

Article

Lightweight Two-Stage RAG Retrieval Model Construction and Optimization for Coal Mine Safety: An Industrial RegTech and Industry 5.0 Perspective

Meng Yang ^{1,*} , Zhonghao Zhao ², Ning Chen ³ and Pengpeng Zhang ⁴

¹ School of Artificial Intelligence, China University of Mining and Technology (Beijing), Beijing 100083, China

² Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China; zhaozhonghao@iie.ac.cn

³ Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing 100871, China; chenning@pku.edu.cn

⁴ Beijing Longsoft Technology Co., Ltd., Beijing 100080, China; zhangpengpeng@longsoft.com.cn

* Correspondence: m.yang@cumt.edu.cn

Abstract

In the field of coal mine safety, traditional large language models (LLMs) face issues such as poor knowledge timeliness, lack of traceable evidence, and susceptibility to hallucinations when handling complex knowledge-intensive tasks. To align with the human-centric principles of Industry 5.0 and meet the strict compliance requirements of Industrial Regulatory Technologies (RegTechs), this study aims to enhance the performance of a Retrieval-Augmented Generation (RAG)-based coal mine safety compliance system. A comprehensive coal mine knowledge dataset (CMK) is constructed to serve as a dynamic RegTech knowledge base. Based on the RAG framework, a two-stage retrieval structure is designed and optimized, transforming the generative AI into a safe, constrained reasoning core. Also, a hard-negative ranking dataset (CMK-R) is developed to support the evaluation of the reranking model and mitigate the risk of semantic hallucinations. The proposed approach adopts Dmeta-embedding-base and BGE-rerank-base as backbone models and applies fine-tuning, teacher–student structured distillation, and Matryoshka Representation Learning (MRL) to improve efficiency and reduce energy consumption. Experimental results demonstrate that the proposed CMS-base model achieves FAHR, MRR@10, and mAP scores of 93.48, 95.59, and 95.63, respectively, indicating that the system can retrieve the correct safety regulation or operational clause at the top ranks with high reliability. Furthermore, the lightweight variant reduces inference time by over 60%, and accuracy degradation remains below 1% even when the vector dimension is compressed to 64, significantly facilitating Edge Intelligence deployments. This robust, human-centric framework provides a scalable industrial AI solution for real-time safety compliance and decision support.



Academic Editor: Ananda Maiti

Received: 7 June 2026

Revised: 30 July 2026

Accepted: 30 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: retrieval-augmented generation; coal mine safety; two-stage retrieval; knowledge distillation; lightweight optimization

1. Introduction

The transition from Industry 4.0 to Industry 5.0 marks a profound paradigm shift in manufacturing and industrial processes. While Industry 4.0 prioritized hyper-automation, Industry 5.0 emphasizes a human-centric, sustainable, and resilient approach, leveraging

technology to empower workers rather than replace them. In high-risk industries such as coal mining, on-site operations demand strict compliance with operational thresholds, procedural sequences, and continuously updated safety regulations. RegTechs are applied to verify compliance in these complex industrial processes, ensuring that industry practices meet basic regulatory requirements and protecting worker safety.

Traditional end-to-end LLMs often rely on implicit parametric knowledge, leading to limited knowledge timeliness and fatal hallucinations in domain-specific tasks. RAG mitigates these limitations by introducing an external, updatable, non-parametric knowledge base to LLMs, allowing traceable evidence and grounding for knowledge-intensive tasks [1,2].

Globally, RAG has attracted considerable research attention in multiple domains. Lewis et al. [1] first formalized RAG as a framework that demonstrated significant improvements in factual consistency for open-domain question answering (QA). Subsequent studies optimized the retrieval component through dense vector models and hybrid retrieval strategies that integrate sparse and dense representations [3,4]. Researchers have explored efficiency-oriented techniques, including late interaction architectures, adaptive index compression, and vector quantization for large-scale retrieval [5,6]. In specialized fields like biomedicine, finance, and law, tailored RAG frameworks have been created to support domain adaptation and to operate effectively under strict latency constraints [7–9].

In the coal mine domain, on-site QA tasks demand strict compliance with operational thresholds, procedural sequences, clause identifiers, and clearly defined responsibility boundaries. Regulations, operational procedures, equipment manuals, and accident reports are frequently updated, posing additional challenges to maintain accuracy and consistency of knowledge. Given these characteristics, the updateability and traceability advantages of RAG make it an essential technical pathway for intelligent QA, risk assessment, and emergency decision support in mining operations. Existing studies focus on building QA systems that relate to regulations, equipment knowledge, and blasting operations. Wang Ying et al. [10] proposed a QA framework based on multi-task reading comprehension centered on coal mine safety regulations, improving adaptability and robustness in professional contexts. Dong Yaxin [11] designed a QA system for coal mine equipment that integrates intent recognition, knowledge retrieval, and response generation, demonstrating the feasibility of knowledge-graph-driven QA in specialized domains. Sun Jiayi et al. [12] developed a knowledge-graph-enhanced QA system for precise retrieval and reasoning over safety procedures and emergency plans. Zhu et al. [13] implemented an end-to-end RAG system that combined dense retrieval and generative responses for coal mine safety QA tasks. Also, the open source CoalQA project [14] integrates multi-source heterogeneous mining knowledge through a Neo4j knowledge graph and LangChain, supporting real-world applications of intelligent QA.

By constructing a domain-specific knowledge base, optimizing retrieval-generation pipelines and improving semantic understanding, RAG-based coal mine safety QA systems are evolving from functional prototypes to reliable, production-ready systems. However, most existing studies depend on open source retrieval models without reproducible or systematic solutions for hybrid retrieval, index compression, risk-aligned reranking, or end-to-end evaluation.

In industrial Edge IoT (IIoT) environments, RAG systems must overcome challenges related to large index sizes, high latency, and high resource consumption. In industrial and mining contexts, the challenges of RAG implementation manifest themselves as measurable and optimizable engineering problems. Under conditions involving millions of long documents, strict domain terminology, and cross-version clause retrieval, systems must maintain both high recall and precision while satisfying low-latency requirements

in production environments. Although uncompressed dense retrieval achieves strong recall, the resulting large index size and high search cost hinder deployment in resource-constrained settings [15,16]. Techniques such as model distillation and dimensionality reduction can effectively compress model size with minimal performance degradation, yet fine-grained parameter tuning is still required to ensure stability across heterogeneous hardware and high-concurrency deployments [17,18]. Large-scale indexes and the high cost of updates continue to be key obstacles to system maintainability, and although multi-vector representations enhance semantic coverage, they also substantially raise GPU memory usage and storage overhead [19–21].

Overall, three gaps are still not adequately addressed. First, existing coal mine safety QA studies mainly focus on functional system construction, while reproducible domain-specific retrieval datasets with detailed quality control are still limited. Second, most RAG systems rely on general-purpose embedding and reranking models, which may not capture safety-critical terminology, numerical thresholds, and cross-version regulatory expressions in coal mine documents. Third, existing RAG studies usually emphasize answer accuracy, while less attention is paid to latency, storage overhead, vector compression, and evidence-grounded refusal mechanisms required for industrial edge deployment.

This paper bridges the gap between sophisticated generative AI and resource-constrained industrial edge environments. The main contributions of this study are as follows:

1. A coal mine safety knowledge dataset (CMK) and a hard-negative reranking dataset (CMK-R) are constructed for domain-specific retrieval and compliance-oriented evaluation. Due to restrictions associated with the original data sources, the complete datasets are not publicly released; instead, detailed dataset statistics, construction procedures, and evaluation settings are provided to ensure reproducibility.
2. A two-stage retrieval framework is developed by fine-tuning Dmeta-embedding-base and BGE-rerank-base on coal mine safety data, improving the retrieval of safety-critical clauses and operational requirements.
3. Knowledge distillation and Matryoshka Representation Learning are incorporated to reduce inference latency and vector storage while maintaining retrieval performance, enabling resource-constrained industrial deployment.
4. A constrained generation and refusal mechanism is designed to ensure that generated answers are grounded in retrieved evidence, preserve safety-critical numerical values, and avoid unsupported compliance suggestions.

2. Related Theories

2.1. Industry 5.0 and Industrial RegTech

Industrial processes are inherently complex and strictly regulated. RegTechs consist of digital applications aimed at simplifying data collection and compliance checking. Within the Industry 5.0 framework, RegTech serves as a macro-level auditing tool and a micro-level operational assistant. Integrating Generative AI into Industrial RegTech requires rigorous safeguards against inaccuracies. Our RAG approach incorporates Explainable AI principles by treating retrieved text chunks as a strict chain of evidence for the generative process.

2.2. Retrieval-Augmented Generation and Embedding Models

RAG architecture connects a semantic retrieval engine with a generative language model. Mainstream industrial RAG systems adopt a two-stage retrieval framework, which consists of an initial bi-encoder for recall and a subsequent cross-encoder for reranking. In the first stage, efficient bi-encoder models such as DPR [4] and Contriever [3] perform large-scale recall. In the second stage, cross-encoder models such as ColBERT [22] rerank

the retrieved candidates to enhance precision and contextual alignment. The proposed system optimizes both stages specifically for the syntactic and semantic nuances of coal mine safety literature.

2.3. Recent Advances in Retrieval-Augmented Generation

Recent RAG frameworks mainly improve retrieval flexibility, reasoning ability, or generation reliability. GraphRAG constructs knowledge graphs and community summaries from private corpora to support global and multi-hop question answering over large document collections [23]. Its strength lies in relational reasoning and corpus-level summarization, but it also introduces additional graph construction and maintenance costs. Self-RAG improves factuality by enabling the language model to decide when to retrieve and to critique its own generations through reflection tokens, which is particularly useful for open-domain generation scenarios [24]. Adaptive RAG dynamically selects retrieval or generation strategies according to query complexity, reducing unnecessary retrieval and balancing cost and answer quality [25]. Corrective RAG further introduces a retrieval evaluator to detect low-quality retrieved passages and trigger corrective retrieval actions, thereby improving robustness against retrieval failures [26].

As summarized in Table 1, GraphRAG focuses on graph-based global reasoning, Self-RAG introduces self-reflection mechanisms for improving factuality, Adaptive RAG dynamically selects retrieval strategies according to query complexity, and Corrective RAG improves robustness through retrieval correction. In contrast, the proposed framework targets safety-critical industrial applications, where retrieval accuracy, evidence traceability, conservative generation, and deployment efficiency must be jointly optimized. Therefore, the framework integrates domain-specific retrieval enhancement, evidence-grounded generation, refusal control, and edge-oriented optimization through knowledge distillation and embedding compression.

Table 1. Comparison between representative RAG frameworks and the proposed framework.

Framework	Retrieval Enhancement	Generation Control	Evidence Grounding	Edge Deployment	Application Scenario
GraphRAG	✓	–	△	–	Global knowledge reasoning and multi-hop summarization
Self-RAG	✓	✓	✓	–	Open-domain factual generation with self-reflection
Adaptive RAG	✓	✓	△	△	Dynamic retrieval strategy selection according to query complexity
Corrective RAG	✓	✓	✓	–	Robust retrieval through low-quality evidence detection and correction
Proposed Framework	✓	✓	✓	✓	Safety-critical industrial compliance with lightweight retrieval and edge deployment

✓: Explicitly supported; △: Partially supported or implementation-dependent; –: Not a primary design objective.

2.4. Model Training and Evaluation Metrics

Training methods for embedding models generally include unsupervised contrastive learning, supervised positive–negative pair construction, and multi-task joint training strategies. Among these, contrastive learning aims to maximize similarity between semantically related text pairs while minimizing similarity between unrelated ones, thus enabling the model to capture semantic structures effectively from unlabeled corpora [27]. In supervised settings, positive and negative pairs can be constructed from question–answer pairs, title–body pairs, or abstract–text pairs, significantly enhancing the model’s ability to capture sentence-level semantic relationships.

For embedding model performance evaluation, common metrics include First Article Hit Rate (FAHR), Mean Reciprocal Rank (MRR), and mean Average Precision (mAP).

3. CMK Dataset

The CMK dataset was systematically crawled from the Coal Mine Safety Production network, yielding 53,821 valid text pairs after deduplication and data cleaning. It spans five primary categories: coal mine knowledge, management, operational procedures, technical measures, and regulations & standards.

3.1. Multi-Level Pagination Crawling

This work focuses on knowledge related to coal mine safety, encompassing multiple aspects such as laws and regulations, mine management, operational procedures, and technical measures. The target data are distributed across multiple directories and pages on the website. A multi-level pagination crawling strategy was employed to systematically acquire the target text information.

Figure 1 presents the four-stage web crawling pipeline. In Step 1, first-level category pages are located from the website's homepage, representing broad coal mine safety themes. In Step 2, each first-level page is parsed to obtain the URLs of second-level subcategories. Step 3 iterates through paginated lists within each subcategory to gather the URLs of relevant target articles. In Step 4, each target article is downloaded, HTML tags are stripped, and the title, category labels, publication information, and main content are extracted and saved as a structured text pair. During post-processing, invalid pages, duplicate URLs, empty articles, and documents with abnormal lengths are filtered out.



Figure 1. Process of multi-level pagination crawling for the CMK dataset.

3.2. Construction of the CMK Dataset

The collected data were stored in the form of text pairs, shown as Figure 2. A text is composed of *id*, *tag1*, *tag2*, *title* and *context*, *tag1* refers to the top-level category and *tag2*

refers to the lower-level category. To ensure data quality, we performed duplicate removal, empty-field filtering, HTML tag cleaning, abnormal-length filtering, and manual sampling inspection. A stratified random subset of 5000 samples, comprising 1000 samples from each category, was independently examined by two domain-aware annotators. Three quality dimensions were considered: category correctness, title–context consistency, and regulatory relevance. The observed agreement rates were 97.38%, 95.76%, and 95.30%, respectively. The corresponding Cohen’s kappa coefficients were $\kappa = 0.95$, $\kappa = 0.92$, and $\kappa = 0.91$, with an average agreement coefficient of $\bar{\kappa} = 0.93$. These results indicate almost perfect inter-annotator agreement. All disagreements were resolved through discussion, and an additional domain expert was consulted when necessary.

```
{ "id": 编号, "tag1": "一级标题", "tag2": "二级标题", "title": "题目", "context": "内容" }
{ "id": 1, "tag1": "操作规程", "tag2": "掘进采煤", "title": "胶带机司机安全技术操作规程", "context": "胶带机司机安全技术操作规程一、适用范围1、本规程适用于综采工作面SSJ1200/2*200、DSJ140/230/3*400可伸缩胶带输送机司机的操作作业。2、本规程适用我矿使用SSJ1200/2*200、DSJ140/230/3*400胶带输送机司机及
```

Figure 2. Storage format showing an example of text pairs in the CMK dataset.

Table 2 presents the statistical properties of the CMK dataset. Coal mine knowledge and management make up the largest share, mirroring the real-world distribution of safety-related information in mining environments. Although regulations and standards constitute a smaller portion, they are particularly crucial because they include safety-critical compliance requirements and numerical limit values. To assess whether CMK is appropriate for retrieval-based tasks, we examined the length distributions of documents and queries. The average document length is about 337 tokens, whereas queries average around 20 tokens. This pattern aligns with real-world coal mine safety consultation use cases, in which users generally pose brief operational questions and expect the system to retrieve comparatively lengthy regulatory excerpts. The dataset was partitioned into training, validation, and test sets using an 8:1:1 split, while preserving a consistent distribution across all subsets.

Table 2. Statistics of the CMK dataset.

Category	Number of Samples	Percentage (%)	Average Context Length (Tokens)	Average Query Length (Tokens)	Main Content Description
Coal Mine Knowledge	15,326	28.48	286	18.6	Basic mining knowledge, terminology, geological and production knowledge
Coal Mine Management	13,482	25.05	315	19.4	Safety management systems, responsibility allocation, inspection procedures
Operational Procedures	10,864	20.18	342	20.1	Standard operating procedures, equipment operation, worker instructions
Technical Measures	8276	15.37	368	21.3	Gas control, ventilation, prevention and control technologies
Regulations & Standards	5873	10.92	395	22.7	National regulations, safety standards, compliance requirements
Total	53,821	100.00	337	20.4	–

4. Two-Stage Retrieval Model

In a single-stage retrieval framework, retrieval results are obtained solely by computing the vector similarity of text embeddings, as illustrated in Figure 3 (steps 1 and 2). The two-stage retrieval framework extends this process by introducing an additional Step 3. After obtaining the Top-K candidate texts from the initial retrieval stage, a reranking

model is applied to reorder these candidates based on relevance, thereby producing the final ranked results.

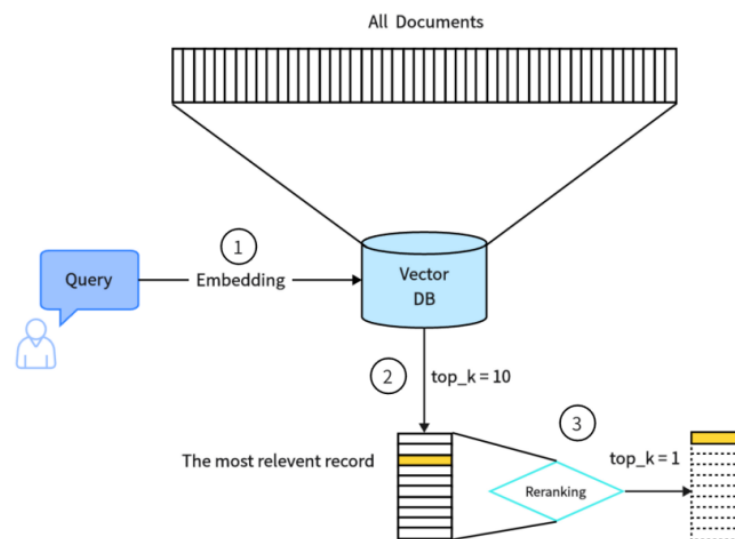


Figure 3. Two-stage retrieval flowchart.

In this study, an embedding model, denoted *CMS-base*, was developed by fine-tuning an open source general-purpose embedding model on the CMK dataset. To enhance efficiency, the model incorporates teacher–student structured distillation and MRL, enabling improved inference performance and reduced storage requirements without compromising retrieval accuracy. Building upon the CMK dataset, a CMK-R dataset was further constructed using the *CMS-base* model to improve the effectiveness of the reranking stage.

All experiments were conducted on a computing platform equipped with an Intel Xeon Gold 5318Y CPU (2.10 GHz) and an NVIDIA A40 GPU. The software environment comprised Python 3.8.11, the PyTorch 1.13.1 deep learning framework, and CUDA 11.7. A fixed-length text segmentation strategy was used, segmenting the corpus into chunks of 300 tokens with a sliding window of 50-tokens to preserve contextual continuity between segments. All vector similarities mentioned in this paper are computed by using cosine similarity. All experiments were repeated with five random seeds. We report the mean and standard deviation of FAHR, MRR@10, and mAP.

4.1. Embedding Model and Green Computing Optimization

The Dmeta-embedding-base model was fine-tuned using Multiple Negatives Symmetric Ranking Loss. To adhere to green computing principles and reduce energy overhead in continuous monitoring scenarios, teacher–student structured distillation was applied, compressing the 12-layer model into smaller variants. To address the bandwidth constraints of IIoT edge devices, MRL, which allows the model to retain critical semantic features at lower dimensions was incorporated. This adaptive capability means that a wearable smart mining lamp can transmit a highly compressed 64-dimensional vector over a constrained network, rather than the standard 768-dimensional vector, drastically reducing latency and network congestion.

4.1.1. Fine-Tuning of the Embedding Model

The Dmeta-embedding-base model is an open source semantic representation model released by Beijing Zhipu AI. It employs a bi-encoder architecture and outputs 768-dimensional dense vectors. The model is specifically designed for text vectorization and semantic retrieval tasks and is suitable for retrieval-augmented generation, semantic

search, similarity computation, and knowledge inference in both Chinese and mixed Chinese–English contexts.

Table 3 shows the hyperparameter configuration used to fine-tune the embedding model. During the fine-tuning process, a Warmup learning rate strategy is applied, with the number of Warmup steps set to 10% of the total training steps. This approach allows the model to start with a relatively low learning rate during the initial training phase to maintain the stability of deeper layers [28]. As training progresses, the learning rate gradually increases to the predefined value, thereby accelerating the convergence of the model and improving the training performance.

Table 3. Fine-tuning parameter settings for the embedding model.

Parameter	Value
Base Model	Dmeta-embedding-base
Train Batch Size	16
Epochs	20
Loss Function	Multiple Negatives Symmetric Ranking Loss
Learning Rate	2×10^{-5}
Scheduler	Warmup Linear Schedule
Warmup Steps	5384
Evaluation Steps	2692

During training, the model was evaluated on the validation set after each epoch. The evaluation metrics included FAHR and MRR@10 (Mean Reciprocal Rank of the Top-10 retrieved texts). Among these, FAHR was adopted as the primary reference metric for model selection to preserve the optimal checkpoint.

As illustrated in Figure 4, both the FAHR and MRR@10 improved consistently between epochs in the validation set. The fine-tuned model achieved its most significant performance gain after the first epoch, followed by gradual and steady improvements in subsequent epochs.

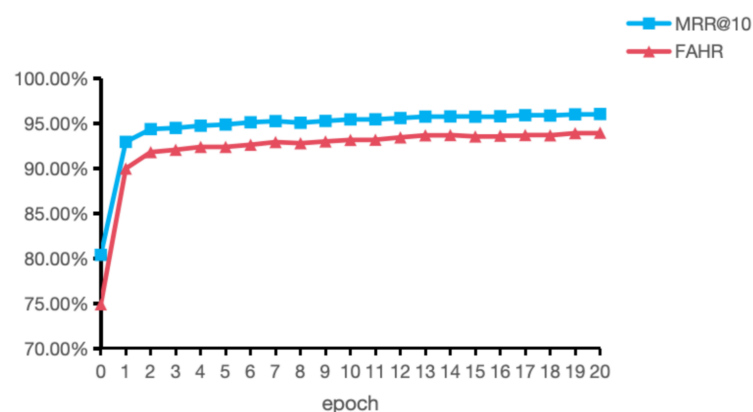


Figure 4. FAHR and MRR@10 trends on the CMK validation dataset during training.

The model obtained after 20 epochs of training was selected as the final model, called the *CMS-base*. A comparative evaluation was conducted between the *CMS-base* model and the original base model on the CMK test dataset. Table 4 presents the results. Compared to the base model, the *CMS-base* achieved notable performance improvements, with FAHR increased by 18.07, MRR@10 increased by 14.89, and mAP improved by 14.64.

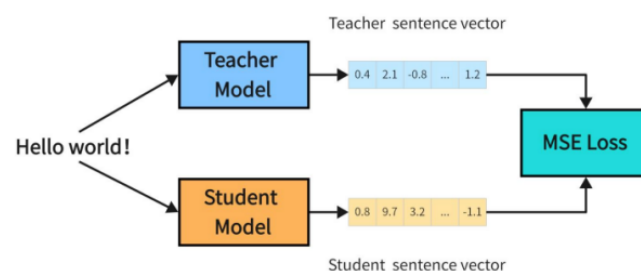
Table 4. Comparison of CMS-base and base model results.

Model	FAHR	MRR@10	mAP
Dmeta-embedding-base	75.41 ± 0.18	80.70 ± 0.21	80.99 ± 0.23
CMS-base	93.48 ± 0.15	95.59 ± 0.13	95.63 ± 0.14

4.1.2. Inference Time Optimization—Model Distillation

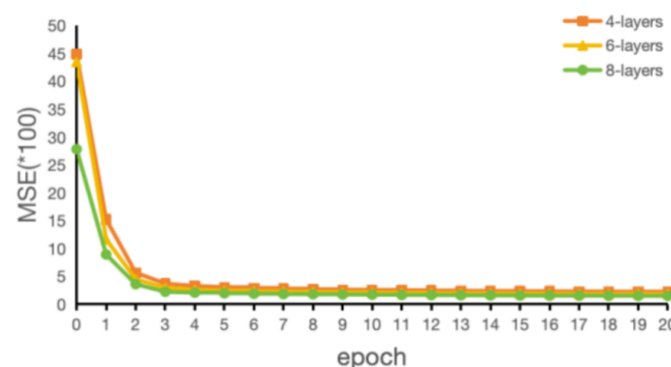
Model distillation is a machine learning model compression technique that is used to transfer knowledge from a large teacher model to a smaller student model. The student model is obtained by retaining a subset of the layers of the teacher model, and it is trained to mimic the teacher by minimizing the mean squared error (MSE) between the vector representations generated by both models, as shown in Equation (1). Here, n denotes the dimensionality of the embedding vectors. v_i and $\hat{v}_i$ represent the corresponding elements of the vectors generated by the student model and teacher model, respectively. By learning to replicate the behavior of the teacher, the student model can achieve relatively high performance, as illustrated in Figure 5.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (v_i - \hat{v}_i)^2 \quad (1)$$

**Figure 5.** Schematic diagram showing the model distillation.

In this study, the *title* and *context* texts from the CMK training dataset were used as training data for the model distillation, while those from the CMK validation dataset were used as validation data. The original CMS-base model contains 12 hidden layers. Based on this, subsets of layers were retained to construct three student models: *CMS_4_layers* with layers [1, 4, 7, 10], *CMS_6_layers* with layers [0, 2, 4, 6, 8, 10], and *CMS_8_layers* with layers [0, 1, 3, 4, 6, 7, 9, 10].

As shown in Figure 6, the MSE values between the student and teacher models continuously decreased and became stable after the third epoch, indicating that the performance of the student models was gradually approaching that of the teacher model.

**Figure 6.** MSE trends during training of student and teacher models.

The performance of the three student models on the CMK test set is shown in Table 5. Compared with the teacher model, the student models achieved a substantial reduction in model size and a significant improvement in inference efficiency, while maintaining retrieval effectiveness within an acceptable range of degradation.

Table 5. Model distillation training results.

Model	Size (Params)	FAHR	MRR@10	mAP	Time Cost
CMS-base	411 M	93.48 $\pm$ 0.15	95.59 $\pm$ 0.13	95.63 $\pm$ 0.14	101.87 s $\pm$ 1.24 s
CMS_8_layers	297 M	93.35 $\pm$ 0.16	95.50 $\pm$ 0.14	95.55 $\pm$ 0.14	71.11 s $\pm$ 0.96 s
CMS_6_layers	241 M	92.70 $\pm$ 0.18	95.02 $\pm$ 0.17	95.07 $\pm$ 0.16	55.64 s $\pm$ 0.83 s
CMS_4_layers	184 M	92.29 $\pm$ 0.21	94.66 $\pm$ 0.20	94.72 $\pm$ 0.20	40.39 s $\pm$ 0.69 s

4.1.3. Vector Storage Optimization with MRL

When the knowledge base contains a large number of documents, the storage of text embeddings constitutes a significant portion of the overall storage overhead. To address this, this work adopts the MRL method, which aims to optimize storage by reducing the dimensionality of text embeddings while preserving their representational capacity as much as possible.

The core idea of MRL is to evaluate the semantic quality of truncated embeddings at different dimensions during training, enabling the model to retain the most critical features for semantic discrimination in lower dimensions. This allows a dynamic trade-off between vector compression and semantic preservation.

CMS-base was used as the base model, with Multiple Negatives Symmetric Ranking Loss as the loss function. The embedding dimension of the CMS-base is 768, and the target lower dimensions were set at 64, 128, 256, and 512. During training, the FAHR metric in the CMK validation set was used as the primary reference to determine the optimal model for checkpoint saving. The trends of FAHR and MRR@10 over 20 epochs are illustrated in Figure 7.

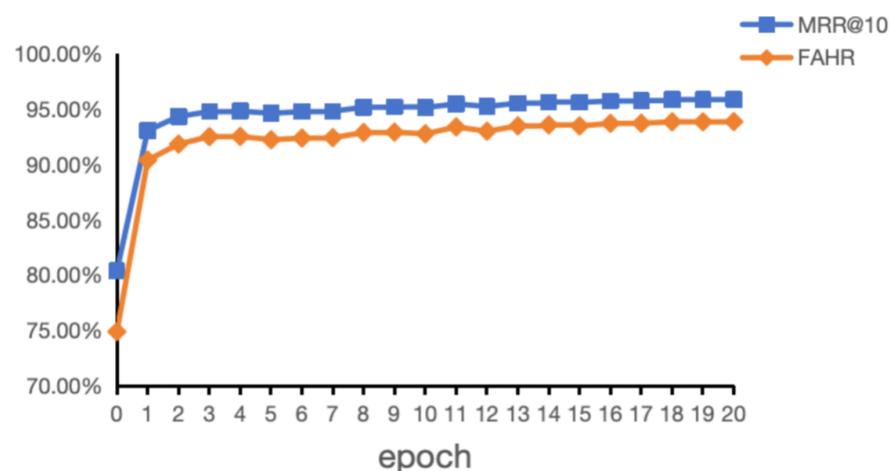


Figure 7. FAHR and MRR@10 trends over 20 epochs during MRL.

The model trained for 20 epochs was selected as the final *CMS-mrl* model. Its performance at different embedding dimensions is presented in Table 6. When the embedding dimension was reduced, the model performance degradation did not exceed 1%, thereby achieving storage optimization within an acceptable range of accuracy loss.

Table 6. Performance metrics for Matryoshka embedding training.

Model	Vector Dim.	FAHR	MRR@10	mAP
CMS-base	768	93.48 $\pm$ 0.15	95.59 $\pm$ 0.13	95.63 $\pm$ 0.14
CMS_mrl_512dim	512	93.02 $\pm$ 0.16	95.31 $\pm$ 0.15	95.34 $\pm$ 0.15
CMS_mrl_256dim	256	93.02 $\pm$ 0.17	95.25 $\pm$ 0.16	95.29 $\pm$ 0.16
CMS_mrl_128dim	128	92.76 $\pm$ 0.18	95.00 $\pm$ 0.18	95.04 $\pm$ 0.19
CMS_mrl_64dim	64	92.57 $\pm$ 0.20	94.81 $\pm$ 0.19	94.87 $\pm$ 0.19

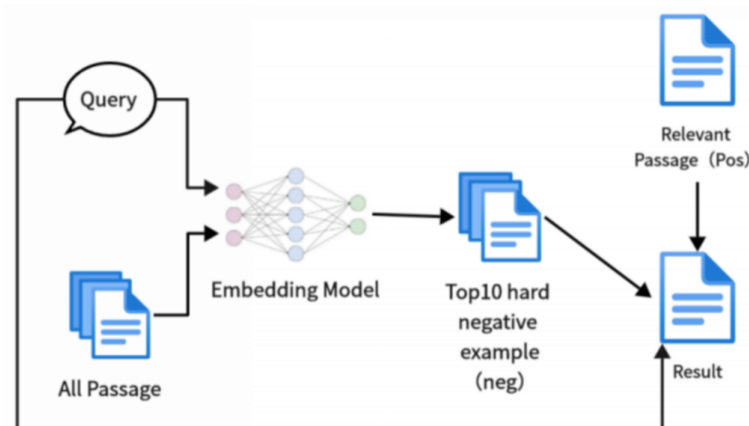
4.2. Reranking Model for Hallucination Mitigation

To ensure the LLM receives only highly relevant context, the CMK-R dataset was built by mining hard negatives from the base model's Top-10 recall list. The BGE-rerank-base model was fine-tuned on this dataset to act as a strict filter before the generative phase.

4.2.1. Construction of the CMK-R Dataset

In the context of embedding-based retrieval, irrelevant texts ranked at the top positions are referred to as hard negatives, while relevant texts are treated as positives. The task of the reranking model is to effectively distinguish positives from hard negatives and re-order the Top- N texts recalled by the retrieval model.

In this study, the *CMS-base* model was applied to the CMK training set to construct the data set for reranking, namely *CMK-R*. For model evaluation, independently partitioned CMK validation and test datasets were used to measure retrieval performance, thereby avoiding overfitting caused by overlap between the CMK-R and training samples. Figure 8 shows the construction process of the CMK-R dataset.

**Figure 8.** Construction process of the CMK-R dataset.

The construction of the CMK-R dataset followed four main steps:

1. **Obtaining initial retrieval results.** For each query, the CMS-base model was used to retrieve the Top-10 candidate texts from the CMK data set.
2. **Positive example annotation.** According to the original annotations, the relevant standard text corresponding to each query was defined as the positive example.
3. **Hard negative mining.** From the Top-10 retrieval results, after excluding the positive example, the top-ranked texts that are semantically similar to the query but actually irrelevant were selected as hard negatives.
4. **Construction of triplet samples (query, positive, hard negative).** Each query was paired with its corresponding positive and hard negative samples to form a training instance. The storage format is {"query": "xxxxx", "pos": ["xxxxx"], "neg": ["xxxxx", "xxxxx", ...]}.

4.2.2. Fine-Tuning of the Reranking Model

In this study, BGE-rerank-base was selected as the base model and fine-tuned on the CMK-R dataset for 10 epochs with a batch size of 64. The FAHR metric on the validation set was used as a reference indicator to determine the optimal checkpoint. As illustrated in Figure 9, the evaluation metrics tended to stabilize after the second epoch. The model obtained after the sixth epoch was adopted as the final *CMS-Reranking* model.

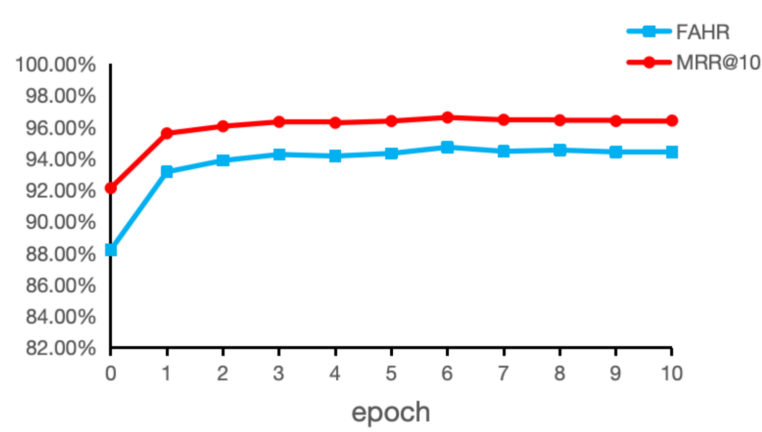


Figure 9. Performance trends of the reranking model during training.

The performance of *CMS-Reranking* was compared with the BGE-rerank-base baseline model in the validation set, and the results are shown in Table 7. Compared to baseline, *CMS-Reranking* achieved improvements of 6.24 in FAHR, 4.45 in MRR@10, and 4.43 in mAP, demonstrating its effectiveness in enhancing reranking performance.

Table 7. Reranking training results.

Model	Training Strategy	FAHR	MRR@10	mAP
BGE-rerank-base	No domain fine-tuning	87.93 ± 0.18	91.75 ± 0.16	91.78 ± 0.17
BGE-rerank-base + random negatives	Random-negative fine-tuning	91.86 ± 0.20	94.72 ± 0.18	94.75 ± 0.18
CMS-Reranking	Hard-negative fine-tuning	94.17 ± 0.14	96.20 ± 0.13	96.21 ± 0.13

4.3. Generative Core and Prompt Engineering

In the generative stage, the LLM is not treated as an open-ended text generator, but as a constrained reasoning core for regulation-grounded industrial assistance. Its primary function is to transform the evidence retrieved by the two-stage retrieval module into concise, structured, and traceable answers for coal mine safety compliance support. To align with the Industrial RegTech objective of this study, the generative module is designed under three principles: *evidence grounding*, *constrained answer generation*, and *explicit source attribution*.

4.3.1. Query Processing and Answer Generation Workflow

The complete generative workflow consists of five steps:

1. **Query normalization.** The user query is cleaned and normalized to reduce wording variation while preserving domain terminology.
2. **First-stage retrieval.** CMS-base retrieves the Top-10 semantically relevant passages from the CMK knowledge base.
3. **Second-stage reranking.** CMS-Reranking re-orders the candidate passages and selects the Top-5 evidence passages for generation.
4. **Prompt construction.** The system prompt, user query, and evidence passages are assembled into the constrained generation template.

5. **Answer generation and output formatting.** The LLM generates the final answer, which is then returned together with its evidence references.

This design is particularly important in coal mine safety scenarios, where incorrect threshold values, incomplete procedural steps, or fabricated regulatory references may lead to severe operational risks. The generative core is used as a context-bounded compliance assistant, whose output is restricted by prompt-level control rules and evidence-aware formatting requirements. The end-to-end pipeline was implemented in the Dify open source framework.

4.3.2. Prompt Engineering Strategy

To reduce hallucinations and improve compliance reliability, a constrained prompt engineering strategy was designed. The prompt consists of four parts: *system instruction*, *task definition*, *retrieved evidence context*, and *answer format constraints*. The complete prompt templates are provided in Appendix A.

1. **System instruction.** The model is assigned the role of a coal mine safety compliance assistant. It is instructed to provide only regulation-grounded answers and to avoid unsupported assumptions.
2. **Task definition.** The user question is reformulated as a compliance-oriented query, emphasizing that the response should identify the relevant rule, threshold, procedure, or operational requirement.
3. **Retrieved evidence context.** The top-ranked text passages returned by the two-stage retrieval module are inserted into the prompt as the only valid evidence source. Each passage is prefixed with an index number to facilitate citation and traceability.
4. **Answer format constraints.** The model is explicitly required to generate a brief and precise answer, cite the corresponding evidence passage or clause when possible, and refuse to answer when the retrieved context is insufficient.

Table 8 summarizes the key components of the proposed prompt design and their corresponding objectives in the generative stage. Compared with unconstrained prompting, this strategy reduces the possibility that the model supplements missing facts using its pre-trained knowledge, which is especially important for industrial questions involving gas concentration thresholds, power-cut conditions, evacuation rules, and operating procedures.

Table 8. Prompt design for the generative core.

Component	Description	Objective
System role	Defines the model as a safety compliance assistant	Restricts open-ended generation
Task prompt	Formulates the query as a regulation-oriented task	Focuses on thresholds, rules, and procedures
Evidence input	Provides retrieved passages as the sole evidence source	Grounds answers in external regulations
Passage indexing	Labels passages with identifiers	Supports traceability and citation
Answer constraint	Requires concise and evidence-based responses	Improves factual consistency
Numeric preservation	Preserves original threshold values and wording	Avoids safety-critical paraphrasing errors
Conflict handling	Prefers more specific regulatory evidence	Reduces ambiguity in generation
Refusal policy	Returns fallback output when evidence is insufficient	Mitigates hallucinations
Source attribution	Retains clause numbers or source-aware references	Enhances explainability and auditability
Brevity control	Limits unnecessary output expansion	Reduces token cost and latency

4.3.3. Constrained Generation and Refusal Mechanism

A core objective of the generative design is to prevent the model from producing plausible but unverified safety instructions. A refusal mechanism was incorporated into the prompt policy. The model is instructed to return a conservative fallback response whenever one of the following conditions occurs:

- no retrieved passage directly answers the question.
- the evidence passages contain conflicting threshold values or ambiguous procedural conditions.
- the retrieved content is descriptive but does not contain an explicit operational rule.

Under these conditions, the model does not generate a speculative answer. Instead, it outputs a standardized response indicating that the current evidence is insufficient and that further manual verification is required. This conservative strategy is consistent with the human-centric principle of Industry 5.0, where the AI system assists rather than replaces human safety judgment.

In addition, the model is instructed to preserve critical numerical values exactly as stated in the retrieved evidence. This is necessary because safety compliance queries often depend on small numeric differences, such as alarm thresholds, power-cut values, or ventilation parameters. Even minor paraphrasing errors may alter the operational meaning of the answer.

4.3.4. Evidence Formatting and Source Attribution

To improve interpretability, the retrieved evidence is normalized before being inserted into the prompt. Specifically, each passage is assigned a unique evidence identifier and ordered according to the final ranking score. When possible, the title or local metadata of the original source is retained to strengthen source awareness. This formatting enables the LLM to generate answers in an evidence-aware manner and simplifies subsequent manual inspection.

For regulation-intensive questions, the model is encouraged to include explicit traceable references such as “according to Passage [2]” or “the retrieved regulation indicates that ...”. If the retrieved text contains clause numbers or article identifiers, these are preserved in the generated answer. In this way, the output is not merely fluent natural language, but a traceable compliance-oriented response that supports Industrial RegTech requirements.

5. Experiments and Results

5.1. Baselines

In this section the CMS base model is compared with several widely used open source embedding models, such as the BM25 algorithm, GTE [29], BGE [30], and others. These models are known for their strong generalization and transferability in natural language processing tasks and have been widely applied to semantic retrieval, question answer systems, and multimodal alignment scenarios. Using FAHR, MRR@10, and mAP as evaluation metrics, experiments were conducted on the CMK test set. The results of each model are summarized in Table 9.

As shown, CMS-base significantly outperforms BM25 and other embedding models in the CMK test set, indicating that domain-specific fine-tuning is critical for coal mine safety retrieval.

Table 9. Performance comparison of various embedding models.

Model	FAHR	MRR@10	mAP
BM25	74.63 ± 0.00	79.80 ± 0.00	80.08 ± 0.00
gte-large-zh	80.14 ± 0.22	84.30 ± 0.25	84.52 ± 0.24
bge-large-zh-v1.5	84.10 ± 0.20	88.06 ± 0.23	88.21 ± 0.21
stella-large-zh-v3-1792d	84.23 ± 0.19	87.80 ± 0.20	87.97 ± 0.21
alime-embedding-large-zh	82.86 ± 0.24	86.55 ± 0.26	86.72 ± 0.25
xiaobu-embedding	86.36 ± 0.17	86.92 ± 0.19	87.08 ± 0.20
acge_text_embedding	73.68 ± 0.28	79.26 ± 0.30	79.58 ± 0.29
CMS-base	93.48 ± 0.15	95.59 ± 0.13	95.63 ± 0.14

5.2. Implementation Details of the Generative Module

To ensure the reproducibility of the proposed RAG framework, the implementation details of the generative module are provided in this section. The generative module is responsible for producing compliance-oriented answers based on retrieved evidence passages.

In our experiments, the large language model receives only the retrieved regulatory evidence and user queries as input. The model is instructed to generate answers strictly based on the provided evidence and to refuse answering when the retrieved information is insufficient, contradictory, or does not contain explicit safety requirements.

The embedding retriever first retrieves the top-10 candidate passages from the CMK knowledge base. These candidates are subsequently processed by the CMS-Reranking module, and the top-5 ranked passages are used as evidence for final answer generation.

The inference configuration and decoding parameters are summarized in Table 10.

Table 10. Implementation details of the generative module.

Item	Setting
LLM	Qwen2.5-7B-Instruct
Deployment	Local inference
Framework	Dify
Temperature	0.0
Top-p	0.9
Maximum generation length	512 tokens
Retrieved candidates	Top-10 passages retrieved by CMS-base
Evidence passages for generation	Top-5 passages after CMS-Reranking
Decoding strategy	Greedy decoding
Refusal trigger	No direct evidence, conflicting evidence, or no explicit operational rule

5.3. Ablation Study

Table 11 presents the ablation results of the proposed lightweight two-stage RAG framework. The test data set consisted of three hundred questions from the coal mine safety examination with standard references and answers. The questions were selected using a stratified sampling strategy and covered the five major categories. Each category contained 60 questions. The expanded evaluation set included numerical-threshold questions, procedural-sequence questions, equipment-specific questions, responsibility and management questions, emergency-response questions, and insufficient-evidence questions. An LLM-as-a-judge protocol was used to evaluate the generated responses in terms of answer correctness and relevance.

The CMS-Reranking module provides the largest individual improvement, increasing answer accuracy from 73.7% to 85.7%. This demonstrates that domain-specific hard-negative optimization is more effective than random-negative training for distinguishing semantically similar safety regulations. A representative comparison of the retrieved passages and generated answers before and after applying CMS-Reranking is provided in Appendix B.

The distilled models maintain competitive answer accuracy while reducing model complexity. Among them, CMS-6-layer achieves 83.7% accuracy, showing that moderate model compression can preserve most retrieval and generation capability. The MRL variants further demonstrate that embedding compression introduces only limited performance degradation, with the 128-dimensional configuration achieving 84.0% accuracy.

The complete framework achieves the best performance with an accuracy of 87.0% and a relevance score of 0.890. Removing prompt constraints decreases accuracy to 81.7%, indicating that evidence-bounded generation and refusal control are essential for preventing unsupported answers in safety-critical scenarios.

Table 11. Ablation study of the proposed lightweight two-stage RAG framework.

Variant	Hard-Negative Reranking	Distillation	MRL	Prompt Constraint	Answer Accuracy	Relevance
CMS-base	×	×	×	×	73.7%	0.821
CMS-base + random-negative reranker	Partial	×	×	×	79.0%	0.843
CMS-base + CMS-Reranking	✓	×	×	×	85.7%	0.881
CMS-8-layer + CMS-Reranking	✓	✓	×	×	85.0%	0.877
CMS-6-layer + CMS-Reranking	✓	✓	×	×	83.7%	0.865
CMS-4-layer + CMS-Reranking	✓	✓	×	×	82.3%	0.853
CMS-mrl-512dim + CMS-Reranking	✓	×	✓	×	85.3%	0.879
CMS-mrl-256dim + CMS-Reranking	✓	×	✓	×	84.7%	0.874
CMS-mrl-128dim + CMS-Reranking	✓	×	✓	×	84.0%	0.867
CMS-mrl-64dim + CMS-Reranking	✓	×	✓	×	83.3%	0.858
Full framework: CMS-6-layer + MRL-128dim + CMS-Reranking + constrained prompt	✓	✓	✓	✓	87.0%	0.890
Full framework without prompt constraint	✓	✓	✓	×	81.7%	0.848
Full framework without refusal mechanism	✓	✓	✓	Partial	84.3%	0.864

5.4. Error Analysis

Although the proposed two-stage RAG framework achieves strong retrieval and generation performance, several failure cases were still observed during the evaluation. To better understand the limitations of the system, we manually analyzed the incorrect cases in the test set and grouped them into five major error types: retrieval miss, reranking error, incomplete generation, evidence conflict, and refusal failure. As shown in Table 12, retrieval miss was the dominant error category, accounting for 30.8% of all incorrect responses. These cases mainly occurred when user queries contained informal expressions, domain abbreviations, or terminology that differed from the wording used in the corresponding regulatory clauses. Although the embedding model was fine-tuned on coal mine safety data, semantic differences between operational language and formal regulatory language remained a major source of retrieval failure. Reranking errors accounted for 23.1% of all errors. These errors typically involved passages that were highly similar in terminology and numerical content but differed in equipment type, installation location, or operational context. Incomplete generation accounted for 20.5% of all errors. In these cases, the correct evidence was successfully retrieved, but the generated response omitted one or more safety requirements. This problem was most common in questions whose reference clauses contained multiple thresholds, sequential actions, or parallel compliance conditions. Checklist-style prompting and explicit completeness constraints may therefore improve the generation of multi-item answers. Evidence conflicts and refusal failures each accounted for 12.8% of all errors. Evidence conflicts primarily resulted from differences among regulatory versions, authority levels, or operational contexts. Refusal failures occurred when the retrieved passages were topically related to the query but did not provide a sufficiently explicit operational rule. These findings indicate that future improvements should incorporate regulatory version metadata, evidence-authority ranking, conflict-resolution rules, and more reliable evidence-sufficiency detection.

Overall, retrieval and reranking errors jointly accounted for 53.9% of all errors. This finding indicates that fine-grained evidence selection remains the main performance bottleneck. Generation- and refusal-related errors accounted for 33.3%, demonstrating that correct retrieval alone does not guarantee a complete and safe compliance

response. The quantitative analysis therefore further supports the combined use of domain-specific retrieval, hard-negative reranking, constrained generation, and conservative refusal mechanisms.

Table 12. Error analysis of the proposed RAG framework.

Error Type	Percentage of Errors	Main Cause	Representative Example	Impact on Safety Compliance	Possible Mitigation
Retrieval miss	30.8%	The user query and regulation text use different terminology, abbreviations, or implicit expressions.	The query uses “gas alarm threshold”, whereas the regulation uses “methane sensor alarm concentration threshold”.	The correct regulatory clause may not appear in the Top-K retrieved candidates, leading to incomplete or incorrect evidence selection.	Introduce domain synonym expansion, query rewriting, and hybrid sparse–dense retrieval strategies.
Reranking error	23.1%	Hard negatives contain highly similar terminology, equipment names, or numerical thresholds, making fine-grained discrimination difficult.	A passage concerning an “upper corner gas sensor” is ranked above the passage for a “return airway gas sensor”.	The LLM may apply a correct threshold to an incorrect equipment type or operational scenario.	Construct equipment-aware, location-aware, and numerical hard negatives for reranking optimization.
Incomplete generation	20.5%	The retrieved evidence contains multiple safety requirements, but the LLM only outputs part of the required information.	The evidence specifies both alarm and power-cut thresholds, but the generated answer reports only the power-cut threshold.	Incomplete safety instructions may mislead operators and omit critical compliance requirements.	Apply checklist-style prompts and explicit multi-item completeness constraints.
Evidence conflict	12.8%	Retrieved passages originate from different regulatory versions, standards, or operational contexts.	One passage provides a general national threshold, while another specifies a stricter enterprise-level threshold.	The model may select an outdated, less authoritative, or contextually inappropriate regulation.	Incorporate version metadata, validity information, authority ranking, and conflict-resolution mechanisms.
Refusal failure	12.8%	Retrieved evidence is semantically related but does not contain an explicit operational requirement.	A descriptive passage about gas monitoring is retrieved, but no specific threshold or action requirement is provided.	The model may generate unsupported compliance advice instead of refusing an insufficiently supported request.	Strengthen evidence sufficiency detection and confidence-based refusal policies.
Total	100.0%	–	39 incorrect cases	–	–

Note: The full framework correctly answered 261 out of 300 questions, resulting in 39 incorrect cases. The percentage of errors is calculated using the 39 incorrect responses as the denominator.

6. Conclusions

In this work, the concepts of Industry 5.0 and Industrial RegTech are operationalized as concrete system design requirements. Human-centricity is embodied in the refusal mechanism and evidence attribution, ensuring that the AI system assists rather than replaces human safety decisions. Sustainability is achieved through model distillation and MRL-based vector compression, which lower inference latency and storage costs. Resilience is ensured by the updatable CMK knowledge base and a two-stage retrieval process, enabling regulatory knowledge to be refreshed without retraining the LLM. RegTech-specific needs are met via traceable evidence passages, preservation of numerical thresholds, and compliance-focused answer formatting.

This study also has several limitations. First, the CMK dataset is primarily built from text-based coal mine safety documents, and does not yet integrate multimodal resources such as equipment schematics, monitoring trend curves, or accident images. Second, the current evaluation relies on offline test sets and coal mine safety examination questions, which may not adequately reflect real-time interactions in underground environments. Third, although a refusal mechanism is incorporated, the system still cannot fully avoid hallucinations when the retrieved evidence is incomplete or outdated. Fourth, the framework has not been deployed in an actual coal mine edge-computing setting characterized by

unstable networks, constrained hardware, and simultaneous users. In future work, we plan to expand the system toward multimodal RAG, regulatory version tracking, and validation through field deployment.

Author Contributions: Conceptualization, M.Y.; methodology, M.Y., Z.Z. and N.C.; validation, P.Z.; formal analysis, M.Y.; data curation, Z.Z.; writing—original draft preparation, M.Y. and Z.Z.; writing—review and editing, M.Y.; funding acquisition, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The Fundamental Research Funds for the Central Universities funded this research OF FUNDER grant number 2024ZKPYZN01 and Beijing Longruan Spatiotemporal Intelligence Technology Co., Ltd.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: Author Pengpeng Zhang is employed by Beijing Longsoft Technology Co., Ltd., Beijing, China. The remaining 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.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BGE	BAAI General Embedding
BM25	Best Matching 25
CMK	Coal Mine Knowledge dataset
CMK-R	Coal Mine Knowledge Reranking dataset
CMS	Coal Mine Safety
CUDA	Compute Unified Device Architecture
DPR	Dense Passage Retrieval
EMNLP	Empirical Methods in Natural Language Processing
FAHR	First Article Hit Rate
GTE	General Text Embeddings
GPU	Graphics Processing Unit
LLM	Large Language Model
mAP	mean Average Precision
MRL	Matryoshka Representation Learning
MRR	Mean Reciprocal Rank
MSE	Mean Squared Error
NLP	Natural Language Processing
QA	Question Answering
RAG	Retrieval-Augmented Generation
RegTech	Regulatory Technologies
TLA	Three Letter Acronym

Appendix A. Prompt Templates for Evidence-Constrained Generation

To improve the reproducibility of the generative stage, this appendix provides the complete prompt templates used in the proposed evidence-constrained RAG framework.

Appendix A.1. System Prompt

You are a coal mine safety compliance assistant. You must answer only according to the retrieved evidence passages. Do not use unsupported prior knowledge. Preserve all numerical thresholds, clause names, and

procedural conditions exactly as stated in the evidence. If the evidence is insufficient, conflicting, or does not contain an explicit operational rule, respond that the current evidence is insufficient and manual verification is required.

Appendix A.2. User Prompt Template

Question: {user_query}

Retrieved evidence:

- [1] {passage_1}
- [2] {passage_2}
- [3] {passage_3}
- [4] {passage_4}
- [5] {passage_5}

Please provide a concise compliance-oriented answer. Cite the evidence passage numbers. If no passage directly supports the answer, refuse to answer.

Appendix A.3. Refusal Template

The retrieved evidence is insufficient to provide a reliable coal mine safety compliance answer. Manual verification of the relevant regulation or operating procedure is required.

Appendix B. Representative Comparison of Retrieval and Answer Generation Results

Question	What are the alarm and power-cut evacuation thresholds for gas concentration in the return airway?
Reference Answer	≥1.0% for alarm; ≥1.5% for power cut and evacuation
CMS-base	Passage 1: When the gas concentration in the return airway reaches 1.5%, non-intrinsically safe power should be cut off automatically and personnel evacuated. . . Passage 2: The alarm value can be set by enterprises as needed, appropriately adjusted according to mine conditions. . . Passage 3: Parameters for methane monitoring in the upper corner differ from those in the return airway,with an example alarm threshold of 0.8%. . . Passage 4: Local fan power cut-off: 1.5% for power cut and evacuation, 1.2% for early warning. . . Passage 5: Gas sensor calibration and log management: weekly function tests, zero calibration before shifts. . . Answer: 1.5% for power cut and evacuation (incorrect)
CMS-base + CMS-Reranking	Passage 1: Return airway parameter table: Alarm threshold ≥ 1.0%; power cut and evacuation threshold ≥ 1.5%. . . Passage 2: The alarm value must not exceed 1.0%; enterprises may set stricter thresholds but not relax them; monitoring system should trigger audio-visual alarms. . . Passage 3: Conditions for power restoration after cut-off: concentration < 1.0% and stable for 10 min before power resumes. . . Passage 4: Handling of false alarms and sensor malfunctions (not related to thresholds). . . Passage 5: Monitoring points: return airway, local fan outlets, etc., should be continuously monitored, alarm parameters consistent with the parameter table. . . Answer: ≥1.0% for alarm; ≥1.5% for power cut and evacuation (correct)

References

- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation: San Diego, CA, USA, 2020; Volume 33, pp. 9459–9474.
- Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; Weston, J. Retrieval augmentation reduces hallucination in conversation. In *Proceedings of the Association for Computational Linguistics (ACL)*, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3784–3803.
- Izacard, G.; Lewis, P.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; Grave, E. Contriever: Unsupervised dense information retrieval with contrastive learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Online, 6–14 December 2021.
- Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 16–20 November 2020.
- Santhanam, K.; Khattab, O.; Saad-Falcon, J.; Potts, C.; Zaharia, M. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*; Association for Computational Linguistics: Seattle, WA, USA, 2022.
- Zhan, J.; Mao, J.; Liu, Y.; Guo, J.; Zhang, M.; Ma, S. Optimizing Dense Retrieval Model Training with Hard Negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1503–1512. [\[CrossRef\]](#)
- Yasunaga, M.; Chen, D.; Liang, P.; Leskovec, J. RAG-based Biomedical QA Systems: Integrating Domain Retrieval with Large Language Models. *Bioinformatics* **2022**, *38*, 3028–3036. [\[CrossRef\]](#) [\[PubMed\]](#)
- Luo, T.; Wang, C.; Zhang, Y. FinRAG: Retrieval-Augmented Financial Question Answering with Domain Adaptation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, Toronto, ON, Canada, 10–12 July 2023; pp. 6123–6135. [\[CrossRef\]](#)
- Talmor, A.; Ravfogel, S.; Goldberg, Y. LegalRAG: Retrieval-Augmented Generation for Case Law Question Answering. *Artif. Intell. Law* **2023**, *31*, 245–263.
- Wang, Y. Research and Implementation of a Coal Mine Safety Regulation-Based Question Answering System. Master's Thesis, China University of Mining and Technology, Xuzhou, China, 2023.
- Dong, Y. Research on Construction of Coal Mine Safety Equipment Knowledge Graph and Intelligent Question Answering Algorithm. Master's Thesis, Xi'an University of Science and Technology, Xi'an, China, 2023.
- Sun, J.; Li, P.; Guan, W.; Tan, J.; Zhao, M.; Yu, H.; Wen, Y.; Tang, H. Knowledge Graph-based Q&A System for Blasting Safety Management in Open-pit Mines. *BLASTING* **2025**, *42*, 188–201.
- Zhu, P.; Chen, W.; Chen, D.; Liu, J.; Liu, Z.; Liu, Y.; Lin, G. An efficient query system for coal mine safety information based on retrieval-augmented language model. In *Advanced Intelligent Computing Technology and Applications*; Huang, D.-S., Zhang, C., Pan, Y., Eds.; Springer Nature: Singapore, 2024; pp. 235–243.
- CoalQA: A RAG-Based QA System for Coal Mine Safety Knowledge. Available online: <https://github.com/yaosenJ/CoalQA> (accessed on 26 July 2024).
- Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Virtual, 19–23 April 2021; pp. 874–880.
- Johnson, J.; Douze, M.; Jegou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **2019**, *7*, 535–547. [\[CrossRef\]](#)
- Kusupati, A.; Bhatt, G.; Rege, A.; Wallingford, M.; Sinha, A.; Ramanujan, V.; Howard-Snyder, W.; Chen, K.; Kakade, S.; Jain, P.; et al. Matryoshka representation learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, 28 November–9 December 2022; pp. 1–15.
- Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv* **2024**, arXiv:2402.03216. [\[CrossRef\]](#)
- Ge, T.; He, K.; Ke, Q.; Sun, J. Optimized product quantization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 744–755. [\[CrossRef\]](#) [\[PubMed\]](#)
- Malkov, Y.A.; Yashunin, D.A. Efficient and robust approximate nearest neighbor search using HNSW. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 824–836. [\[CrossRef\]](#) [\[PubMed\]](#)
- Izacard, G.; Petroni, F.; Hosseini, L.; De Cao, N.; Riedel, S.; Grave, E. A memory-efficient baseline for open-domain question answering. *arXiv* **2020**, arXiv:2012.15156. [\[CrossRef\]](#)
- Khattab, O.; Zaharia, M. ColBERT: Efficient and effective passage search via late interaction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2020; Volume 43, pp. 39–48.

23. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130.
24. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
25. Jeong, S.; Baek, J.; Cho, S.; Hwang, S.J.; Park, J.C. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), Mexico City, Mexico, 16–21 June 2024.
26. Yan, S.Q.; Gu, J.C.; Zhu, Y.; Ling, Z.H. Corrective Retrieval Augmented Generation. *arXiv* **2024**, arXiv:2401.15884.
27. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6894–6910.
28. Gotmare, A.; Keskar, N.S.; Xiong, C.; Socher, R. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv* **2018**, arXiv:1810.13243.
29. Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv* **2023**, arXiv:2308.03281.
30. Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; Nie, J.Y. C-Pack: Packaged resources to advance general Chinese embedding. *arXiv* **2023**, arXiv:2309.07597.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.