Received: 03 May 2026; **Revised:** 18 October 2026; **Accepted:** 29 October 2026; **Published:** 20 November 2026
Archs Sci. (2026) Volume 76, Issue 5 Pages 38-50, Paper ID 2026504.
https://doi.org/10.68304/as/76504

Design of an Interactive Intelligent QA System for Power Knowledge Base Based on Language Model Combined with Intelligent Retrieval

Chuangye Wang[1], Tongxu Liu[1], Yu Wang[2], Degui Qiu[3] and Shilin Shen[1,*]
[1]Bengbu Power Supply Company, State Grid Anhui Electric Power Co., Ltd., Bengbu 233000, Anhui, China
[2]Electric Power Research Institute, State Grid Anhui Electric Power Co., Ltd., Hefei 230601, Anhui, China
[3]State Grid Fengyang County Electric Power Supply Company, Fengyang 233100, Anhui, China
Correspondence: ahssss11111@163.com

Abstract To address challenges in the power sector, such as heterogeneous knowledge sources, complex domain-specific semantic relationships, the limitations of traditional retrieval methods in handling multi-turn interactions, and the tendency of large language models (LLMs) to generate factual hallucinations—this paper proposes an interactive intelligent question-answering framework for power-domain knowledge bases. This framework integrates multi-agent knowledge construction, context-aware processing, and adaptive retrieval-augmented generation (RAG). First, an ontology-constrained, multi-agent collaborative knowledge extraction mechanism is designed; through entity localization, attribute fusion, and evidence tracing, a hybrid vector-graph knowledge base is constructed using FAISS and Neo4j. Second, the framework incorporates structured dialogue memory, context-aware query reformulation, and a two-tier keyword enhancement strategy to facilitate entity coreference resolution, semantic completion, and the propagation of domain-specific constraints across multi-turn interactions. Building on this, a multi-path retrieval mechanism is developed, incorporating semantic coverage, terminology anchoring, and document structure priors, which utilizes a lightweight language model to dynamically adjust retrieval weights and optimizes evidence ranking through cross-path consensus feedback. Finally, by integrating evidence-constrained generation, factual consistency verification, and uncertainty assessment, the framework enhances the traceability and reliability of the generated answers. The proposed framework is evaluated through the design of a multi-type power-domain QA benchmark, alongside comparative experiments, ablation studies, and robustness tests, assessing performance across dimensions such as retrieval quality, answer accuracy, multi-turn consistency, and computational efficiency.

Index Terms Large language models; power-sector knowledge base; retrieval-augmented generation; multi-agent collaboration; knowledge graphs; adaptive hybrid retrieval; multi-turn interactive question answering

I. Introduction

As the global energy landscape accelerates its transition toward a low-carbon, cleaner, and smarter paradigm, a new type of power system is emerging, characterized by the large-scale integration of new energy sources, the coordinated control of distributed energy resources, and the digital management of power equipment. The increasing complexity of power system operational models, equipment types, and business scenarios has created a greater need for efficient access to specialized knowledge, effective management of technical information, and intelligent decision-support tools. Throughout processes such as power generation, dispatch and operation, equipment maintenance, fault diagnosis, and safety management, technical personnel frequently need to consult a wide range of professional resources—including industry standards, equipment manuals, operational protocols, maintenance logs, and historical fault case studies. This knowledge is typically scattered across various formats, such as PDF documents, structured databases, technical reports, and internal corporate knowledge management platforms. These data sources are often heterogeneous, dense with technical terminology, characterized by complex semantic relationships and hierarchical document structures, and subject to frequent updates. While traditional knowledge retrieval systems—based on keyword matching and rule-based reasoning—can handle information queries under specific, well-defined conditions, they often fall short when faced with complex natural language questions, cross-document knowledge associations, or sustained interactive technical consultations. Limitations include a restricted capacity for semantic understanding, inefficient information retrieval, and a lack of flexibility in how knowledge is utilized. This is particularly critical in scenarios

involving equipment operating conditions, the applicability of technical standards, or troubleshooting procedures, where the accuracy and traceability of retrieval results directly impact the reliability of technical judgments. Consequently, research into high-precision, explainable, and trustworthy interactive question-answering methods tailored for specialized power-industry knowledge holds significant value for the field of knowledge engineering. Furthermore, such research provides essential technical support for the digital transformation of knowledge services within the power sector, as well as for professional training and decision-making assistance.

In recent years, large language models (LLMs) have demonstrated robust capabilities in tasks such as natural language understanding, text generation, knowledge reasoning, and multi-turn interaction, offering a new technical pathway for developing a next generation of intelligent knowledge service systems for the power sector. Majumder et al. [1] provide a systematic discussion on the opportunities and limitations of applying LLMs within the power and energy domain, highlighting that domain-specific data, specialized tools, and external knowledge augmentation are crucial for realizing reliable, intelligent energy applications. Ruan et al. [2] further analyze the potential security risks introduced by applying LLMs to power systems, emphasizing the necessity for reliability verification and risk control. Although general-purpose LLMs can generate coherent responses based on user input, their specialized knowledge is primarily derived from parametric memory acquired during the pre-training phase; consequently, they may fail to encompass the full spectrum of power industry standards, equipment specifications, and continuously evolving operational knowledge. Furthermore, when addressing technical inquiries, these models are prone to "hallucinations"—generating factually incorrect information, miscitations, or technical recommendations that do not align with actual operating conditions. This is particularly critical when dealing with voltage levels, equipment ratings, protection protocols, or specific versions of technical standards; in such contexts, a linguistically coherent response does not guarantee the accuracy of the technical conclusion. Therefore, effectively integrating the semantic understanding capabilities of LLMs with authoritative, external power-domain knowledge has emerged as a key research direction for developing reliable, intelligent question-answering systems for the power sector.

Retrieval-Augmented Generation (RAG) offers a representative technical solution for mitigating issues such as outdated model knowledge, a lack of domain-specific expertise, and factual hallucinations by integrating external knowledge retrieval into the language model generation process. The RAG framework, proposed by Lewis et al. [3], combines a parametric language model with a non-parametric knowledge store, enabling the generative model to leverage externally retrieved evidence to perform knowledge-intensive tasks. Compared to generating answers by relying solely on model parameters, RAG allows for the incorporation of domain-specific knowledge without the need to retrain the underlying language model, while also providing traceable sources for the generated content. However, traditional RAG approaches typically employ fixed-size text chunking, single-vector retrieval, and straightforward context concatenation. Since the retrieval process relies primarily on semantic similarity between the query and text segments, it often fails to adequately account for the entity relationships, hierarchical structure, and technical applicability conditions inherent in specialized power-domain knowledge. When a user's query involves multiple equipment entities, complex causal relationships, or information spanning across multiple documents, the retrieval system may return content that is semantically similar but contextually mismatched in terms of technical requirements—thereby compromising the factual consistency of the generated answers. Consequently, simply increasing vector similarity thresholds or expanding the number of retrieved segments cannot fundamentally resolve the challenges of knowledge organization and evidence selection in complex power-domain question-answering scenarios.

To further enhance the ability of RAG systems to comprehend structured knowledge and complex relationships, researchers have begun incorporating knowledge graphs, hierarchical document representations, and adaptive retrieval strategies into the knowledge-augmented generation process. Edge et al. [4] proposed GraphRAG, which improves the system's capacity to analyze global-level questions across large-scale document collections by constructing entity-relationship graphs and generating community-level knowledge summaries. Guo et al. [5] introduced LightRAG, which combines graph-structured indexing with a two-tier knowledge retrieval mechanism to enhance both the efficiency of retrieving complex semantic relationships and the capability for knowledge updates. Meanwhile, Asai et al. [6] proposed Self-RAG, which bolsters the factual reliability of language models through a synergistic mechanism involving retrieval, generation, and self-reflection; similarly, Yan et al. [7] introduced Corrective RAG, which employs retrieval quality assessment and error-correction strategies to mitigate the impact of erroneous evidence on answer generation. These studies demonstrate that knowledge graph augmentation, dynamic retrieval control, and generative feedback can improve the performance of traditional RAG systems in tasks involving complex knowledge. However, general-purpose RAG methods, typically designed for open-domain question answering or broad document comprehension, often employ retrieval strategies and knowledge verification mechanisms that may not satisfy the rigorous, domain-specific constraints of the power industry. Knowledge in the power sector is characterized not only by dependencies between entities and relationships but also by a complex array of conditions, including equipment models, voltage levels, operating environments, numerical units, versions of regulations, and the specific scope of technical provisions. Without explicit modeling of these conditions, retrieved information—even if semantically relevant—may fail to provide valid evidence for addressing specific operational issues within the power sector.

Given the specific characteristics of knowledge in the power

sector, research has begun to explore domain-specific RAG frameworks. Wang et al. [8] proposed Meta-RAG, which enhances question-answering capabilities for power-related regulatory documents through metadata extraction, hybrid retrieval, and domain-model adaptation; they also constructed the EleQA dataset for the power domain. Shen et al. [9] introduced HG-RAG, which incorporates a hierarchical graph structure into the retrieval of power documents, specifically addressing complex entity relationships within long documents, cross-document associations, and knowledge provenance. Furthermore, Shen et al. [10] proposed a multi-path adaptive retrieval framework guided by a lightweight language model. This approach integrates power-term normalization, section-structure awareness, semantic retrieval, term anchoring, and structural priors, while dynamically adjusting fusion weights based on retrieval confidence and path consistency. Meanwhile, TSAC-RAG [11] establishes a constraint-aware retrieval mechanism—accounting for clause types, numerical units, section hierarchies, and applicability conditions within power technical standards to improve the accuracy of matching normative knowledge evidence. Furthermore, PQRAG [12] integrates query-adaptive graph retrieval with traceable chains of evidence, offering a novel approach for linking standards, equipment constraints, and mitigation measures within the context of power quality knowledge reasoning. These efforts demonstrate that the alignment of technical terminology, the organization of structured knowledge, graph-based relationship retrieval, and dynamic evidence fusion have emerged as key technical directions for RAG research in the power sector, providing a robust foundation for developing complex knowledge-based question-answering systems for the power industry.

Despite significant progress in existing research, intelligent question-answering (QA) for the power sector—involving multi-source, heterogeneous knowledge and continuous interactive scenarios—still faces several unresolved challenges. First, it remains difficult to simultaneously ensure semantic integrity and structural consistency during the knowledge construction process. Specialized power-sector knowledge often spans multiple document segments and technical clauses; traditional text segmentation tends to disrupt the associations between fault phenomena, equipment attributes, and operational conditions. Furthermore, single-pass knowledge extraction using LLMs can lead to issues such as entity duplication, omitted relationships, and conflicting attributes. Even when knowledge graphs are employed to organize entity relationships, the absence of robust ontological constraints, precise node localization, and links to original evidentiary sources makes it challenging to guarantee the accuracy and traceability of knowledge integration across multiple documents. Furthermore, multi-granularity knowledge matching and the coordination of specialized constraints within complex queries remain challenging. While semantic vector retrieval can capture similarities in natural language expressions, it may overlook critical device parameters; conversely, keyword-based retrieval can match technical terminology but struggles to interpret abstract business intent. Graph-based retrieval can establish knowledge associations, yet its effectiveness hinges on the quality of the knowledge graph construction and the accuracy of entity localization. Although existing research has proposed multi-path retrieval and dynamic fusion mechanisms, further systematic study is required to determine how to effectively integrate continuously updated dialogue states, specialized constraints, and verifiable evidence into the retrieval decision-making process.

To address the aforementioned challenges, this paper proposes an intelligent QA framework for power-domain knowledge bases. The framework integrates multi-agent collaborative knowledge construction, context-aware multi-turn interaction, and adaptive RAG, with a focus on the structured organization of domain knowledge, the propagation of contextual constraints, and the utilization of verifiable evidence. Grounded in external professional knowledge, the framework employs dynamic query understanding as a bridge and multi-path evidence retrieval as the core mechanism for knowledge acquisition; it further incorporates fact-checking of generated answers to facilitate continuous, multi-turn interactions. Unlike existing approaches that optimize for a single retrieval strategy or specific document type, this framework emphasizes the interplay between the quality of knowledge construction, the state of historical dialogue, and the validity of current evidence. This enables the system to dynamically organize knowledge based on evolving professional criteria within user queries, while ensuring that verifiable source evidence is preserved throughout the answer-generation process.

II. Proposed Methodology

To address the limitations of traditional power-domain knowledge base QA systems, such as challenges in integrating multi-source heterogeneous knowledge, interpreting specialized terminology, maintaining contextual coherence, and ensuring the reliability of retrieval results, this paper proposes an interactive, intelligent QA framework for power-domain knowledge bases [11]. This framework integrates LLMs, multi-agent knowledge extraction, context-augmented memory, and adaptive multi-path retrieval. By constructing a hybrid knowledge base that synergistically represents both structured knowledge and unstructured semantic information, the framework enables the efficient organization and dynamic retrieval of specialized power-domain knowledge. Furthermore, it incorporates dialogue state tracking and query rewriting mechanisms to enhance the model's capacity to interpret sequential inquiries and implicit semantics. During the knowledge acquisition phase, a multi-path retrieval approach—incorporating semantic coverage, specialized term anchoring, and document structure priors—is employed, with a lightweight language model utilized to adaptively adjust retrieval weights [10]. Finally, by synthesizing retrieved evidence with prompt-based constraints, the framework generates answers that include traceable knowledge sources.

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

A. LLM-Driven Power Knowledge Question Answering Framework

Figure 1 illustrates the overall architecture of the interactive intelligent question-answering system for the power domain proposed in this paper. The framework employs RAG technology, integrating the natural language understanding and generation capabilities of large language models with an external, domain-specific knowledge base for the power industry. It establishes a comprehensive technical workflow encompassing data preprocessing, knowledge vectorization, semantic retrieval, knowledge augmentation, and answer generation. Unlike traditional question-answering models that rely solely on the internal parametric knowledge of a language model, this system utilizes an externally maintained, updatable knowledge store. This provides a factual basis, grounded in specific power-related documents for professional inquiries, thereby mitigating risks such as knowledge obsolescence and the generation of unsubstantiated information.

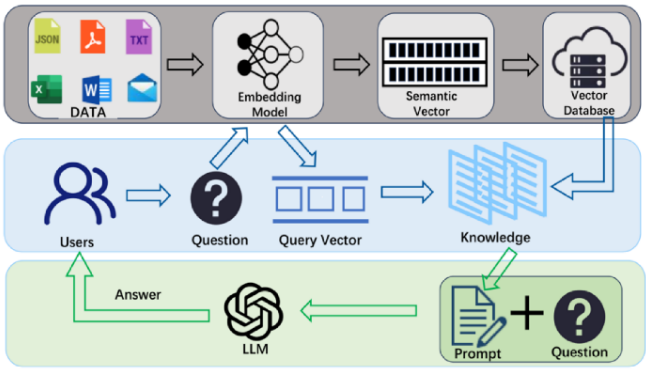


Figure 1: Overall process

During the knowledge acquisition phase, the system supports input in various file formats—including JSON, PDF, TXT, Word, and Excel—encompassing heterogeneous data sources such as technical manuals for power equipment, operation and maintenance protocols, fault diagnostic records, technical standards, and historical operational reports. For each file type, the system employs specialized parsing methods to extract text, tabular structures, section information, and domain-specific metadata. Subsequently, a structure-aware chunking strategy is applied to segment long documents into knowledge fragments that maintain semantic integrity [10]. Unlike fixed-length text segmentation, this approach prioritizes the preservation of clause boundaries, the contextual parameters of equipment, and the logical connections between fault causes and their corresponding remedial measures, all of which are critical in technical documentation for the power sector.

Let the original collection of power-related documents be $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, where N is the total number of documents. Following format parsing and structural segmentation, a collection of knowledge fragments $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ is generated. Each knowledge fragment is associated with metadata m_j, such as equipment type, voltage level, document source, and applicable version—and its vector representation is defined as follows:

$$\mathbf{z}_j = \frac{f_{\text{emb}}(c_j \oplus \rho(m_j); \theta_e)}{\|f_{\text{emb}}(c_j \oplus \rho(m_j); \theta_e)\|_2 + \varepsilon}, \tag{1}$$

where $f_{\text{emb}}(\cdot; \theta_e)$ denotes a text embedding model parameterized by θ_e, $\rho(\cdot)$ represents a metadata encoding function, $\oplus$ signifies the operation for combining text and metadata representations, and ε is a stability constant included to prevent division by zero. This approach incorporates domain-specific attributes from the power industry into the semantic representations, ensuring that knowledge fragments not only encompass natural language semantic information but also retain auxiliary features essential for subsequent domain-specific constraint matching. For elements requiring strict matching—such as equipment models, numerical values, and standard versions—the original metadata is preserved independently, thereby avoiding potential mismatches that could arise from relying solely on approximate vector-based retrieval.

Subsequently, the knowledge fragment vectors are stored in a vector database, while maintaining their index relationships with the original documents and structured knowledge nodes. When a user submits a query q_t, the system generates a representation of the query using the same embedding model and retrieves initial candidate knowledge based on semantic similarity within the vector space.

$$s_{\text{sem}}(q_t, c_j) = \frac{f_{\text{emb}}(q_t)^T \mathbf{z}_j}{\|f_{\text{emb}}(q_t)\|_2 + \varepsilon}, \tag{2}$$

$$\mathcal{E}_t^{(0)} = \underset{c_j \in \mathcal{C}}{\text{TopK}} \left[s_{\text{sem}}(q_t, c_j) \right], \tag{3}$$

where $\mathcal{E}_t^{(0)}$ denotes the set of candidate evidence retrieved via initial semantic search, and K represents the number of candidate knowledge items. Vector-based retrieval provides the foundational semantic recall capability for the entire framework, while the final retrieval results will be obtained in the following section through multi-path fusion and re-ranking based on specialized constraints.

During the answer generation phase, the system organizes the retrieved and filtered power-domain knowledge snippets, the current query, and the historical dialogue context into a structured prompt, which is then fed into a large language model. Let $\widetilde{q}_t$ denote the context-augmented query, M_t the current dialogue state, and $\mathcal{E}_t$ the final set of retrieved evidence; the answer generation process can then be expressed as:

$$P_\phi(a_t \mid \widetilde{q}_t, M_t, \mathcal{E}_t) = \prod_{\ell=1}^{|a_t|} P_\phi\big(a_{t,\ell} \mid a_{t,<\ell}, \text{Prompt}(\widetilde{q}_t, M_t, \mathcal{E}_t)\big), \tag{4}$$

where a_t represents the answer generated by the system, ϕ denotes the parameters of the large language model, and $a_{t,<\ell}$ signifies the tokens generated prior to the current position. By

incorporating retrieval results into the generation context, the model can leverage external, domain-specific knowledge—in this case, regarding the power sector—to address user queries, rather than relying solely on the implicit knowledge acquired during the pre-training phase.

B. Multi-Agent Collaborative Power Knowledge Base Construction

Figure 2 further illustrates the multi-agent collaborative approach for constructing a power-domain knowledge base, as proposed in this paper. Given that knowledge in the power sector typically exists in the form of unstructured technical documents and semi-structured professional records, conventional methods—such as standard text chunking and vector-based storage, often fail to fully capture the complex interrelationships between equipment entities, operational conditions, fault phenomena, and technical specifications. To address this, we have designed a knowledge construction framework consisting of four stages: data preprocessing, multi-agent knowledge extraction, entity localization and attribute fusion, and hybrid knowledge storage. This framework transforms unstructured power-related text into a structured knowledge network, enriched with semantic relationships, domain-specific constraints, and traceability to the original source material [13], [14].

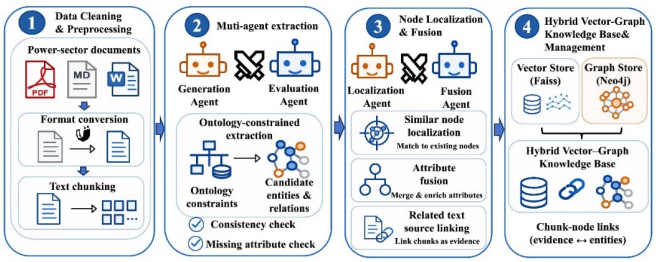


Figure 2: Multi-agent collaborative construction of the hybrid vector–graph knowledge base

In the first phase, the system converts various technical power-industry documents—such as PDFs, Markdown files, and Word documents—into a unified format and encoding. It then performs character normalization, filters out redundant content, parses tables, and identifies document sections. For materials like equipment maintenance procedures, operational manuals, and fault analysis reports, the system employs a structure-aware text segmentation method that preserves document titles, section numbering, and contextual information.

Within power-industry documentation, a single technical parameter may correspond to different requirements depending on factors such as rated voltage, equipment model, or operating environment. Consequently, each text segment must include not only the core content but also metadata—specifically the section, source file, applicable conditions, and document version. This metadata serves as a foundation for subsequent knowledge extraction and is essential for knowledge retrieval and the verification of the sources from which answers are derived.

In the second phase, this paper introduces a "Generation Agent" and an "Evaluation Agent" to construct a multi-agent knowledge extraction mechanism consisting of generation, verification, and refinement.

The Generation Agent is responsible for identifying domain-specific entities, attributes, and semantic relationships within the power sector, mapping elements such as equipment names, fault types, operational parameters, and remedial measures into candidate knowledge triples. The Evaluation Agent, guided by predefined domain ontologies for the power industry, performs consistency checks, validates the legitimacy of relationships, and identifies missing attributes within these candidate results. In cases where candidate results exhibit semantic ambiguity or incomplete attributes, the Evaluation Agent provides refinement suggestions, prompting the Generation Agent to re-extract or supplement the relevant knowledge [14].

Let Ω denote the ontology for the power sector, and let $\widehat{\mathcal{T}}$ be the set of candidate knowledge triples, where each candidate triple is represented as $z = (h, r, t)$, corresponding to the head entity, relation, and tail entity, respectively. The objective of multi-agent collaborative extraction is defined as follows:

$$\begin{aligned}\mathcal{T}^* = \arg\max_{\mathcal{T}\subseteq\widehat{\mathcal{T}}} \Big\{ & \sum_{z\in\mathcal{T}} \big[\log P_{\text{gen}}(z \mid c, \Omega) \\ & + \lambda_e \log P_{\text{eval}}(z \mid c, \Omega)\big] \\ & - \lambda_c \sum_{z_i, z_j \in \mathcal{T}} \text{Viol}_\Omega(z_i, z_j) \\ & - \lambda_m \,\text{Missing}(\mathcal{T}) \Big\}, \end{aligned} \tag{5}$$

where P_{gen} represents the credibility of a knowledge candidate provided by the generative agent, P_{eval} denotes the evaluative agent's assessment of the candidate's validity, Viol_Ω signifies the degree of violation of ontological constraints, and Missing refers to essential knowledge attributes that should have been extracted but remain uncovered. The parameters λ_e, λ_c, and λ_m serve to weight the contributions of these respective objectives. In practical implementation, this optimization can be approximated through methods such as scoring candidate triples, applying constraint-based filtering, and employing a finite number of refinement iterations by the agents.

In this way, the system can not only extract fundamental entity relationships, such as "transformer-contains-winding" but also represent specialized semantic relationships, such as "voltage sag, may affect, sensitive equipment" and "protective measures—applicable to—specific operating conditions." Furthermore, it employs evidence from the original documents to verify these relationships, thereby preventing the generation of unsubstantiated connections.

The third stage primarily addresses issues such as redundant representations of the same entity across different documents, the matching of knowledge nodes, and the fusion of attributes from diverse sources. In this paper, we design a "Localization

Agent" and a "Fusion Agent" to facilitate the automated identification of existing entities within the knowledge graph and the incremental integration of new knowledge [5].

The node-localization agent determines whether a newly extracted entity can be mapped to an existing knowledge node, based on the entity's name, semantic features, professional attributes, and local relational structure. Let x be the newly extracted entity, and let $\mathcal{V}$ be the set of nodes in the existing knowledge graph; the optimal matching target for x is represented as follows:

$$v^* = \arg\max_{v \in \mathcal{V}} \big[\alpha\, \mathrm{Sim}_{\mathrm{sem}}(x, v) + \beta\, \mathrm{Sim}_{\mathrm{attr}}(x, v) + \gamma\, \mathrm{Sim}_{\mathrm{rel}}(x, v) - \delta\, \mathrm{Conflict}(x, v)\big]. \tag{6}$$

Among them, $\mathrm{Sim}_{\mathrm{sem}}$ represents entity semantic similarity, $\mathrm{Sim}_{\mathrm{attr}}$ represents device attribute similarity, $\mathrm{Sim}_{\mathrm{rel}}$ represents entity neighborhood relationship consistency, and $\mathrm{Conflict}$ represents the degree of entity attribute or professional constraint conflict. Only when the matching result reaches the preset credibility threshold and there are no unacceptable hard attribute conflicts will the system map the new entity to the existing node. Otherwise, a new knowledge node will be created or submitted for manual verification.

The fusion agent further integrates entity attributes and relational evidence derived from various text segments. To balance the stability of existing knowledge with the utility of newly acquired information, this paper proposes a gated knowledge representation fusion mechanism:

$$\begin{aligned} \mathbf{z}_v &= \sum_{c_j \in \mathcal{L}(v)} \alpha_{vj} \mathbf{z}_j, \\ \mathbf{g}_v &= \sigma(W_g[\mathbf{h}_v; \mathbf{z}_v; \mathbf{o}_v] + \mathbf{b}_g), \\ \mathbf{h}_v^+ &= \mathbf{g}_v \odot \mathbf{h}_v + (1 - \mathbf{g}_v) \odot \mathbf{z}_v, \end{aligned} \tag{7}$$

where $\mathcal{L}(v)$ denotes the set of textual evidence snippets associated with node v, α_{vj} represents the normalized evidence weight, $\mathbf{h}_v$ is the existing node representation, $\mathbf{o}_v$ is the encoded ontological attribute, $\mathbf{g}_v$ is a gating vector, and $\odot$ denotes element-wise multiplication. This mechanism enables the system to adjust knowledge fusion weights based on the degree of consistency among existing knowledge, source evidence, and ontological attributes. It should be noted that this vector-based fusion is intended to refine knowledge representations; it cannot serve as a substitute for the explicit recording and verification of raw numerical values, versioning, or conflicting attributes.

C. Context-Aware Multi-Turn Dialogue and Query Reformulation

Figure 3 illustrates the context-aware, multi-turn interactive question-answering mechanism proposed in this paper. Traditional power-domain question-answering systems typically employ an independent query-processing approach, treating each user input as a discrete, isolated question. In practice, however, during technical consultations or inquiries related to operation and maintenance, users often engage in a continuous dialogue, progressively describing equipment anomalies, clarifying operational conditions, or refining their search criteria. Subsequent questions may involve anaphora (references to previous terms), ellipses (omitted information), and implicit semantics; consequently, retrieval mechanisms that rely solely on the current input often struggle to accurately discern the user's true intent.

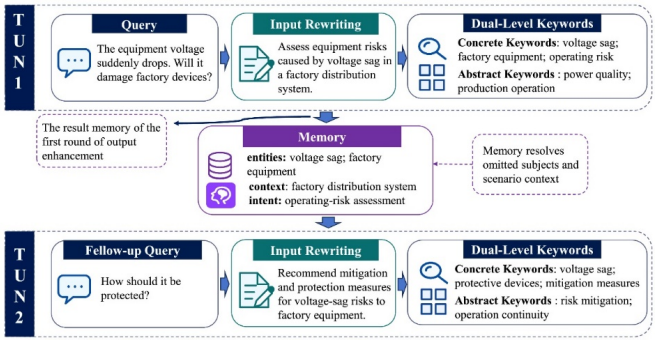


Figure 3: Context-aware multi-turn dialogue modeling and dual-level keyword-enhanced query reformulation

First, we construct a Dialogue Memory Module tailored for the power sector, designed to record key entities, technical terminology, operational scenarios, user intents, and relevant constraints throughout multi-turn interactions. Unlike approaches that simply concatenate the entire history of dialogue text, this method transforms the dialogue history into a structured memory representation, thereby reducing the interference of irrelevant historical information on the current query [15].

Let q_t denote the user input at turn t, M_{t-1} the memory state from the previous turn, and a_{t-1} and $\mathcal{E}_{t-1}$ the verified system response and its associated evidence from the previous turn, respectively. The current dialogue state can then be represented as:

$$M_t = \mathrm{Update}_{\psi}\big(M_{t-1}, \mathrm{Parse}(q_t), \mathrm{Verify}(a_{t-1}, \mathcal{E}_{t-1})\big), \tag{8}$$

$$M_t = \{\mathcal{V}_t^d, \mathcal{C}_t^d, \mathcal{I}_t^d, \mathcal{P}_t^d\}, \tag{9}$$

where $\mathcal{V}_t^d$ denotes the set of power-related entities involved in the dialogue, $\mathcal{C}_t^d$ represents the currently applicable device attributes and operating conditions, $\mathcal{I}_t^d$ signifies the user's query intent, and $\mathcal{P}_t^d$ encompasses verified knowledge along with its source information. Update_{ψ} serves as the dialogue memory update function, with ψ representing its parameters. By storing entities, constraints, and intents separately, the system can selectively inherit historical information during subsequent interactions, avoiding the need to reprocess the entire dialogue content. Generated content that has not been validated by evidence should not be directly incorporated into the factual memory.

As shown in Figure 3, in the first round of the dialogue, the user asks, "The equipment voltage suddenly drops. Will it damage factory devices?" The system first identifies key

entities and concepts—such as "voltage sag" and "factory equipment"—and categorizes the inquiry as an assessment of risks to electrical operations. This establishes an initial memory state encompassing voltage sags, factory power distribution systems, and operational risks to equipment [16].

When the user follows up in the second round by asking, "How should it be protected?" the system is able to leverage the state from the previous round to identify the implied object of protection and the relevant context. This prevents the query from being misconstrued as a generic question about protection that lacks a specific equipment-related context.

To enhance the semantic integrity of context-dependent queries, this paper introduces a Query Reformulation Module. By integrating the current query with historical memory, this module generates a comprehensive query that incorporates explicit entities and operational scenarios.

Query reformulation requires not only filling in omitted information but also ensuring that the historical context does not introduce conditions inconsistent with the current query. For instance, when a user explicitly modifies parameters such as voltage level, equipment model, or fault type, the system must prioritize the most recent, valid conditions while discarding the corresponding outdated constraints.

Let the candidate rewritten query be $\widehat{q}_t$; then, the optimal query reformulation process can be modeled as:

$$\begin{aligned}\widetilde{q}_t = \arg\max_{\widehat{q}_t} \big\{ & \log P_\psi(\widehat{q}_t \mid q_t, M_t) \\ & + \lambda_s \,\mathrm{SemPreserve}(\widehat{q}_t, q_t) \\ & - \lambda_c \,\mathrm{ConstraintLoss}(\widehat{q}_t, M_t) \big\},\end{aligned} \tag{10}$$

where P_ψ denotes the conditional probability assigned by the query-rewriting model to a candidate expression; $\mathrm{SemPreserve}$ measures the semantic consistency between the rewritten query and the user's original intent; and $\mathrm{ConstraintLoss}$ quantifies the degree of conflict between the rewritten result and the current valid conversational constraints. The parameters λ_s and λ_c control the relative importance of semantic preservation and constraint consistency.

For the second round of questions in Figure 3, the system can rewrite "How should it be protected?" as "How can protective measures mitigate voltage-sag risks for equipment in a factory distribution system?" This rewritten version incorporates contextual elements—specifically voltage sags, protective measures, and factory distribution systems—thereby enhancing the precision and relevance of subsequent retrieval.

It is important to emphasize that the purpose of query rewriting is not to alter the user's original intent, but rather to supplement it with necessary background information while preserving the core semantic meaning. In particular, one must avoid making unsubstantiated assumptions—such as presuming that a piece of equipment has already failed or that a specific protective measure is guaranteed to be effective.

Given that queries in the power sector involve both specific equipment terminology and abstract operational semantics, this paper proposes a dual-level keyword extraction mechanism. This mechanism constructs separate sets of concrete keywords and abstract keywords, thereby enabling a complementary approach that bridges fine-grained entity retrieval with high-level semantic matching.

Concrete keywords primarily encompass equipment names, electrical parameters, fault phenomena, protective devices, and operating conditions; abstract keywords, by contrast, describe concepts such as power quality, operational risks, fault diagnosis, risk mitigation, and system reliability.

Let $\mathcal{K}_t^c$ and $\mathcal{K}_t^a$ denote the sets of concrete and abstract keywords, respectively; these are then jointly encoded to construct an augmented query representation.

$$(\mathcal{K}_t^c, \mathcal{K}_t^a) = \mathrm{DualExtract}(\widetilde{q}_t, M_t, \Omega), \tag{11}$$

$$\begin{aligned}\mathbf{h}_t^q = W_q \big[& f_{\mathrm{emb}}(\widetilde{q}_t); \mathrm{Pool}_{k\in\mathcal{K}_t^c}(\mathbf{e}_k); \\ & \mathrm{Pool}_{k\in\mathcal{K}_t^a}(\mathbf{e}_k) \big] + \mathbf{b}_q,\end{aligned} \tag{12}$$

where $\mathrm{DualExtract}$ denotes a dual-layer keyword extraction function that incorporates domain ontologies;$\mathbf{e}_k$ represents the keyword embedding vector; Pool signifies a set-level semantic aggregation operation; and W_q and $\mathbf{b}_q$ are the transformation parameters for query representation.

In the example shown in Figure 3, the first round of questioning allows for the extraction of specific keywords—such as "voltage sag" and "factory equipment"—as well as abstract keywords like "power quality" and "operating risk." Following semantic reconstruction, the second round of questioning enables the identification of specific keywords such as "protective devices" and "mitigation measures," while also incorporating high-level concepts like "risk mitigation" and "operation continuity."

This dual-layer semantic representation enables the system to perform retrieval across varying levels of knowledge granularity: on one hand, it pinpoints specific equipment and technical terms via specialized entities and terminology; on the other, it leverages abstract concepts to broaden the scope to include relevant risk analysis, preventive strategies, and equipment protection knowledge.

D. Adaptive Multi-Route Retrieval and Evidence-Grounded Answer Generation

Figure 4 illustrates the adaptive multi-path intelligent retrieval framework proposed in this paper. Traditional RAG systems primarily rely on vector similarity between queries and documents to retrieve candidate knowledge. While this approach can accommodate a certain degree of semantic variation, it often falls short in the power sector, where challenges include specialized terminology, hierarchical technical specifications, and complex operational conditions—frequently resulting in incomplete knowledge matching or the retrieval of evidence with limited applicability.

To address these limitations, this paper proposes a multi-path retrieval method that integrates semantic coverage, specialized terminology anchoring, and document structural priors. A Small Language Model (SLM) is employed to dynamically determine the contribution weights for each retrieval path. Furthermore, a consensus feedback mechanism—based

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

on retrieval confidence, cross-path consistency, and the coverage of specialized operational conditions—is introduced to facilitate the adaptive ranking and filtering of candidate knowledge.

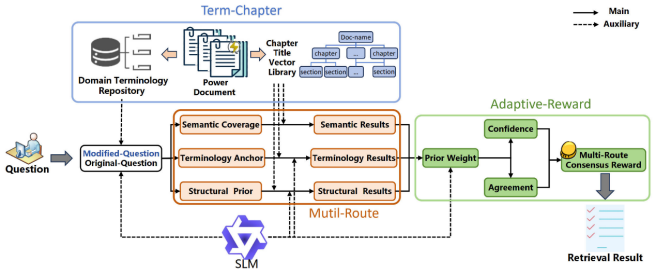


Figure 4: SLM-guided adaptive multi-route retrieval

To enhance the coverage and precision of knowledge retrieval in the power domain, this paper proposes a candidate knowledge recall mechanism consisting of three complementary pathways.

The first pathway, Semantic Coverage Retrieval, employs embedding models to calculate the semantic similarity between an augmented query and knowledge fragments; it is primarily designed to capture variations in natural language expression and latent semantic connections between specialized concepts.

The second pathway, Terminology-Anchored Retrieval, identifies technical terms, entity names, and their corresponding ontological concepts within a query to precisely locate key technical entities—thereby preventing general semantic similarity from obscuring critical details regarding equipment conditions.

The third pathway, Structural-Prior Retrieval, leverages document titles, hierarchical section structures, and knowledge graph relationships to identify the specific knowledge topics and technical provisions relevant to a given query.

Let the three retrieval paths be denoted as s, τ, and h; their candidate knowledge scoring functions are defined as follows:

$$s_k(\widetilde{q}_t, d) = \begin{cases} \mathrm{CosSim}(\mathbf{h}_t^q, \mathbf{z}_d), & k = s, \\ \mathrm{BM25}(\mathcal{K}_t^c \cup \mathcal{K}_t^a, d) \\ \quad + \lambda_o \, \mathrm{OntoAlign}(\widetilde{q}_t, d), & k = \tau, \\ \mathrm{ChapterMatch}(\widetilde{q}_t, d) \\ \quad + \lambda_g \, \mathrm{GraphProx}(\widetilde{q}_t, d), & k = h, \end{cases} \tag{13}$$

where $\mathbf{z}_d$ denotes the vector representation of a document evidence segment d; $\mathrm{OntoAlign}$ represents the degree of ontological term matching between the query and candidate knowledge; $\mathrm{ChapterMatch}$ indicates the relevance of the document's chapter structure to the query topic; and $\mathrm{GraphProx}$ signifies the relational proximity within the knowledge graph. The parameters λ_o and λ_g control the respective contributions of ontological constraints and graph-structural information.

Each of the three paths generates a set of candidate knowledge items; these are subsequently merged based on knowledge segment identifiers and provenance links. For results retrieved via knowledge graph nodes, the system first fetches the associated original document segments, thereby unifying the outputs from different retrieval paths into consistent evidence units suitable for answer generation.

To address the limitations of fixed-weight retrieval strategies—which struggle to adapt to diverse types of queries—this paper introduces an SLM-guided dynamic weight prediction mechanism. This mechanism dynamically estimates the suitability of various retrieval paths for a given task by analyzing the semantic features of the user's query, the distribution of technical terminology, the conversational context, and the quality of candidate retrieval results.

For instance, when a query centers on specific device models, regulatory standards, or precise electrical parameters, term-anchored and structure-based retrieval methods may prove highly effective; conversely, for vaguely described fault phenomena or open-ended technical questions, semantic retrieval may yield a broader range of relevant knowledge candidates.

Let the current query representation be denoted by $\mathbf{h}_t^q$, the dialogue memory by $\mathbf{m}_t$, and the statistical features of the retrieval path by $\mathbf{u}_t$; then, the initial weights for multi-path retrieval are defined as:

$$\mathbf{r}_t = g_\theta([\mathbf{h}_t^q; \mathbf{m}_t; \mathbf{u}_t]), \tag{14}$$

$$w_{t,k} = \frac{\exp(r_{t,k}/T)}{\sum_{j \in \{s,\tau,h\}} \exp(r_{t,j}/T)}, \tag{15}$$

whereg_θ denotes the SLM-driven weight prediction module, $\mathbf{r}_t$ represents the predicted weights for various retrieval paths, T is the Softmax temperature parameter, and $w_{t,k}$ is the initial weight for the k-th retrieval path, subject to the constraint $\sum_k w_{t,k} = 1$.

To ensure that the dynamic weights accurately reflect the true contribution of each retrieval path, the training phase leverages query samples with annotated evidence to construct a ranking-based supervisory signal; the weight prediction module is then optimized by comparing the quality of evidence retrieved via different paths. During the inference phase, the trained module is used to predict weights, without requiring the ground-truth answers to the test questions.

III. Experiments and Results

A. Experimental Datasets and Data Preparation

1) Dataset Construction

To ensure that the experiments encompass various types of knowledge within the power sector, we adopt a hybrid approach that combines publicly available power-related question-answering datasets with a custom-built multi-turn interactive test set.

As summarized in Table 1, the public datasets are primarily drawn from EleQA and PowerGridQA. EleQA, which derives its content from technical standards and industry regulations, is well-suited for evaluating a model's grasp of specialized

terminology, its ability to retrieve specific clauses from standards, and its capacity to align requirements with regulatory constraints. PowerGridQA [17] covers theoretical knowledge of power systems, operational incidents, and reasoning scenarios, making it effective for assessing a model's understanding of equipment operation, fault analysis, and technical concepts.

Given that existing public question-answering datasets may not adequately cover issues such as cross-turn condition inheritance and knowledge conflicts, we propose to construct two additional test sets: Power-MultiTurnQA and Power-ConflictQA. The former evaluates interactive question-answering capabilities through techniques like sequential questioning, entity coreference, contextual ellipsis, and modifications to operational conditions. The latter assesses system reliability by introducing scenarios such as inconsistencies between different versions, conflicts in device attributes, interference from irrelevant documents, and missing evidence.

Table 1: Description of the datasets used for power-domain question answering evaluation.

Dataset	Knowledge domain	Dataset size	Main evaluation purpose
EleQA	Power standards and regulations	19,560 QA pairs	Domain knowledge retrieval
PowerGridQA	Power system dynamics and operations	28,852 QA pairs	Knowledge reasoning and answer accuracy
Power-MultiTurnQA	Context-dependent power-domain dialogue	125 dialogues, planned	Multi-turn reasoning and context consistency
Power-ConflictQA	Knowledge conflicts and insufficient evidence	600 samples, planned	Robustness and factual verification

In practical experiments, data should be deduplicated based on the original document source and knowledge entities. A document-level partitioning strategy should be employed to separate the training, validation, and test sets, ensuring that highly similar content—such as that found within the same technical standard or equipment manual—does not appear across different data splits. For technical specifications that include versioning over time, a separate test set should be constructed based on the effective dates of the knowledge, allowing for the evaluation of the model's retrieval applicability when faced with version updates.

2) Evaluation Task Design

Table 2 summarizes the planned query allocation. Based on the semantic characteristics of inquiries in the power sector, the primary evaluation tasks are categorized into four types: fact-based Q&A, regulation and rule-based Q&A, equipment fault-related Q&A, and multi-turn interactive Q&A.

Table 2: Planned evaluation tasks for the power-domain QA benchmark.

Task	Description	Evaluation queries
Fact QA	Equipment parameters, terminology and definitions	500
Rule QA	Technical standards, operating rules and constraints	500
Fault QA	Fault causes, symptoms and maintenance knowledge	500
Multi-turn QA	Context inheritance, follow-up queries and condition changes	500
Total		2,000

For all types of test questions, it is necessary to specify reference answers, relevant evidentiary excerpts, and applicable domain-specific constraints. In the case of multi-turn dialogue questions, it is also required to document historical entities, changes in conditions, and information to be carried over to the current turn.

During the expert annotation phase, it is recommended that two evaluators with expertise in the power sector independently review the question-answer samples, with a subsequent re-evaluation process for any instances where their annotations diverge. For categorical evaluation metrics, Cohen's Kappa may be used to measure inter-annotator agreement, thereby minimizing the influence of subjective factors on the experimental results.

B. Baseline Methods

To comprehensively verify the effectiveness of the proposed method, representative traditional retrieval, knowledge graph-enhanced, and retrieval feedback methods are selected as baselines [18].

The primary methods used for comparison include:

1) LLM-only: Generate answers directly using the base language model, without employing external knowledge retrieval.
2) BM25-RAG: Use BM25 for text retrieval, and take the retrieved results as input for the model.
3) Dense-RAG: Perform knowledge retrieval based on text embedding vectors and semantic similarity.
4) Hybrid-RAG: Fuse BM25 and vector retrieval results using fixed weights.
5) GraphRAG: Use a graph structure to organize entity relationships and associated knowledge.
6) LightRAG: Employs a graph-structure augmentation and two-layer knowledge retrieval strategy.
7) Self-RAG: Optimize response quality through adaptive retrieval and a generation-reflection mechanism.
8) CRAG: Perform knowledge filtering and search error correction based on search quality feedback.
9) Proposed: A framework for multi-agent knowledge construction, context awareness, and adaptive multi-path retrieval fusion, as proposed in this paper.

C. Overall Performance Comparison

Based on the illustrative simulation data in Table 3, the proposed method exhibits higher illustrative scores across all evaluation metrics. In terms of retrieval recall, the proposed method achieves 89.0%—an improvement of 15.0 and 5.0 percentage points over Dense-RAG and GraphRAG, respectively—highlighting the potential advantages of combining adaptive multi-path retrieval with specialized terminology matching, document structure priors, and graph-based retrieval.

In terms of answer accuracy, the proposed method achieves a score of 92.2%, representing improvements of 9.0 and 3.7 percentage points over Dense-RAG and GraphRAG, respectively. This performance aligns with the design objectives of the hybrid knowledge organization and multi-stage evidence-filtering mechanisms. By incorporating a power-industry-specific ontology and equipment-attribute constraints

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

into the retrieval process, the system is able to more precisely filter knowledge segments relevant to a user's query. Regarding factual credibility, the proposed method attains an evidence-faithfulness score of 90.9%—outperforming Dense-RAG and GraphRAG by 11.4 and 5.4 percentage points, respectively—while reducing the "unsupported claim rate" to 5.9%. These illustrative results suggest that integrating evidence-chain organization, constraint-based validation, and verification of generated outputs can effectively minimize the inclusion of information not supported by the source documents. Furthermore, the proposed method scores 90.0% in multi-turn conversational consistency and 93.1% in adherence to domain-specific constraints, demonstrating the efficacy of its context-memory and constraint-aware mechanisms in handling complex, sequential Q&A tasks.

Table 3: Illustrative overall performance comparison of power-domain question answering methods (%).

Method	Recall@5	Answer Acc.	Faithfulness	Multi-turn Cons.	Constraint Acc.	Unsupported Claims
LLM-only	—	68.1	62.0	56.4	65.4	25.0
BM25-RAG	71.8	79.3	75.1	66.0	77.4	17.9
Dense-RAG	74.0	83.2	79.5	72.0	81.0	14.8
Hybrid-RAG	79.1	85.1	82.1	75.0	84.5	12.5
GraphRAG	84.0	88.5	85.5	83.0	87.5	9.8
LightRAG	85.1	89.0	86.6	84.2	88.3	8.9
Self-RAG	81.5	86.8	86.8	79.2	85.9	9.2
CRAG	82.0	87.3	87.0	79.8	86.6	8.6
Proposed	89.0	92.2	90.9	90.0	93.1	5.9

As shown by the example data in Table 4, the proposed method consumes fewer tokens for retrieved evidence across all three types of question-answering (QA) tasks compared with both Dense-RAG and GraphRAG. In fact-based QA, the median token count for the proposed method is 165.8K—a reduction of approximately 49.4% and 31.5% relative to Dense-RAG and GraphRAG, respectively—suggesting that the adaptive retrieval strategy effectively limits the inclusion of redundant knowledge fragments in the generation context. In rule-based QA, the proposed method's median token count is 54.2K, representing a reduction of about 31.9% and 20.8% compared to the two baseline methods. This aligns with the design objective of leveraging technical term anchoring and structural priors (based on document sections) to improve the precision of locating standard provisions. For multi-turn interaction tasks, the median token count for the proposed method drops to 61.5K—approximately 35.1% and 25.3% lower than that of Dense-RAG and GraphRAG, respectively. The use of structured dialogue memory and query rewriting likely reduces the need for redundant retrieval caused by irrelevant historical information. Regarding response latency, the proposed method exhibits a median latency of 2.18 seconds; while this is slightly lower than the 2.36 seconds observed for GraphRAG, it remains higher than the 1.42 seconds recorded for Dense-RAG. This indicates that reducing the number of retrieved tokens does not necessarily result in the lowest possible inference latency, as processes such as multi-path retrieval, graph-based searching, and evidence verification entail their own additional computational costs.

Table 4: Illustrative comparison of retrieved evidence tokens and response latency.

Method	Fact QA Tokens (K)	Rule QA Tokens (K)	Multi-turn QA Tokens (K)	Median Latency (s)	P95 Latency (s)
Dense-RAG	327.4	79.6	94.7	1.42	2.31
GraphRAG	241.9	68.4	82.3	2.36	4.10
Proposed	165.8	54.2	61.5	2.18	3.74

Figure 5 compares the performance of the proposed interactive intelligent question-answering framework for the power domain against two baseline methods—Dense-RAG and GraphRAG—across four metrics: Answer Accuracy, Evidence Faithfulness, Multi-turn Consistency, and Retrieval Recall@5. The semi-transparent scatter points in the figure represent the performance distribution across various test samples or repeated experiments, while the diamond markers indicate the average performance for each method. As shown, the proposed framework demonstrates a clear performance advantage across all four evaluation metrics.

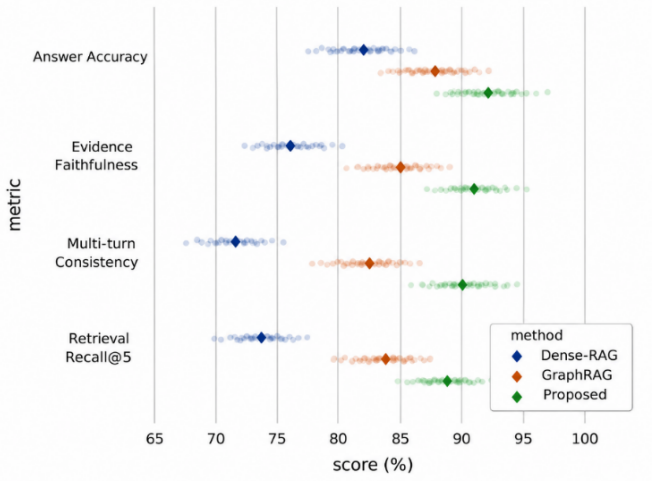


Figure 5: Comparative performance evaluation of different retrieval-augmented generation methods

In terms of Answer Accuracy, the average scores for Dense-RAG, GraphRAG, and the proposed framework are approximately 82%, 88%, and 92%, respectively; this suggests that integrating domain-specific terminology anchoring, knowledge graph associations, and adaptive multi-path retrieval effectively enhances the accuracy of responses to complex questions within the power sector. Regarding Evidence Faithfulness, the three methods achieved scores of approximately 76%, 85%, and 91%, respectively, indicating that mechanisms such as multi-agent knowledge extraction, structured knowledge linking, and evidence-constrained generation help ensure greater consistency between the generated responses and the original professional documents, thereby reducing the occurrence of content lacking evidentiary support. For Multi-turn Consistency, the proposed framework reached approximately 90%—significantly higher than the 72% achieved by Dense-RAG and the 83% by GraphRAG—highlighting the effectiveness of context-memory modeling, query semantic recon-

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

struction, and two-tier keyword extraction in handling entity coreference, elliptical expressions, and sequential follow-up questions.

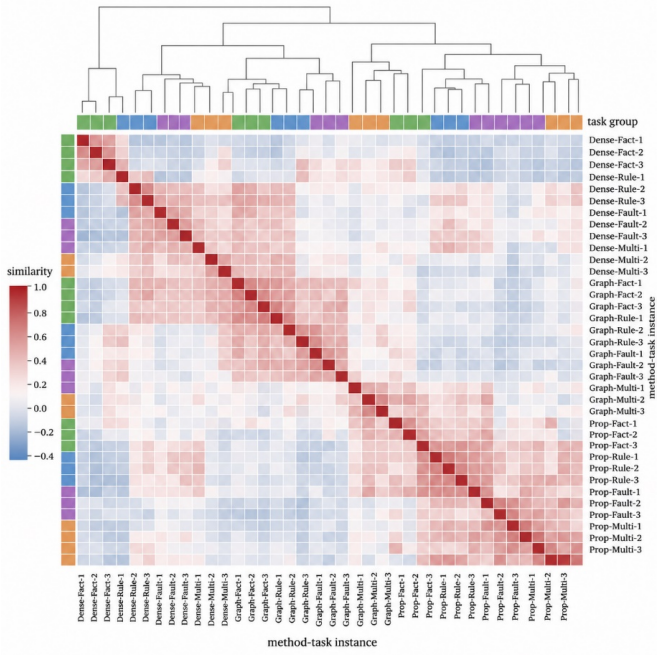


Figure 6: Hierarchical clustering and similarity analysis of performance patterns

Figure 6 illustrates the performance similarities and hierarchical clustering characteristics of the proposed interactive intelligent QA method for power-domain knowledge bases, compared against two baseline methods: Dense-RAG and GraphRAG. The figure employs a visualization combining hierarchical clustering with a similarity heatmap to analyze four types of tasks: fact-based QA (Fact), QA concerning power regulations and rules (Rule), equipment fault diagnosis (Fault), and multi-turn interactive QA (Multi). For each method, three illustrative experimental instances were considered for every task category, resulting in a total of 36 method-task combinations. The colors in the heatmap represent the degree of similarity between different experimental instances: red indicates a strong positive similarity, while blue signifies a lower or negative similarity. The deep red regions along the diagonal correspond to the perfect self-similarity of each instance. As can be observed from the figure, different QA methods exhibit a degree of clustering. The similarity distribution for Dense-RAG is relatively dispersed, with notable instances of low similarity appearing across certain task categories; this suggests that QA frameworks relying solely on semantic vector retrieval may show significant variations in their representational features or performance patterns when handling specialized rule matching, complex fault-knowledge associations, or multi-turn dialogues. In contrast, GraphRAG displays more consistent, positive similarity across various task categories, indicating that the entity associations and structured relational information provided by knowledge graphs help enhance the consistency of knowledge representation across different types of tasks.

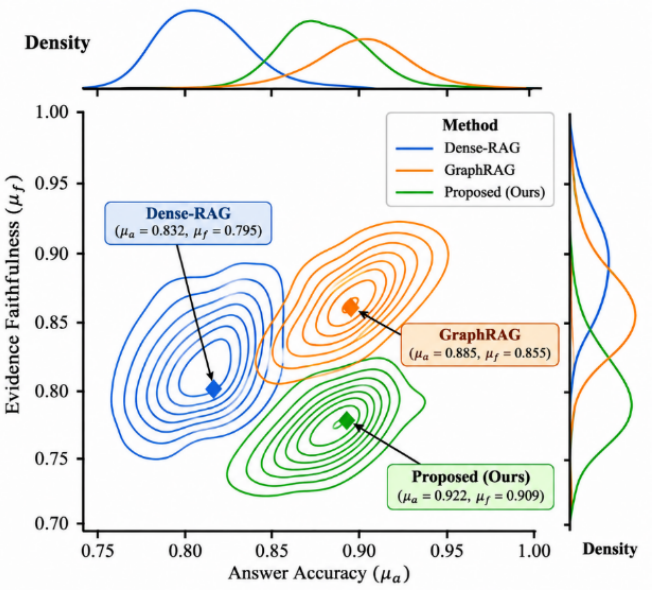


Fig. 6. Joint distribution of *answer accuracy* and *evidence faithfulness* for different methods in the power-domain question answering task. The contour lines and marginal curves show the performance distribution of each method, and the diamonds with annotations indicate the mean values (μ_a, μ_f).

Figure 7: Joint probability density distributions of answer accuracy and evidence faithfulness for different retrieval-augmented generation methods

Figure 7 illustrates the joint probability density distribution of Dense-RAG, GraphRAG, and the interactive intelligent Q&A framework for power-domain knowledge bases proposed in this paper (referred to as "Proposed"), evaluated across two key metrics: Answer Accuracy and Evidence Faithfulness. The performance distribution profiles for each method are visualized using two-dimensional Kernel Density Estimation (KDE). Marginal probability density curves—displayed along the top and right axes—further highlight the concentration of performance and the distributional differences among the methods across various test samples. Diamond markers indicate the average performance for each respective approach. As shown in the figure, Dense-RAG achieves an average answer accuracy of 83.2% and an evidence faithfulness score of 79.5%, with its probability density concentrated primarily in the lower performance range. This indicates that traditional RAG frameworks—which rely solely on semantic vector retrieval—face inherent limitations when it comes to matching complex, domain-specific knowledge in the power sector and maintaining factual grounding. GraphRAG reaches 88.5% in answer accuracy and 85.5% in evidence faithfulness—an improvement of 5.3 and 6.0 percentage points, respectively, over Dense-RAG—highlighting the value of entity-relationship modeling and structured knowledge retrieval (via knowledge graphs) for enhancing performance in complex power-related question answering. In contrast, the method proposed in this paper achieves an answer accuracy of 92.2% and an evidence faithfulness score of 90.9%. These results represent

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

improvements of 9.0 and 11.4 percentage points over Dense-RAG, and 3.7 and 5.4 percentage points over GraphRAG, suggesting that the proposed framework could offer superior overall performance under the illustrated conditions in terms of both answer correctness and the consistency of retrieved evidence.

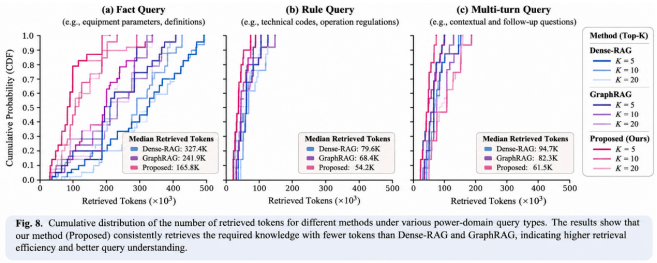


Fig. 8. Cumulative distribution of the number of retrieved tokens for different methods under various power-domain query types. The results show that our method (Proposed) consistently retrieves the required knowledge with fewer tokens than Dense-RAG and GraphRAG, indicating higher retrieval efficiency and better query understanding.

Figure 8: Comparative analysis of cumulative retrieval token distributions

Figure 8 illustrates the number of retrieved tokens and their cumulative distribution characteristics for the proposed interactive intelligent question-answering framework for the power domain, compared against two baseline methods—Dense-RAG and GraphRAG—across three types of tasks: fact-based queries (Fact Query), power-related regulations and rules (Rule Query), and multi-turn interactive queries (Multi-turn Query). In the figure, the horizontal axis represents the number of tokens retrieved during the search process ($\times 10^3$), while the vertical axis represents the cumulative probability (Cumulative Distribution Function, CDF). The stepped curves in different colors correspond to the three retrieval methods, with the analysis further accounting for variations in the distribution when the Top-K retrieval scale is set to 5, 10, and 20. Overall, the method proposed in this paper demonstrates a clear advantage in retrieval efficiency across three types of power-domain question-answering tasks. Its cumulative distribution curve consistently skews toward lower token consumption, indicating that the framework can acquire candidate knowledge using a relatively compact retrieval context. For fact-based question-answering tasks, the median token counts for retrieval are 327.4K for Dense-RAG, 241.9K for GraphRAG, and 165.8K for the proposed method—representing reductions of approximately 49.4% and 31.5% compared to the two baselines, respectively. This suggests that the adaptive retrieval mechanism—which integrates domain-specific terminology anchoring, semantic coverage, and document structure priors—effectively minimizes redundant information during fact-based knowledge queries. In tasks involving procedures and regulations, the median token counts for the three methods are 79.6K, 68.4K, and 54.2K, respectively. The proposed method achieves reductions of approximately 31.9% and 20.8% compared to Dense-RAG and GraphRAG, demonstrating that a hybrid vector-graph knowledge base—constructed using power-domain ontological constraints, hierarchical section indexing, and a multi-agent approach—enhances the precision of retrieving specialized procedures and technical clauses, while reducing context irrelevant to the specific query.

IV. Conclusion

To address challenges in the power sector, such as heterogeneous knowledge sources, complex domain-specific semantic relationships, limitations in understanding multi-turn interactions, and difficulties in ensuring the reliability of generated answers, this paper proposes a trustworthy intelligent question-answering framework. This framework integrates multi-agent knowledge construction with context-aware, adaptive RAG. First, a hybrid vector-graph knowledge base—built using FAISS and Neo4j—is developed through ontology-constrained, multi-agent collaborative extraction, entity localization, and attribute fusion, enabling the structured organization of multi-source power-domain knowledge and the traceability of evidence. Second, the framework employs structured dialogue memory, query reformulation, and a dual-layer keyword enhancement mechanism to improve semantic integrity and ensure consistency with domain-specific constraints during complex, multi-turn Q&A. Furthermore, a lightweight language model is utilized to dynamically integrate semantic coverage, terminology anchoring, and document structure-based retrieval. The framework also incorporates retrieval consensus feedback, evidence re-ranking, and fact-consistency verification to enhance the precision of knowledge acquisition and the traceability of generated answers. Finally, an experimental evaluation system—covering retrieval quality, answer accuracy, multi-turn consistency, robustness, and computational efficiency, is established to validate the individual contributions and synergistic effects of the framework's components.

Author Contributions

All authors contributed equally to the conception, development, and preparation of this manuscript. All authors reviewed and approved the final version of the manuscript and agreed to its submission for publication.

Conflict of Interest

The authors declare that they have no known financial or non-financial conflicts of interest that could have influenced the research presented in this manuscript.

Funding Information

This research was supported by the Science and Technology Project of State Grid Anhui Electric Power Company under Grant No. 5212L0250002.

Data Availability

The datasets used and analyzed during the present study are available from the corresponding author upon reasonable request.

References

[1] Majumder, S., Dong, L., Doudi, F., Cai, Y., Tian, C., Kalathil, D., Ding, K., Thatte, A. A., Li, N., & Xie, L. (2024). Exploring the capabilities and

limitations of large language models in the electric energy sector. *Joule, 8*(6), 1544–1549.

[2] Ruan, J., Liang, G., Zhao, H., Liu, G., Sun, X., Qiu, J., Xu, Z., Wen, F., & Dong, Z. Y. (2024). Applying large language models to power systems: Potential security threats. *IEEE Transactions on Smart Grid, 15*(3), 3333–3336.

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates.

[4] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). *From local to global: A graph RAG approach to query-focused summarization* [Preprint]. arXiv. https://arxiv.org/abs/2404.16130

[5] Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2025). LightRAG: Simple and fast retrieval-augmented generation. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 10746–10761). Association for Computational Linguistics.

[6] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

[7] Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2026). Corrective retrieval augmented generation. In Y. Zhang, R. Xu, A. Purwarianti, T. Rong, & J. Suzuki (Eds.), *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026* (pp. 2991–3003). Association for Computational Linguistics.

[8] Wang, H., Wei, J., Jing, H., Song, H., & Xu, B. (2026). Meta-RAG: A metadata-driven retrieval-augmented generation framework for the power industry. *Computer Engineering, 52*(2), 383–392.

[9] Shen, Z., Cai, X., Ni, B., Meng, Z., Huang, Z., & Yu, T. (2026). HG-RAG: Hierarchical graph-enhanced retrieval-augmented generation for power systems. *Electronics, 15*(7), Article 1445.

[10] Shen, Z., Guo, Q., Jiang, L., Huang, J., Yu, Z., Cai, X., Pang, H., & Yu, T. (2026). Multi-route search and adaptive fusion for power QA with small language model guidance. *Algorithms, 19*(5), Article 378.

[11] Zhou, W., Chen, H., Lu, X., Lin, S., Feng, Y., & Cai, Q. (2026). TSAC-RAG: Type- and structure-aware constraint retrieval for question answering over power-system standards. In *2026 6th Power System and Green Energy Conference (PSGEC)* (pp. 93–98). IEEE.

[12] Xue, Z., Jiang, H., Zhang, Y., & Guo, Z. (2026). Query-adaptive graph retrieval-augmented generation for traceable power-quality knowledge reasoning. *Energy and AI, 26*, Article 100909.

[13] Chen, Q., Wu, J., Li, Q., Peng, G., Gong, Z., Zhang, X., Huang, Y., & Yang, B. (2026). A comprehensive review of multi-modal data fusion-driven systematic knowledge construction technologies for HVDC operation and maintenance. *Processes, 14*(15), Article 2486.

[14] Liu, K., Dong, Y., Pu, O., Li, J., Liu, X., Bao, Z., Zhang, S., & Ye, J. (2026). Towards dynamic knowledge graph-based RAG system optimisation for power system large language models. *International Journal of Information and Communication Technology, 27*(90), 27–48.

[15] Ghafari, M., Sami, A., Rezapour, H., Saravani Ghayour, S., Lin, F., Basaran, K., Cristian Lazaroiu, G., & Siano, P. (2025). A comprehensive review on the application of large language models (LLMs) in power systems. *IEEE Access, 13*, 209450–209486.

[16] Qiao, X., Tang, Y., Qiao, L., Li, B., Yong, Q., Meng, L., Li, Y., & Yu, A. (2026). Research on an intelligent WeChat-based Q&A system for electricity services using large language and speech models. In *2026 IEEE International Conference on Power, Electronics and Green Energy (ICPEGE)* (pp. 1605–1609). IEEE.

[17] Hannaan, A., et al. (2026). A structured question–answer dataset for large language models in power system dynamics and operations. *IEEE Access, 14*, 8391–8407. https://doi.org/10.1109/ACCESS.2026.3652625

[18] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019