Renewable Energy and Power Quality Journal (RE&PQJ)

ISSN 2172-038 X, Volume No.22, September 2024

Research on LLM Method for Knowledge-Assisted Generation of Power Standards

Xiaoxuan Fan[1,*], Sai Zhang[1], Xiao Liang[1], Zhihao Wang[1], Tao Peng[2]

[1]State Grid Laboratory of Grid Advanced Computing and Applications, China Electric Power Research Institute Co., Ltd., Beijing, China

[2]State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China

[*]Corresponding author email: fanxiaoxuan@geiri.sgcc.com.cn

Abstract. With the rapid development of large language models, their application in various industries is becoming more and more common, especially in the power sector. Electricity power standard knowledge, as an industry specification for electric power equipment, engineering construction and operation management, can guide product production and engineering practice to improve production efficiency and quality level. In order to help solve the problem of low efficiency of electric power standard knowledge supply and assist the generation of electric power standard knowledge, we proposes an electric standard knowledge auxiliary generation method based on retrieval enhancement generation, which uses the retrieval model and evaluation model to provide large language model with accurate external knowledge, and then uses the generation model to provide a more efficient and accurate solution for the generation of standard knowledge in the electric industry. The experimental results indicate that the approach introduced in this paper can effectively generate accurate electricity-related knowledge and holds significant practical value.

Key words. Large language model, Retrieval, Power standards knowledge, Knowledge-assisted, Knowledge generation

1. Introduction

Standards knowledge documents in the power sector are voluminous and dispersed due to the wide range of technical requirements, operating procedures and safety norms, and often contain redundant, repetitive and contradictory information, making it difficult for practitioners to efficiently access and accurately understand them. The complexity not only increases the difficulty of implementation, but may also lead to deviations in the application of standards, affecting the safety, reliability and efficiency of the power system [1,2], as well as hindering technological innovation and international cooperation.

Assisted generation of knowledge about standards in the power sector can effectively improve the accessibility and applicability of standards, ensure consistency in their understanding and implementation, enhance the safety, efficiency and compatibility of power systems [3-6], and promote technological innovation and international cooperation. Knowledge of power standards often covers multiple aspects of equipment design, installation, operation, and maintenance, and the amount of information is large and complex [7]. By assisting in the generation of this standards knowledge, core concepts and key information can be organized and distilled into clear, structured, easy to understand and implement standardized documents. This process can eliminate information redundancy and contradictions, make the standard terms clearer, and help power industry practitioners accurately master and apply these standards, thus reducing operational errors and safety accidents caused by understanding bias [8].

Standard knowledge-assisted generation technologies in the electric power field are rapidly developing, mainly through the use of advanced technologies such as artificial intelligence, machine learning and big data analysis to automate the organization, extraction and optimization of complex standard documents [9]. These technologies are able to distill key concepts and core information from huge documents, generate structured and easy to understand standardized documents, significantly improve the efficiency of practitioners in accessing and applying standards, and reduce the deviation in understanding and implementation, thus enhancing the safety and operational efficiency of the power system. In addition, these assistive generation tools enable real-time updating and dynamic maintenance of standards documentation, ensuring that practitioners are always using the most up-to-date and accurate information, and supporting continuous technological innovation. However, these techniques require large amounts of high-quality data to train models, and data in the power sector may be insufficient or of questionable quality, affecting the accuracy and comprehensiveness of the documents generated. In addition, algorithms may be biased and discriminatory, resulting in less than objective and unbiased documents being generated.

In the current research on knowledge-assisted generation methods, common methods include knowledge graph-based generation models, knowledge distillation techniques, and knowledge-guided generation using pre-trained language models such as Generative Pre-Trained Transformer (GPT) and BERT. The main shortcomings of these methods are: the construction and update frequency of knowledge graphs are low, which may result in the generation of content using outdated or inaccurate information; knowledge distillation may lose some details, resulting in a decline in the quality and richness of the generated results; although pre-trained models can capture a lot of knowledge, they still have limitations in specific context understanding and knowledge integration, which may result in the generated content lacking relevance and coherence. These challenges limit the effectiveness and reliability of knowledge-assisted generation methods in practical applications.

With the recent development of large language model technology, there are great advantages of using large language models (LLM) for standard knowledge-assisted generation in the power domain. Large language models typically have stronger language comprehension and generation capabilities, and can handle complex semantic information and contexts to generate more accurate and fluent documents [10-13]. In addition, large language models can learn knowledge from large-scale data, and therefore can cover a wider range of domains and standard content to generate more comprehensive documents. However, there are some drawbacks to using large language models. First, these models usually require a lot of computational resources and time for training and tuning, and thus are expensive to develop and deploy. In addition, due to the large number of parameters in large models, their explanatory and interpretable nature is low, and it may be difficult to understand the specific reasons and logic of the results they generate, making it more difficult to review and validate documents. At the same time, LLMs may be affected by the bias of training data and missing data, leading to errors or bias in the generated documents.

The advantage of using retrieval-enhanced generation of large language modeling techniques for assisted generation of standard knowledge in the electric power domain lies in combining the advantages of LLM and information retrieval. The technique is able to utilize the semantic comprehension and generation capabilities of large-scale language models while combining them with information retrieval techniques to retrieve relevant information from massive amounts of data based on user-supplied queries or keywords, thereby generating more accurate and targeted documents [14]. This approach not only improves the quality and accuracy of documents, but also enhances efficiency as it can quickly locate and extract content relevant to the query, avoiding unnecessary information overload. In addition, the retrieval-enhanced large language model technology is able to make real-time adjustments and optimizations based on user feedback and needs, thus continuously improving the applicability and user satisfaction of the generated documents [15-16].

This paper presents a traditional retrieval-augmented generation (RAG) method based on retrieval and evaluation models. RAG combines retrieval and generation capabilities, introducing external knowledge for text sequence generation tasks. Secondly, frequent real-time retrieval may lead to system performance degradation. First, the LLM is fine-tuned using multi-task learning, and the keywords of the question are analyzed to generate corresponding results. When retrieval is necessary, relevant paragraph blocks are extracted from external authoritative knowledge documents, and an evaluation model is used to compute the relevance scores between these paragraph blocks and the question. Ultimately, the retrieval result with the highest relevance score is chosen, combined with the question as input to the LLM, to generate the final answer. Compared to traditional methods, using large language models for knowledge-assisted generation offers significant advantages, including strong contextual understanding and rich knowledge representation, resulting in more coherent and relevant content. Additionally, the flexibility and adaptability of these models allow for quick responses to diverse input needs. The quality and diversity of generated results are generally higher, reducing redundancy and repetitiveness. Furthermore, this method simplifies the knowledge integration process by directly invoking relevant knowledge to enhance generation effectiveness, making it more prominent in practical applications.

The paper is structured as follows: chapter 2 is an introduction to the related work; chapter 3 is an introduction to the RAG algorithm in this chapter; chapter 4 is the experimental part of the paper, and finally, the paper is concluded.

2. Related Work

In recent years, open-architecture knowledge acquisition tools have been widely adopted in power systems, and knowledge engineering techniques have been introduced in many applications to solve scientific and engineering problems in the field of electric power [17-19]. However, the traditional informatization engineering products are still in the simple application of electric power data and knowledge and lack a comprehensive grasp of the knowledge system [3-4]. In order to accelerate the realization of business synergy and data coherence in China's electric power companies, experts and scholars have in recent years proposed technical routes and application cases based on domain knowledge mapping, an emerging cognitive methodology, in many fields such as electric power dispatching, operation and inspection, and marketing.

In power dispatching, the mainstream work of expert scholars still focuses on the combined application of natural language processing (NLP), automatic speech recognition (ASR) and DKG technologies [20]. In power operation and inspection, experts and scholars are centered on power equipment, and the research work related to domain knowledge graph is relatively deep in the segmented business points [21,22]. In terms of power marketing, the technology combination based on ASR,

NLP and DKG realizes the overall enhancement of intelligent retrieval, intelligent Q&A, and active outbound calling capabilities in the power customer service business [23]. However, knowledge graphs in the electric power domain still suffer from differences in the standardization and consistency of knowledge representation, and data from different sources may adopt different standards and formats, making it difficult to integrate and interact effectively.

To address the above problems, with the development of big language modeling in recent years is expected to be effectively solved. Due to its increasingly excellent capability in language understanding and analysis, many researchers are also trying to use retrieval enhancement based big language modeling techniques. The current mainstream retrieval enhancement methods mainly follow the process of retrieval followed by generation, e.g., given a problem, the model first retrieves a set of relevant documents from an external knowledge base. Then, its internal knowledge is combined with these retrieved documents to generate an answer. Research on this process falls into three main categories: improved retrievers [24] or readers [25] or training combining these two components [26-28]. However, RAG techniques require the need for continuous real-time retrieval, which increases the computational and time costs of the system and reduces the speed and efficiency of the generated text.

In our paper, we first analyze the result analysis of the large language model in generating electric power standard knowledge, determine when retrieval is needed in the process of generating the result, call the pre-trained retrieval model on demand by designing the reasoning process to provide external knowledge for the large language model, use the pre-trained evaluation model to evaluate the scores of the retrieved external knowledge, and select the results with the highest scores as inputs to the LLM, which assists in the electric power standard knowledge Auxiliary Generation of Electricity Standard Knowledge.

3. Methodology

Figure 1 illustrates the overall design framework of this paper. We first employ a multi-task learning model to pre-train the large language model (LLM), using both original and standardized data for the training and testing of sub-tasks. The fine-tuning process for each task utilizes the Low-Rank Adaptation (LoRA) method to tailor the LLM for tasks related to power standard knowledge. LoRA is a fine-tuning approach for training large language models that reduces model complexity, thereby enhancing training efficiency and improving the model's generalization ability.

Subsequently, during the execution of RAG, the data is input into the pre-trained LLM to generate power knowledge. Based on the generation results, we determine whether external knowledge retrieval is necessary. If the generated results meet the requirements, they are directly output; otherwise, the retrieval process is initiated. During retrieval, the program searches for documents related to power standard knowledge and combines multiple retrieval results with the questions, which are then re-input into the LLM for question-answer generation. Finally, the generated results are scored and ranked, with the top-ranked results selected as the final generated knowledge for the question.

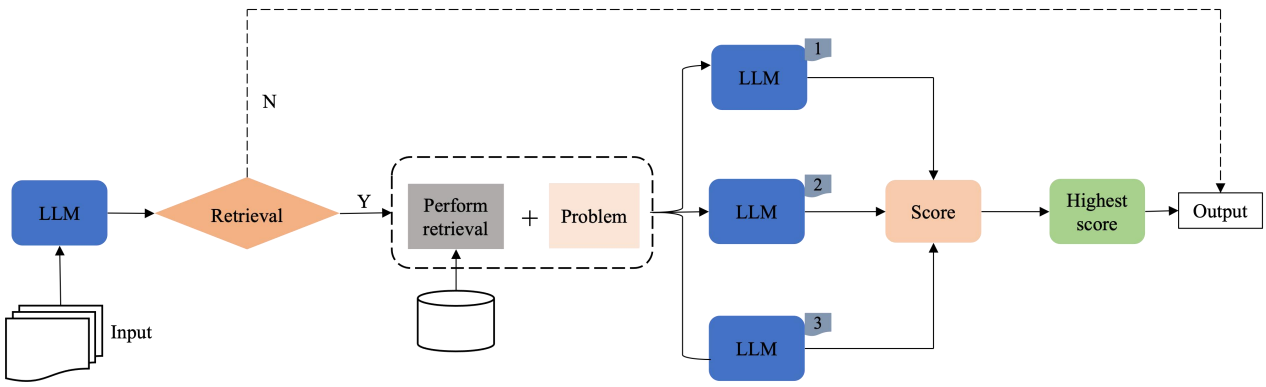


Figure 1. The overall framework flow of our method.

A. Fine-tuning the Knowledge of Electrical Standards LLM

The basic definition of Multi-task Learning (MTL) refers to simultaneously learning multiple tasks from different domains and enhancing generalization ability through the domain information of specific tasks. This paper refers to refers to the fact that given m learning tasks, which are related but not identical to each other or a subset of them, it helps to improve the learning of a particular model by using the knowledge contained in all m tasks. The LLM fine-tuning framework based on multi-task learning designed in this paper is shown in Figure 2.

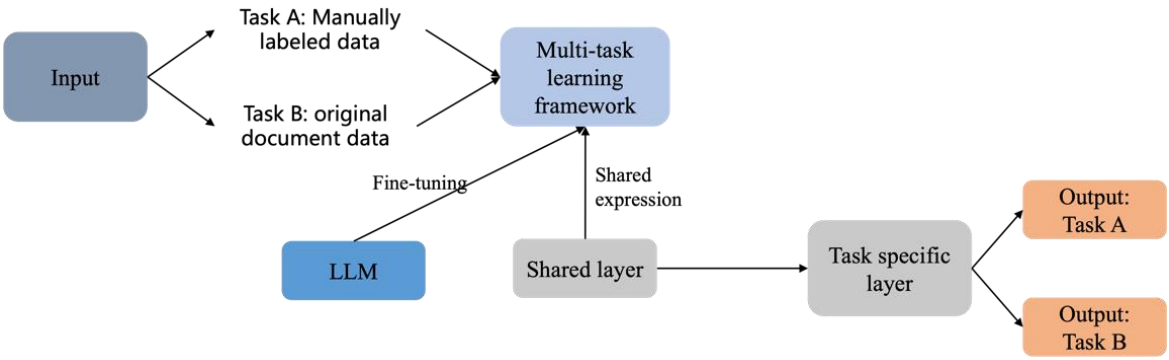


Figure 2. The diagram of large model fine-tuning based on multi-task learning.

The advantage of MTL is its ability to enhance the contextual understanding and adaptability of LLM. By processing multiple related tasks through simultaneous training, data resources can be utilized more efficiently and the impact of data scarcity on model performance can be reduced. In addition, MTL can facilitate knowledge transfer and migration between tasks by sharing the underlying feature representations, thus improving the generalization ability and robustness of the model. The MTL are mainly classified into hard-parameter shared learning and soft-parameter shared learning. In our paper, we use the soft-parameter shared model structure to design the dual-task learning, where one of the tasks is to predict the model of entities and relationships in the electric power standard documents, and the other task is to predict answer statements based on questions or prompts. The two tasks share parameters at the bottom level except for their own unique parameters, but at the top level they have their own parameters. Compared with the structure of hard parameter sharing, soft parameter sharing has more relaxed constraints and can learn to get better policy models, whose task prediction model is shown in Figure 3.

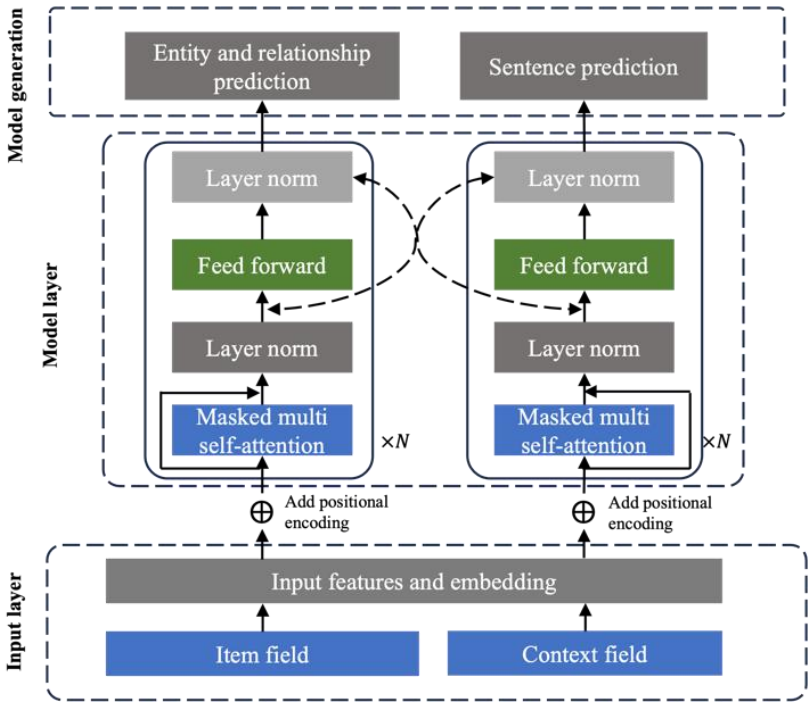


Figure 3. The Large model training and prediction graph based on multi-task learning.

Multi-task learning trains multiple related tasks together with a total loss function that is a weighted summation of the loss functions of each task:

$$L_{total} = \sum_k w_k(t) L_k \tag{1}$$

As opposed to the direct summation of the losses, this loss function weights the losses for each task. This approach allows us to manually adjust the level of importance of each task, and the weight w_k will be adjusted according to the stage of learning of different tasks, the difficulty of learning, and even the effect of learning. Here k denotes the task index, t denotes that it is currently at step t of network training, L_k denotes the input, and w_k should be chosen in such a way that it balances the training of each task, so that each task receives a beneficial boost. The purpose of multi-task learning is to find the optimal

parameter of the model, which is said to be Pareto-optimal if any change in that parameter leads to an increase in the loss function for a given task. Pareto optimality implies that the loss of each task is relatively small and the performance of one task cannot be sacrificed for the performance improvement of another. Therefore, the optimization of the loss function needs to consider the weight of the loss function of each task to regulate the balance of the multi-task learning process, reduce the conflict between different tasks, and then improve the effect of multi-task learning.

The specific design method of the loss function is introduced. In our paper, we use the method of Gradient Normalization for the calculation of the loss function, because Gradient Normalization considers both the magnitude of the loss and the training speed of different tasks, so it is effective in its implementation. Gradient Normalization defines the Gradient Loss, which measures how good the weights $w_k(t)$ of the loss for each task are, Gradient Loss is a function of the weights $w_k(t)$. The weight $w_k(t)$ for each task is a variable and w is also updated by gradient descent. In order to represent the Gradient Loss as a function about the weights of the losses, some variables are first defined to measure the magnitude of the losses of the tasks,

$$G_w^{(k)}(t) = \left\| \nabla_w w_k(t) L_k(t) \right\|_2 \tag{2}$$

$$\overline{G}_w(t) = E_{task}\left[G_w^k(t) \right] \tag{3}$$

Where, $G_w^{(k)}(t)$ is the value of the gradient normalization of task k , which is the product of the weight $w_k(t)$ of task k and $ossL_k(t)$. The $L2$ paradigm of the gradient over the parameter w . $G_w^{(k)}(t)$ measures the magnitude of the sub-task $loss$; and $\overline{G}_w(t)$ is the value of the global gradient normalization, i.e., the expectation of the normalized value of the gradient for all tasks, which can be be realized by averaging all $G_w^{(k)}(t)$. In addition, this paper defines some variables to measure the learning speed of the tasks,

$$\tilde{L}_k(t) = L_k(t) / L_k(0) \tag{4}$$

$$r_k(t) = \tilde{L}_k(t) / E_{task}\left[\tilde{L}_k(t) \right] \tag{5}$$

$L_k(0)$ and $L_k(t)$ represent the loss at step 0 and step t of subtask k , respectively. $\tilde{L}_k(t)$ measures the training familiarity of the inverse of subtask k to a certain extent, and the larger $\tilde{L}_k(t)$ is, the slower the network is trained. $E_{task}\left[\tilde{L}_k(t) \right]$ denotes the expectation of the training speed of the inverse of each subtask, and $r_k(t)$ is the relative inverse training speed of the task. training speed, and a larger value indicates that task k is trained slower over all task retraining. The final Gradient Loss is:

$$L_{grad}(t, w_k(t)) = \sum_k \left| G_w^{(k)}(t) - \overline{G}_w(t) \times \left[r_k(t) \right]^{\alpha} \right|_1 \tag{6}$$

Where, α denotes the hyperparameter and $\overline{G}_w(t) \times r_k(t)$ denotes the ideal gradient normalized value. After calculating the Gradient Loss, $w_k(t)$ is updated by the following function:

$$w_k(t+1) = w_k(t) + \lambda * Gradient(GL, w_k(t)) \tag{7}$$

GL refers to Gradient Loss, and λ is an empirical hyperparameter.

B. Inference

In this section, we first define the variables required for inference, the input is x, and y is the multiple text paragraphs generated by the model $y = [y_1, y_2, \cdots, y_T]$, where y_t denotes the sequence of tokens of the t paragraph generated, and in this paper, we have also designed fields such as Fully supported, Partially supported, and so on, respectively, to express the model-generated answer's trustworthiness. Algorithm 1 is the process of reasoning based on external knowledge for LLM designed in this paper. For each x and the previous generated result $\boldsymbol{y}_{<t}$, the retrieval model checks the token to confirm whether the LLM needs to be retrieved. If retrieval is not required, the LLM predicts the next output segment. If retrieval is required, the model generates: an evaluation token for assessing the relevance of the retrieved article, the next response segment, and a comment token for assessing whether the information in the response segment is supported by the article. Finally, all evaluation scores are summed up as the final evaluation of the currently generated paragraph, with the specific algorithmic flow shown in Algorithm 1.

Algorighm 1 LLM Inference

Require: Generate LM M , Retriever R , Large-scale passage collections $\{d_1, d_2, \cdots, d_N\}$

1: **Input:** input prompt x and preceding generation $y_{<t}$, **Output:** next output segment y_t

2: M predict the token f_{ret} for whether retrieval is required of given $(x, y_{<t})$

3: if f_{ret} == Yes then

4: Retrieve relevant text passages D using R given (x, y_{t-1})

5: M predicts s_{isrel} given x , d and y_t given x , d , $y_{<t}$ for each $d \in D$

6: M predicts s_{issup} and s_{isuse} given x , y_t and d for each $d \in D$

7: Rank y_t based on s_{isrel} , s_{issup} , s_{isuse}

8: else if f_{ret} == No then

M_{gen} predicts y_t given x

M_{gen} *predicts* s_{isuse} *given* x , y_t

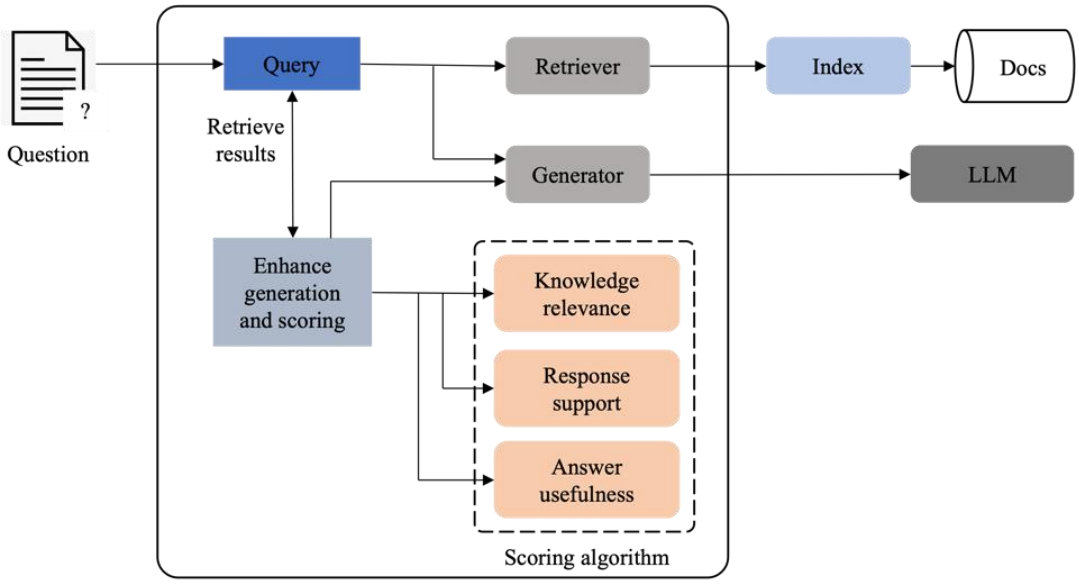


Figure 4. Retrieve the sorting chart.

C. Retrieval

In this section, we have designed efficient retrieval and judging algorithms from power standard knowledge documents of data sources as shown in Figure 4.

Retrieval algorithm

The main steps in the process of retrieving knowledge documents for power standards include: constructing an accurate semantic space, aligning query and documents and aligning the retriever and LLM.

Constructing accurate semantic spaces

During text generation using the LLM, task-specific fine-tuning of the Embedding model is crucial to ensure that the model understands the user's problem in terms of content relevance, and a model without fine-tuning may not be able to adequately meet the needs of the task of assisted generation of electric power standard knowledge. The Embedding model is fine-tuned by segmenting and chunking the standard knowledge documents in the electric power domain and embedding them into the semantic space together with query. The dataset used for Embedding model fine-tuning usually contains three main elements: questions (query), corpus and related documents. The model uses the query to identify relevant documents in the corpus. Then, the effectiveness of the Embedding model is measured based on the retrieval results. In this paper, we fine-tune the Embedding model by managing a corpus of standardized knowledge in the electricity domain, so that it is closely aligned with the task of assisting in the generation of electricity standardized knowledge.

Aligning query and documentation

The Retriever can encode both query and document using a single Embedding model, or use separate models for each. In addition, the original query may suffer from imprecise wording and lack of semantic information. Therefore, it is crucial to align the semantic space of the user query with the semantic space of the document. In this paper, we use rewriting query to maintain this alignment. Rewriting query means aligning the semantics of the query and the document by combining the original query with additional instructions.

D. Critic Model

The evaluation model is trained to generate tokens for evaluating the quality of the retrieved passages and the output of a given question. The initial model of the evaluation model can be any pre-trained LM, here the same model is chosen as the generator LM, whose goal is to maximize the expected value of the training dataset by maximizing the likelihood, which is based on the logarithm of the conditional probability of certain "reflection tokens".

$$\max_{c} \mathbb{E}_{((x,y),r)\sim D_{critic}} \log p_c(r \mid x,y), r \text{ for reflection tokens}. \quad (8)$$

E. Text Generation

The output of LLM is in fact the process of predicting the next token according to the hints and looping until all of them are completed. The prediction process goes through a series of complex operations and neural network processing, and finally outputs a list containing multiple possible next tokens and their probabilities, from which LLM selects the highest green token to output. In the retrieval of the power standard knowledge document, if the token [Fully supported] appears in the output at one time, then it means that the field Fully supported has the highest probability of outputting the token at the time of LLM reasoning. Therefore, when evaluating the score of the output, it can be calculated based on the probability of these tokens. In this paper, the results generated by LLM are judged from three aspects: knowledge relevance, response support, and response validity. The knowledge relevance is calculated as the ratio of the probability of the relevant token to the sum of the probabilities of the two tokens of the type,

$$s_{rel} = \frac{p_{rel}}{p_{rel} + p_{irrel}} \quad (9)$$

The corresponding support is calculated as the proportion of the probability of a "fully supported" token to the sum of the probabilities of the three types of tokens of this type, plus the proportion of the probability of a "partially supported" token to the sum of the probabilities of the three types of tokens of this type. which is multiplied by a weight of 0.5 to facilitate the calculation,

$$s_{sup} = \frac{p_{full\,\sup}}{S} + 0.5 \times \frac{p_{part\,\sup}}{2S} \quad (10)$$

$$S = \sum_{t \in \{full\,\sup, past\,\sup, no\}} p_t \quad (11)$$

The calculation of response validity is done by multiplying the probability of the five types of tokens of this type as a proportion of the total probability by the corresponding weights (ranging from -1 to 1, respectively) and then summing them up as follows,

$$s_{use} = \sum_{i}^{5} w_i \frac{p_i}{S}, w = \{-1, -0.5, 0, 0.5, 1\} \quad (12)$$

$$S = \sum_{t \in \{1,2,3,4,5\}} p_t \quad (13)$$

The final judgment score for an outcome generated by LLM is calculated as,

$$s_{final} = s_{rel} + s_{\sup} + s_{use} \quad (14)$$

where, the model selects the highest generated result of s_{final} as the final retrieved answer.

The generator model is trained using a corpus modified by tokens during the training phase of the generative model, during which the retrieved text blocks are masked for loss computation. The objective function describes the log-likelihood of maximizing the probability of M on the output y and the associated external knowledge r given the input x . The objective function is the log-likelihood of maximizing the probability of M on the output y and the associated external knowledge r .

$$\max_{M} \mathbb{E}_{(x,y,r)} \sim D_{gen} \log p_M(y, r \mid x) \quad (15)$$

4. Experiments

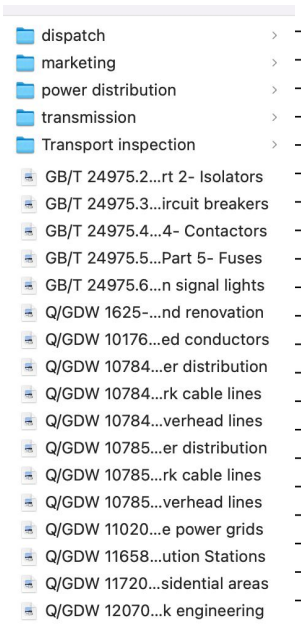
The experimental part of our paper includes preprocessing of the dataset, experiments comparing LLMs in multitask learning, and experimental analysis of answer generation using external information.

A. Datasets

Data collection and pre-processing

The dataset used in this paper mainly comes from two sources: first, a portion of the dataset from the State Grid Corporation of China's Grid Laboratory (NGLD). This dataset serves as a key reference for the company's formulation of medium- and long-term technical standard planning, as well as for the implementation of technical standards in production, operations, and management. It plays a crucial role in promoting the active standardization of both domestic and foreign advanced standards across various units. The dataset includes a total of 1,347 enterprise standards, 103 group standards, 2,235 industry standards, 2,597 national standards, and 368 international standards. The data is stored in document format, comprising 123 standardized knowledge documents that cover various aspects of the power sector, as shown in Figure 5.

Additionally, the comparative experimental dataset utilizes a publicly available web dataset called PopQA. This is a large-scale open-domain question-answering (QA) dataset that contains 14,000 entity-centered QA pairs. Each question is generated using templates by extracting knowledge tuples from Wikidata. Each question includes annotations for the original subject entity, object entity, and their relationship type, along with monthly page views from Wikipedia. The evaluation and generation models are fine-tuned using a multi-task learning approach.



2-3-19	2	GB/T 10623-2008
2-3-20	2	GB/T 11383-1989
2-3-21	2	GB/T 11457-2006
2-3-22	2	GB/T 11464-2013
2-3-23	2	GB 11643-1999
2-3-24	2	GB 11714-1997
2-3-25	2	GB/T 11804-2005
2-3-26	2	GB/T 11943-2008
2-3-27	2	GB/T 12212-2012
2-3-28	2	GB/T 12250-2005
2-3-29	2	GB/T 12402-2000
2-3-30	2	GB/T 12404-1997
2-3-31	2	GB/T 12405-2008
2-3-32	2	GB/T 12406-2022
2-3-33	2	GB/T 12407-2008
2-3-34	2	GB/T 12408-1990
2-3-35	2	GB/T 12409-2009
2-3-36	2	GB/T 12903-2008
2-3-37	2	GB/T 12905-2019
2-3-38	2	GB/T 12936-2007
2-3-39	2	GB/T 131-2006
2-3-40	2	GB/T 13306-2011

Figure 5. Standard Knowledge Document Dataset.

Since the standard knowledge in the electric power domain is widely derived from structured data such as traditional electric power knowledge engineering system, expert experience knowledge base, and semi-/unstructured data such as electric power standards, systems, laws, regulations, as well as the experience of experts and technicians, which involves a number of business domains, the dataset used in this paper can be classified into two parts of general and specialized knowledge according to the difference in the degree of reuse. Among them, some of the names of power equipment, voltage level, capacity and unit organization structure and other information in the distribution, transmission, marketing, safety and quality and other standard knowledge documents are in demand, which is called the general knowledge of the electric power field; for example, customer service in the field of customer name, electricity, electricity price and other information is relatively specialized in scheduling, operation and inspection, safety and quality and other standard knowledge involved in the probability of very small, or even will not appear at all, which is the special knowledge of the electric power field. The information is specialized in the field of electric power.

B. Performance of Large Models Trained by Multi-Task Learning vs. Single Large Model Training

In this paper, two Nvidia 4090 graphics cards are used for multi-card training using the DeepSeed command, and a single card with 24G of video memory is used for training the multi-task learning model. In this paper, in order to verify that the proposed multitask learning framework has significant effect on model fine-tuning, after considering and comparing the pre-trained large model we finally decided to use is llama2-Chinese-chat, which can be obtained from Hugging Face website. In this paper, we use single-task training and multi-task learning trained models for comparison experiments, in which the single-task model and multi-task learning model are analyzed from the results of sentence comprehension, entity and relation extraction, respectively.

In the comparison experiments of semantic understanding, two design several electric power domain problems to analyze the performance of single-task model and multi-task model on the same problems, and the results are shown in Table 1. In addition, a total of 15 topics were designed from three aspects of power distribution, transmission, and marketing, five topics from each aspect, respectively, to count the degree of semantic understanding of the single-task model and multi-task model. A response was judged to be correct if it contained more than 2⁄3 of the keywords of the exact answer. In the comparison experiments of entities and relations, this paper counts the percentage of the number of extractions of entities and relations to Groundtruth by the single-task model and the multi-task model respectively. The specific experimental results are shown in Table 1, where Groundtruth is a combination of expert experience and traditional algorithmic models to extract entities and relations from the dataset.

Table 1. The performance of single task model and multi-task training model on the same problem.

Sentence Comprehension		
Question 1: What is the insulation separation of conductors? **Question 2:** What is an intelligent coal-fired system?	Single task LLM	**Answer 1:** The insulation separation of conductors refers to the measures taken in the power system to prevent direct contact between conductors, in order to prevent accidental flow of current or failure of electrical equipment. **Answer 2:** Intelligent coal-fired system is a system that utilizes advanced information technology and automation control technology to improve the operational efficiency of coal-fired boiler systems, reduce energy consumption, and reduce pollution emissions.
	Multi task LLM	**Answer 1:** Insulation separation usually uses insulation materials or structures to isolate conductors from each other, preventing short circuits or arc faults caused by current between conductors. **Answer 2:** By comprehensively utilizing information technology, network technology, control technology, and related professional technologies, it integrates intelligent control systems, online monitoring and diagnosis, coal storage equipment and facilities, intelligent fuel management, etc., achieving an intelligent organic combination for the purpose of controlling the scientific and rational flow of coal.

Table 1 shows that the answer to the question "Question 1: What is the insulation of a conductor?" is not specific enough for the single-task model. The answer of the single-task model is not specific enough and does not include the keyword "insulation material", but the answer of the multi-task model is more specific and includes the keyword "insulation material" at the same time. For "Question 2: What is an intelligent coal combustion system?" The answer of the single-task model has the same problem and lacks the keywords "intelligent control system, intelligent fuel management", while the answer of the multi-task model is not comprehensive and specific but contains most of the keywords, so the results in Table 1 show that the model trained by multi-task learning can give a good answer to specific electric power problems in terms of sentence comprehension. on being able to give more specific answers to specific power questions. From Table 2, it can be seen that the single-task model gives generally less favorable answers than the results given by the multitask learning model in answering the 15 questions designed for the three areas of power distribution, transmission, and marketing. By in the extraction of entities and relationships from the electricity standard documents, the single-task model effect is still not as good as the performance of the multi-task learning model. In summary, the multi-task learning model outperforms the single-task model both in the understanding of contextual semantics and in the extraction of entities and relationships, therefore, the fine-tuned large language model of the multi-task learning framework designed in this paper can be effective for the assisted generation of knowledge of power standards.

Table 2. The comparative experiment on language understanding, entity and relationship extraction.

Tasks	power distribution	transmission	marketing	Entity extraction	Relationship extraction
Single Task LLM	5	5	6	0.678	0.659
Multi task learning for LLM	7	8	9	0.812	0.809

C. Experimental Results

In this section, we conduct comparative experiments based on NGLD (power standard knowledge documents) and the open-domain question-answering dataset PopQA using different LLMs (large language models). In the generation process of power standard knowledge, we establish two baselines to compare the model's performance with and without external retrieval, assessing the accuracy of the generated results. The large language models used include Llama2-13B, alpaca13B-lora, and Qwen1.5-14B-Chat. Additionally, we validate the framework proposed in this paper using Fine-tuning and retrieval methods based on the two datasets, with experimental results shown in Table 3. Since the text descriptions of power standard knowledge typically appear in short text form, we focus on conducting short text experiments with the LLMs. The evaluation criterion for the models is based on the inclusion of Groundtruth in the generated text to calculate accuracy, rather than strict text matching. Furthermore, some parameter settings in this study are as follows: Number of retrievals: 5, similarity threshold: 0.7, generation temperature: 0.7, maximum generation length: 150 tokens, learning rate: 1e-5, batch size: 32.

Table 3. Comparative experimental results of answer-assisted generation.

	LLM	NGLD	PopQA
Baselines without retrieval	Llama2-13B	18.2	21.3
	alpaca13B-lora	20.1	23.4
	Qwen1.5-14B-Chat	29.7	24.3
	Llama-3.1-8B-Instruct	32.1	29.8
Baselines with retrieval	Llama2-13B	44.8	44.2
	alpaca13B-lora	46.9	48.4
	Qwen1.5-14B-Chat	50.8	49.7
	Llama-3.1-8B-Instruct	52.1	50.7
Fine tuning and retrieval	Our method-7B	57.2	57.9
	Our method-13B	59.5	59.1

In Table 3, the "baseline without retrieval" refers to the use of LLMs, including Llama2-13B, alpaca13B-lora, Qwen1.5-14B-Chat and Llama-3.1-8B-Instruct, where the model relies solely on its internal knowledge base to generate answers to questions without utilizing external knowledge. The "baseline with retrieval" indicates the process where the LLM generates answers to questions and retrieves information from external documents as needed to assist in generating those answers.

From Table 3, it can be observed that when using the same LLM, there is a significant difference in the accuracy of the answers generated by the LLM between using external retrieval knowledge and not using external knowledge. For example, in the baseline experiments, the performance of the answer generation improved by 26.6%, 26.8%, 21.1% and 20% respectively for Llama2-13B, alpaca13B-lora, Qwen1.5-14B-Chat and Llama-3.1-8B-Instruct on the NGLD dataset when employing the retrieval process. On the PopQA dataset, the performance improvements for answer generation were 22.9%, 25.0%, 25.4% and 20.9% respectively with the retrieval process. Moreover, the proposed method improved the accuracy of the best results using retrieval in the baseline by an average of 6.2% and 7.8%. Additionally, fine-tuning the pre-trained retrieval and evaluation models when using external knowledge can further enhance the accuracy of the final answer generation. In summary, the proposed auxiliary generation method effectively assists LLMs in generating answers to questions.

5. Conclusions

The main work of this paper is to analyze the existing problems of LLM in assisting the generation of power standard knowledge, and use the RAG method to assist the generation of power standard knowledge, thereby improving the accuracy of answer generation. Different from the traditional RAG method, this paper first fine-tunes the retrieval model and evaluation model, determines the retrieval requirements of LLM through the reasoning process, uses the retrieval model to retrieve external power documents, and selects the retrieval paragraphs with the highest scores through the evaluation model to provide accurate external knowledge for LLM. Experiments have verified that the method proposed in this paper can effectively assist the generation of power standard knowledge.

References

[1] J.P. Burde, T. Wilhelm. Teaching electric circuits with a focus on potential differences. Physical Review Physics Education Research. 2020, 16(2), 020153. DOI: 10.1103/PhysRevPhysEducRes.16.020153

[2] K. Altmeyer, S. Kapp, M. Thees, S. Malone, J. Kuhn, et al. The use of augmented reality to foster conceptual knowledge acquisition in STEM laboratory courses-Theoretical background and empirical results. British Journal of Educational Technology. 2020, 51(3), 611-628. DOI: 10.1111/bjet.12900

[3] I.L. Alberts, L. Mercolli, T. Pyka, G. Prenosil, K. Shi, et al. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? European Journal of Nuclear Medicine and Molecular Imaging. 2023, 50(6), 1549-1552. DOI: 10.1007/s00259-023-06172-w

[4] O. Friha, M.A. Ferrag, B. Kantarci, B. Cakmak, A. Ozgun, et al. Llm-based edge intelligence: A comprehensive survey on architectures, applications, security and trustworthiness. IEEE Open Journal of the Communications Society. 2024, 5, 5799-5856. DOI: 10.1109/OJCOMS.2024.3456549

[5] N. Firoozeh, A. Nazarenko, F. Alizon, B. Daille. Keyword extraction: Issues and methods. Natural Language Engineering. 2020, 26(3), 259-291. DOI: 10.1017/S1351324919000457

[6] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, et al. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering. 2024, 36(7), 3580-3599. DOI: 10.1109/TKDE.2024.3352100

[7] K. Janowicz, S. Gao, G. McKenzie,Y. Hu, B. Bhaduri. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science. 2020, 34(4), 625-636. DOI: 10.1080/13658816.2019.1684500

[8] T. Nguyen Quang, T.O. Nguyen. Language Knowledge-Assisted in Topology Construction for

Skeleton-Based Action Recognition. Proceedings of the 12th International Symposium on Information and Communication Technology. 2023, 443-449. DOI: 10.1145/3628797.3629008

[9] J.K. Kim, M. Chua, M. Rickard, A. Lorenzo. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. Journal of Pediatric Urology. 2023, 19(5), 598-604. DOI: 10.1016/j.jpurol.2023.05.018

[10] C. Peng, F. Xia, M. Naseriparsa, F. Osborne. Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review. 2023, 56(11), 13071-13102. DOI: 10.1007/s10462-023-10465-9

[11] R. Venkatakrishnan, E. Tanyildizi, M.A. Canbaz. Semantic interlinking of Immigration Data using LLMs for Knowledge Graph Construction. WWW '24: Companion Proceedings of the ACM on Web Conference 2024. 2024, 605-608. DOI: 10.1145/3589335.3651557

[12] L.Y. Yang, C. Lv, X. Wang, J. Qiao, W.P. Ding, et al. Collective entity alignment for knowledge fusion of power grid dispatching knowledge graphs. IEEE/CAA Journal of Automatica Sinica. 2022, 9(11), 1990-2004. DOI: 10.1109/JAS.2022.105947

[13] L.P. Meyer, C. Stadler, J. Frey, N. Radtke, K. Junghanns, et al. Llm-assisted knowledge graph engineering: Experiments with chatgpt. First Working conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow. Wiesbaden: Springer Fachmedien Wiesbaden. 2023, 103-115. DOI: 10.1007/978-3-658-43705-3_8

[14] J. Varas, B.V. Coronel, I. Villagrán, G. Escalona, R. Hernandez, et al. Innovations in surgical training: exploring the role of artificial intelligence and large language models (LLM). Revista do Colégio Brasileiro de Cirurgiões. 2023, 50, e20233605. DOI: 10.1590/0100-6991e-20233605-en

[15] Y.C. Zhuang, Y. Yu, K. Wang, H. Sun, C. Zhang. Toolqa: A dataset for llm question answering with external tools. Advances in Neural Information Processing Systems. 2023, 36, 50117-50143. DOI: 10.48550/arXiv.2306.13304

[16] G. Sun, W. Jin, S. Xu, L.X. Yang, W.J. Song. Research on Constructing a Knowledge Graph for Risk-Aware Electricity Marketing Events. 3rd International Conference on Digital Economy and Computer Application (DECA 2023). Atlantis Press. 2023, 677-688. DOI: 10.2991/978-94-6463-304-7_71

[17] R.K. Mishra, H. Raj, S. Urolagin, J. Jothi, N. Nawaz. Cluster-based knowledge graph and entity-relation representation on tourism economical sentiments. Applied Sciences. 2022, 12(16), 8105. DOI: 10.3390/app12168105

[18] J. Chen, Y. Geng, Z. Chen, J.Z. Pan, Y. He, et al. Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. Pro-ceedings of the IEEE. 2023, PP(99), 1-33. DOI: 10.1109/JPROC.2023.3279374

[19] A.D. Pace, A. Tommasel, H.C. Vazquez. The JavaScript Package Selection Task: A Comparative Experiment Using an LLM-based Approach. CLEI Electronic Journal. 2023, 27(2), 4: 1-4: 19. DOI: 10.19153/cleiej.27.2.4

[20] A. Biswas, W. Talukdar. Guardrails for trust, safety, and ethical development and deployment of Large Language Models (LLM). Journal of Science & Technology. 2023, 4(6), 55-82. DOI: 10.55662/JST.2023.4605

[21] J.K. Kim, M. Chua, M. Rickard, A. Lorenzo. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. Journal of Pediatric Urology. 2023, 19(5), 598-604. DOI: 10.1016/j.jpurol.2023.05.018

[22] J. Vizcarra, S. Haruta, M. Kurokawa. Representing the Interaction between Users and Products via LLM-assisted Knowledge Graph Construction. 2024 IEEE 18th International Conference on Semantic Computing (ICSC). IEEE. 2024, 231-232. DOI: 10.1109/ICSC59802.2024.00043

[23] S. Kernan Freire, C. Wang, M. Foosherian, S. Wellsandt, S. Ruiz-Arenas, et al. Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. Frontiers in Artificial Intelligence. 2024, 7, 1293084. DOI: 10.3389/frai.2024.1293084

[24] L. Xia, C. Li, C. Zhang, S. Liu, P. Zheng. Leveraging error-assisted fine-tuning large language models for manufacturing excellence. Robotics and Computer-Integrated Manufacturing. 2024, 88, 102728. DOI: 10.1016/j.rcim.2024.102728

[25] Y. Wang, N. Lipka, R.A. Rossi, A. Siu. Knowledge graph prompting for multi-document question answering. Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(17), 19206-19214. DOI: 10.1609/aaai.v38i17.29889

[26] J. Chen, G. Lu, Z. Pan, Y. Tao. Research review of the knowledge graph and its application in power system dispatching and operation. Frontiers in Energy Research. 2022, 10, 896836. DOI: 10.3389/fenrg.2022.896836

[27] M.J. Zhang, N. Xu, J.H. Hu, Y.F. Wang, C. Li, et al. Knowledge graph construction and intelligent question answering for transformer operation and maintenance. Journal of Global Energy Interconnection. 2020, 3(6), 607-617. DOI: 10.19705/j.cnki.issn2096-5125.2020.06.008

[28] A. Fensel, Z. Akbar, E. Kärle, C. Blank, P. Pixner, et al. Knowledge graphs for online marketing and sales of touristic services. Information. 2020, 11(5), 253. DOI: 10.3390/info11050253