapt to risk management needs. The first layer is the Risk Type (RT) layer, which categorizes risks into four types based on adverse event data; the second layer is the Risk Element (RE) layer, serving as a hub connecting standards and risks; the third layer is the Risk Text (Rt) layer, providing traceable semantic support to risk elements through standard documents and descriptions of adverse events. Through this method of dividing knowledge units, the systematic deconstruction of complex professional knowledge is achieved. An example of the standard-risk system architecture is shown in Figure 2.

Information Extraction

In the task of information extraction, this study formalizes prompt-based text processing methods as a structured mapping process: given the input text space $\mathcal{X}$ and the output key value space $\mathcal{K} \times \mathcal{V}$, there is a mapping function $f : \mathcal{X} \rightarrow P(\mathcal{K} \times \mathcal{V})$ that converts the original text into a discrete set of key-value pairs. Where $(k_i, v_i) \in f(x)$ satisfies

 https://doi.org/10.2147/RMHP.S571156

Figure 2 Three-tier architecture of knowledge modeling.

the injective constraints $k_i \neq k_j \Rightarrow v_i \neq v_j$, ensuring that each key $k_i i \in \mathcal{K}$ uniquely identifies a specific semantic unit (such as an entity or relationship) in the text, and its corresponding value $vi \in \mathcal{V}$ encodes the normalized representation of the unit guided by prompts. This formal framework can be mathematically characterized by:

$$f(x) = \{(k_i, v_i) \mid k_i = \phi(x, p_i), v_i = \psi(x, p_i)\}_{i=1}^{n} \tag{1}$$

In the formula, $\phi : \mathcal{X} \times \mathcal{P} \rightarrow \mathcal{K}$ is the key generation function based on the prompt template $p_i \in \mathcal{P}$, and $\psi : \mathcal{X} \times \mathcal{P} \rightarrow \mathcal{V}$ is the value derivation function, both achieve dynamic optimization through the parameter tuning θ_p via prompts. This configurable input paradigm enables dynamic adaptive optimization for down-stream tasks through the selective replacement mechanism of task introduction com-ponents. Taking the extraction of text segments from this article as an example, semantic index keys are used for structured organization, the value range stores the instance of the quintuples in the form of (s, p, o, t_s, t_o). Within this framework, the entity pairs $(s, o) \in T$ originate from the discrete representation of text fragments following deep semantic analysis, while the relationship $p \in R$ is restricted to a predefined set of ontology relations.

The baseline prompt template designed in this study is divided into three parts: Task Blueprint, which focuses on and outlines specific issues in the context of healthcare risk management; Model Input, which is used to process the unstructured standard documents and adverse event datasets in this work; and Prompt Optimization, which includes three key configurations across task and model dimensions—task objectives, prompt reasoning, and examples, as illustrated in Figure 3.

In order to maximize the extraction effect of Deepseek on data in this field, this paper designs and compares six types of prompt thinking guidance extraction: TwoStep CoT-zeroshot, TwoStep CoT-fewshot, EndtoEnd-zeroshot, EndtoEnd-fewshot, SecondRound dialogue-zeroshot, and SecondRound dialogue- fewshot. TwoStep CoT extraction transforms the model into a controllable "input-thought process-output" paradigm by embedding a thought framework, explicitly guiding LLMs' reasoning process on the cognitive path of healthcare risk management. End-to-end extraction achieves the coordinated optimization of risk entity recognition and association relationship classification through shared representation learning and joint decoding strategies, to a certain extent avoiding the error accumulation and propagation caused by pipeline extraction, and also alleviating the problem of entity boundaries constrained by relationship types in text. SecondRound dialogue interactive dialogue extraction lies in the fact that the input and output of each round of dialogue rely on the medical risk information clearly defined in the previous interactions, strengthening the strong relevance of the context. The specific differences are shown in Table 1.

https://doi.org/10.2147/RMHP.S571156

[TASK BLUEPRINT]

You are an expert in medical electrical equipment standard documentation and adverse event risk analysis management for entity relationship extraction. Extract structured quintuplets from the given text, and the output format should be： {“subject”: “”, “subject_type”:””,”predicate”:””,”object”:””,”object_type”:””}

[MODEL INPUT]

Input: In medical settings, medical errors and adverse events are recorded and analyzed in different ways, technically classified as distinct 'types of risk events'.。

[PROMPT OPTIMIZATION]

[Task Objectives]

1.Reference set of relationships R = {lead to, trigger, relate to, cause, include, preventive measures} and similar functional words.

2. Separate each output result with a "," and strictly follow the format specified by the user, without any additional output.

[Prompt Thinking]

3. First extract the entities from the text for identification, and then extract the relationships within the text.

[Examples]

4. Example Reference:

Text: <Excessive motor noise detected during use>

Output: ("subject": "during use", "subject_type": "process", "predicate": "trigger condition", "object": "excessive motor noise", "object_type": "phenomenon")

Figure 3 Quintuple extraction baseline prompt template.

Knowledge Integration

In the process of knowledge graph construction, entity merging and disambiguation are key steps in addressing the diversity of entity references and semantic ambiguities in multi-source heterogeneous data. We achieve entity merging and disambiguation as well as relationship aggregation by constructing a multi-level entity representation system and similarity computation model. The system adopts a three-dimensional index structure $\mathcal{I} = \{\mathcal{T} \times \mathcal{N} \rightarrow \mathcal{E}\}$, where $\mathcal{T}$ represents the entity type set, $\mathcal{N}$ represents the entity name set, and $\mathcal{E}$ represents the entity identification set, through the mapping function $f : \mathcal{T} \times \mathcal{N} \rightarrow \mathcal{E}$ implements type - quick retrieval of names to entities. In the field of entity similarity calculation, the system proposes a comprehensive similarity measurement model:

$$S = \alpha S_n + \beta S_t \tag{2}$$

Table 1 An Overview of the Structural Differences in Each Prompt Thinking Pattern

Prompt Optimization	Differentiation Structure Overview
TwoStep CoT-zeroshot	I. Do not provide any examples. II. Please complete the task according to the following chain of thought: “Step 1: Deconstructing the Risk Logic → Step 2: Extracting the Quintuples”. The reasoning process must reflect the causal relationship and logical progression of the medical risk scenario.
TwoStep CoT-fewshot	I.Provide 0–8 examples. II. Please complete the task according to the following chain of thought: “Step 1: Deconstructing the Risk Logic → Step 2: Extracting the Quintuples”. The reasoning process must reflect the causal relationship and logical progression of the medical risk scenario.
EndtoEnd-zeroshot	I. Do not provide any examples II. Synchronously identify the entities in this text and the semantic relationships between these entities.
EndtoEnd- fewshot	I. Provide 0–8 examples II. Synchronously identify the entities in this text and the semantic relationships between these entities.

(*Continued*)

Table 1 (Continued).

Prompt Optimization	Differentiation Structure Overview
SecondRound dialogue-zeroshot	I.Do not provide any examples II. First round of dialogue: Preliminary extraction of entity relationship quintuples from the text; Second round of dialogue: Combine the output of the previous round with the text again, and optimize and correct the output entity relationship quintuple according to the three dimensions of entity accuracy verification, relationship rationality judgment, and missing content supplementation.
SecondRound dialogue-fewshot	I. Provide 0–8 examples. II. First round of dialogue: Preliminary extraction of entity relationship quintuples from the text; Second round of dialogue: Combine the output of the previous round with the text again, and optimize and correct the output entity relationship quintuple according to the three dimensions of entity accuracy verification, relationship rationality judgment, and missing content supplementation.

Among them, S_n represents the name similarity component. Obtained by editing the weighted average of distance (formula 3) and sequence matching degree (formula 4) weighted average, Thus, we derive Formula 5.

$$L(s_1, s_2) = 1 - d(s_1, s_2)/max(|\ s_1\ |, |\ s_2\ |) \tag{3}$$

$$SM(s_1, s_2) = SequenceMatcher(s_1, s_2) \tag{4}$$

$$S_n = L(s_1, s_2) + SM(s_1, s_2)/2 \tag{5}$$

S_t is the type matching component, when the entity type is the same, the value is 1, otherwise it is 0; αandβ are the weight coefficients, with values of 0.85 and 0.15, respectively. When the comprehensive similarity S exceeds the threshold $\theta =$ 0.85, the system determines that the two entities are the same referent and performs the fusion operation, adding the new entity name to the alias set $A(e)$ of the existing entity, and updating the mapping relationship $\mathcal{M} : \mathcal{N} \times \mathcal{T} \rightarrow \mathcal{E}$.

During the entity disambiguation phase, the system adopts a three-tier processing architecture of candidate generation, ranking, and decision-making. First, a candidate entity set $\mathcal{C} = \{e_i \mid S(n, t, e_i) \geq \theta, i = 1, 2, \ldots, k\}$ is generated through type filtering and similarity calculation, where $k = 5$ represents the maximum number of candidates, The candidate entities are ranked based on their similarity scores, and the entity with the highest similarity is selected as the target entity. If all candidate entities have similarity scores below the threshold, a new entity is created and the index structure is updated. In terms of relational aggregation, the system adopts the MERGE operation semantics of the graph database to define the relational aggregation function $\mathcal{R} : \mathcal{E} \times \mathcal{P} \times \mathcal{E} \rightarrow \mathcal{G}$, where $\mathcal{P}$ represents the set of predicates, and $\mathcal{G}$ represents the graph structure. For each input tuple (s, p, o, t_s, t_o), the system first obtains the standard entity ID $s' = \mathcal{D}(s, t_s)$and$o' = \mathcal{D}(o, t_o)$ to through entity disambiguation, Then execute the $\mathcal{R}(s', p, o')$ operation, this operation automatically detects and merges the same relationships between the same entity pairs that already exist in the graph, and attaches timestamp attributes to the newly created relationships to form a temporal knowledge graph structure. The entire processing process adopts a batch transaction mechanism, and the data is processed in chunks through sliding window technology, with a processing scale of $\mathcal{B} = 200$ records per batch, ensuring the stability and efficiency of the system when processing large-scale data.

Knowledge Storage

The final step in building a knowledge graph is knowledge storage. This article uses the open source Neo4j graph database developed based on Java for storage. First, we define the node types and relationship types by designing an optimized graph model. Then, we use the Cypher query language to write a data import script, convert the pre-processed entity data into nodes in the graph database, and convert the semantic associations between entities into edges connecting nodes. This allows for visualization of the graph, allowing users to intuitively explore the complex relationship network between entities.

Experiments and Results

Datasets and Foundational LLMs

This thesis divides 14 collected standard documents and adverse event reports from infant incubators into an experimental subset and an implementation subset using a binary strategy. The experimental subset was annotated using expert-defined standards and independently by two research assistants, with a Kappa coefficient of 0.85 among annotators to ensure the high quality of the baseline dataset for evaluating the performance of different prompting strategies. The implementation subset contains approximately 6000 text units, used for constructing a full knowledge graph under the optimal strategy, forming a closed-loop verification system of "experimental validation - practical application". Specific data partitioning and processing scale statistics are shown in Table 2.

This thesis selects Deepseek as the foundational large language model for automatic labeling, opting for the Deepseek-r1 API.[33] This API supports a context length of 128K (ie, 128,000 tokens). The extended context window allows for the processing of longer text inputs, making it suitable for applications that require comprehensive understanding of lengthy texts.

Results of Prompt Optimization Method Extraction

We compared the performance of different prompt optimization templates on a subset of experiments. Figure 4 shows the output results of each prompt under the single-instance condition. Furthermore, this study investigated the impact of the number of examples on entity relation extraction using the controlled variable method, as shown in Figure 5. The figure shows that the six prompting strategies performed better in terms of precision, recall, and F1 score when the number of examples was small than when there were zero examples. Preliminary observations indicate that TwoStep CoT-fewshot performed best among the six strategies.

To ensure the statistical robustness of this conclusion and to quantify the differences between different strategies, we further supplemented the statistical significance test. We conducted three independent replicate experiments for each of the six cue strategies, with samples randomly drawn from the experimental subset for each experiment. We calculated the mean, standard deviation, and 95% confidence interval (calculated using a t-distribution with 2 degrees of freedom) of the F1 score for each strategy. We further employed paired t-tests to compare the performance differences between the optimal strategy and each of the other strategies. The results are shown in Table 3.

As shown in the table, the TwoStep CoT-fewshot strategy has the highest mean F1 score of 0.871, and its 95% confidence interval [84.1, 90.2] lower bound is generally higher than the upper bound of other strategies, indicating that this advantage has statistical robustness. Therefore, TwoStep CoT-fewshot can be considered the optimal prompting strategy for entity relation extraction in unstructured medical electrical equipment standard documents and adverse events of infant incubators, providing a reliable foundation for subsequent knowledge graph construction and risk-based intelligent question answering.

Table 2 Statistics on Data Partitioning and Processing Scale

Data Source	Total Number of Text Units	Experiment Subset	Implementation Subset	Experimental Subset Labeled	Perform Subset Extraction to Extract. (Divisors)
Standard documents (14 copies)	4237	850 (number)/20.1% (percentage)	3387 (quantity)/ 79.9% (percentage)	4011 (entity)/2950 (relation)	17274 entity/12860 (relation)
Adverse event reporting	2083	415 (number)/19.9% (percentage)	1668 (quantity)/ 80.1% (percentage)	1980 entity /1458 (relation)	8462 entity/6305 (relation)
Total	6320	1265 (number)/ 20.0% (percentage)	5055 (quantity)/ 80.0% (percentage)	5991 entity /4408 (relation)	25736 entity/19165 (relation)

Figure 4 Optimized templates and output results for different prompts.

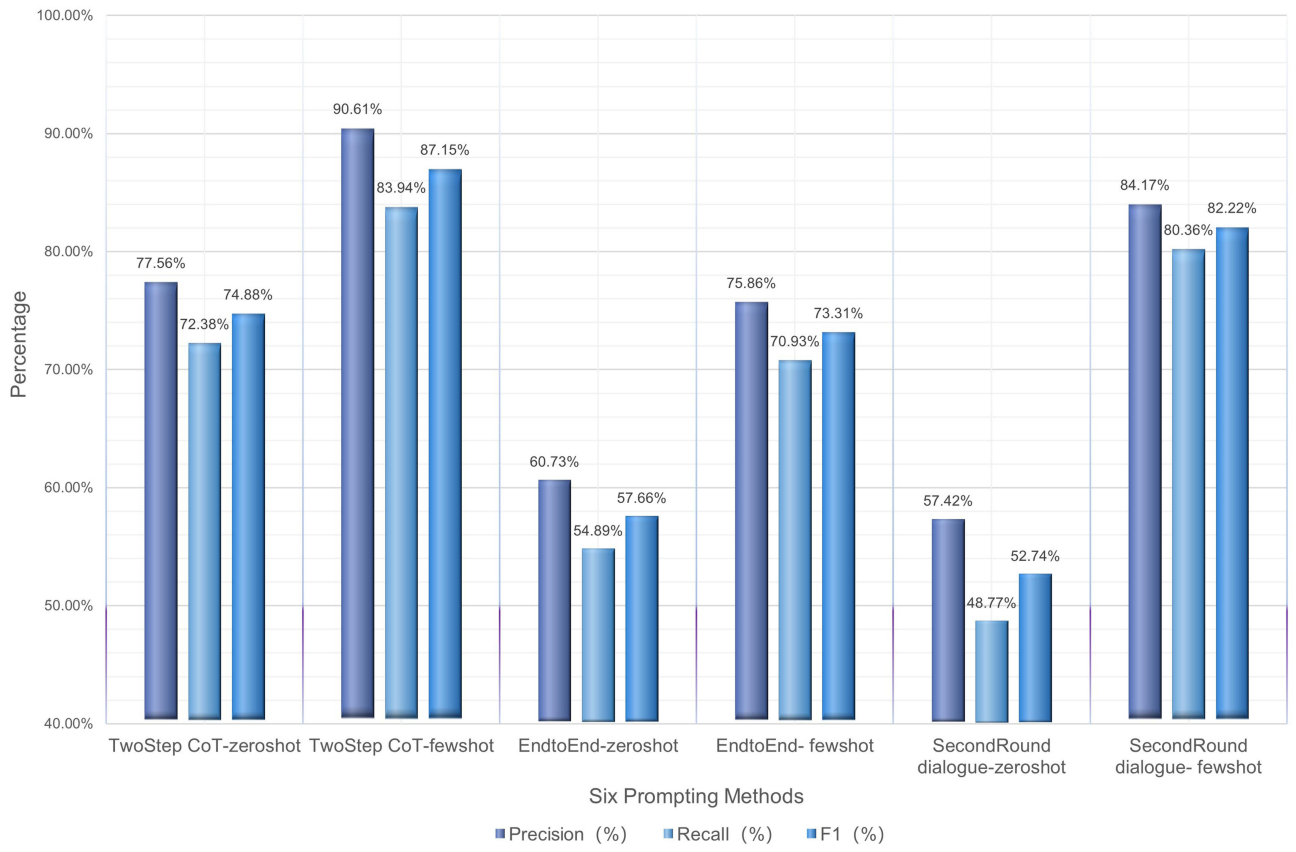

Figure 5 Precision, recall, and F1 value of the six prompting thinking methods under zero-shot and few-shot examples.

Performance Comparison of Different Models

To comprehensively evaluate the relative advantages of this method, we conducted a systematic comparative analysis. Considering the scarcity of labeled data in the field of medical device risk management, we comprehensively compared this method with traditional methods reported in the literature and tested the performance of other mainstream LLMs under the same experimental conditions. It is important to note that in the comparative analysis in Table 4, the performance data of traditional methods mainly comes from reports in relevant literature on similar medical information extraction tasks, while the comparative data of other LLMs were obtained from actual operation under the same experimental conditions to ensure the scientific rigor and impartiality of the comparison.

Table 3 Statistical Significance Analysis of the Six Prompting Strategies

Prompt Strategy	F1 Mean	Standard Deviation	95% Confidence Interval	Comparison with the Optimal Strategy
TwoStep CoT-fewshot	87.15%	1.2%	[84.1, 90.2]	–
TwoStep CoT-zeroshot	74.88%	1.5%	[71.3, 78.5]	p<0.05
EndtoEnd-fewshot	73.31%	1.4%	[69.9, 76.7]	p<0.05
EndtoEnd-zeroshot	57.66%	1.8%	[54.0, 61.3]	p<0.05
SecondRound dialogue-fewshot	82.22%	1.3%	[79.0, 85.4]	p<0.05
SecondRound dialogue-zeroshot	52.74%	2.0%	[48.7, 56.8]	p<0.05

Table 4 Performance Characteristics Comparison of Different Information Extraction Methods in Medical Text Tasks

Representation Method	Label Data Requirements	Typical F1 Value Range (Reported in Literature/Obtained Experimentally)	Calculation/ Deployment Costs	Main Applicable Scenarios
BiLSTM-CRF	Large-scale annotation	0.83–0.89 (obtained from literature)	Medium	Field fixed
BioBERT	Large-scale annotation	0.85–0.92 (obtained from literature)	Higher	Abundant annotation resources and fixed domains
DeepSeek + TwoStep CoT (ours)	Few samples	0.871 (obtained experimentally)	Medium	Low-resource, general-purpose rapid adaptation
GLM-4 + TwoStep CoT	Few samples	0.839 (obtained experimentally)	Medium	General domain adaptation
Qwen3 + TwoStep CoT	Few samples	0.826 (obtained experimentally)	Medium	General domain adaptation

Comprehensive comparisons show that the proposed method, which integrates LLMs and the prompting engineering paradigm, achieves the best F1 score in entity relation extraction under limited annotation resources, approaching the performance of models like BioBERT and BiLSTM-CRF, which require large-scale annotation. Under the same prompting strategy, DeepSeek also demonstrates superior domain adaptability compared to other mainstream LLMs. Therefore, this method offers unparalleled advantages over traditional methods in reducing annotation dependence and enabling rapid deployment, providing a feasible and efficient path for automated knowledge construction in resource-constrained, highly specialized scenarios such as medical device risk regulation.

Graph Visualization

The standard-risk-oriented knowledge graph contains 24,106 standard documents and adverse event nodes, as well as 18,053 relationship instances. Using "infant incubator" as the search core, it links risk management elements related to medical electrical equipment, including temperature requirements, components, alarm triggering mechanisms, verification requirements, inspections under certain circumstances, and corresponding standard documents, as shown in Figure 6.

In terms of adverse events, for the case entity "fan failure", the graph provides the possible causes of fan failure, the symptoms after the failure, and the subsequent solutions, as shown in Figure 7. This study has achieved knowledge expansion and practical implementation. It not only links with medical electrical standard documents to provide a basis for compliance attribution and responsibility definition of risk failures, but also provides certain practical value for risk management with the support of adverse event data.

Q&A System Development

To support efficient retrieval and intelligent services in healthcare risk management, this paper constructs a question-and-answer system oriented towards adverse events, standards and regulations, and risk associations, as shown in Figure 8. This system not only effectively addresses the knowledge gap between standard provisions for infant incubator temperatures and clinical risks at the application level, intuitively demonstrating how a graph-based question-and-answer system accurately responds to professional queries, but also theoretically analyzes how standards and adverse events are embedded in the risk management process, forming a relatively mature risk control framework.

The question-answering system was evaluated using a manual verification mechanism, and the question set used for testing was also reviewed by the aforementioned experts. The system-generated answers underwent blind review by experts and research assistants, and were scored on a five-point scale based on three dimensions: whether the answer accurately reflects the knowledge in the atlas, whether the answer directly answers the question, and whether the answer is clearly and

https://doi.org/10.2147/RMHP.S571156

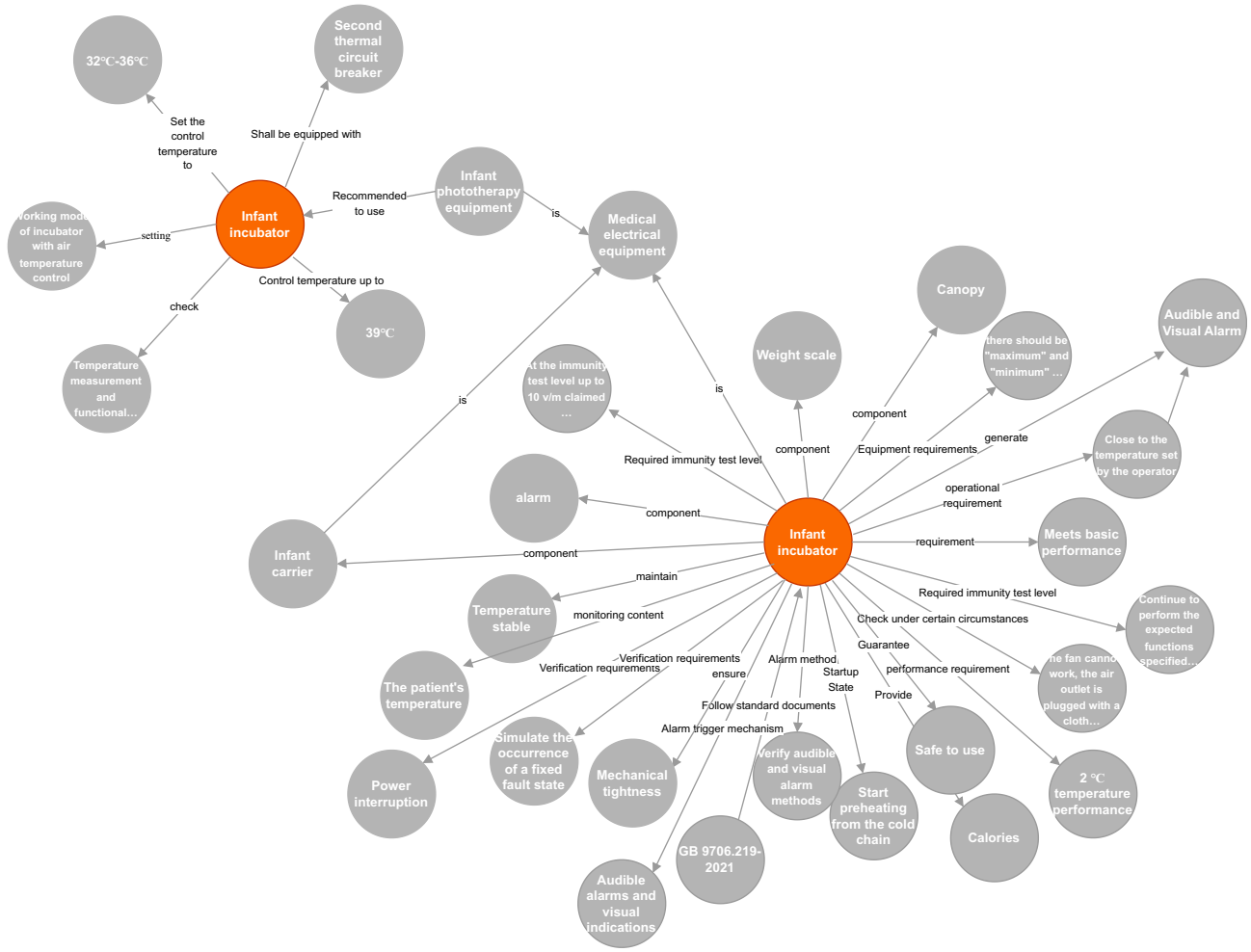

Figure 6 Search results for infant incubator.

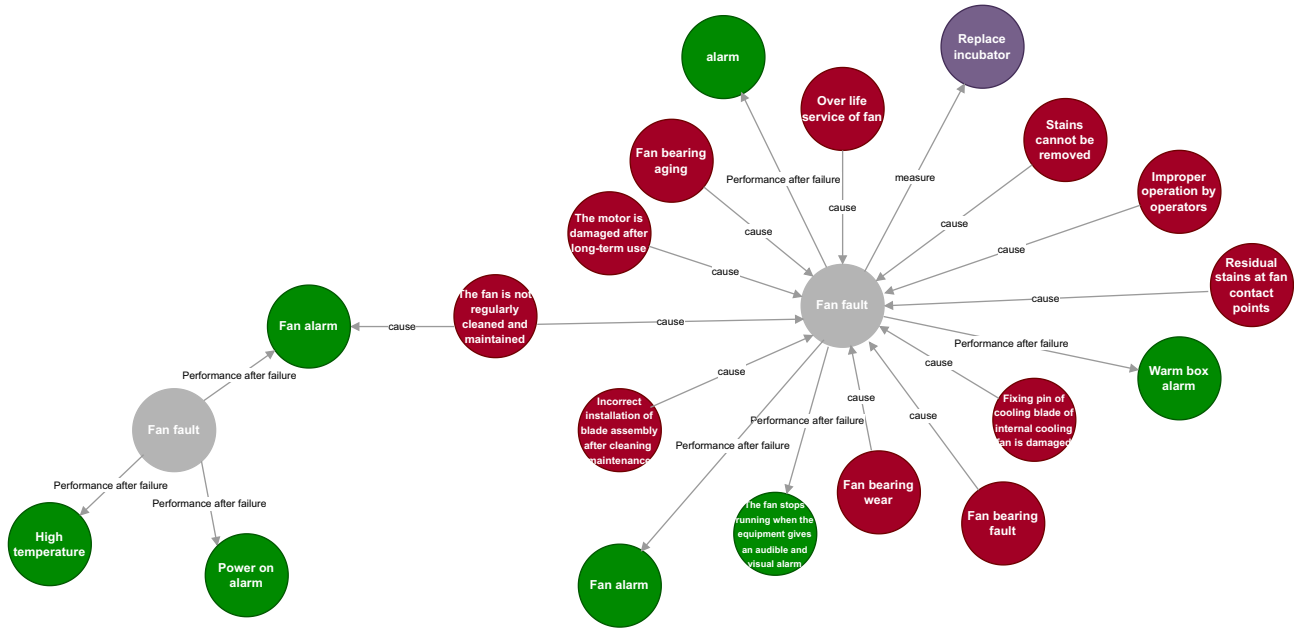

Figure 7 Causes of fan failure (indicated in red), symptoms (indicated in green), and solutions (indicated in purple).

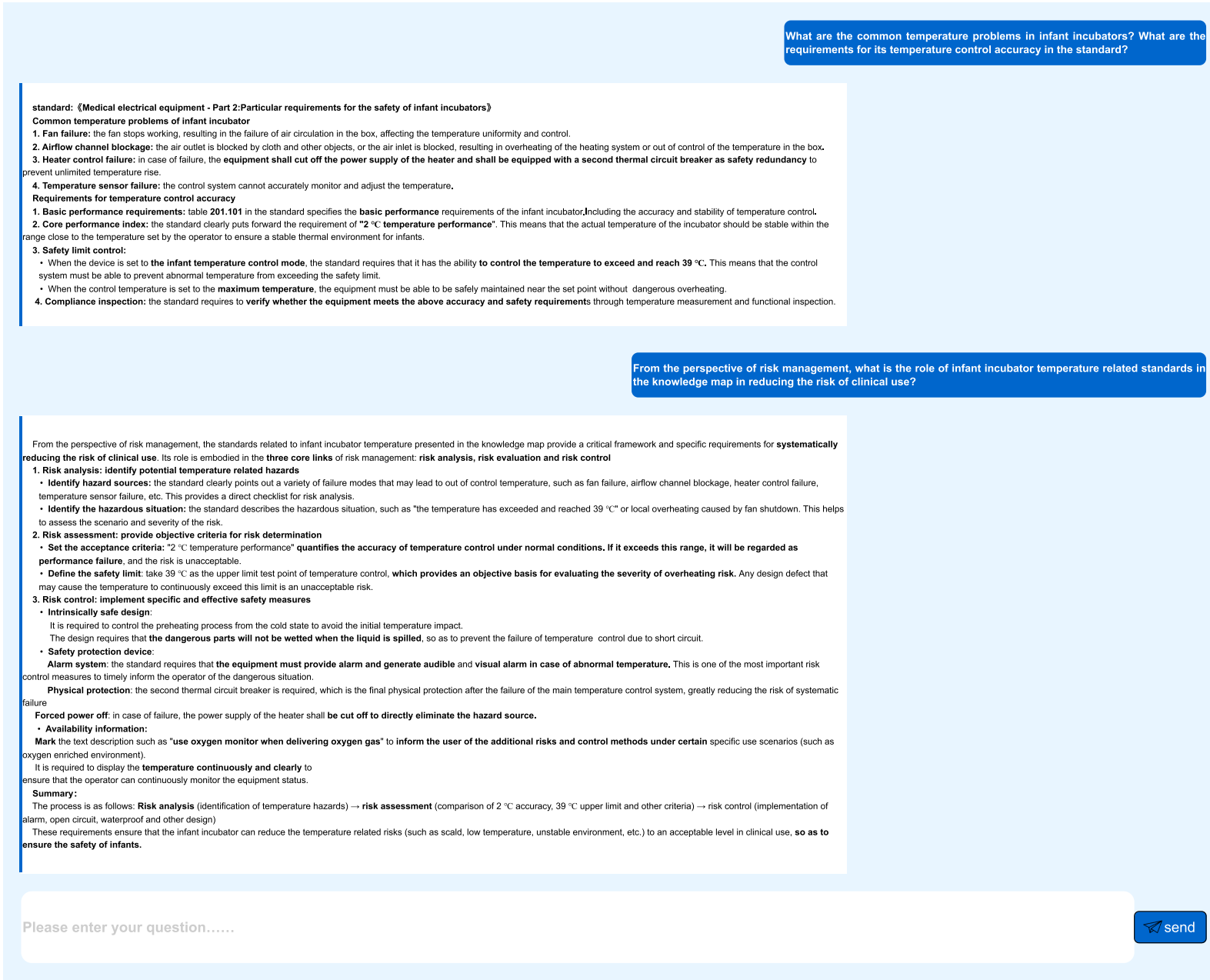

Figure 8 Example of a question and answer diagram.

unambiguously worded. The accuracy of the final report was based on the consistency of the scores from two reviewers (an average score > 4 was considered correct). This provides intelligent solutions for medical device regulation and risk control, further solidifying the foundation for the application of artificial intelligence in healthcare risk management.

Analysis and Discussion

The Influence of Different Prompt Optimization Methods on Extraction Performance

Overall, the differences in the effectiveness of the six entity relationship extraction methods are limited by the number of examples and the adaptability of the technical framework to the semantic characteristics of medical text. In the zero-sample scenario, all methods showed a common problem of insufficient semantic understanding due to the lack of example guidance. The most obvious problem was the SecondRound dialogue. Because its progressive dialogue method lacked an effective correction basis in the absence of examples, the path diverged and the error was transmitted, resulting in low accuracy. The overall analysis indicates that the reason lies in the strong coupling of semantic associations between medical electrical equipment standards and adverse event texts, which makes it difficult for LLMs to accurately capture such associative features in a zero-shot setting. In contrast, the introduction of few-shot examples effectively enhances the model's understanding of medical semantics, confirming the conclusion that "example-guided learning can improve the accuracy of entity-relation recognition". To further examine the performance differentiation among methods, TwoStep CoT leads across all three metrics. Analysis shows that its advantage stems from the distributed reasoning paradigm's adaptation to the progressive structure of this paper, from standard specifications to adverse event risk analysis. This approach not only avoids the high-dimensional interaction errors inherent in end-to-end synchronous modeling but also reduces the error propagation between steps in multi-round dialogues. At the same time, under the constraints of limited samples, stage-wise reasoning allows for a greater focus on the core elements of the text, thereby mitigating the cumulative effect of errors.

Performance Evaluation of Q&A Systems

To evaluate the effectiveness of the constructed question-answering system, we discussed with experts and compiled a set of simple queries and multi-hop queries containing standard professional terminology definitions, common adverse event failure manifestations, and solutions. This was used as a benchmark to test the performance of the question-answering system, as shown in Table 5.

Analysis found that static single-hop questions yield the best answers, as they can be directly matched with the entity-attribute pairs in the knowledge graph with a unique and clear reasoning path. In contrast, complex multi-hop questions or those requiring explanatory analysis necessitate dynamically linking these relationships. If the knowledge graph's coverage of such causal risk relationships is insufficient, or if there is ambiguity in inferring associations with adverse events, the answers may be biased. 'Despite these limitations, the overall accuracy still indicates that this question-answering system sufficiently meets the requirements for risk relevance and adverse event response in healthcare risk management, it is demonstrating significant application value.

Reliability Considerations and Mitigation Strategies for LLM-Generated Content

Although LLMs demonstrate strong semantic understanding and relational reasoning capabilities in few-shot scenarios, their generated content still carries the risks of "illusion" and "semantic drift," which are particularly critical in medical device risk management. To ensure the reliability and practicality of the knowledge graph, this study implements a three-layer mitigation mechanism: First, a two-step thought chain reasoning framework is used to limit the semantic space of LLM output, avoiding the introduction of irrelevant or fictitious entities. Second, entity disambiguation and relation aggregation algorithms are used to merge conflicting or redundant information, filter out obvious outliers, and improve knowledge consistency. Third, in the prompting strategy evaluation stage, domain experts validate the sampling results (Kappa=0.85), and the evaluation results are used to select the optimal prompting strategy, indirectly constraining the generation quality. The aforementioned mechanisms can reduce the error rate to some extent, but they still cannot completely eliminate the inherent uncertainty of LLMs. Future research will explore introducing uncertainty quantification mechanisms and real-time human feedback loops to further enhance the system's reliability in critical risk scenarios.

Table 5 Example Table of Q&A Task Performance

Question	Retrieval of Triads	Generate Answers
How is the accessible part defined? (Factual)	{"predicate": "definition", "subject": "accessible part", "object": "the parts of the electrical equipment other than the applied parts touched during standard testing"}	The accessible part refers to the parts of electrical equipment that can be touched by standard testing methods, excluding the applied parts, ensuring safe access and maintenance operations.
What are the common causes of faults in equipment? (Explanatory)	{"predicate": "lead to", "subject": " poor contact", "object": " unable to detect water level"}, {"predicate": "lead to", "subject": "the equipment has exceeded its service life", "object": "the water tank is leaking"}, {"predicate": " lead to", "subject": "the bearings of the fan have rusted and are unable to rotate due to liquid ingress", "object": "fan malfunction alarm"}	Common causes of equipment failure include: poor contact leading to an inability to detect water levels, which affects the monitoring system; equipment that has exceeded its service life may cause leakage in the water tank, resulting in resource wastage and potential damage; liquid entering the internal bearings of the fan and causing rust may trigger a fan fault alarm, leading to equipment shutdown.
What problems can arise from the aging of fans and blue light panels? (Application)	{"predicate": "lead to", "subject": "the fan has deteriorated and damaged after prolonged use", "object": "loud voice"}, {"predicate": "lead to", "subject": "the blue light panel of the equipment has exceeded its service life", "object": "aging leads to a decrease in brightness"}	The prolonged use of the fan may lead to aging and damage, resulting in increased noise levels, while the aging of the equipment's blue light panel beyond its service life may lead to a decrease in brightness.

The aforementioned technical mitigation strategies aim to improve the internal consistency of knowledge graphs. However, when such AI systems are placed in decision-making scenarios involving medical device regulation, their application inevitably touches upon deeper ethical dimensions: the principles of human oversight and ultimate responsibility must be upheld. In matters involving security alerts or compliance reviews, AI systems must serve as supplementary evidence and be subject to the final judgment and confirmation of domain experts to ensure clear accountability in decision-making. Meanwhile, algorithmic fairness, transparency, and explainability are the cornerstones of building regulatory trust. To avoid injustices arising from systemic misjudgments of specific devices or groups due to data bias, it is also necessary to enhance regulators' trust in AI-assisted decision-making by combining visualization and reasoning pathways. This study is only a preliminary exploration of the technology. Its mature application from an ethical perspective depends on interdisciplinary collaboration and social-technical dialogue, in order to ensure that technological innovation truly serves the fundamental goal of protecting patient safety and public health and well-being.

Limitation Analysis

While this study validated the effectiveness of the proposed framework in the category of infant incubator medical devices, certain limitations remain. First, the data source sample in this study is relatively limited, lacking diversity in terms of population, region, and ethnicity, which may introduce reporting bias and affect the model's generalization ability. Furthermore, current validation is limited to a single medical device. Future research will focus on collaborating with hospitals and institutions to conduct diverse tests on different data samples, extending the proposed validation framework to various medical device platforms to assess its broad applicability. Secondly, the evaluation in this study primarily relies on internally partitioned datasets and expert validation. Although cross-validation was employed, benchmark testing on external datasets is lacking. Future work will focus on validating the generalization ability of this framework on publicly available corpora of adverse events related to medical devices or cross-device standard datasets. Finally, LLMs have certain limitations in their fine-grained understanding of technical terminology, their entity and relation extraction is somewhat redundant, and their static storage characteristics are difficult to adapt to the dynamic revision requirements of standard documents, inevitably affecting the timeliness of risk warnings. Future research will focus on improving the model's ability to understand technical terminology in a fine-grained manner and its ability to quickly adapt to dynamic standards, thereby enhancing the system's practicality and timeliness.

Conclusion

This thesis proposes to construct a standard-risk-oriented knowledge graph by integrating large language models and prompting engineering. Using adverse events in infant incubators as a typical case, it completes the full-process verification from standard text parsing to intelligent regulatory response, solving the dual dilemma of scarcity of high-quality labeled data and fragmented semantic association in this field. By using the T-E-t (Type-Element-text) three-layer knowledge model, a progressive deconstruction and association of standard clauses and adverse event data is achieved. Six prompt optimization schemes are designed and compared to reveal the most suitable entity relationship extraction method for text in this field. This automated extraction process saves most of the time cost compared to traditional manual annotation, and provides a low-cost and reusable technical paradigm for the digital transformation of medical electrical equipment standards and the application of adverse events. Meanwhile, the graph-based question-and-answer system has shifted the focus of correlation analysis from manual sorting to intelligent retrieval. This facilitates knowledge sharing in cross-industry regulatory collaboration, enabling grassroots regulators to quickly understand the risk logic behind standards. It also provides implementable technical modules for the digital health regulatory ecosystem, proactive post-market surveillance systems for medical devices, and AI-enhanced clinical safety analysis, driving the transformation of risk management from experience-driven to data-driven approaches.

Although this study uses an infant incubator as a validation case, the core of the proposed framework lies in deconstructing the semantic relationship between "risk factors - standard terms - failure manifestations," rather than relying on specific device parameters. When expanding to other medical devices in the future, the main work will focus on collecting data from the new domain (corresponding standards and adverse event reports) and preprocessing it using the same procedures. The core extraction and construction process can be referenced and reused. Meanwhile, this study explores a new path at the level of regulatory science, which is expected to empower regulatory agencies to achieve three major transformations: first, from passive response to proactive monitoring, by automatically linking standards and adverse events to provide early warning of potential risks; second, from experience-based reliance to data-driven approaches, providing structured knowledge support for the review of medical devices and the formulation of inspection priorities; and third, from information silos to knowledge collaboration, providing reusable technical templates for building a unified medical device risk knowledge base and promoting the sharing of cross-departmental regulatory experience.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grant no. 12302417), the Shanghai University of Technology Professional Degree Graduate Practice Base Project. (grant no. 2025-009), and the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (grant no. 24CGA51).

Disclosure

All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhang ZH, Rao W. Key risks and development strategies for China's high-end medical equipment innovations. *Risk Manag Healthc Policy*. 2021;14:3037–3056. doi:10.2147/RMHP.S306907
2. Shi FC, Shi JY, Zhao Y, Zheng Y. Text extraction and structuring of standard maintenance documents for metallurgical continuous casting equipments. In: *18th International Conference on Neural Networks (ISNN); July 11–14, 2024*. Weihai, China: Springer; 2024.
3. Xia TW, Dai ZX, Huang Z, et al. Establishment of technical standard database for surface engineering construction of oil and gas field. *Processes*. 2023;11(10):2831. doi:10.3390/pr11102831
4. Dao N, Quesada L, Hassan SM, et al. Generative artificial intelligence for automated data extraction from unstructured medical text. *JAMIA Open*. 2025;8(5):ooaf097. doi:10.1093/jamiaopen/ooaf097
5. Aquino YS, Rogers WA, Jacobson SLS, et al. Defining change: exploring expert views about the regulatory challenges in adaptive artificial intelligence for healthcare. *Health Policy Technol*. 2024;13(3):100892. doi:10.1016/j.hlpt.2024.100892
6. Cui HJ, Lu JY, Xu R, et al. A review on knowledge graphs for healthcare: resources, applications, and promises. *J Biomed Inform*. 2025;168:104861. doi:10.1016/j.jbi.2025.104861
7. Yu HL, Cao H, Dong CX, et al. A Joint entity and relation extraction model driven by fine-grained feature extraction. *Int J Pattern Recogn*. 2025;39(9):2554009. doi:10.1142/S0218001425540096

8. Cho SH, Lee M, Lee WH, et al. Mastitis classification in dairy cows using weakly supervised representation learning. *Agriculture*. 2024;14(11):2084. doi:10.3390/agriculture14112084
9. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–1240. doi:10.1093/bioinformatics/btz682
10. Gleb D, Timur I, Konstantin K, et al. The classification of short scientific texts using pretrained BERT model. In: *31st Medical Informatics in Europe Conference (MIE); May 29–31, 2021; Electr Network*. Netherlands: IOS PRESS; 2021.
11. Luo L, Yang ZH, Cao MY, et al. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *J Biomed Inform*. 2020;103:103384. doi:10.1016/j.jbi.2020.103384
12. Kirpalani A. Global ChatGPT interest across healthcare and education access. *Health Policy Technol*. 2025;14(5):101061. doi:10.1016/j.hlpt.2025.101061
13. Giannilias T, Papadakis A, Nikolaou N, Zahariadis T. Classification of Hacker's posts based on zero-shot, few-shot, and fine-tuned LLMs in environments with constrained resources. *Future Internet*. 2025;17(5):207. doi:10.3390/fi17050207
14. Wang L, Chen X, Deng XW, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digit Med*. 2024;7(1):41. doi:10.1038/s41746-024-01029-4
15. Zhang CH, Lei XM, Xia Y, Sun LM. Automatic bridge inspection database construction through hybrid information extraction and large language models. *Dev Built Environ*. 2024;20:100549. doi:10.1016/j.dibe.2024.100549
16. Wang L, Ma YY, Bi WS, et al. An entity extraction pipeline for medical text records using large language models: ana-lytical study. *J Med Internet Res*. 2024;26:e54580. doi:10.2196/54580
17. Chen YB, Zhang BL, Li SR, et al. Prompt robust large language model for Chinese medical named entity recognition. *Inform Process Manag*. 2025;62(5):104189. doi:10.1016/j.ipm.2025.104189
18. Carl N, Haggenmüller S, Wies C, et al. Evaluating interactions of patients with large language models for medical information. *BJU Int*. 2025;135(6):1010–1017. doi:10.1111/bju.16676
19. Li XJ, Chen SH, Meng MQ, et al. Research progress and implications of the application of large language model in shared decision-making in China's healthcare field. *Front Public Health*. 2025;13:1605212. doi:10.3389/fpubh.2025.1605212
20. He WT, Ma HJ, Li SH, Dong H, Zhang HX, Feng J. Using augmented small multimodal models to guide large language models for multimodal relation extraction. *Appl Sci*. 2023;13(22):12208. doi:10.3390/app132212208
21. Matheny ME, Yang J, Smith JC, et al. Enhancing postmarketing surveillance of medical products with large language models. *JAMA Network Open*. 2024;7(8):e2428276. doi:10.1001/jamanetworkopen.2024.28276
22. Li H, Yang RZ, Xu SS, Xiao Y, Zhao HY. Intelligent checking method for construction schemes via fusion of knowledge graph and large language models. *Buildings*. 2024;14(8):2502. doi:10.3390/buildings14082502
23. Chen Y, Fan QX, Yuan XP, Zhang QH, Dong Y. PGD-GP: a Chinese named entity recognition model for constructing food safety standard knowledge graph. *IEEE Trans Multimedia*. 2025;27:2836–2847. doi:10.1109/TMM.2024.3373249
24. Liu WL, Yang YX, Tu XY, Wang W. ERSDMM: a standard digitalization modeling method for emergency response based on knowledge graph. *Sustainability*. 2022;14(22):14975. doi:10.3390/su142214975
25. Chen Y, Lu GY, Wang K, Chen S, Duan CF. Knowledge graph for safety management standards of water conservancy construction engineering. *Automat Constr*. 2024;168(Pt B):105873. doi:10.1016/j.autcon.2024.105873
26. Park J, Choo S. Generative AI prompt engineering for educators: practical strategies. *J Spec Educ Technol*. 2025;40(3):411–417. doi:10.1177/01626434241298954
27. Zhu J, Feng PC, Lu JW, Fang BW, Yang HS. ZeROf-Offload: forward-gradient scheme for efficient full parameter fine-tuning of billion-scale language models. *Mach Learn*. 2024;5(4):45054.
28. Li ZX, Yan CY, Cao Y, et al. Evaluating performance of large language models for atrial fibrillation management using different prompting strategies and languages. *Sci Rep*. 2025;15(1):19028. doi:10.1038/s41598-025-04309-5
29. Chen LY, Liu J, Duan YT, Wang RZ. KG-prompt: interpretable knowledge graph prompt for pre-trained language models. *Knowl-Based Syst*. 2025;311:113118. doi:10.1016/j.knosys.2025.113118
30. Soman K, Rose PW, Morris JH, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*. 2024;40(9). doi:10.1093/bioinformatics/btae560
31. Xu TH, Gu YX, Xue MT, et al. Knowledge graph construction for heart failure using large language models with prompt engineering. *Front Comput Neurosc*. 2024;18:1389475. doi:10.3389/fncom.2024.1389475
32. Kan ZG, Feng LH, Yin ZY, Qiao LB, Qiu XP, Li DS. A composable generative framework based on prompt learning for various information extraction tasks. *IEEE Trans Big Data*. 2023;9(4):1238–1251. doi:10.1109/TBDATA.2023.3278977
33. Xiong LL, Wang HF, Chen X, et al. DeepSeek: paradigm shifts and technical evolution in large AI models. *IEEE-CAA J Automatic*. 2025;12(5):841–858. doi:10.1109/JAS.2025.125495