

Fusion of Large Language Model and Knowledge Graph for Intelligent Power-System Q&A

Zijie Meng, Xinlei Cai*, Lizhou Jiang, Kai Dong, Bingzhao Zhu, Junhong Guo
Power Dispatching and Controlling Center of Guangdong Power Grid Company Limited,
Guangzhou, Guangdong 510000, China

Abstract—Intelligent question-answering (Q&A) models are needed for modern power systems that can provide accurate, reliable, and explainable answers to engineers and operators. Conventional Q&A systems cannot always understand the question semantically or combine domain knowledge, and stand-alone Large Language Models (LLMs) can generate hallucinated or non-grounded answers. To tackle these challenges, this research introduces a fusion model that combines the power of LLMs with a Power Knowledge Graph (PKG) for intelligent power-system question answering. The proposed approach integrates LLMs' ability to understand language with that of structured knowledge representation by PKGs. First, z-score normalization is used to normalize the operational parameters, such as voltage, current, temperature, and load. After that, Bidirectional Encoder Representations from Transformers (BERT) is used to obtain semantic relations and domain-specific elements from textual data related to the power system. The PKG stores fault, operational, and maintenance knowledge, and a knowledge-enhanced retrieval mechanism extracts relevant information and adds it to the LLM prompt to boost the accuracy of reasoning and diminish hallucination. The model is assessed with the following tasks: Fault Type Identification, Environmental and Operational Causation, and Maintenance and Recovery Decision. The proposed Python-based implementation achieves BERT Score F1 values of 71.5%, 69.2%, and 67.8%, Faithfulness Scores of 64.8%, 61.5%, and 59.7%, and Hallucination Rates of 18.6%, 21.4%, and 24.1%, respectively. The results validate the proposed model's efficiency in intelligent power-system decision support through semantic consistency, factual accuracy, and context relevance.

Keywords—Large language model; intelligent question answering; knowledge graph reasoning; smart grid, power systems; knowledge management

I. INTRODUCTION

Power systems are now highly interconnected and intelligent infrastructures enabling large-scale electricity generation, transmission, distribution, and consumption [1]. The complexity of power systems operations has risen dramatically due to the addition of renewable energy resources, distributed generation, smart meters, and automated control technologies [2]. Therefore, it is essential for power engineers and system operators to have efficient knowledge services that can deliver accurate information to ensure system operation, equipment maintenance, and fault diagnosis [3]. Intelligent question-answering (Q&A) systems have been identified as a key tool for enhancing knowledge accessibility and aiding in decision-making in power systems [4].

Smart grids are being scaled up, and a wide variety of power equipment (generators, transformers, substations, transmission lines, and protection devices) is recording large amounts of operational and maintenance data [5]. The efficient use of this knowledge is critical for ensuring system reliability, maintaining continuous operation, and streamlining the process of fault recovery [6]. Traditional power-system Q&A approaches are mostly based on rules for retrieval and/or keyword-matching, but these have a known lack of semantic understanding, poor contextual reasoning, and limited adaptability to complex technical questions [7]. The constraints limit their ability to provide clear and understandable answers in dynamic operational contexts.

With recent advancements in LLMs, performance on natural language understanding and reasoning in context has proven promising for intelligent question-answering applications [8]. However, conventional LLMs may generate hallucinated responses, lack specific domain knowledge, have poor explainability, and are thus unsuitable for critical applications in power systems. Fortunately, a PKG can effectively represent the knowledge associated with power systems, which includes information on equipment, operating procedures, maintenance history, and fault occurrence [9]. Power System Faults Dataset. The proposed methodology enhances answer correctness, consistency, and explainability of the Q&A system by combining PKG with LLMs.

With the development of intelligent power system fault diagnosis and Q&A, knowledge graph-based models have been increasingly applied to improve structured knowledge retrieval and reasoning. An intelligent model for fault retrieval based on knowledge graphs was proposed by [10] for power communication equipment, with F1-scores of 81.7% and 87%, respectively, with multi-layer encoding and graph convolution. It demonstrated that it could effectively represent the semantics of a graph, but it had a heavy reliance on fault-specific data quality requirements. Table I presents a comparison of existing power-system fault analysis and decision-support methods.

Based on this motivation, this research suggests that LLM-PKG can be fused to realize intelligent power-system Question and Answer (Q&A). The model utilizes Knowledge-Enhanced Retrieval techniques to retrieve relevant entities, relationships, and operational records from the PKG and to incorporate them into the reasoning process of the LLM. Experimental assessment shows increased accuracy of answers, relevance to context, explanation, and answer consistency, and reduction of hallucination.

*Corresponding author.

TABLE I. COMPARISON OF EXISTING METHODS FOR POWER-SYSTEM FAULT ANALYSIS AND KNOWLEDGE-BASED DECISION SUPPORT

Ref	Main Objective	Method	Results	Limitation
[11]	Fault analysis in Industrial IoT	Multi-Entity Knowledge Joint Extraction (MEKJE) with jointly trained models	F1-score: 78.6%	Limited generalization
[12]	Knowledge utilization for recommendation systems	Cross-domain knowledge graph embedding	Improved prediction accuracy	Scalability and computational complexity
[13]	Automatic electric-power fault knowledge extraction and Q&A	BERT-BiLSTM-CRF + electric power fault knowledge graph	Improved Chinese entity extraction and Q&A	Limited multilingual adaptability
[14]	Equipment fault-diagnosis named entity recognition	RoBERTa-BiLSTM-CRF (RBC)	F1-score: 94.98%	Limited dataset
[15]	Power-system fault detection and reasoning	Random Forest + knowledge reasoning	Accuracy: 90%	Small-scale IEEE 14-bus simulation limits real-world applicability
[16]	Power knowledge retrieval and inference	RPKR-GNN with GNN, user-intent fusion, and knowledge inference	Improved power-knowledge retrieval and reasoning	Limited explainability and scalability
[17]	Reliable LLM-based decision support for smart-grid operations	Locally hosted LLMs	Local LLMs outperformed standard LLMs in identifying unauthorized operations.	LLM hallucination and insufficient domain-specific knowledge grounding
[18]	Improve retrieval-augmented generation (RAG) accuracy and retrieval efficiency for complex queries.	Query complexity classification, keyword retrieval, semantic rules, and natural language inference (NLI)-based relevance checking.	On HotpotQA, accuracy improved by 9.4% and F1 by 16.14% over Adaptive-RAG.	Depends on keyword rules, custom training data, and NLI quality.
[19]	Improve retrieval and generation for complex multi-hop queries.	Domain-specific Knowledge Graph (KG) + Large Language Model (LLM) using entity extraction, path finding, query decomposition, semantic path ranking, and context aggregation.	Outperformed RAG and GraphRAG, especially for 5–6-hop queries, with strong directness.	Limited generalizability across domains.
[20]	Improve multi-hop reasoning over knowledge graphs.	Question-Directed Reasoning with Relation-Aware Graph Attention Network (QRGAT) using relation-aware GAT and reasoning graphs.	Improved Hits@1 by 2.3% on WebQuestionsSP and 3.6% on ComplexWebQuestions over NSM+h.	Computationally complex for multi-layer graph reasoning.
[21]	Improve online power-grid fault diagnosis and decision support.	Time-series density maps + Residual Network (ResNet) for equipment and fault classification.	Accurately identifies faulty equipment and fault type for online diagnosis.	Mainly designed for power-grid fault data; generalization may be limited.

Based on these investigations, this research observes that knowledge graphs provide structured knowledge representation and reasoning, whereas LLMs provide strong semantic understanding and natural-language generation. However, without combination, the reliable, interpretable, and consistent knowledge base still fails. The knowledge extraction approach in examination [16] uses the powerful ability of Graph neural networks to retrieve the most related facts from large corpora, the knowledge retrieval method also fails due to problems with explanation and scalability. Investigation [17] presents an LLM-based system, the smart grid knowledge management approach, but because LLMs can hallucinate, which brings down trust and reliability, and no domain knowledge is grounded in the models. Therefore, combining LLM and PKG is a potential solution for creating accurate, explainable, and intelligent power system Q&A systems. From an electrical engineering point of view, the proposed framework should support structured power-system knowledge and provide intelligent decision support. In contrast to existing question-answering systems based on languages, the proposed approach captures information about electrical operating conditions, equipment state, fault events, environmental conditions, and maintenance actions in the form of a structured PKG. These electrical relationships are then applied as proof in the framework for fault interpretation and the generation of maintenance-related responses. Thus, the contribution can be seen not only as LLM-based Q&A but as an intelligent information and decision-support model for power system fault analysis.

The proposed LLM-PKG combines structured power knowledge from the domain, while avoiding reliance on text retrieval or graph querying. Conventional RAG only retrieves the most relevant textual passages, while the proposed scheme retrieves structured triples representing equipment, operating variables, environmental conditions, faults, equipment states, and maintenance actions in the power system. Retrieval is based on BERT for power-domain entity identification and a one-hop neighborhood expansion performed in the power domain, followed by ranking candidates based on semantic similarity and graph-based electrical relationship preservation. The retrieval process differs from general GraphRAG methods by first using BERT to identify entities in the power domain, then applying the one-hop neighborhood expansion in the power domain, and subsequently ranking the candidate triples by semantic similarity while also retaining the graph-based electrical relationship information. The PKG is not meant to be the only system supporting a Q&A component of a knowledge graph; it is meant to be used as a fact-based system to provide evidence for an LLM that provides a natural language response in context. To this end, the contribution involved a task-specific combination of power-domain entity/relation extraction, structured retrieval of PKGs, relevance training of the search results provided to the LLM, and knowledge-guided LLM language generation of fault and cause analysis, as well as of maintenance/recovery options.

II. POWER-SYSTEM PROBLEM FORMULATION

A. Electrical Operating Variables

Electrical, environmental, and equipment-condition variables such as voltage, current, power load, temperature, wind speed, fault duration, downtime, component health, and maintenance status determine the power-system operating state. Voltage and current define an electrical operating condition, and power load defines the level of loading of the power equipment or network. The temperature indicates thermal stress and overheating situations, while wind speed and weather conditions indicate environmental disturbance which may lead to equipment problems. Equipment condition and recovery information include component health, fault duration, downtime, and maintenance status. The proposed PKG and knowledge-grounded LLM for fault diagnosis and decision support rely on these variables as the electrical evidence.

B. Electrical Fault-Causality Representation

Faults in power systems are expressed in terms of causality relationships between electrical operating conditions, environmental conditions, equipment situations, events resulting in a potential fault, and maintenance needs. For instance, if the power load is increased, this can lead to higher current and thermal stresses; this might increase the temperature and reduce the material performance. Adverse weather, like windstorms or thunderstorms, can also add to electrical equipment and so increase the chances of fault occurrence. These dependencies are represented as triples of entity–relation–entity in the proposed PKG, which enables the LLM to rely on electrical evidence to deduce the relation, in addition to language-based evidence.

C. Electrical Fault-State Representation

The operating condition of the power system, along with the relevant electrical, environmental, and component-health parameters, is mapped in a fault-state vector to characterize the condition systematically and reliably for fault identification, as presented in Eq. (1):

$$X = [V, I, P, T, W, H, D_f, D_o] \quad (1)$$

where, V denotes the voltage, I is the current, P is the power load, T is the temperature, W is the wind speed/weather condition, H is the component health, D_f is the fault duration, and D_o is the downtime.

The electrical fault state is represented in Eq. (2) as a function of the observed operating state and structured power-system knowledge:

$$F = f(X, G) \quad (2)$$

where, F denotes the identified fault state, and G represents the PKG. For operator decision support, the generated response is formulated by Eq. (3):

$$Y = g(Q, X, G) \quad (3)$$

where, Q is the user query, and Y is the knowledge-grounded engineering response.

III. METHODOLOGY

The proposed methodology consists of sequential phases for developing an intelligent power-system Q&A model by integrating an LLM with a PKG. The model utilizes the Power System Faults Dataset as the foundational knowledge source and extends it through structured knowledge representation and retrieval-enhanced reasoning. As illustrated in Fig. 1, the Overall Architecture of the proposed LLM–PKG model for Intelligent Power-System Q&A.

A. Data Collection

The proposed model is based on the Power System Faults Dataset (<https://www.kaggle.com/datasets/ziya07/power-system-faults-dataset>) that is obtained from Kaggle, which is used as a main knowledge source for intelligent power-system Q&A. There are 506 fault records, including operational, environmental, maintenance, and recovery data, such as voltage, current, power load, temperature, wind speed, weather conditions, component health, fault duration, and downtime. 80% of the records are used for training and 20% for testing. To ensure that the data partition is reproducible, the value 42, which is the same for all runs, is used as a fixed random seed. The evaluation looks at three Q&A tasks: Fault Type Identification, Environmental and Operational Causation, and Maintenance and Recovery Decision. Table II presents the categories of the dataset for structured fault reasoning: Fault Type Identification, Environmental and Operational Causation, and Maintenance and Recovery Decision. Sample Q&A pairs are presented in Table III for the purpose of replicably providing a benchmark of performance for the provided fault detection and decision support to be further compared and studied.

TABLE II. POWER SYSTEM FAULT DATASET CATEGORY OVERVIEW

Category	Key Features Used	Fault Focus
Fault Type Identification	Voltage, Current, Power Load, Component Health	Line Breakage, Transformer Failure, Overheating
Environmental & Operational Causation	Temperature, Wind Speed, Weather Condition, Fault Duration	Weather-induced faults and operational triggers
Maintenance & Recovery Decision	Maintenance Status, Component Health, Downtime, Duration	Recovery urgency and maintenance prioritization

TABLE III. REPRESENTATIVE Q&A PAIRS FROM THE POWER FAULT DATASET

Category	Question	Answer
Fault Diagnosis	What fault occurred in record F002?	The recorded fault was Transformer Failure, with faulty component health and 5 hours of downtime.
Environmental and Operational Analysis	What factors contributed to the overheating fault in F003?	The fault was associated with high temperature (35°C) and windstorm conditions, leading to component overheating.
Maintenance and Recovery Decision	What maintenance action is required for a transformer failure under faulty health status?	Immediate inspection and corrective maintenance are required to restore system stability and reduce downtime.

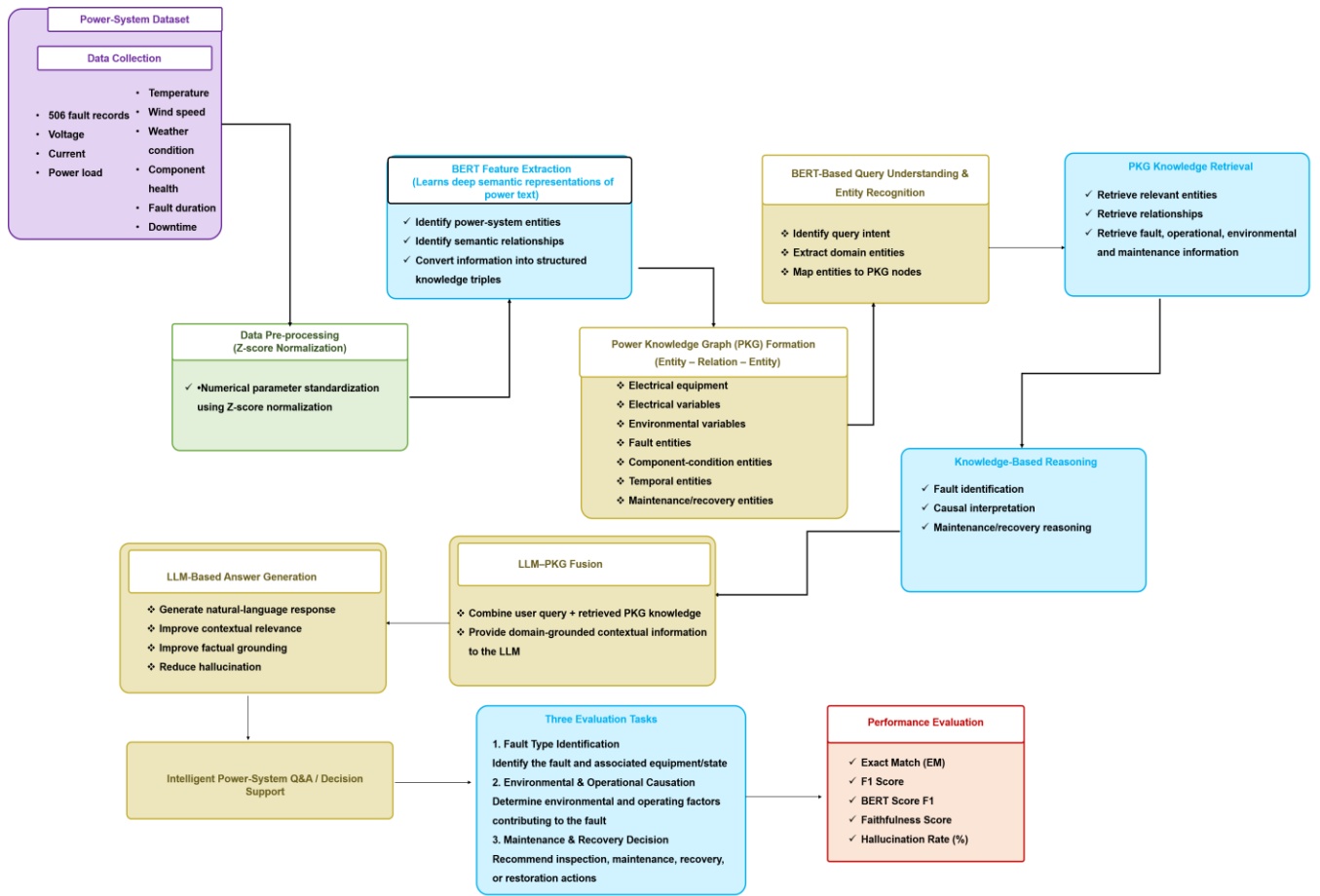


Fig. 1. Overall architecture of the proposed LLM-PKG fusion model for intelligent power-system Q&A.

The question-answer dataset is created from structured information in the 506 power-system fault records. For each of the following categories of semantic reasoning, questions were designed: fault diagnosis, electrical-state analysis, environmental and operational causation, maintenance and recovery, fault-duration and operational assessment, and LLM-based decision support. Each cluster of reasoning has the appropriate number of pairs of questions: 87, 85, 84, 83, 83, and 84 pairs of questions for the six thinking clusters. Each question has a fault record and structured PKG evidence associated. Answers to the ground truth questions are produced directly from the factual data attributes and relationships from the source records, such as downtime, maintenance status, fault type, equipment/component condition, electrical variables, environmental variables, fault duration, answers from the recovery requirements, and answers from the source records. The questions created are evaluated against the source record to make sure there is enough supporting information for each question to derive a deterministic source answer. Before assessing the question, the question is excluded from the examination if it has been duplicated or is unclear or not presented with the source record. The final reference answers are thus based on the original power-system records and associated PKG facts. These questions are coupled together in a question-answer set, which is subsequently applied in evaluating the accuracy of lexical matching, semantic similarity, factual grounding, and hallucination.

B. Data Pre-processing

This technique comes in handy for preprocessing and making heterogeneous numerical representations uniform before entity and relation extraction. Operating variables like voltage, current, power load, temperature, wind velocity, fault length, and time outage could range and have significantly varying numerical scales between different types of sensors. While these parameters can be utilized in the subsequent knowledge extraction process, their local scales can be different, so to scale down influence differences, they should be normalized using z-scores to convert them to a common scale, which allows for comparison of the same systems and easier representation of entities and relations in the power system. The transformation is defined in Eq. (4), where a normalized fault parameter is calculated based on its original value, the mean, and the standard deviation of the fault records. This normalization process helps maintain consistency among data and offers a stable numeric foundation for building the proposed PKG:

$$X = \frac{Z - \mu}{\sigma} \quad (4)$$

where, X is the normalized fault parameter value, and Z is the original operational parameter value, μ is the mean value of the same fault parameter, and σ is the standard deviation of the same fault parameter for all fault records. This normalization helps to maintain data consistency and to analyze and extract

knowledge from the faults that may occur within the proposed LLM-PKG system.

C. Power Knowledge Graph Construction

The PKG contains different interacting elements and relationships among the elements of electrical equipment, operating conditions, environment, fault events, and maintenance actions. The main entities are electrical equipment, fault types, electrical parameters, environmental conditions, component health types, maintenance types, and recovery conditions as shown in Table IV. These entities have causal, operational, spatial, temporal, and maintenance dependencies, known as relations. Power-System LLM-PKG Data Fusion Architecture integrating multi-modal operational data, Knowledge Graphs, and LLM-based reasoning is depicted in Fig. 2.

TABLE IV. ENTITY CATEGORIES AND REPRESENTATIVE ENTITIES IN THE POWER KNOWLEDGE GRAPH

Entity category	Representative entities
Electrical equipment	Transformer, transmission line, substation, relay, generator
Electrical variables	Voltage, current, power load
Environmental variables	Temperature, wind speed, weather
Fault entities	Line breakage, transformer failure, overheating
Condition entities	Component health, maintenance status
Temporal entities	Fault duration, downtime
Decision entities	Inspection, corrective maintenance, recovery

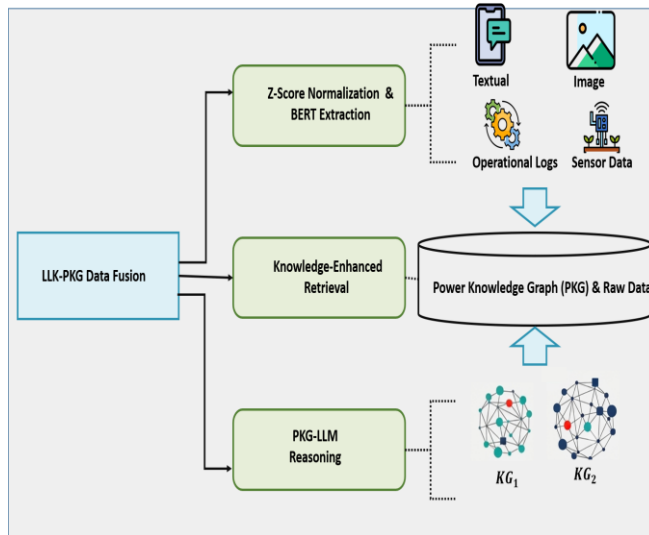


Fig. 2. Power-system LLM-PKG data fusion architecture.

The PKG is built based on the power-system fault records that have been normalized and deduplicated. The 506 normalized and deduplicated power-system fault records are used to build the PKG. It produces 1,248 distinct nodes, 1,836 entity-relationship-entity triples, 1,836 specific entity relationships, and 24 specific relation types. The types of nodes are electrical equipment, electrical variable, environmental,

fault type, component, temporal, or maintenance or recovery decision. Relations express causal, operational, spatial, temporal, and maintenance dependencies. BERT will encode user query as a list of domain entities relevant to that query, mapping domain entities to the corresponding PKG nodes. The neighborhood system is then executed in a one-hop fashion to create potential candidate triples that are related to the extracted nodes. To rank candidate triples, the query representation is compared with representations of the triples, based on cosine similarity; thereafter, any duplicate triples are filtered out, and the top 5 triples are provided to the LLM as instructive and factual context. The process is dynamically linked to the query understanding process, the graph retrieval process, and the process of generating answers from the knowledge base.

The extracted entities are sorted based on their functional roles in the domain of power systems. Electrical equipment entities are physical objects like transformers, lines, substations, relays, and generators, while electrical and environmental variables are measurable operating conditions. Fault, condition, temporal, or decision entities offer more information on abnormal equipment health, fault duration, downtime, inspections, maintenance, and recovery. Finally, those entity pairs are translated into subject–relation–object triples and added to the PKG. This structured representation allows for enabling the connection of semantically relevant but different textual information by providing explicit semantic relations instead of handling only textual language. The resulting PKG can then be used as a knowledge layer when mapping queries to the graph and for evidence retrieval in the following step.

D. BERT-Based Power Entity and Relation Extraction

BERT is employed for extracting power-system entities and their semantic relationships from textual power-domain datasets. BERT is selected due to its bidirectional contextual learning capability, which enables effective understanding of technical terminology, operational dependencies, and fault-related semantics in power-system documents. The BERT input representation is composed of three embedding layers:

- 1) *Token embedding*: It transforms power-system textual tokens into dense vector representations. These embeddings capture domain-specific terms such as transformers, relays, substations, and fault types.
- 2) *Input*: The input to the BERT model consists of a linear sequence of tokens derived from the pre-processed power-system corpus. Special tokens [CLS] and [SEP] are added to mark the beginning and separation of input sequences.
- 3) *Segment embedding*: It distinguishes multiple sentence segments, which is useful for identifying inter-sentence relationships in fault reports and maintenance logs.
- 4) *Position embedding*: It preserves the sequential order of words in power-system text, enabling the model to understand the structural dependencies among entities and actions.

Using contextual embeddings, BERT extracts power-system entities and relations, converting fault, location, operational, environmental, maintenance, and temporal information into structured knowledge triples for PKG.

The semantic extraction module extracts power-system entities and relations from textual descriptions of the power system and user queries using BERT-base-uncased. The contextual BERT representation is then input to a token-classification output layer to perform entity classification. After the identification of entities, entity pairs are candidates in the considered textual context and then categorized based on the power-system relation inventory defined. The extracted entity–relational information is coded as structured triples (head entity, relation, tail entity) and added to the PKG. The same BERT representation process is used during the query processing stage for domain entity identification and to aid in their mapping to the respective nodes in the PKG. The final answer generation is then left to BERT. It is designed to deliver meaning to natural-language queries and to link them to the structured knowledge of the power system. The LLM then gets the evidence retrieved by the PKG and produces the completed natural-language result.

E. User Query Processing and Entity Recognition

In this phase, the proposed model receives natural language queries related to power-system faults, operational conditions, and maintenance decisions. The query is processed using BERT-based semantic analysis to understand user intent, capture contextual information, and identify domain-specific entities and relationships. The extracted entities are then aligned with the corresponding nodes in the PKG, enabling accurate knowledge retrieval and supporting subsequent reasoning and response generation.

F. Contextual Query Models

To illustrate the query understanding process, representative contextual Q&A examples derived from the power fault dataset are presented in Fig. 3 and Fig. 4. These examples demonstrate how fault records are transformed into structured contextual inputs for semantic interpretation. The contextual representation enables BERT to capture fault-specific dependencies, operational conditions, environmental influences, and maintenance-related information from user queries. This process supports effective entity recognition and improves the alignment between user queries and Power Knowledge Graph entities for accurate knowledge retrieval and reasoning.

Context 1: Fault record F002 reports a Transformer Failure at coordinates (34.056, -118.245) under rainy weather conditions. The transformer operated at 1800 V, 180 A, and 45 MW load with a recorded temperature of 28°C. The component health status was marked as Faulty, with a fault duration of 3 hours and a total downtime of 5 hours. Maintenance history indicates the component was previously serviced.

Question: Which operational parameter indicates that the transformer failure was likely associated with equipment degradation rather than thermal overload?

Answer: The moderate operating temperature (28°C) combined with the Faulty health status suggests the failure was caused by internal equipment degradation rather than excessive thermal stress.

Question: Why does the downtime exceed the fault duration in F002?

Answer: The longer downtime (5 hrs) compared to the fault duration (3 hrs) indicates additional time was required for fault isolation, transformer inspection, and restoration of grid connectivity.

Fig. 3. Contextual query examples for power fault analysis for fault diagnosis and operational analysis.

Context 2: Fault F003 records an Overheating event at (34.0525, -118.244) during windstorm conditions. The system operated at 2100 V, 230 A, and 55 MW load. The temperature reached 35°C, and the component health was classified as Overheated. The fault persisted for 4 hours, resulting in 6 hours of downtime.

Question: Which combination of parameters most strongly indicates thermal stress as the primary cause of failure?

Answer: The elevated temperature (35°C), high power load (55 MW), and Overheated component condition collectively indicate thermal stress as the dominant fault-driving factor.

Question: How could windstorm conditions contribute to overheating despite not being a direct thermal source?

Answer: Windstorm conditions can introduce mechanical stress and cooling instability, reducing heat dissipation efficiency and increasing the probability of sustained thermal accumulation in the component.

Fig. 4. Contextual query examples for power fault analysis for environmental and thermal fault reasoning.

G. LLM-PKG for Intelligent Power-System Question Answering

The proposed LLM-PKG model integrates structured knowledge retrieval with language generation for intelligent power-system question answering. After query entity recognition, relevant knowledge triples are retrieved from the PKG and combined with the user query as contextual input to the LLM. This hybrid retrieval-reasoning mechanism improves answer accuracy, contextual relevance, and explainability, and reduces hallucinations through domain-grounded response generation. The proposed model employs Llama-3.1-8B-Instruct (8 billion parameters), a decoder-only Transformer LLM, for knowledge-grounded answer generation, while BERT-base-uncased is used for entity and relation extraction and query understanding. The LLM is used in pretrained form without additional domain-specific pretraining. For each query, BERT identifies relevant entities, which are mapped to PKG nodes. A two-stage graph-neighborhood and semantic retrieval method is then applied, using cosine similarity between query and triple embeddings. The top 5 relevant triples ($K=5$) are retrieved and inserted into a structured prompt containing the task instruction, user query, and retrieved PKG evidence. The PKG constructed from the 506 fault records contains 1,248 nodes, 1,836 edges, 24 relation types, and 1,836 unique triples. Since the LLM is used without fine-tuning, training epochs and learning rate are not applicable to the LLM. During inference, the temperature is set to 0.2, with a maximum generation length of 256 tokens. The implementation uses Python 3.11, PyTorch 2.3.1, and Transformers 4.44.2, and experiments are performed on an NVIDIA RTX 4090 (24 GB), Intel Core i9-13900K CPU, and 64 GB RAM. A fixed random seed of 42 is used to ensure reproducibility.

The fusion process includes four consecutive steps: Query interpretation, evidence retrieval in the graph database, contextual fusion, and response generation. The first is the BERT query processor, which detects user intents, entities, and relationships. Second, the identified entities are mapped onto corresponding nodes in PKG, and their graph neighborhoods are searched to get related knowledge triples. Third, recovered triples are reduced and structured into contextual evidence with the original query. Lastly, the query-evidence context is given to the LLM for answer generation, where the instruction is tiled.

Finally, the query-evidence context is presented to the LLM for answer generation, with an instruction tiled around it. This design decouples the structured knowledge retrieval from the natural-language generation and enables the LLM to rely on explicitly available evidence from the power system for response generation. Language understanding and generation in LLMs should therefore remain unchanged, with the aim of enhancing the factual grounding via the fusion mechanism.

1) Power Knowledge Graph (PKG) retrieval and reasoning: After query entity recognition, the extracted entities are mapped to their corresponding nodes in the PKG to retrieve relevant structured knowledge. The retrieval process explores entity–relation connections associated with fault events, operational parameters, environmental conditions, and maintenance records. This is followed by semantic reasoning on extracted knowledge triples to infer fault causes, operating dependencies, and maintenance needs. This process enhances context understanding, ensuring the generated response is based on knowledge in power systems. The constructed PKG contains 1,248 unique entities (nodes), 24 relation types, and 1,836 unique entity–relation–entity triples represented by 1,836 directed edges. The graph was constructed from the 506 power-system fault records after entity normalization and duplicate removal. The entities include electrical equipment, electrical variables, environmental conditions, fault types, component conditions, temporal information, and maintenance/recovery decisions. For each query, the identified entities are mapped to corresponding PKG nodes, and their connected relations and triples are retrieved to provide domain-specific evidence for LLM-based reasoning.

To ensure a repeatable knowledge-retrieval approach, the suggested LLM–PKG strategy involves the semantic retrieval phase based on a query, the neighborhood expansion in the retrieved graph, and the relevance-based triple selection. The recognized entities and the representation of the query are fed into the retrieval system after the query is processed using BERT. Both the textual description and the semantic embedding of each PKG entity and each PKG relation are used. Both the query embedding and the embeddings for candidate entity–relation–entity triples are compared with cosine similarity in Eq. (5):

$$S(q, t_i) = \frac{q \cdot t_i}{\|q\| \|t_i\|} \quad (5)$$

where, (q) denotes the user-query representation and t_i denotes the semantic representation of the i – th knowledge triple. A higher cosine-similarity value indicates greater semantic relevance between the query and the corresponding PKG triple. There are two steps in the retrieval process. The entities in the query are first matched to the review nodes corresponding to the entities in PKG, and candidate triples are extracted that have a connection with these nodes by a 1-hop graph traversal. Second, the candidate triples are ordered by the cosine-similarity score. Only the top-K (K is set by the user) highest-ranked triples are stored in order to avoid irrelevant information being passed to the LLM. The proposed implementation assumes that $K=5$; that is, it supplies at most

five knowledge triples to the LLM that are relevant to the same query. All triples are left in if fewer than 5 satisfy the relevance criterion. Semantic similarity is the main criterion for determining the final selection score, leaving the graph relationship criterion for power-system reasoning unchanged. So, the triple is chosen if it meets both conditions: 1) it is connected with at least one query-relevant entity or its graph neighborhood and 2) it meets a high enough similarity score compared to other candidate triples. Triples are deleted unless they represent duplicates, and semantically redundant triples are removed prior to the final top-(K) selection. This approach involves minimizing the contextual data provided to the LLM while keeping the electrical, environmental, fault, and maintenance connections pertinent to the query.

The retrieval process can be summarized as follows: 1) receive the natural-language query; 2) recognize domain entities and query intent with BERT; 3) match the recognized entities to the PKG nodes; 4) conduct a one-hop neighborhood expansion to get candidate triples; 5) produce semantic representations for the query and candidate triples; 6) calculate the similarity of the query with candidate triples based on cosine similarity; 7) rank the candidate triples by the cosine similarity score in descending order; 8) filter duplicate or redundant triples; 9) use the top five triples as the knowledge context input to the LLM to generate the answer. The retrieved triples are then appended to the original query to follow a prompt template as outlined in 3.7.2. This retrieval mechanism acts as a direct link between understanding the query, the graph-based retrieval of evidence, and reasoning with the LLM. For instance, if a query about transformer overheating is issued, the overload retrieval will return triples related to the transformer equipment along with temperature, overheating, component health, weather, and maintenance actions, but not unrelated power-system entities such as circuit breakers or an index. This results in the LLM only being given a subset of the evidence relevant to each particular query, thereby increasing the relevance of the material and minimizing the likelihood of unsupported generation.

Only the graph evidence most closely related to the semantic content of the query should be retained during the retrieval stage. Once entity recognition is complete, the retrieved anchors are nodes of the corresponding PKG. Their neighbouring triples give the evidence for the candidates in the form of descriptions of equipment states, fault conditions, environmental influences, temporal characteristics, and maintenance actions. Finally, the retrieved evidence is ranked by semantic similarity between the query representation and candidate triples. Top-ranked triples are kept as knowledge in context, because redundant triples are meant to be stripped away to reduce repetition of context and unnecessary details. This retrieval approach minimizes the information that the LLM receives in the graph and focuses the response on the most relevant facts within the world of the power system.

2) Large Language Model (LLM)-based answer generation: The retrieved knowledge triples after retrieval and inference are supplemented with the user query as input for generating answers by the LLM. The LLM is able to produce accurate, context-aware, and explainable responses based on the natural language query combined with the structured PKG

context. This knowledge-guided generation enhances semantic understanding and mitigates hallucination by embedding domain-specific evidence. The proposed model uses Llama-3.1-8B-Instruct, with 8 billion parameters, as the LLM for knowledge-grounded answer generation. The model follows a decoder-only Transformer architecture and is used in its pretrained, instruction-tuned form without additional fine-tuning on the power-system dataset. The 506 power-system fault records are used to construct the PKG and generate the knowledge-grounded Q&A context, rather than being used for additional LLM pretraining. BERT is used separately for power-system entity/relation extraction and query understanding. The top 5 relevant PKG triples are retrieved using semantic similarity and appended to the query as contextual evidence. The instruction requires the LLM to answer using the retrieved power-system facts and avoid unsupported information. Since the LLM is not further fine-tuned, training epochs, learning rate, and fine-tuning method are not applicable. During inference, the temperature is set to 0.2, the maximum generated length is 256 tokens, and deterministic decoding is used to improve reproducibility. Thus, the LLM functions as the final knowledge-grounded response generator, while the PKG supplies the domain-specific factual evidence.

To evaluate the generated answers, the 20% test set is compared to the corresponding reference answers in the Power System Faults Dataset. Exact Match (EM) indicates verbatim accuracy of the generated answer and the reference. F1 is used to assess the accuracy of generated answers on a token-by-token basis. BERT Score F1 is the version of BERT that compares answers based on contextual representations. Faithfulness measures how consistent the responses are with the retrieved PKG evidence, while Hallucination Rate measures the percentage of responses that include information that is unsupported or incorrect. All comparison variants use the same testing subset, reference answers, retrieval settings, and evaluation criteria as are utilized for the proposed LLM-PKG model. The scores obtained are then averaged across the assessed test set on each of the three Q&A scenarios, quantifying the correctness of the answers, semantic similarity, factual grounding, and answering behavior of hallucinated answers.

Llama-3.1-8B-Instruct (8B parameters), pretrained in an instruction-tuned fashion, has been used for the implementation without fine-tuning. BERT-base-uncased is used alone for entity and relation extraction, and for query understanding. Power-System does not, therefore, update the parameters of the LLM, but rather its records are mapped to structured, evidence-of-expertise power-system instances, as well as question-answer instances associated with knowledge. Given a query, BERT finds entities that are relevant to the query, aligns the entities by mapping them to the nodes in PKG, and expands the one-hop neighborhood for formulating candidate triples. Corresponding to the query representation, a group of top 5 ($K=5$) triples is returned, with the highest-ranking one according to cosine similarity between the query and the corresponding triple representations. The LLM is provided with a structured prompt that consists of the following three parts: 1) an instruction that

asks the LLM to respond in a relevant way that can be supported with evidence; 2) the original user query; and 3) the retrieved PKG triples. The answer must be limited to the information provided in the Power System Supplies and does not include statements of facts that are not supported by the information provided. To evaluate a model for inference, they set the temperature to 0.2 and the maximum generation length to 256, and set the decoding method to deterministic to be reproducible. No further fine-tuning, optimization, or further training epochs are applied, both on the LLM and during the preprocessing. It is implemented on an NVIDIA RTX 4090 (24 GB memory), an Intel Core i9-13900k (CPU), and 64 GB of RAM with Python 3.11, PyTorch 2.3.1, and Transformers 4.44.2. In the experimental implementation, a random seed of 42 is fixed throughout the implementation. The cross-entropy training loss is not applied to the LLM in the proposed experimental setting, as is the norm with the LLM in isolation. The generation stage is thus handled as inference, where a pre-trained instruction-tuned model is conditioned on the retrieved PKG evidence. The cross-entropy loss function is used to optimize the LLM during training, measuring the difference between the probabilities of the tokens generated by the LLM and the actual target sequence as shown in Eq. (6):

$$K = -\sum_{i=1}^m \log P(x_i|y, x_1, x_2, \dots, x_{i-1}) \quad (6)$$

where, K denotes the cross-entropy loss; m is the total answer length, y represents the knowledge-augmented input (user query + PKG facts), x_i is the i^{th} generated token, $\log$ denotes the logarithmic function measuring prediction confidence, and $\log P(x_i|y, x_1, x_2, \dots, x_{i-1})$ is the conditional probability of generating the next answer token.

H. Electrical Fault Diagnosis and Decision-Support

This retrieved knowledge in turn is leveraged to make three engineering decisions: fault identification, causal interpretation, and maintenance/recovery recommendation. The query is initially converted into electrical bodies and working parameters. The relevant PKGs are then retrieved to look for supporting evidence. This LLM then uses the retrieved electrical evidence and the user question to formulate the final response. The response obtained is then compared with the fault information and the decision requirements. The proposed model thus couples machine measurements and equipment conditions to structured fault knowledge and natural-language decision support.

IV. RESULTS

The experimental evaluation aimed to determine the language-based reasoning capability and assess the electrical fault-diagnosis capability of the proposed LLM-PKG model. This evaluation focuses on fault identification; causal factors associated with the environment and operation; and decision-making on maintenance and recovery. Moreover, measures of semantic similarity and knowledge-grounding, the performance in electrical fault diagnosis is assessed with the application of classification-oriented measures. The Power System Faults Dataset is then statistically analysed and validated to assess its structural quality and appropriateness for intelligent power-system question answering. The pairs of Q&A were then analyzed in a rule-based system using Python, where the results

were grouped into three categories: Fault Type Identification, which includes fault detection, category, and location; Environmental and Operational Causation, which classifies environmental and operational factors that cause faults; and Maintenance and Recovery Decision, which identifies

corrective actions, maintenance planning, and restoration of systems. The distribution of the dataset by category, shown in Table V and Fig. 5, is a well-balanced distribution that supports the different aspects of the fault reasoning and decision-making process.

TABLE V. DISTRIBUTION OF QUESTION CATEGORIES ACROSS THE POWER SYSTEM FAULTS DATASET

Question Category	Fault Type Identification (%)	Environmental & Operational Causation (%)	Maintenance & Recovery Decision (%)
Diagnostic Reasoning	38.2	12.4	8.6
Causal Analysis	14.5	41.3	9.2
Operational Assessment	21.8	28.6	14.1
Maintenance Planning	6.4	4.8	39.5
Recovery Evaluation	5.1	3.9	21.4
Decision Support	14.0	9.0	7.2

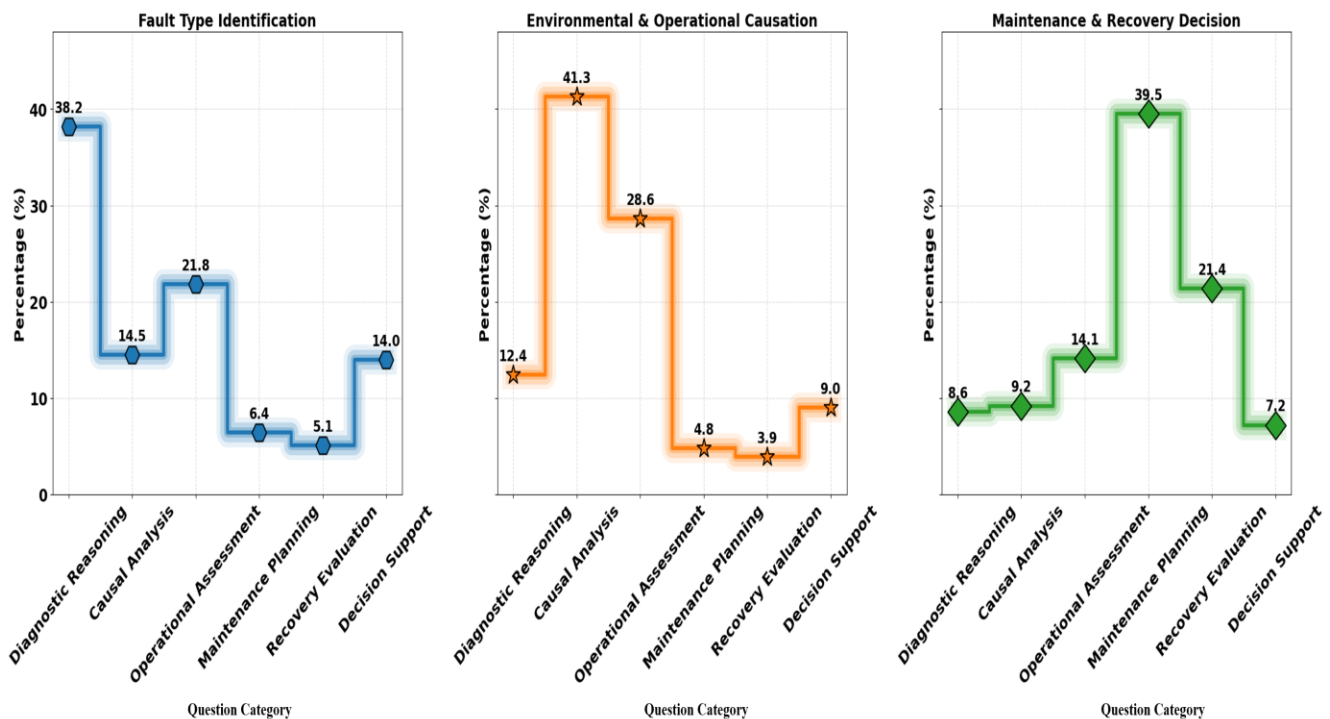


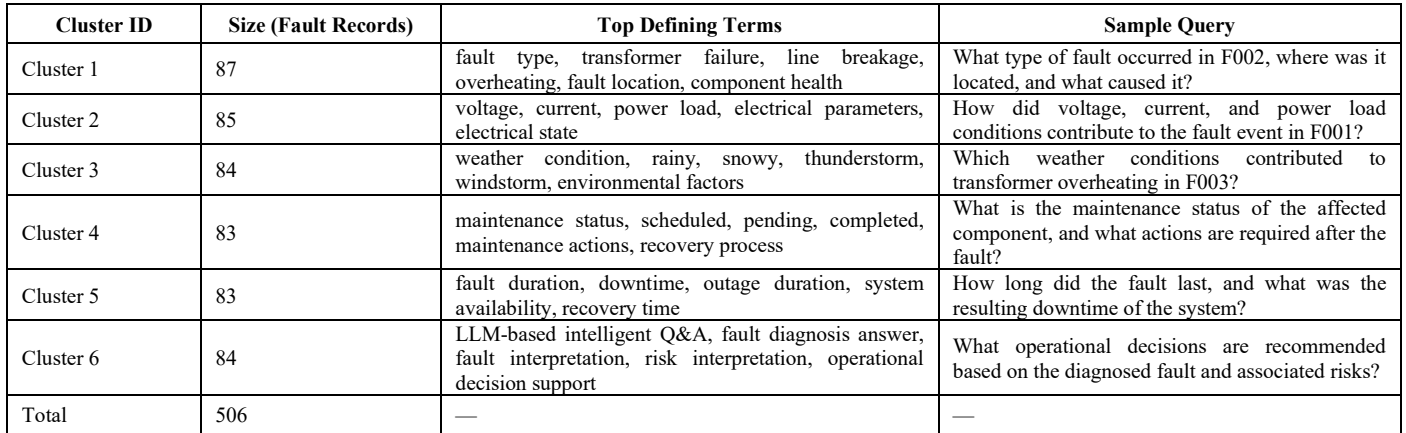
Fig. 5. Category-wise distribution of power-system fault reasoning tasks.

The dataset exhibits a well-balanced distribution across six defined reasoning categories. The questions generated from the texts were clustered into six clusters of semantic reasoning, based on the dominant information content and reasoning requirements of queries. There are a total of 87, 85, 84, 83, 83, and 84 pairs of Q&A generated in the six clusters, respectively, for a total of 506 pairs across all six clusters. Overall cluster proportions varied from 16.27% to 17.26%, suggesting a reasonably even balance of types of questions within the clusters that focused on fault diagnosis, electrical-state analysis, environmental and operational causes, maintenance and recovery, duration of fault, and operational decision support using LLM.

Table VI summarizes the properties of the dataset, and Fig. 6 outlines the Semantic Network of the principal entities related

to power system faults. Power system fault diagnosis, electrical system state analysis, power system condition evaluation, power system maintenance management, power system fault loss assessment, and fault analysis on decisions supported by LLM are the six main categories of questions on which the Power System Faults Dataset displays an even distribution. Such a distribution of question types allows rich reasoning abilities for fault diagnosis, maintenance planning, fault repair, system restoration, and operation decisions in the power system context. The generated question set was grouped into six semantic reasoning clusters according to the dominant information and reasoning requirements of the queries. The cluster sizes were relatively comparable, indicating a balanced distribution of question types.

*(IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 17, No. 9, 2026*



2) *Token-level F1-score*: Evaluates partial lexical overlap between generated and reference answers. Precision (P), recall (R), and F1 are calculated as given in Eq. (8)-(10):

$$P = \frac{|T_{\dot{y}} \cap T_y|}{|T_{\dot{y}}|} \quad (8)$$

$$R = \frac{|T_{\dot{y}} \cap T_y|}{|T_y|} \quad (9)$$

$$F1 = \frac{2PR}{P+R} \quad (10)$$

where, T_y and T_r represent the predicted and reference token sets, giving partial partial credit when the response contains only part of the expected information.

3) *BERTScore*: Measures contextual semantic similarity using the RoBERTa-large pretrained checkpoint as the contextual encoder. Pairwise cosine similarities between candidate token x_i and reference token y_j are calculated. The final BERTScore is reported as Eq. (11):

$$F1_{BS} = \frac{2P_{BS}R_{BS}}{P_{BS}+R_{BS}} \quad (11)$$

This accommodates technically correct power-system answers that use alternative wording while preserving semantic meaning. $2P_{BS}$ specifies the BERTScore precision. R_{BS} denotes BERTScore recall. $F1_{BS}$ represents the harmonic mean of precision and recall, indicating the overall contextual semantic similarity between the predicted and reference answers.

4) *PKG-Based Faithfulness Score (FS)*: Serves as the primary measure of factual grounding for the LLM-PKG architecture. Generated answers are decomposed into atomic claims ($C = \{c_1, c_2, \dots, c_m\}$) and compared against retrieved PKG facts ($K = \{k_1, k_2, \dots, k_r\}$). Claim-level faithfulness is calculated using Eq. (12).

$$FS = \frac{N_{supported}}{N_{claims}} \times 100 \quad (12)$$

environmental factors, or maintenance actions are directly supported by PKG evidence.

5) *Hallucination Rate (HR)*: Quantifies unsupported content generated by the LLM shown in Eq. (13):

$$HR = \frac{U}{A} \times 100 \quad (13)$$

where, U is the number of unsupported factual answer units and A is the total number of factual answer units. A lower rate indicates better grounding.

The Faithfulness Score (FS) represents the ratio of factual claims, stated in the answer generated by the model, that are backed up by the retrieved evidence/items from the PKG. Each answer is broken down into individual pieces of factual information (fingers) such as: fault type, equipment condition, environmental factor, operational condition, or recommended maintenance action. A claim is faithful if there exists evidence in the retrieved PKG triples or the related source record to the claim. The Faithfulness Score is therefore equal to the ratio of the number of claims that get supported and the total number of factual claims.

The Hallucination Rate is the percentage of factual statements that cannot be substantiated with the evidence from the PKG retrieved, or the corresponding source record. There is a possibility that one or more variables of the unsupported claim may be incorrect: the fault classified as the fault type which is not the true cause; the operating condition which is not the true cause; an environmental cause not observed in the evidence; a maintenance recommendation which is not supported by the available power-system data. Therefore, a lower HR means a more solid basis of facts.

In addition to automated grounding measures, domain-expert assessment is used because numerical and semantic text measures are not sufficient to determine if the response is technically correct with respect to decision-making in power systems. Each answer needs to be evaluated by a specialist for its factual accuracy, relevance to the specific electrical circumstance for which the inquiry was made, consistency among the retrieved evidence, and engineering suitability of the suggested actions. The expert assessment considers and reports

on linguistic similarity without also reporting on technical content and validity of the domain-specific content, which are reported separately from it. The results of all three tasks are shown in Table VII.

TABLE VII. DISTRIBUTION OF QUESTION CATEGORIES ACROSS LLM-PKG-BASED POWER SYSTEM FAULT DATASET

Metric	Fault Type Identification	Environmental & Operational Causation	Maintenance & Recovery Decision
EM	62.4%	58.7%	55.9%
F1 Score	68.1%	65.3%	63.4%
BERT Score F1	71.5%	69.2%	67.8%
Faithfulness Score	64.8%	61.5%	59.7%
Hallucination Rate	18.6%	21.4%	24.1%

The proposed model gets the best score in Fault Type Identification because fault types have a structured and clear pattern; its performance decreases slightly in Environmental and Operational Causation due to a complex system because of the dependency on various factors (e.g., weather, load). The scores in Maintenance and Recovery Decision tasks are lower as the need for high-order deduction across various system characteristics increases; the consistently good results in BERTScore F1 scores (ranging from 67.8% to 71.5%) confirm a high semantic similarity to ground truth; Faithfulness scores in these tasks confirm the excellent effect of PKG fact grounding. The hallucination rate (ranging from 18.6% to 24.1%) shows a relatively low rate thanks to knowledge-graph fusion. To assess the contribution of the main elements of the suggested LLM-PKG fusion framework, an ablation study was carried out. The whole model was contrasted with variations that did not include PKG retrieval and reasoning, BERT-based semantic extraction, Z-score normalization, or an LLM-only baseline. The findings in Table VIII show that eliminating these elements increases hallucination while decreasing answer accuracy, fidelity, and semantic similarity. Specifically, a greater performance loss results from the elimination of PKG retrieval and reasoning, underscoring the need for domain-grounded knowledge for trustworthy power-system Q&A and decision support.

TABLE VIII. ABLATION ANALYSIS OF LLM-PKG COMPONENTS FOR POWER-SYSTEM Q&A

Model / Variant	Z-score	BERT Extraction	PKG Retrieval & Reasoning	LLM Fusion	EM (%)	F1 (%)	BERTScore F1 (%)	Faithfulness (%)	Hallucination (%)
Rule-based Q&A	X	X	X	X	43.5	49.1	52.8	55.6	34.2
PKG-only retrieval	X	X	✓	X	46.8	52.4	56.7	57.2	31.6
LLM-only baseline	X	X	X	X	51.2	57.6	61.5	53.8	28.1
Conventional RAG	✓	✓	X	✓	56.7	62.5	66.1	59.3	23.7
Without PKG retrieval/reasoning	✓	✓	X	✓	54.6	60.8	64.3	56.9	25.3
Without BERT extraction	✓	X	✓	✓	58.9	64.2	67.8	60.7	21.8
Without Z-score normalization	X	✓	✓	✓	59.8	65.4	69.1	61.9	20.7
Full LLM-PKG	✓	✓	✓	✓	62.4	68.1	71.5	64.8	18.6

The advancement of intelligent power system fault diagnosis and Q&A systems leverages KG-based reasoning as a means to

enrich structure information retrieval and enhance explainability. Numerous works prove that KG-augmented

models achieve high accuracy in the task of fault retrieval and entity-relation extraction from the domain-specific query based on KG, but struggle with generalization capability, dependence on labeled data quality, and scalability issues in real-world large, complex power grid scenarios [14]. KG embedding and KG-reasoning models can further obtain satisfactory results regarding prediction performance; however, this comes with computational expense and lower efficiency in massive knowledge contexts [13].

Smart grid operation with LLMs makes considerable improvements in reasoning efficiency but is deficient in hallucination and shallow-grounded structured domain knowledge to guarantee safe decision-making. For this reason, the LLM-PKG fusion model is designed by fusing context-language modeling with structured knowledge graphs that enhance the factual grounding of predictions and their inferential logic correctness. Extensive experiments verify the model's robust effectiveness across all benchmarks that reflect semantic understanding, strengthen knowledge grounding, boost response faithfulness, and mitigate hallucination. The obtained results validate and prove the effectiveness of the proposed framework, LLM-PKG, in providing truthful and trustworthy answers to intelligent question answering in a power system with context. The model performs well in fault type identification with fixed pattern and relatively poor performance in complex logical inference with dependent variables as multiple factors, respectively. Several limitations also existed in terms of limitations of data scale, extraction noise in the construction of PKG, and insufficient temporal inference abilities. Future directions involve real-time and scalable applications, graph learning, and multi-modality. The proposed model is designed to be a tool for the decision-making process of power-system engineers and operators. It offers a natural language user interface to which one can pose questions on conditions for the electrical operation, on faults, on the environment, and on maintenance. The PKG now structures electrical evidence, and the LLM is used to interpret and generate the evidence to be formulated into a response. The model thus does not stand alone as a protection or control tool, but is designed to be advisory in nature.

V. CONCLUSION

With the gradual development of the power system, the research and application of intelligent Q&A models have become more and more important in the field of power system fault diagnosis and effective operation decision support. Traditional methods are difficult to reason with and lack structured reasoning ability, and standalone LLMs have limitations in factually grounded reasoning and hallucination. To overcome these limitations, the proposed LLM-PKG fusion model combines z-score normalization, BERT-based entity and relation extraction, and Power Knowledge Graph-guided retrieval to boost knowledge-driven reasoning. The experimental results for all tasks (Fault Type Identification, Environmental and Operational Causation, and Maintenance and Recovery Decision) are 71.5%, 69.2%, and 67.8%, respectively, with the Hallucination Rate of <25% for all of them (18.6%, 21.4%, and 24.1%). The results demonstrate that structured graph knowledge, alongside LLM reasoning, has a substantial impact on the semantic consistency, factual

correctness, and explainability of questions and answers within a power system. The electrical fault-analysis experiments illustrate how the model can link operating measurements to environmental conditions, structure, and equipment states, and incorporate fault knowledge and natural-language decision support. The model would serve as a practical basis for intelligent fault diagnosis and smart grid decision support, and future research involves real-time data integration, temporal reasoning, and the scaling of graph learning for widespread use.

ACKNOWLEDGMENT

Supported by the Science and Technology Projects of China Southern Power Grid Corporation, Project Number : 036000KC24080003 (GDKJXM20240652).

REFERENCES

- [1] Abdelkader S, Amissah J, Abdel-Rahim O. Virtual power plants: an in-depth analysis of their advancements and importance as crucial players in modern power systems. *Energy Sustain Soc* 2024;14(1):52. <https://doi.org/10.1186/s13705-024-00483-y>
- [2] Zeb N, Zulkifli SA, Sepeeh MS, et al. Generalized Unified Power Flow Controller for Low Oscillation Damping for Smart Grid Using Adaptive Centralized Control. *J Electr Eng Technol* 2026;21(3):2207–2222. <https://doi.org/10.1007/s42835-026-02590-6>
- [3] Rosati R, Romeo L, Cecchini G, et al. From knowledge-based to big data analytic model: a novel IoT and machine learning-based decision support system for predictive maintenance in Industry 4.0. *J Intell Manuf* 2023;34(1):107–121. <https://doi.org/10.1007/s10845-022-01960-x>
- [4] Gao M, Li M, Ji T, et al. Key technologies of intelligent question-answering system for power system rules and regulations based on improved BERTserini algorithm. *Processes* 2023;12(1):58. <https://doi.org/10.3390/pr12010058>
- [5] Gupta R, Chaturvedi KT. Adaptive energy management of big data analytics in smart grids. *Energies* 2023;16(16):6016. <https://doi.org/10.3390/en16166016>
- [6] Yang S. Transformer model optimization method for multi-modal data fusion. *Int J Adv Comput Sci Appl* 2025;16(7). <https://doi.org/10.14569/IJACSA.2025.0160708>
- [7] Zhang P, Li J, Chen S, et al. Multi-feature fusion for relation extraction using entity types and word dependencies. *Int J Adv Comput Sci Appl* 2023;14(7). <https://doi.org/10.14569/IJACSA.2023.0140731>
- [8] Kosasih EE, Margaroli F, Gelli S, et al. Towards knowledge graph reasoning for supply chain risk management using graph neural networks. *Int J Prod Res* 2024;62(15):5596–5612. <https://doi.org/10.1080/00207543.2022.2100841>
- [9] Kim J, Podlasek A, Shidara K, et al. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Sci Rep* 2025;15(1):39426. <https://doi.org/10.1038/s41598-025-22940-0>
- [10] Qu H, Zhang Y, Liang K, et al. A knowledge-graph-driven method for intelligent decision-making on power communication equipment faults. *Electronics* 2023;12(18):3939. <https://doi.org/10.3390/electronics12183939>
- [11] Liang K, Zhou B, Zhang Y, et al. A multi-entity knowledge joint extraction method of communication equipment faults for Industrial IoT. *Electronics* 2022;11(7):979. <https://doi.org/10.3390/electronics11070979>
- [12] Liu J, Huang W, Li T, et al. Cross-domain knowledge graph chiasmal embedding for multi-domain item-item recommendation. *IEEE Trans Knowl Data Eng* 2022;35(5):4621–4633. <https://doi.org/10.1109/TKDE.2022.3151986>
- [13] Meng F, Yang S, Wang J, et al. Creating a knowledge graph of electric power equipment faults based on a BERT-BiLSTM-CRF model. *J Electr Eng Technol* 2022;17(4):2507–2516. <https://doi.org/10.1007/s42835-022-01032-3>
- [14] Gao F, Zhang L, Wang W, et al. Named entity recognition for equipment fault diagnosis based on RoBERTa-wwm-ext and deep learning

- integration. Electronics 2024;13(19):3935. <https://doi.org/10.3390/electronics13193935>
- [15] Li C, Wang B. A knowledge graph method towards power system fault diagnosis and classification. Electronics 2023;12(23):4808. <https://doi.org/10.3390/electronics12234808>
- [16] Hou R, Zhang Y, Ou Q, et al. Recommendation method of power knowledge retrieval based on graph neural network. Electronics 2023;12(18):3922. <https://doi.org/10.3390/electronics12183922>
- [17] Imanifard A, Majidi B, Shamisa A. SmartGridAgent: an educational framework for reliable digital twin-based smart grid workforce training with locally hosted LLMs. Smart Grids Sustain Energy 2025;10(2):41. <https://doi.org/10.1007/s40866-025-00274-0>
- [18] Huang J, Wang M, Cui Y, et al. Layered query retrieval: An adaptive framework for retrieval-augmented generation in complex question answering for large language models. Applied Sciences 2024;14(23):11014. <https://doi.org/10.3390/app142311014>
- [19] Osipjan A, Khorashadizadeh H, Kessel AL, et al. Graphtrace: A modular retrieval framework combining knowledge graphs and large language models for multi-hop question answering. Computers 2025;14(9):382. <https://doi.org/10.3390/computers14090382>
- [20] Zhang G, Liu J, Zhou G, et al. Question-directed reasoning with relation-aware graph attention network for complex question answering over knowledge graph. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2024;32:1915–1927. <https://doi.org/10.1109/TASLP.2024.3375631>
- [21] Zhang X, Ding R, Wang Z, et al. Power grid fault diagnosis model based on the time series density distribution of warning information. International Journal of Electrical Power & Energy Systems 2023;146:108774. <https://doi.org/10.1016/j.ijepes.2022.108774>